

so that  $z/|z|$  is a complex number of absolute value 1. Now any complex number  $a = x + iy$  with  $1 = |a| = x^2 + y^2$  can be written in the form

$$a = (\cos \theta, \sin \theta) = \cos \theta + i \sin \theta$$

for some number  $\theta$ . Thus every nonzero complex number  $z$  can be written

$$z = r(\cos \theta + i \sin \theta)$$

for some  $r > 0$  and some number  $\theta$ . The number  $r$  is unique (it equals  $|z|$ ), but  $\theta$  is not unique; if  $\theta_0$  is one possibility, then the others are  $\theta_0 + 2k\pi$  for  $k$  in  $\mathbf{Z}$ —any one of these numbers is called an **argument** of  $z$ . Figure 3 shows  $z$  in terms of  $r$  and  $\theta$ . (To find an argument  $\theta$  for  $z = x + iy$  we may note that the equation

$$x + iy = z = |z|(\cos \theta + i \sin \theta)$$

means that

$$\begin{aligned} x &= |z| \cos \theta, \\ y &= |z| \sin \theta. \end{aligned}$$

So, for example, if  $x > 0$  we can take  $\theta = \arctan y/x$ ; if  $x = 0$ , we can take  $\theta = \pi/2$  when  $y > 0$  and  $\theta = 3\pi/2$  when  $y < 0$ .)

Now the product of two nonzero complex numbers

$$\begin{aligned} z &= r(\cos \theta + i \sin \theta), \\ w &= s(\cos \phi + i \sin \phi), \end{aligned}$$

is

$$\begin{aligned} z \cdot w &= rs(\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi) \\ &= rs[(\cos \theta \cos \phi - \sin \theta \sin \phi) + i(\sin \theta \cos \phi + \cos \theta \sin \phi)] \\ &= rs[\cos(\theta + \phi) + i \sin(\theta + \phi)]. \end{aligned}$$

Thus, the absolute value of a product is the product of the absolute values of the factors, while the sum of any argument for each of the factors will be an argument for the product. For a nonzero complex number

$$z = r(\cos \theta + i \sin \theta)$$

it is now an easy matter to prove by induction the following very important formula (sometimes known as De Moivre's Theorem):

$$z^n = |z|^n(\cos n\theta + i \sin n\theta), \text{ for any argument } \theta \text{ of } z.$$

This formula describes  $z^n$  so explicitly that it is easy to decide just when  $z^n = w$ :

#### THEOREM 2

Every nonzero complex number has exactly  $n$  complex  $n$ th roots.

More precisely, for any complex number  $w \neq 0$ , and any natural number  $n$ , there are precisely  $n$  different complex numbers  $z$  satisfying  $z^n = w$ .

PROOF Let

$$w = s(\cos \phi + i \sin \phi)$$

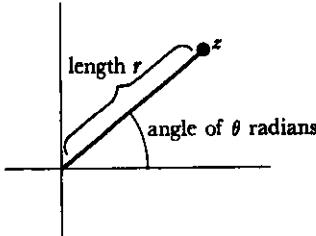


FIGURE 3

for  $s = |w|$  and some number  $\phi$ . Then a complex number

$$z = r(\cos \theta + i \sin \theta)$$

satisfies  $z^n = w$  if and only if

$$r^n(\cos n\theta + i \sin n\theta) = s(\cos \phi + i \sin \phi),$$

which happens if and only if

$$\begin{aligned} r^n &= s, \\ \cos n\theta + i \sin n\theta &= \cos \phi + i \sin \phi. \end{aligned}$$

From the first equation it follows that

$$r = \sqrt[n]{s},$$

where  $\sqrt[n]{s}$  denotes the positive real  $n$ th root of  $s$ . From the second equation it follows that for some integer  $k$  we have

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

Conversely, if we choose  $r = \sqrt[n]{s}$  and  $\theta = \theta_k$  for some  $k$ , then the number  $z = r(\cos \theta + i \sin \theta)$  will satisfy  $z^n = w$ . To determine the number of  $n$ th roots of  $w$ , it is therefore only necessary to determine which such  $z$  are distinct. Now any integer  $k$  can be written

$$k = nq + k'$$

for some integer  $q$ , and some integer  $k'$  between 0 and  $n - 1$ . Then

$$\cos \theta_k + i \sin \theta_k = \cos \theta_{k'} + i \sin \theta_{k'}.$$

This shows that every  $z$  satisfying  $z^n = w$  can be written

$$z = \sqrt[n]{s} (\cos \theta_k + i \sin \theta_k) \quad k = 0, \dots, n - 1.$$

Moreover, it is easy to see that these numbers are all different, since any two  $\theta_k$  for  $k = 0, \dots, n - 1$  differ by less than  $2\pi$ . ■

In the course of proving Theorem 2, we have actually developed a method for finding the  $n$ th roots of a complex number. For example, to find the cube roots of  $i$  (Figure 4) note that  $|i| = 1$  and that  $\pi/2$  is an argument for  $i$ . The cube roots of  $i$  are therefore

$$1 \cdot \left[ \cos \frac{\pi}{6} + i \sin \frac{\pi}{6} \right],$$

$$1 \cdot \left[ \cos \left( \frac{\pi}{6} + \frac{2\pi}{3} \right) + i \sin \left( \frac{\pi}{6} + \frac{2\pi}{3} \right) \right] = \cos \frac{5\pi}{6} + i \sin \frac{5\pi}{6},$$

$$1 \cdot \left[ \cos \left( \frac{\pi}{6} + \frac{4\pi}{3} \right) + i \sin \left( \frac{\pi}{6} + \frac{4\pi}{3} \right) \right] = \cos \frac{3\pi}{2} + i \sin \frac{3\pi}{2}.$$

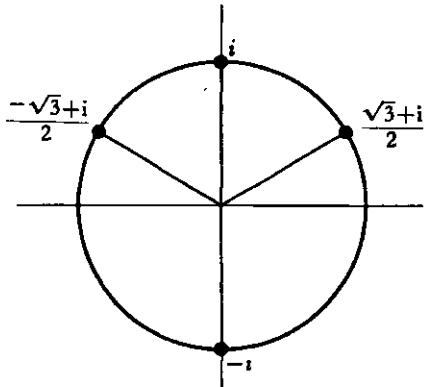


FIGURE 4

Since

$$\begin{aligned}\cos \frac{\pi}{6} &= \frac{\sqrt{3}}{2}, & \sin \frac{\pi}{6} &= \frac{1}{2}, \\ \cos \frac{5\pi}{6} &= -\frac{\sqrt{3}}{2}, & \sin \frac{5\pi}{6} &= \frac{1}{2}, \\ \cos \frac{3\pi}{2} &= 0, & \sin \frac{3\pi}{2} &= -1,\end{aligned}$$

the cube roots of  $i$  are

$$\frac{\sqrt{3}+i}{2}, \quad \frac{-\sqrt{3}+i}{2}, \quad -i.$$

In general, we cannot expect to obtain such simple results. For example, to find the cube roots of  $2 + 11i$ , note that  $|2 + 11i| = \sqrt{2^2 + 11^2} = \sqrt{125}$  and that  $\arctan \frac{11}{2}$  is an argument for  $2 + 11i$ . One of the cube roots of  $2 + 11i$  is therefore

$$\begin{aligned}\sqrt[3]{125} \left[ \cos \left( \frac{\arctan \frac{11}{2}}{3} \right) + i \sin \left( \frac{\arctan \frac{11}{2}}{3} \right) \right] \\ = \sqrt{5} \left[ \cos \left( \frac{\arctan \frac{11}{2}}{3} \right) + i \sin \left( \frac{\arctan \frac{11}{2}}{3} \right) \right].\end{aligned}$$

Previously we noted that  $2 + i$  is also a cube root of  $2 + 11i$ . Since  $|2 + i| = \sqrt{2^2 + 1^2} = \sqrt{5}$ , and since  $\arctan \frac{1}{2}$  is an argument of  $2 + i$ , we can write this cube root as

$$2 + i = \sqrt{5}(\cos \arctan \frac{1}{2} + i \sin \arctan \frac{1}{2}).$$

These two cube roots are actually the same number, because

$$\frac{\arctan \frac{11}{2}}{3} = \arctan \frac{1}{2}$$

(you can check this by using the formula in Problem 15-9), but this is hardly the sort of thing one might notice!

The fact that every complex number has an  $n$ th root for all  $n$  is just a special case of a very important theorem. The number  $i$  was originally introduced in order to provide a solution for the equation  $x^2 + 1 = 0$ . The *Fundamental Theorem of Algebra* states the remarkable fact that this one addition automatically provides solutions for all other polynomial equations: every equation

$$z^n + a_{n-1}z^{n-1} + \cdots + a_0 = 0 \quad a_0, \dots, a_{n-1} \text{ in } \mathbf{C}$$

has a complex root!

In the next chapter we shall give an almost complete proof of the Fundamental Theorem of Algebra; the slight gap left in the text can be filled in as an exercise (Problem 26-5). The proof of the theorem will rely on several new concepts which come up quite naturally in a more thorough investigation of complex numbers.

## PROBLEMS

1. Find the absolute value and argument of each of the following.

- (i)  $3 + 4i$ .
- (ii)  $(3 + 4i)^{-1}$ .
- (iii)  $\overline{(1+i)^5}$ .
- (iv)  $\sqrt[7]{3+4i}$ .
- (v)  $|3+4i|$ .

2. Solve the following equations.

- (i)  $x^2 + ix + 1 = 0$ .
- (ii)  $x^4 + x^2 + 1 = 0$ .
- (iii)  $x^2 + 2ix - 1 = 0$ .
- (iv)  $\begin{cases} ix - (1+i)y = 3, \\ (2+i)x + iy = 4 \end{cases}$ .
- (v)  $x^3 - x^2 - x - 2 = 0$ .

3. Describe the set of all complex numbers  $z$  such that

- (i)  $\bar{z} = -z$ .
- (ii)  $\bar{z} = z^{-1}$ .
- (iii)  $|z - a| = |z - b|$ .
- (iv)  $|z - a| + |z - b| = c$ .
- (v)  $|z| < 1 - \text{real part of } z$ .

- 4. Prove that  $|z| = |\bar{z}|$ , and that the real part of  $z$  is  $(z + \bar{z})/2$ , while the imaginary part is  $(z - \bar{z})/2i$ .
- 5. Prove that  $|z + w|^2 + |z - w|^2 = 2(|z|^2 + |w|^2)$ , and interpret this statement geometrically.
- 6. What is the pictorial relation between  $z$  and  $\sqrt{i} \cdot z\sqrt{-i}$ ? Hint: Which line goes into the real axis under multiplication by  $\sqrt{-i}$ ?
- 7. (a) Prove that if  $a_0, \dots, a_{n-1}$  are *real* and  $a + bi$  (for  $a$  and  $b$  real) satisfies the equation  $z^n + a_{n-1}z^{n-1} + \dots + a_0 = 0$ , then  $a - bi$  also satisfies this equation. (Thus the nonreal roots of such an equation always occur in pairs, and the number of such roots is even.)  
(b) Conclude that  $z^n + a_{n-1}z^{n-1} + \dots + a_0$  is divisible by  $z^2 - 2az + (a^2 + b^2)$  (whose coefficients are real).
- \*8. (a) Let  $c$  be an integer which is not the square of another integer. If  $a$  and  $b$  are integers we define the **conjugate** of  $a + b\sqrt{c}$ , denoted by  $\bar{a} + \bar{b}\sqrt{c}$ , as  $a - b\sqrt{c}$ . Show that the conjugate is well defined by showing that a number can be written  $a + b\sqrt{c}$ , for integers  $a$  and  $b$ , in only one way.  
(b) Show that for all  $\alpha$  and  $\beta$  of the form  $a + b\sqrt{c}$ , we have  $\bar{\bar{\alpha}} = \alpha$ ;  $\bar{\alpha} = \alpha$  if

and only if  $\alpha$  is an integer;  $\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}$ ;  $\overline{-\alpha} = -\bar{\alpha}$ ;  $\overline{\alpha \cdot \beta} = \bar{\alpha} \cdot \bar{\beta}$ ; and  $\overline{\alpha^{-1}} = (\bar{\alpha})^{-1}$  if  $\alpha \neq 0$ .

- (c) Prove that if  $a_0, \dots, a_{n-1}$  are integers and  $z = a + b\sqrt{c}$  satisfies the equation  $z^n + a_{n-1}z^{n-1} + \dots + a_0 = 0$ , then  $\bar{z} = a - b\sqrt{c}$  also satisfies this equation.
- 9. Find all the 4th roots of  $i$ ; express the one having smallest argument in a form that does not involve any trigonometric functions.
- \*10. (a) Prove that if  $\omega$  is an  $n$ th root of 1, then so is  $\omega^k$ .  
 (b) A number  $\omega$  is called a **primitive  $n$ th root** of 1 if  $\{1, \omega, \omega^2, \dots, \omega^{n-1}\}$  is the set of all  $n$ th roots of 1. How many primitive  $n$ th roots of 1 are there for  $n = 3, 4, 5, 9$ ?  
 (c) Let  $\omega$  be an  $n$ th root of 1, with  $\omega \neq 1$ . Prove that  $\sum_{k=0}^{n-1} \omega^k = 0$ .
- \*11. (a) Prove that if  $z_1, \dots, z_k$  lie on one side of some straight line through 0, then  $z_1 + \dots + z_k \neq 0$ . Hint: This is obvious from the geometric interpretation of addition, but an analytic proof is also easy: the assertion is clear if the line is the real axis, and a trick will reduce the general case to this one.  
 (b) Show further that  $z_1^{-1}, \dots, z_k^{-1}$  all lie on one side of a straight line through 0, so that  $z_1^{-1} + \dots + z_k^{-1} \neq 0$ .
- \*12. Prove that if  $|z_1| = |z_2| = |z_3|$  and  $z_1 + z_2 + z_3 = 0$ , then  $z_1, z_2$ , and  $z_3$  are the vertices of an equilateral triangle. Hint: It will help to assume that  $z_1$  is real, and this can be done with no loss of generality. Why?

# CHAPTER 26 COMPLEX FUNCTIONS

You will probably not be surprised to learn that a deeper investigation of complex numbers depends on the notion of functions. Until now a function was (intuitively) a rule which assigned real numbers to certain other real numbers. But there is no reason why this concept should not be extended; we might just as well consider a rule which assigns complex numbers to certain other complex numbers. A rigorous definition presents no problems (we will not even accord it the full honors of a formal definition): a function is a collection of pairs of complex numbers which does not contain two distinct pairs with the same first element. Since we consider real numbers to be certain complex numbers, the old definition is really a special case of the new one. Nevertheless, we will sometimes resort to special terminology in order to clarify the context in which a function is being considered. A function  $f$  is called **real-valued** if  $f(z)$  is a real number for all  $z$  in the domain of  $f$ , and **complex-valued** to emphasize that it is not necessarily real-valued. Similarly, we will usually state explicitly that a function  $f$  is defined on [a subset of]  $\mathbf{R}$  in those cases where the domain of  $f$  is [a subset of]  $\mathbf{R}$ ; in other cases we sometimes mention that  $f$  is defined on [a subset of]  $\mathbf{C}$  to emphasize that  $f(z)$  is defined for complex  $z$  as well as real  $z$ .

Among the multitude of functions defined on  $\mathbf{C}$ , certain ones are particularly important. Foremost among these are the functions of the form

$$f(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0,$$

where  $a_0, \dots, a_n$  are complex numbers. These functions are called, as in the real case, polynomial functions; they include the function  $f(z) = z$  (the “identity function”) and functions of the form  $f(z) = a$  for some complex number  $a$  (“constant functions”). Another important generalization of a familiar function is the “absolute value function”  $f(z) = |z|$  for all  $z$  in  $\mathbf{C}$ .

Two functions of particular importance for complex numbers are  $\operatorname{Re}$  (the “real part function”) and  $\operatorname{Im}$  (the “imaginary part function”), defined by

$$\begin{aligned}\operatorname{Re}(x + iy) &= x, \\ \operatorname{Im}(x + iy) &= y,\end{aligned}\quad \text{for } x \text{ and } y \text{ real.}$$

The “conjugate function” is defined by

$$f(z) = \bar{z} = \operatorname{Re}(z) - i \operatorname{Im}(z).$$

Familiar real-valued functions defined on  $\mathbf{R}$  may be combined in many ways to produce new complex-valued functions defined on  $\mathbf{C}$ —an example is the function

$$f(x + iy) = e^y \sin(x - y) + ix^3 \cos y.$$

The formula for this particular function illustrates a decomposition which is always possible. Any complex-valued function  $f$  can be written in the form

$$f = u + iv$$

for some real-valued functions  $u$  and  $v$ —simply define  $u(z)$  as the real part of  $f(z)$ , and  $v(z)$  as the imaginary part. This decomposition is often very useful, but not always; for example, it would be inconvenient to describe a polynomial function in this way.

One other function will play an important role in this chapter. Recall that an *argument* of a nonzero complex number  $z$  is a (real) number  $\theta$  such that

$$z = |z|(\cos \theta + i \sin \theta).$$

There are infinitely many arguments for  $z$ , but just one which satisfies  $0 \leq \theta < 2\pi$ . If we call this unique argument  $\theta(z)$ , then  $\theta$  is a (real-valued) function (the “argument function”) on  $\{z \in \mathbf{C} : z \neq 0\}$ .

“Graphs” of complex-valued functions defined on  $\mathbf{C}$ , since they lie in 4-dimensional space, are presumably not very useful for visualization. The alternative picture of a function mentioned in Chapter 4 can be used instead: we draw two copies of  $\mathbf{C}$ , and arrows from  $z$  in one copy, to  $f(z)$  in the other (Figure 1).

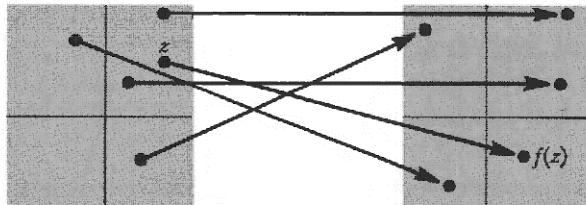
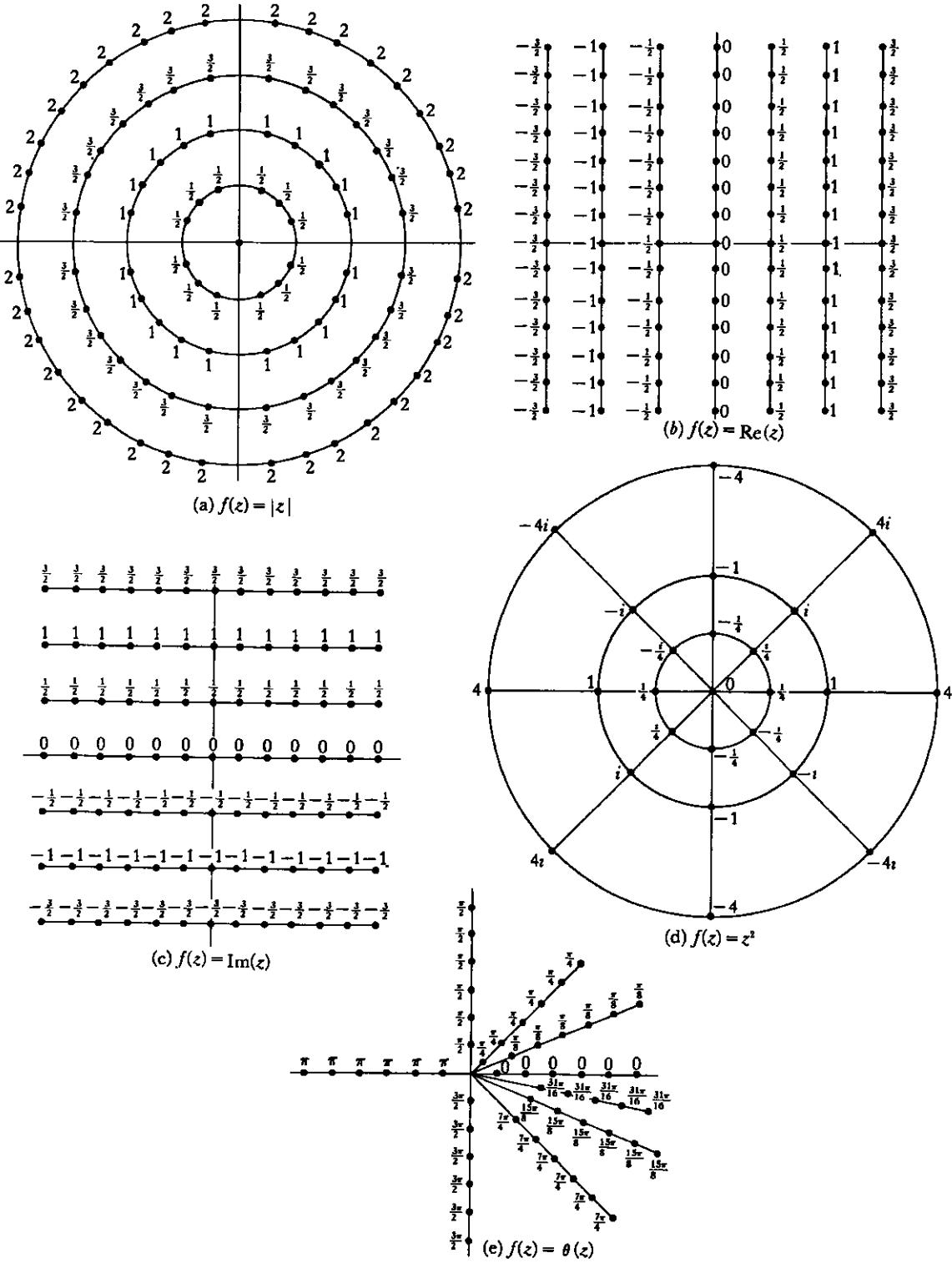


FIGURE 1

The most common pictorial representation of a complex-valued function is produced by labeling a point in the plane with the value  $f(z)$ , instead of with  $z$  (which can be estimated from the position of the point in the picture). Figure 2 shows this sort of picture for several different functions. Certain features of the function are illustrated very clearly by such a “graph.” For example, the absolute value function is constant on concentric circles around 0, the functions  $\operatorname{Re}$  and  $\operatorname{Im}$  are constant on the vertical and horizontal lines, respectively, and the function  $f(z) = z^2$  wraps the circle of radius  $r$  twice around the circle of radius  $r^2$ .

Despite the problems involved in visualizing complex-valued functions in general, it is still possible to define analogues of important properties previously defined for real-valued functions on  $\mathbf{R}$ , and in some cases these properties may be easier to visualize in the complex case. For example, the notion of limit can be defined as follows:

$\lim_{z \rightarrow a} f(z) = l$  means that for every (real) number  $\varepsilon > 0$  there is a (real) number  $\delta > 0$  such that, for all  $z$ , if  $0 < |z - a| < \delta$ , then  $|f(z) - l| < \varepsilon$ .



**FIGURE 2**

Although the definition reads precisely as before, the interpretation is slightly different. Since  $|z - w|$  is the distance between the complex numbers  $z$  and  $w$ , the equation  $\lim_{z \rightarrow a} f(z) = l$  means that the values of  $f(z)$  can be made to lie inside any given circle around  $l$ , provided that  $z$  is restricted to lie inside a sufficiently small circle around  $a$ . This assertion is particularly easy to visualize using the “two copy” picture of a function (Figure 3).

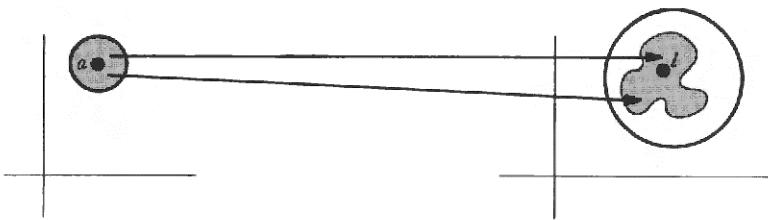


FIGURE 3

Certain facts about limits can be proved exactly as in the real case. In particular,

$$\lim_{z \rightarrow a} c = c,$$

$$\lim_{z \rightarrow a} z = a,$$

$$\lim_{z \rightarrow a} [f(z) + g(z)] = \lim_{z \rightarrow a} f(z) + \lim_{z \rightarrow a} g(z),$$

$$\lim_{z \rightarrow a} f(z) \cdot g(z) = \lim_{z \rightarrow a} f(z) \cdot \lim_{z \rightarrow a} g(z),$$

$$\lim_{z \rightarrow a} \frac{1}{g(z)} = \frac{1}{\lim_{z \rightarrow a} g(z)}, \quad \text{if } \lim_{z \rightarrow a} g(z) \neq 0.$$

The essential property of absolute values upon which these results are based is the inequality  $|z + w| \leq |z| + |w|$ , and this inequality holds for complex numbers as well as for real numbers. These facts already provide quite a few limits, but many more can be obtained from the following theorem.

**THEOREM 1** Let  $f(z) = u(z) + i v(z)$  for real-valued functions  $u$  and  $v$ , and let  $l = \alpha + i\beta$  for real numbers  $\alpha$  and  $\beta$ . Then  $\lim_{z \rightarrow a} f(z) = l$  if and only if

$$\lim_{z \rightarrow a} u(z) = \alpha,$$

$$\lim_{z \rightarrow a} v(z) = \beta.$$

**PROOF** Suppose first that  $\lim_{z \rightarrow a} f(z) = l$ . If  $\varepsilon > 0$ , there is  $\delta > 0$  such that, for all  $z$ ,

$$\text{if } 0 < |z - a| < \delta, \text{ then } |f(z) - l| < \varepsilon.$$

The second inequality can be written

$$|[u(z) - \alpha] + i[v(z) - \beta]| < \varepsilon,$$

or

$$[u(z) - \alpha]^2 + [v(z) - \beta]^2 < \varepsilon^2.$$

Since  $u(z) - \alpha$  and  $v(z) - \beta$  are both real numbers, their squares are positive; this inequality therefore implies that

$$[u(z) - \alpha]^2 < \varepsilon^2 \quad \text{and} \quad [v(z) - \beta]^2 < \varepsilon^2,$$

which implies that

$$|u(z) - \alpha| < \varepsilon \quad \text{and} \quad |v(z) - \beta| < \varepsilon.$$

Since this is true for all  $\varepsilon > 0$ , it follows that

$$\lim_{z \rightarrow a} u(z) = \alpha \quad \text{and} \quad \lim_{z \rightarrow a} v(z) = \beta.$$

Now suppose that these two equations hold. If  $\varepsilon > 0$ , there is a  $\delta > 0$  such that, for all  $z$ , if  $0 < |z - a| < \delta$ , then

$$|u(z) - \alpha| < \frac{\varepsilon}{2} \quad \text{and} \quad |v(z) - \beta| < \frac{\varepsilon}{2},$$

which implies that

$$\begin{aligned} |f(z) - l| &= |[u(z) - \alpha] + i[v(z) - \beta]| \\ &\leq |u(z) - \alpha| + |i| \cdot |v(z) - \beta| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

This proves that  $\lim_{z \rightarrow a} f(z) = l$ . ■

In order to apply Theorem 1 fruitfully, notice that since we already know the limit  $\lim_{z \rightarrow a} z = a$ , we can conclude that

$$\lim_{z \rightarrow a} \operatorname{Re}(z) = \operatorname{Re}(a),$$

$$\lim_{z \rightarrow a} \operatorname{Im}(z) = \operatorname{Im}(a).$$

A limit like

$$\lim_{z \rightarrow a} \sin(\operatorname{Re}(z)) = \sin(\operatorname{Re}(a))$$

follows easily, using continuity of  $\sin$ . Many applications of these principles prove such limits as the following:

$$\lim_{z \rightarrow a} \bar{z} = \bar{a},$$

$$\lim_{z \rightarrow a} |z| = |a|,$$

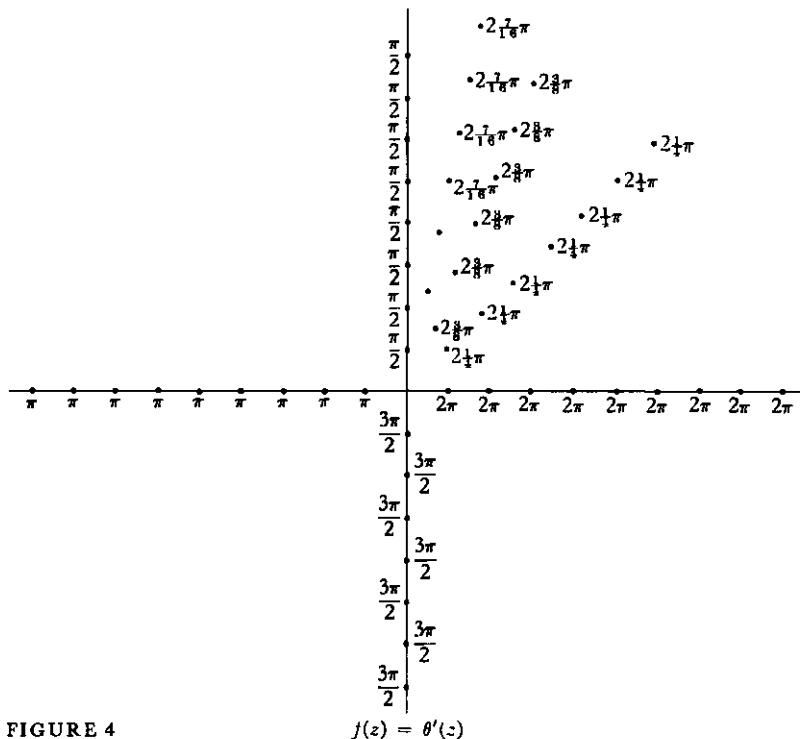
$$\lim_{(x+iy) \rightarrow a+bi} e^y \sin x + ix^3 \cos y = e^b \sin a + ia^3 \cos b.$$

Now that the notion of limit has been extended to complex functions, the notion of continuity can also be extended:  $f$  is **continuous at  $a$**  if  $\lim_{z \rightarrow a} f(z) = f(a)$ , and

$f$  is **continuous** if  $f$  is continuous at  $a$  for all  $a$  in the domain of  $f$ . The previous work on limits shows that all the following functions are continuous:

$$\begin{aligned}f(z) &= a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0, \\f(z) &= \bar{z}, \\f(z) &= |z|, \\f(x+iy) &= e^y \sin x + ix^3 \cos y.\end{aligned}$$

Examples of discontinuous functions are easy to produce, and certain ones come up very naturally. One particularly frustrating example is the “argument function”  $\theta$ , which is discontinuous at all nonnegative real numbers (see the “graph” in Figure 2). By suitably redefining  $\theta$  it is possible to change the discontinuities; for example (Figure 4), if  $\theta'(z)$  denotes the unique argument of  $z$  with  $\pi/2 \leq \theta'(z) < 5\pi/2$ , then  $\theta'$  is discontinuous at  $ai$  for every nonnegative real number  $a$ . But, no matter how  $\theta$  is redefined, some discontinuities will always occur.



**FIGURE 4**

The discontinuity of  $\theta$  has an important bearing on the problem of defining a “square-root function,” that is, a function  $f$  such that  $(f(z))^2 = z$  for all  $z$ . For real numbers the function  $\sqrt{\phantom{x}}$  had as domain only the nonnegative real numbers. If complex numbers are allowed, then every number has two square roots (except 0, which has only one). Although this situation may seem better, it is in some ways worse; since the square roots of  $z$  are complex numbers, there is no clear criterion for selecting one root to be  $f(z)$ , in preference to the other.

One way to define  $f$  is the following. We set  $f(0) = 0$ , and for  $z \neq 0$  we set

$$f(z) = \sqrt{|z|} \left( \cos \frac{\theta(z)}{2} + i \sin \frac{\theta(z)}{2} \right).$$

Clearly  $(f(z))^2 = z$ , but the function  $f$  is discontinuous, since  $\theta$  is discontinuous. As a matter of fact, it is impossible to find a continuous  $f$  such that  $(f(z))^2 = z$  for all  $z$ . In fact, it is even impossible for  $f(z)$  to be defined for all  $z$  with  $|z| = 1$ . To prove this by contradiction, we can assume that  $f(1) = 1$  (since we could always replace  $f$  by  $-f$ ). Then we claim that for all  $\theta$  with  $0 \leq \theta < 2\pi$  we have

$$(*) \quad f(\cos \theta + i \sin \theta) = \cos \frac{\theta}{2} + i \sin \frac{\theta}{2}.$$

The argument for this is left to you (it is a standard type of least upper bound argument). But  $(*)$  implies that

$$\begin{aligned} \lim_{\theta \rightarrow 2\pi} f(\cos \theta + i \sin \theta) &= \cos \pi + i \sin \pi \\ &= -1 \\ &\neq f(1), \end{aligned}$$

even though  $\cos \theta + i \sin \theta \rightarrow 1$  as  $\theta \rightarrow 2\pi$ . Thus, we have our contradiction. A similar argument shows that it is impossible to define continuous “ $n$ th-root functions” for any  $n \geq 2$ .

For continuous complex functions there are important analogues of certain theorems which describe the behavior of real-valued functions on closed intervals. A natural analogue of the interval  $[a, b]$  is the set of all complex numbers  $z = x + iy$  with  $a \leq x \leq b$  and  $c \leq y \leq d$  (Figure 5). This set is called a **closed rectangle**, and is denoted by  $[a, b] \times [c, d]$ .

If  $f$  is a continuous complex-valued function whose domain is  $[a, b] \times [c, d]$ , then it seems reasonable, and is indeed true, that  $f$  is bounded on  $[a, b] \times [c, d]$ . That is, there is some real number  $M$  such that

$$|f(z)| \leq M \quad \text{for all } z \text{ in } [a, b] \times [c, d].$$

It does not make sense to say that  $f$  has a maximum and a minimum value on  $[a, b] \times [c, d]$ , since there is no notion of order for complex numbers. If  $f$  is a real-valued function, however, then this assertion does make sense, and is true. In particular, if  $f$  is any complex-valued continuous function on  $[a, b] \times [c, d]$ , then  $|f|$  is also continuous, so there is some  $z_0$  in  $[a, b] \times [c, d]$  such that

$$|f(z_0)| \leq |f(z)| \quad \text{for all } z \text{ in } [a, b] \times [c, d];$$

a similar statement is true with the inequality reversed. It is sometimes said that “ $f$  attains its maximum and minimum modulus on  $[a, b] \times [c, d]$ .”

The various facts listed in the previous paragraph will not be proved here, although proofs are outlined in Problem 5. Assuming these facts, however, we can

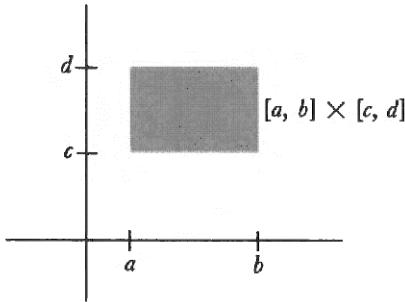


FIGURE 5

now give a proof of the Fundamental Theorem of Algebra, which is really quite surprising, since we have not yet said much to distinguish polynomial functions from other continuous functions.

**THEOREM 2 (THE FUNDAMENTAL THEOREM OF ALGEBRA)**

Let  $a_0, \dots, a_{n-1}$  be any complex numbers. Then there is a complex number  $z$  such that

$$z^n + a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \dots + a_0 = 0.$$

**PROOF** Let

$$f(z) = z^n + a_{n-1}z^{n-1} + \dots + a_0.$$

Then  $f$  is continuous, and so is the function  $|f|$  defined by

$$|f|(z) = |f(z)| = |z^n + a_{n-1}z^{n-1} + \dots + a_0|.$$

Our proof is based on the observation that a point  $z_0$  with  $f(z_0) = 0$  would clearly be a minimum point for  $|f|$ . To prove the theorem we will first show that  $|f|$  does indeed have a smallest value on the *whole complex plane*. The proof will be almost identical to the proof, in Chapter 7, that a polynomial function of even degree (with real coefficients) has a smallest value on all of  $\mathbf{R}$ ; both proofs depend on the fact that if  $|z|$  is large, then  $|f(z)|$  is large.

We begin by writing, for  $z \neq 0$ ,

$$f(z) = z^n \left( 1 + \frac{a_{n-1}}{z} + \dots + \frac{a_0}{z^n} \right),$$

so that

$$|f(z)| = |z|^n \cdot \left| 1 + \frac{a_{n-1}}{z} + \dots + \frac{a_0}{z^n} \right|.$$

Let

$$M = \max(1, 2n|a_{n-1}|, \dots, 2n|a_0|).$$

Then for all  $z$  with  $|z| \geq M$ , we have  $|z^k| \geq |z|$  and

$$\frac{|a_{n-k}|}{|z^k|} \leq \frac{|a_{n-k}|}{|z|} \leq \frac{|a_{n-k}|}{2n|a_{n-k}|} = \frac{1}{2n},$$

so

$$\left| \frac{a_{n-1}}{z} + \dots + \frac{a_0}{z^n} \right| \leq \left| \frac{a_{n-1}}{z} \right| + \dots + \left| \frac{a_0}{z^n} \right| \leq \frac{1}{2},$$

which implies that

$$\left| 1 + \frac{a_{n-1}}{z} + \dots + \frac{a_0}{z^n} \right| \geq 1 - \left| \frac{a_{n-1}}{z} + \dots + \frac{a_0}{z^n} \right| \geq \frac{1}{2}.$$

This means that

$$|f(z)| \geq \frac{|z|^n}{2} \quad \text{for } |z| \geq M.$$

In particular, if  $|z| \geq M$  and also  $|z| \geq \sqrt[n]{2|f(0)|}$ , then

$$|f(z)| \geq |f(0)|.$$

Now let  $[a, b] \times [c, d]$  be a closed rectangle (Figure 6) which contains  $\{z : |z| \leq \max(M, \sqrt[4]{2|f(0)|})\}$ , and suppose that the minimum of  $|f(z)|$  on  $[a, b] \times [c, d]$  is attained at  $z_0$ , so that

$$(1) \quad |f(z_0)| \leq |f(z)| \quad \text{for } z \text{ in } [a, b] \times [c, d].$$

It follows, in particular, that  $|f(z_0)| \leq |f(0)|$ . Thus

$$(2) \quad \text{if } |z| \geq \max(M, \sqrt[4]{2|f(0)|}), \text{ then } |f(z)| \geq |f(0)| \geq |f(z_0)|.$$

Combining (1) and (2) we see that  $|f(z_0)| \leq |f(z)|$  for all  $z$ , so that  $|f|$  attains its minimum value on the whole complex plane at  $z_0$ .

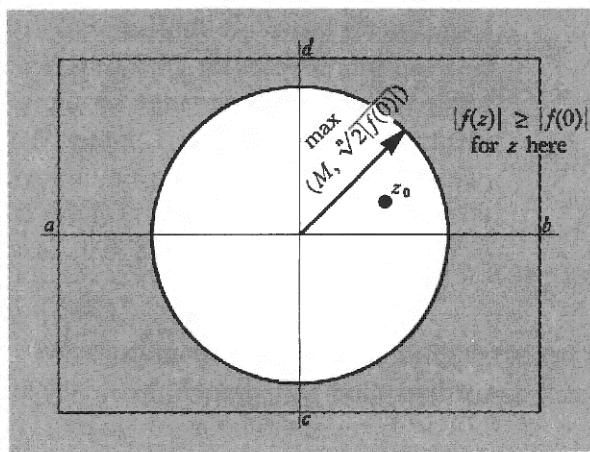


FIGURE 6

To complete the proof of the theorem we now show that  $f(z_0) = 0$ . It is convenient to introduce the function  $g$  defined by

$$g(z) = f(z + z_0).$$

Then  $g$  is a polynomial function of degree  $n$ , whose minimum absolute value occurs at 0. We want to show that  $g(0) = 0$ .

Suppose instead that  $g(0) = \alpha \neq 0$ . If  $m$  is the smallest positive power of  $z$  which occurs in the expression for  $g$ , we can write

$$g(z) = \alpha + \beta z^m + c_{m+1} z^{m+1} + \cdots + c_n z^n,$$

where  $\beta \neq 0$ . Now, according to Theorem 25-2 there is a complex number  $\gamma$  such that

$$\gamma^m = -\frac{\alpha}{\beta}.$$

Then, setting  $d_k = c_k \gamma^k$ , we have

$$\begin{aligned} |g(\gamma z)| &= |\alpha + \beta \gamma^m z^m + d_{m+1} z^{m+1} + \cdots + d_n z^n| \\ &= |\alpha - \alpha z^m + d_{m+1} z^{m+1} + \cdots| \\ &= \left| \alpha \left( 1 - z^m + \frac{d_{m+1}}{\alpha} z^{m+1} + \cdots \right) \right| \\ &= \left| \alpha \left( 1 - z^m + z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right) \right| \\ &= |\alpha| \cdot \left| 1 - z^m + z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right|. \end{aligned}$$

This expression, so tortuously arrived at, will enable us to reach a quick contradiction. Notice first that if  $|z|$  is chosen small enough, we will have

$$\left| \frac{d_{m+1}}{\alpha} z + \cdots \right| < 1.$$

If we choose, from among all  $z$  for which this inequality holds, some  $z$  which is *real and positive*, then

$$\left| z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right| < |z^m| = z^m.$$

Consequently, if  $0 < z < 1$  we have

$$\begin{aligned} \left| 1 - z^m + z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right| &\leq |1 - z^m| + \left| z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right| \\ &= 1 - z^m + \left| z^m \left[ \frac{d_{m+1}}{\alpha} z + \cdots \right] \right| \\ &< 1 - z^m + z^m \\ &= 1. \end{aligned}$$

This is the desired contradiction: for such a number  $z$  we have

$$|g(\gamma z)| < |\alpha|,$$

contradicting the fact that  $|\alpha|$  is the minimum of  $|g|$  on the whole plane. Hence, the original assumption must be incorrect, and  $g(0) = 0$ . This implies, finally, that  $f(z_0) = 0$ . ■

Even taking into account our omission of the proofs for the basic facts about continuous complex functions, this proof verified a deep fact with surprisingly little work. It is only natural to hope that other interesting developments will arise if we pursue further the analogues of properties of real functions. The next obvious step is to define derivatives: a function  $f$  is **differentiable at  $a$**  if

$$\lim_{z \rightarrow 0} \frac{f(a+z) - f(a)}{z} \text{ exists,}$$

in which case the limit is denoted by  $f'(a)$ . It is easy to prove that

$$\begin{aligned} f'(a) &= 0 && \text{if } f(z) = c, \\ f'(a) &= 1 && \text{if } f(z) = z, \\ (f + g)'(a) &= f'(a) + g'(a), \\ (f \cdot g)'(a) &= f'(a)g(a) + f(a)g'(a), \\ \left(\frac{1}{g}\right)'(a) &= \frac{-g'(a)}{[g(a)]^2} && \text{if } g(a) \neq 0, \\ (f \circ g)'(a) &= f'(g(a)) \cdot g'(a); \end{aligned}$$

the proofs of all these formulas are exactly the same as before. It follows, in particular, that if  $f(z) = z^n$ , then  $f'(z) = nz^{n-1}$ . These formulas only prove the differentiability of rational functions however. Many other obvious candidates are *not* differentiable. Suppose, for example, that

$$f(x + iy) = x - iy \quad (\text{i.e., } f(z) = \bar{z}).$$

If  $f$  is to be differentiable at 0, then the limit

$$\lim_{(x+iy) \rightarrow 0} \frac{f(x+iy) - f(0)}{x+iy} = \lim_{(x+iy) \rightarrow 0} \frac{x-iy}{x+iy}$$

must exist. Notice however, that

$$\text{if } y = 0, \text{ then } \frac{x-iy}{x+iy} = 1,$$

and

$$\text{if } x = 0, \text{ then } \frac{x-iy}{x+iy} = -1;$$

therefore this limit cannot possibly exist, since the quotient has both the values 1 and -1 for  $x + iy$  arbitrarily close to 0.

In view of this example, it is not at all clear where other differentiable functions are to come from. If you recall the definitions of  $\sin$  and  $\exp$ , you will see that there is no hope at all of generalizing these definitions to complex numbers. At the moment the outlook is bleak, but all our problems will soon be solved.

## PROBLEMS

1. (a) For any real number  $y$ , define  $\alpha(x) = x + iy$  (so that  $\alpha$  is a complex-valued function defined on  $\mathbf{R}$ ). Show that  $\alpha$  is continuous. (This follows immediately from a theorem in this chapter.) Show similarly that  $\beta(y) = x + iy$  is continuous.  
 (b) Let  $f$  be a continuous function defined on  $\mathbf{C}$ . For fixed  $y$ , let  $g(x) = f(x + iy)$ . Show that  $g$  is a continuous function (defined on  $\mathbf{R}$ ). Show similarly that  $h(y) = f(x + iy)$  is continuous. Hint: Use part (a).
2. (a) Suppose that  $f$  is a continuous real-valued function defined on a closed rectangle  $[a, b] \times [c, d]$ . Prove that if  $f$  takes on the values  $f(z)$  and  $f(w)$

for  $z$  and  $w$  in  $[a, b] \times [c, d]$ , then  $f$  also takes all values between  $f(z)$  and  $f(w)$ . Hint: Consider  $g(t) = f(tz + (1 - t)w)$  for  $t$  in  $[0, 1]$ .

- \*(b) If  $f$  is a continuous complex-valued function defined on  $[a, b] \times [c, d]$ , the assertion in part (a) no longer makes any sense, since we cannot talk of complex numbers between  $f(z)$  and  $f(w)$ . We might conjecture that  $f$  takes on all values on the line segment between  $f(z)$  and  $f(w)$ , but even this is false. Find an example which shows this.

3. (a) Prove that if  $a_0, \dots, a_{n-1}$  are any complex numbers, then there are complex numbers  $z_1, \dots, z_n$  (not necessarily distinct) such that

$$z^n + a_{n-1}z^{n-1} + \cdots + a_0 = \prod_{i=1}^n (z - z_i).$$

- (b) Prove that if  $a_0, \dots, a_{n-1}$  are *real*, then  $z^n + a_{n-1}z^{n-1} + \cdots + a_0$  can be written as a product of linear factors  $z+a$  and quadratic factors  $z^2+az+b$  all of whose coefficients are real. (Use Problem 25-7.)

4. In this problem we will consider only polynomials with real coefficients. Such a polynomial is called a **sum of squares** if it can be written as  $h_1^2 + \cdots + h_n^2$  for polynomials  $h_i$  with real coefficients.

- (a) Prove that if  $f$  is a sum of squares, then  $f(x) \geq 0$  for all  $x$ .

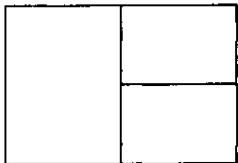
- (b) Prove that if  $f$  and  $g$  are sums of squares, then so is  $f \cdot g$ .

- (c) Suppose that  $f(x) \geq 0$  for all  $x$ . Show that  $f$  is a sum of squares. Hint: First write  $f(x) = x^k g(x)$ , where  $g(x) \neq 0$  for all  $x$ . Then  $k$  must be even (why?), and  $g(x) > 0$  for all  $x$ . Now use Problem 3(b).

5. (a) Let  $A$  be a set of complex numbers. A number  $z$  is called, as in the real case, a **limit point** of the set  $A$  if for every (real)  $\varepsilon > 0$ , there is a point  $a$  in  $A$  with  $|z - a| < \varepsilon$  but  $z \neq a$ . Prove the two-dimensional version of the Bolzano-Weierstrass Theorem: If  $A$  is an infinite subset of  $[a, b] \times [c, d]$ , then  $A$  has a limit point in  $[a, b] \times [c, d]$ . Hint: First divide  $[a, b] \times [c, d]$  in half by a vertical line as in Figure 7(a). Since  $A$  is infinite, at least one half contains infinitely many points of  $A$ . Divide this in half by a horizontal line, as in Figure 7(b). Continue in this way, alternately dividing by vertical and horizontal lines.



(a)

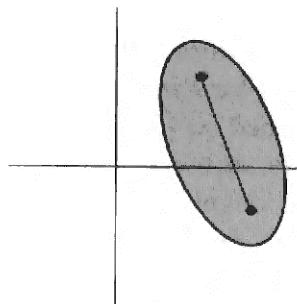


(b)

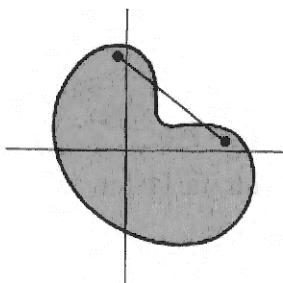
FIGURE 7

(The two-dimensional bisection argument outlined in this hint is so standard that the title “Bolzano-Weierstrass” often serves to describe the method of proof, in addition to the theorem itself. See, for example, H. Petard, “A Contribution to the Mathematical Theory of Big Game Hunting,” *Amer. Math. Monthly*, 45 (1938), 446–447.)

- (b) Prove that a continuous (complex-valued) function on  $[a, b] \times [c, d]$  is bounded on  $[a, b] \times [c, d]$ . (Imitate Problem 22-31.)
- (c) Prove that if  $f$  is a real-valued continuous function on  $[a, b] \times [c, d]$ , then  $f$  takes on a maximum and minimum value on  $[a, b] \times [c, d]$ . (You can use the same trick that works for Theorem 7-3.)



(a) a convex subset of the plane



(b) a nonconvex subset of the plane

FIGURE 8

- \*6. The proof of Theorem 2 cannot be considered to be completely elementary because the possibility of choosing  $\gamma$  with  $\gamma^m = -\alpha/\beta$  depends on Theorem 25-2, and thus on the trigonometric functions. It is therefore of some interest to provide an elementary proof that there is a solution for the equation  $z^n - c = 0$ .

- Make an explicit computation to show that solutions of  $z^2 - c = 0$  can be found for any complex number  $c$ .
- Explain why the solution of  $z^n - c = 0$  can be reduced to the case where  $n$  is odd.
- Let  $z_0$  be the point where the function  $f(z) = z^n - c$  has its minimum absolute value. If  $z_0 \neq 0$ , show that the integer  $m$  in the proof of Theorem 2 is equal to 1; since we can certainly find  $\gamma$  with  $\gamma^1 = -\alpha/\beta$ , the remainder of the proof works for  $f$ . It therefore suffices to show that the minimum absolute value of  $f$  does not occur at 0.
- Suppose instead that  $f$  has its minimum absolute value at 0. Since  $n$  is odd, the points  $\pm\delta, \pm\delta i$  go under  $f$  into  $-c \pm \delta^n, -c \pm \delta^n i$ . Show that for small  $\delta$  at least one of these points has smaller absolute value than  $-c$ , thereby obtaining a contradiction.

7. Let  $f(z) = (z - z_1)^{m_1} \cdots (z - z_k)^{m_k}$ .

- Show that  $f'(z) = (z - z_1)^{m_1} \cdots (z - z_k)^{m_k} \cdot \sum_{\alpha=1}^k m_\alpha (z - z_\alpha)^{-1}$ .
- Let  $g(z) = \sum_{\alpha=1}^k m_\alpha (z - z_\alpha)^{-1}$ . Show that if  $g(z) = 0$ , then  $z_1, \dots, z_k$  cannot all lie on the same side of a straight line through  $z$ . Hint: Use Problem 25-11.
- A subset  $K$  of the plane is **convex** if  $K$  contains the line segment joining any two points in it (Figure 8). For any set  $A$ , there is a smallest convex set containing it, which is called the **convex hull** of  $A$  (Figure 9); if a

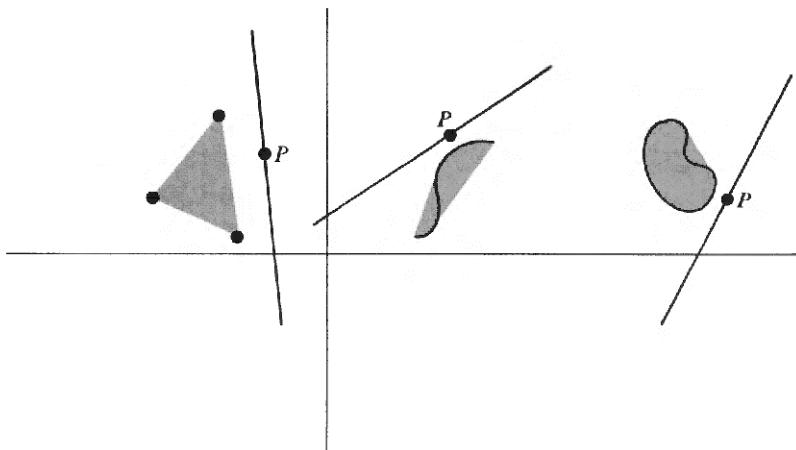


FIGURE 9

point  $P$  is not in the convex hull of  $A$ , then all of  $A$  is contained on one side of some straight line through  $P$ . Using this information, prove that the roots of  $f'(z) = 0$  lie within the convex hull of the set  $\{z_1, \dots, z_k\}$ . Further information on convex sets will be found in reference [19] of the Suggested Reading.

8. Prove that if  $f$  is differentiable at  $z$ , then  $f$  is continuous at  $z$ .
- \*9. Suppose that  $f = u + iv$  where  $u$  and  $v$  are real-valued functions.
  - (a) For fixed  $y_0$  let  $g(x) = u(x + iy_0)$  and  $h(x) = v(x + iy_0)$ . Show that if  $f'(x_0 + iy_0) = \alpha + i\beta$  for real  $\alpha$  and  $\beta$ , then  $g'(x_0) = \alpha$  and  $h'(x_0) = \beta$ .
  - (b) On the other hand, suppose that  $k(y) = u(x_0 + iy)$  and  $l(y) = v(x_0 + iy)$ . Show that  $l'(y_0) = \alpha$  and  $k'(y_0) = -\beta$ .
  - (c) Suppose that  $f'(z) = 0$  for all  $z$ . Show that  $f$  is a constant function.
10. (a) Using the expression
 
$$f(x) = \frac{1}{1+x^2} = \frac{1}{2i} \left( \frac{1}{x-i} - \frac{1}{x+i} \right),$$
 find  $f^{(k)}(x)$  for all  $k$ .
 (b) Use this result to find  $\arctan^{(k)}(0)$  for all  $k$ .

# CHAPTER 27

# COMPLEX POWER SERIES

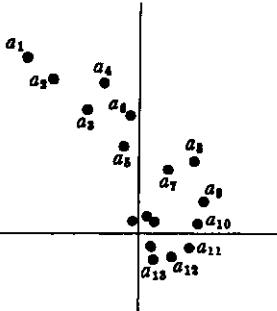


FIGURE 1

If you have not already guessed where differentiable complex functions are going to come from, the title of this chapter should give the secret away: we intend to define functions by means of infinite series. This will necessitate a discussion of infinite sequences of complex numbers, and sums of such sequences, but (as was the case with limits and continuity) the basic definitions are almost exactly the same as for real sequences and series.

An **infinite sequence** of complex numbers is, formally, a complex-valued function whose domain is  $\mathbb{N}$ ; the convenient subscript notation for sequences of real numbers will also be used for sequences of complex numbers. A sequence  $\{a_n\}$  of complex numbers is most conveniently pictured by labeling the points  $a_n$  in the plane (Figure 1).

The sequence shown in Figure 1 converges to 0, “convergence” of complex sequences being defined precisely as for real sequences: the sequence  $\{a_n\}$  **converges** to  $l$ , in symbols

$$\lim_{n \rightarrow \infty} a_n = l,$$

if for every  $\varepsilon > 0$  there is a natural number  $N$  such that, for all  $n$ ,

$$\text{if } n > N, \text{ then } |a_n - l| < \varepsilon.$$

This condition means that any circle drawn around  $l$  will contain  $a_n$  for all sufficiently large  $n$  (Figure 2); expressed more colloquially, the sequence is eventually inside any circle drawn around  $l$ .

Convergence of complex sequences is not only defined precisely as for real sequences, but can even be reduced to this familiar case.

**THEOREM 1** Let

$$a_n = b_n + i c_n \quad \text{for real } b_n \text{ and } c_n,$$

and let

$$l = \beta + i \gamma \quad \text{for real } \beta \text{ and } \gamma.$$

Then  $\lim_{n \rightarrow \infty} a_n = l$  if and only if

$$\lim_{n \rightarrow \infty} b_n = \beta \quad \text{and} \quad \lim_{n \rightarrow \infty} c_n = \gamma.$$

**PROOF** The proof is left as an easy exercise. If there is any doubt as to how to proceed, consult the similar Theorem 1 of Chapter 26. ■

The **sum** of a sequence  $\{a_n\}$  is defined, once again, as  $\lim_{n \rightarrow \infty} s_n$ , where

$$s_n = a_1 + \cdots + a_n.$$

Sequences for which this limit exists are **summable**; alternatively, we may say that the infinite series  $\sum_{n=1}^{\infty} a_n$  **converges** if this limit exists, and **diverges** otherwise. It is unnecessary to develop any new tests for convergence of infinite series, because of the following theorem.

**THEOREM 2** Let

$$a_n = b_n + i c_n \quad \text{for real } b_n \text{ and } c_n.$$

Then  $\sum_{n=1}^{\infty} a_n$  converges if and only if  $\sum_{n=1}^{\infty} b_n$  and  $\sum_{n=1}^{\infty} c_n$  both converge, and in this case

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n + i \left( \sum_{n=1}^{\infty} c_n \right).$$

**PROOF** This is an immediate consequence of Theorem 1 applied to the sequence of partial sums of  $\{a_n\}$ . ■

There is also a notion of absolute convergence for complex series: the series  $\sum_{n=1}^{\infty} a_n$  **converges absolutely** if the series  $\sum_{n=1}^{\infty} |a_n|$  converges (this is a series of real numbers, and consequently one to which our earlier tests may be applied). The following theorem is not quite so easy as the preceding two.

**THEOREM 3** Let

$$a_n = b_n + i c_n \quad \text{for real } b_n \text{ and } c_n.$$

Then  $\sum_{n=1}^{\infty} a_n$  converges absolutely if and only if  $\sum_{n=1}^{\infty} b_n$  and  $\sum_{n=1}^{\infty} c_n$  both converge absolutely.

**PROOF** Suppose first that  $\sum_{n=1}^{\infty} b_n$  and  $\sum_{n=1}^{\infty} c_n$  both converge absolutely, i.e., that  $\sum_{n=1}^{\infty} |b_n|$  and  $\sum_{n=1}^{\infty} |c_n|$  both converge. It follows that  $\sum_{n=1}^{\infty} |b_n| + |c_n|$  converges. Now,

$$|a_n| = |b_n + i c_n| \leq |b_n| + |c_n|.$$

It follows from the comparison test that  $\sum_{n=1}^{\infty} |a_n|$  converges (the numbers  $|a_n|$  and  $|b_n| + |c_n|$  are real and nonnegative). Thus  $\sum_{n=1}^{\infty} a_n$  converges absolutely.

Now suppose that  $\sum_{n=1}^{\infty} |a_n|$  converges. Since

$$|a_n| = \sqrt{b_n^2 + c_n^2},$$

it is clear that

$$|b_n| \leq |a_n| \quad \text{and} \quad |c_n| \leq |a_n|.$$

Once again, the comparison test shows that  $\sum_{n=1}^{\infty} |b_n|$  and  $\sum_{n=1}^{\infty} |c_n|$  converge. ■

Two consequences of Theorem 3 are particularly noteworthy. If  $\sum_{n=1}^{\infty} a_n$  converges absolutely, then  $\sum_{n=1}^{\infty} b_n$  and  $\sum_{n=1}^{\infty} c_n$  also converge absolutely; consequently  $\sum_{n=1}^{\infty} b_n$  and  $\sum_{n=1}^{\infty} c_n$  converge, by Theorem 23-5, so  $\sum_{n=1}^{\infty} a_n$  converges by Theorem 2. In other words, absolute convergence implies convergence. Similar reasoning shows that any rearrangement of an absolutely convergent series has the same sum. These facts can also be proved directly, without using the corresponding theorems for real numbers, by first establishing an analogue of the Cauchy criterion (see Problem 13).

With these preliminaries safely disposed of, we can now consider **complex power series**, that is, functions of the form

$$f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n = a_0 + a_1(z-a) + a_2(z-a)^2 + \dots$$

Here the numbers  $a$  and  $a_n$  are allowed to be complex, and we are naturally interested in the behavior of  $f$  for complex  $z$ . As in the real case, we shall usually consider power series centered at 0,

$$f(z) = \sum_{n=0}^{\infty} a_n z^n;$$

in this case, if  $f(z_0)$  converges, then  $f(z)$  will also converge for  $|z| < |z_0|$ . The proof of this fact will be similar to the proof of Theorem 24-6, but, for reasons that will soon become clear, we will not use all the paraphernalia of uniform convergence and the Weierstrass  $M$ -test, even though they have complex analogues. Our next theorem consequently generalizes only a small part of Theorem 24-6.

**THEOREM 4** Suppose that

$$\sum_{n=0}^{\infty} a_n z_0^n = a_0 + a_1 z_0 + a_2 z_0^2 + \dots$$

converges for some  $z_0 \neq 0$ . Then if  $|z| < |z_0|$ , the two series

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

$$\sum_{n=1}^{\infty} n a_n z^{n-1} = a_1 + 2a_2 z + 3a_3 z^2 + \dots$$

both converge absolutely.

**PROOF** As in the proof of Theorem 24-6, we will need only the fact that the set of numbers  $a_n z_0^n$  is bounded: there is a number  $M$  such that

$$|a_n z_0^n| \leq M \quad \text{for all } n.$$

We then have

$$\begin{aligned} |a_n z^n| &= |a_n z_0^n| \cdot \left| \frac{z}{z_0} \right|^n \\ &\leq M \left| \frac{z}{z_0} \right|^n, \end{aligned}$$

and, for  $z \neq 0$ ,

$$\begin{aligned} |n a_n z^{n-1}| &= \frac{1}{|z|} n |a_n z_0^n| \cdot \left| \frac{z}{z_0} \right|^n \\ &\leq \frac{M}{|z|} n \left| \frac{z}{z_0} \right|^n. \end{aligned}$$

Since the series  $\sum_{n=0}^{\infty} |z/z_0|^n$  and  $\sum_{n=1}^{\infty} n |z/z_0|^n$  converge, this shows that both  $\sum_{n=0}^{\infty} a_n z^n$  and  $\sum_{n=1}^{\infty} n a_n z^{n-1}$  converge absolutely (the argument for  $\sum_{n=1}^{\infty} n a_n z^{n-1}$  assumed that  $z \neq 0$ , but this series certainly converges for  $z = 0$  also). ■

Theorem 4 evidently restricts greatly the possibilities for the set

$$\left\{ z : \sum_{n=0}^{\infty} a_n z^n \text{ converges} \right\}.$$

For example, the shaded set  $A$  in Figure 3 cannot be the set of all  $z$  where  $\sum_{n=0}^{\infty} a_n z^n$  converges, since it contains  $z$ , but not the number  $w$  satisfying  $|w| < |z|$ .

It seems quite unlikely that the set of points where a power series converges could be anything except the set of points inside a circle. If we allow “circles of radius 0” (when the power series converges only at 0) and “circles of radius  $\infty$ ” (when the power series converges at all points), then this assertion is true (with one complication which we will soon mention); the proof requires only Theorem 4 and a knack for good organization.

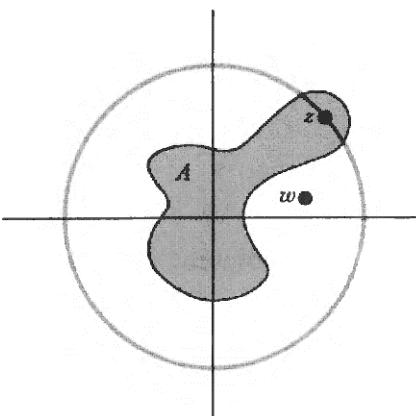


FIGURE 3

**THEOREM 5** For any power series

$$\sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \dots$$

one of the following three possibilities must be true:

(1)  $\sum_{n=0}^{\infty} a_n z^n$  converges only for  $z = 0$ .

(2)  $\sum_{n=0}^{\infty} a_n z^n$  converges absolutely for all  $z$  in  $\mathbf{C}$ .

(3) There is a number  $R > 0$  such that  $\sum_{n=0}^{\infty} a_n z^n$  converges absolutely if  $|z| < R$  and diverges if  $|z| > R$ . (Notice that we do not mention what happens when  $|z| = R$ .)

**PROOF** Let

$$S = \left\{ x \text{ in } \mathbf{R} : \sum_{n=0}^{\infty} a_n w^n \text{ converges for some } w \text{ with } |w| = x \right\}.$$

Suppose first that  $S$  is unbounded. Then for any complex number  $z$ , there is a number  $x$  in  $S$  such that  $|z| < x$ . By definition of  $S$ , this means that  $\sum_{n=0}^{\infty} a_n w^n$  converges for some  $w$  with  $|w| = x > |z|$ . It follows from Theorem 4 that  $\sum_{n=0}^{\infty} a_n z^n$  converges absolutely. Thus, in this case possibility (2) is true.

Now suppose that  $S$  is bounded, and let  $R$  be the least upper bound of  $S$ . If  $R = 0$ , then  $\sum_{n=0}^{\infty} a_n z^n$  converges only for  $z = 0$ , so possibility (1) is true. Suppose, on the other hand, that  $R > 0$ . Then if  $z$  is a complex number with  $|z| < R$ , there is a number  $x$  in  $S$  with  $|z| < x$ . Once again, this means that  $\sum_{n=0}^{\infty} a_n w^n$  converges for some  $w$  with  $|z| < |w|$ , so that  $\sum_{n=0}^{\infty} a_n z^n$  converges absolutely. Moreover, if  $|z| > R$ , then  $\sum_{n=0}^{\infty} a_n z^n$  does not converge, since  $|z|$  is not in  $S$ . ■

The number  $R$  which occurs in case (3) is called the **radius of convergence** of  $\sum_{n=0}^{\infty} a_n z^n$ . In cases (1) and (2) it is customary to say that the radius of convergence is 0 and  $\infty$ , respectively. When  $0 < R < \infty$ , the circle  $\{z : |z| = R\}$  is called the **circle of convergence** of  $\sum_{n=0}^{\infty} a_n z^n$ . If  $z$  is outside the circle, then, of course,

$\sum_{n=0}^{\infty} a_n z^n$  does not converge, but actually a much stronger statement can be made: the terms  $a_n z^n$  are not even bounded. To prove this, let  $w$  be any number with  $|z| > |w| > R$ ; if the terms  $a_n z^n$  were bounded, then the proof of Theorem 4 would show that  $\sum_{n=0}^{\infty} a_n w^n$  converges, which is false. Thus (Figure 4), inside the circle of

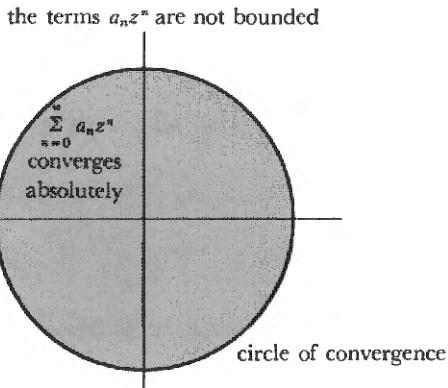


FIGURE 4

convergence the series  $\sum_{n=0}^{\infty} a_n z^n$  converges in the best possible way (absolutely) and outside the circle the series diverges in the worst possible way (the terms  $a_n z^n$  are not bounded).

What happens *on* the circle of convergence is a much more difficult question. We will not consider that question at all, except to mention that there are power series which converge everywhere on the circle of convergence, power series which converge nowhere on the circle of convergence, and power series that do just about anything in between. (See Problem 5.)

Algebraic manipulations on complex power series can be justified just as in the real case. Thus, if  $f(z) = \sum_{n=0}^{\infty} a_n z^n$  and  $g(z) = \sum_{n=0}^{\infty} b_n z^n$  both have radius of convergence  $\geq R$ , then  $h(z) = \sum_{n=0}^{\infty} (a_n + b_n) z^n$  also has radius of convergence  $\geq R$  and  $h = f + g$  inside the circle of radius  $R$ . Similarly, the Cauchy product  $h(z) = \sum_{n=0}^{\infty} c_n z^n$ , for  $c_n = \sum_{k=0}^n a_k b_{n-k}$ , has radius of convergence  $\geq R$  and  $h = fg$  inside the circle of radius  $R$ . And if  $f(z) = \sum_{n=0}^{\infty} a_n z^n$  has radius of convergence  $> 0$  and  $a_0 \neq 0$ , then we can find a power series  $\sum_{n=0}^{\infty} b_n z^n$  with radius of convergence  $> 0$  which represents  $1/f$  inside its circle of convergence.

But our real goal in this chapter is to produce differentiable functions. We therefore want to generalize the result proved for real power series in Chapter 24, that a function defined by a power series can be differentiated term-by-term inside the circle of convergence. At this point we can no longer imitate the proof of Chapter 24, even if we were willing to introduce uniform convergence, because no analogue of Theorem 24-3 seems available. Instead we will use a direct argument (which could also have been used in Chapter 24). Before beginning the proof, we notice that at least there is no problem about the convergence of the series produced by term-by-term differentiation. If the series  $\sum_{n=0}^{\infty} a_n z^n$  has radius of convergence  $R$ , then Theorem 4 immediately implies that the series  $\sum_{n=1}^{\infty} n a_n z^{n-1}$  also converges for  $|z| < R$ . Moreover, if  $|z| > R$ , so that the terms  $a_n z^n$  are unbounded,

then the terms  $na_n z^{n-1}$  are surely unbounded, so  $\sum_{n=1}^{\infty} na_n z^{n-1}$  does not converge.

This shows that the radius of convergence of  $\sum_{n=1}^{\infty} na_n z^{n-1}$  is also exactly  $R$ .

**THEOREM 6** If the power series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

has radius of convergence  $R > 0$ , then  $f$  is differentiable at  $z$  for all  $z$  with  $|z| < R$ , and

$$f'(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}.$$

**PROOF** We will use another “ $\varepsilon/3$  argument.” The fact that the theorem is clearly true for polynomial functions suggests writing

$$\begin{aligned} (*) \quad & \left| \frac{f(z+h) - f(z)}{h} - \sum_{n=1}^{\infty} n a_n z^{n-1} \right| = \left| \sum_{n=0}^{\infty} a_n \frac{((z+h)^n - z^n)}{h} - \sum_{n=1}^{\infty} n a_n z^{n-1} \right| \\ & \leq \left| \sum_{n=0}^{\infty} a_n \frac{((z+h)^n - z^n)}{h} - \sum_{n=0}^N a_n \frac{((z+h)^n - z^n)}{h} \right| \\ & \quad + \left| \sum_{n=0}^N a_n \frac{((z+h)^n - z^n)}{h} - \sum_{n=1}^N n a_n z^{n-1} \right| \\ & \quad + \left| \sum_{n=1}^N n a_n z^{n-1} - \sum_{n=1}^{\infty} n a_n z^{n-1} \right|. \end{aligned}$$

We will show that for any  $\varepsilon > 0$ , each absolute value on the right side can be made  $< \varepsilon/3$  by choosing  $N$  sufficiently large and  $h$  sufficiently small. This will clearly prove the theorem.

Only the first term in the right side of  $(*)$  will present any difficulties. To begin with, choose some  $z_0$  with  $|z| < |z_0| < R$ ; henceforth we will consider only  $h$  with  $|z+h| \leq |z_0|$ . The expression  $((z+h)^n - z^n)/h$  can be written in a more convenient way if we remember that

$$\frac{x^n - y^n}{x - y} = x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + y^{n-1}.$$

Applying this to

$$\frac{(z+h)^n - z^n}{h} = \frac{(z+h)^n - z^n}{(z+h) - z},$$

we obtain

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

Since

$$|(z+h)^{n-1} + z(z+h)^{n-2} + \cdots + z^{n-1}| \leq n|z_0|^{n-1},$$

we have

$$\left| a_n \frac{((z+h)^n - z^n)}{h} \right| \leq n|a_n| \cdot |z_0|^{n-1}.$$

But the series  $\sum_{n=1}^{\infty} n|a_n| \cdot |z_0|^{n-1}$  converges, so if  $N$  is sufficiently large, then

$$\sum_{n=N+1}^{\infty} n|a_n| \cdot |z_0|^{n-1} < \frac{\varepsilon}{3}.$$

This means that

$$\begin{aligned} & \left| \sum_{n=0}^{\infty} a_n \frac{((z+h)^n - z^n)}{h} - \sum_{n=0}^N a_n \frac{((z+h)^n - z^n)}{h} \right| \\ &= \left| \sum_{n=N+1}^{\infty} a_n \frac{((z+h)^n - z^n)}{h} \right| \leq \sum_{n=N+1}^{\infty} \left| a_n \frac{((z+h)^n - z^n)}{h} \right| \\ &\leq \sum_{n=N+1}^{\infty} n|a_n| \cdot |z_0|^{n-1} < \frac{\varepsilon}{3}. \end{aligned}$$

In short, if  $N$  is sufficiently large, then

$$(1) \quad \left| \sum_{n=0}^{\infty} a_n \frac{((z+h)^n - z^n)}{h} - \sum_{n=0}^N a_n \frac{((z+h)^n - z^n)}{h} \right| < \frac{\varepsilon}{3},$$

for all  $h$  with  $|z+h| \leq |z_0|$ .

It is easy to deal with the third term on the right side of (\*): Since  $\sum_{n=1}^{\infty} na_n z^{n-1}$  converges, it follows that if  $N$  is sufficiently large, then

$$(2) \quad \left| \sum_{n=1}^{\infty} na_n z^{n-1} - \sum_{n=1}^N na_n z^{n-1} \right| < \frac{\varepsilon}{3}.$$

Finally, choosing an  $N$  such that (1) and (2) are true, we note that

$$\lim_{h \rightarrow 0} \sum_{n=0}^N a_n \frac{((z+h)^n - z^n)}{h} = \sum_{n=1}^N na_n z^{n-1},$$

since the polynomial function  $g(z) = \sum_{n=0}^N a_n z^n$  is certainly differentiable. Therefore

$$(3) \quad \left| \sum_{n=0}^N \frac{a_n ((z+h)^n - z^n)}{h} - \sum_{n=1}^N na_n z^{n-1} \right| < \frac{\varepsilon}{3}.$$

for sufficiently small  $h$ .

As we have already indicated, (1), (2), and (3) prove the theorem. ■

Theorem 6 has an obvious corollary: a function represented by a power series is infinitely differentiable inside the circle of convergence, and the power series is its Taylor series at 0. It follows, in particular, that  $f$  is continuous inside the circle of convergence, since a function differentiable at  $z$  is continuous at  $z$  (Problem 26-8).

The continuity of a power series inside its circle of convergence helps explain the behavior of certain Taylor series obtained for real functions, and gives the promised answers to the questions raised at the end of Chapter 24. We have already seen that the Taylor series for the function  $f(z) = 1/(1+z^2)$ , namely,

$$1 - z^2 + z^4 - z^6 + \dots,$$

converges for real  $z$  only when  $|z| < 1$ , and consequently has radius of convergence 1. It is no accident that the circle of convergence contains the two points  $i$  and  $-i$  at which  $f$  is undefined. If this power series converged in a circle of radius greater than 1, then (Figure 5) it would represent a function which was continuous in that circle, in particular at  $i$  and  $-i$ . But this is impossible, since it equals  $1/(1+z^2)$  inside the unit circle, and  $1/(1+z^2)$  does not approach a limit as  $z$  approaches  $i$  or  $-i$  from inside the unit circle.

The use of complex numbers also sheds some light on the strange behavior of the Taylor series for the function

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Although we have not yet defined  $e^z$  for complex  $z$ , it will presumably be true that if  $y$  is real and unequal to 0, then

$$f(iy) = e^{-1/(iy)^2} = e^{1/y^2}.$$

The interesting fact about this expression is that it becomes large as  $y$  becomes small. Thus  $f$  will not even be continuous at 0 when defined for complex numbers, so it is hardly surprising that it is equal to its Taylor series only for  $z = 0$ .

The method by which we will actually define  $e^z$  (as well as  $\sin z$  and  $\cos z$ ) for complex  $z$  should by now be clear. For real  $x$  we know that

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots, \\ e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots \end{aligned}$$

For complex  $z$  we therefore define

$$\begin{aligned} \sin z &= z - \frac{z^3}{3!} + \frac{z^5}{5!} - \dots, \\ \cos z &= 1 - \frac{z^2}{2!} + \frac{z^4}{4!} + \dots, \\ \exp(z) &= e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \dots \end{aligned}$$

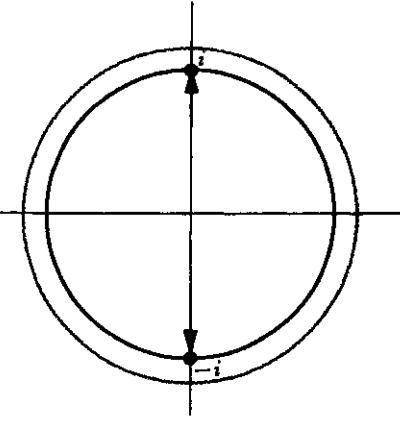


FIGURE 5

Then  $\sin'(z) = \cos z$ ,  $\cos'(z) = -\sin z$ , and  $\exp'(z) = \exp(z)$  by Theorem 6. Moreover, if we replace  $z$  by  $iz$  in the series for  $e^z$ , and make a rearrangement of the terms (justified by absolute convergence), something particularly interesting happens:

$$\begin{aligned} e^{iz} &= 1 + iz + \frac{(iz)^2}{2!} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \frac{(iz)^5}{5!} + \dots \\ &= 1 + iz - \frac{z^2}{2!} - \frac{iz^3}{3!} + \frac{iz^4}{4!} + \frac{iz^5}{5!} + \dots \\ &= \left(1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \dots\right) + i \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} + \dots\right), \end{aligned}$$

so

$$e^{iz} = \cos z + i \sin z.$$

It is clear from the definitions (i.e., the power series) that

$$\begin{aligned} \sin(-z) &= -\sin z, \\ \cos(-z) &= \cos z, \end{aligned}$$

so we also have

$$e^{-iz} = \cos z - i \sin z.$$

From the equations for  $e^{iz}$  and  $e^{-iz}$  we can derive the formulas

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i},$$

$$\cos z = \frac{e^{iz} + e^{-iz}}{2}.$$

The development of complex power series thus places the exponential function at the very core of the development of the elementary functions—it reveals a connection between the trigonometric and exponential functions which was never imagined when these functions were first defined, and which could never have been discovered without the use of complex numbers. As a by-product of this relationship, we obtain a hitherto unsuspected connection between the numbers  $e$  and  $\pi$ : if in the formula

$$e^{iz} = \cos z + i \sin z$$

we take  $z = \pi$ , we obtain the remarkable result

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

(More generally,  $e^{2\pi i/n}$  is an  $n$ th root of 1.)

With these remarks we will bring to a close our investigation of complex functions. And yet there are still several basic facts about power series which have not been mentioned. Thus far, we have seldom considered power series centered at  $a$ ,

$$f(z) = \sum_{n=0}^{\infty} a_n(z - a)^n,$$

except for  $a = 0$ . This omission was adopted partly to simplify the exposition. For power series centered at  $a$  there are obvious versions of all the theorems in this chapter (the proofs require only trivial modifications): there is a number  $R$  (possibly 0 or “ $\infty$ ”) such that the series  $\sum_{n=0}^{\infty} a_n(z - a)^n$  converges absolutely for  $z$  with  $|z - a| < R$ , and has unbounded terms for  $z$  with  $|z - a| > R$ ; moreover, for all  $z$  with  $|z - a| < R$  the function

$$f(z) = \sum_{n=0}^{\infty} a_n(z - a)^n$$

has derivative

$$f'(z) = \sum_{n=1}^{\infty} n a_n (z - a)^{n-1}.$$

It is less straightforward to investigate the possibility of representing a function as a power series centered at  $b$ , if it is already written as a power series centered at  $a$ . If

$$f(z) = \sum_{n=0}^{\infty} a_n(z - a)^n$$

has radius of convergence  $R$ , and  $b$  is a point with  $|b - a| < R$  (Figure 6), then it is true that  $f(z)$  can also be written as a power series centered at  $b$ ,

$$f(z) = \sum_{n=0}^{\infty} b_n(z - b)^n$$

(the numbers  $b_n$  are necessarily  $f^{(n)}(b)/n!$ ); moreover, this series has radius of convergence at least  $R - |b - a|$  (*it may be larger*).

We will *not* prove the facts mentioned in the previous paragraph, and there are several other important facts we shall not prove. For example, if

$$f(z) = \sum_{n=0}^{\infty} a_n(z - a)^n \quad \text{and} \quad g(z) = \sum_{n=0}^{\infty} b_n(z - b)^n,$$

and  $g(b) = a$ , then we would expect that  $f \circ g$  can be written as a power series centered at  $b$ . All such facts could be proved now without introducing any basic new ideas, but the proofs would not be as easy as the proofs about sums, products and reciprocals of power series. The possibility of changing a power series centered at  $a$  into one centered at  $b$  is quite a bit more involved, and the treatment of  $f \circ g$  requires still more skill. Rather than end this section with a *tour de force* of computations, we will instead give a preview of “complex analysis,” one of the most beautiful branches of mathematics, where all these facts are derived as straightforward consequences of some fundamental results.

Power series were introduced in this chapter in order to provide complex functions which are differentiable. Since these functions are actually infinitely differentiable, it is natural to suppose that we have therefore selected only a very special

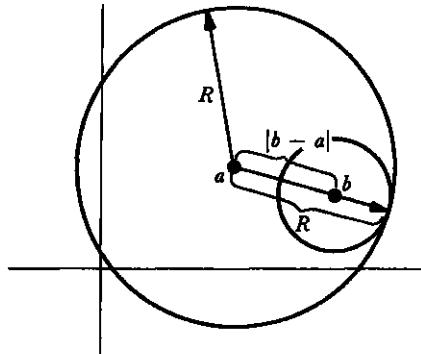


FIGURE 6

collection of differentiable complex functions. The basic theorems of complex analysis show that this is not at all true:

*If a complex function is defined in some region  $A$  of the plane and is differentiable in  $A$ , then it is automatically infinitely differentiable in  $A$ . Moreover, for each point  $a$  in  $A$  the Taylor series for  $f$  at  $a$  will converge to  $f$  in any circle contained in  $A$  (Figure 7).*

These facts are among the first to be proved in complex analysis. It is impossible to give any idea of the proofs themselves—the methods used are quite different from anything in elementary calculus. If these facts are granted, however, then the facts mentioned before can be proved very easily.

Suppose, for example, that  $f$  and  $g$  are functions which can be written as power series. Then, as we have shown,  $f$  and  $g$  are differentiable—it then follows from easy general theorems that  $f + g$ ,  $f \cdot g$ ,  $1/g$  and  $f \circ g$  are also differentiable. Appealing to the results from complex analysis, it follows that they can be written as power series.

We already know how to compute the power series for  $f + g$ ,  $f \cdot g$  and  $1/g$  from those for  $f$  and  $g$ . It is also easy to guess how one would compute an expression for  $f \circ g$  as a power series in  $(z - b)$  when we are given the power series expansions

$$\begin{aligned} f(z) &= \sum_{n=0}^{\infty} a_n(z - a)^n \\ g(z) &= \sum_{k=0}^{\infty} b_k(z - b)^k, \end{aligned}$$

with  $a = g(b) = b_0$ , so that

$$g(z) - a = \sum_{k=1}^{\infty} b_k(z - b)^k.$$

First of all, we know how to compute the power series

$$(g(z) - a)^l = \left( \sum_{k=1}^{\infty} b_k(z - b)^k \right)^l,$$

and this power series will begin with  $(z - b)^l$ . Consequently, the coefficient of  $z^n$  in

$$f(g(z)) = \sum_{l=0}^{\infty} a_l(g(z) - a)^l$$

can be calculated as a finite sum, involving only coefficients arising from the first  $n$  powers of  $g(z) - a$ .

Similarly, if

$$f(z) = \sum_{n=0}^{\infty} a_n(z - a)^n$$

has radius of convergence  $R$ , then  $f$  is differentiable in the region  $A = \{z : |z - a| < R\}$ . Thus, if  $b$  is in  $A$ , it is possible to write  $f$  as a power series centered at  $b$ ,

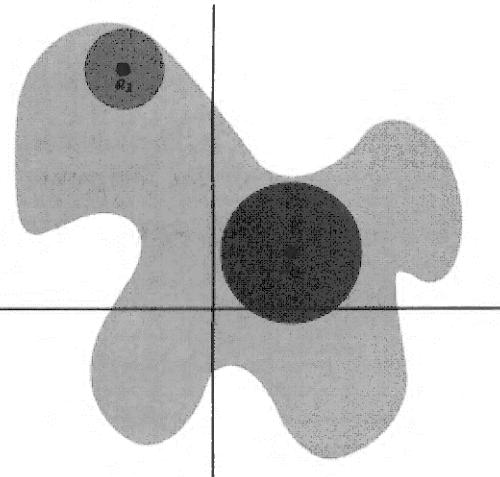


FIGURE 7

which will converge in the circle of radius  $R = |b - a|$ . The coefficient of  $z^n$  will be  $f^{(n)}(b)/n!$ . This series may actually converge in a larger circle, because  $\sum_{n=0}^{\infty} a_n(z - a)^n$  may be the series for a function differentiable in a larger region than  $A$ . For example, suppose that  $f(z) = 1/(1 + z^2)$ . Then  $f$  is differentiable, except at  $i$  and  $-i$ , where it is not defined. Thus  $f(z)$  can be written as a power series  $\sum_{n=0}^{\infty} a_n z^n$  with radius of convergence 1 (as a matter of fact, we know that  $a_{2n} = (-1)^n$  and  $a_k = 0$  if  $k$  is odd). It is also possible to write

$$f(z) = \sum_{n=0}^{\infty} b_n (z - \frac{1}{2})^n,$$

where  $b_n = f^{(n)}(\frac{1}{2})/n!$ . We can easily predict the radius of convergence of this series: it is  $\sqrt{1 + (\frac{1}{2})^2}$ , the distance from  $\frac{1}{2}$  to  $i$  or  $-i$  (Figure 8).

As an added incentive to investigate complex analysis further, one more result will be mentioned, which lies quite near the surface, and which will be found in any treatment of the subject.

For real  $z$  the values of  $\sin z$  always lie between  $-1$  and  $1$ , but for complex  $z$  this is not at all true. In fact, if  $z = iy$ , for  $y$  real, then

$$\sin iy = \frac{e^{i(iy)} - e^{-i(iy)}}{2i} = \frac{e^{-y} - e^y}{2i}.$$

If  $y$  is large, then  $\sin iy$  is also large in absolute value. This behavior of  $\sin$  is typical of functions which are defined and differentiable on the whole complex plane (such functions are called *entire*). A result which comes quite early in complex analysis is the following:

*Liouville's Theorem: The only bounded entire functions are the constant functions.*

As a simple application of Liouville's Theorem, consider a polynomial function

$$f(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0,$$

where  $n > 1$ , so that  $f$  is not a constant. We already know that  $f(z)$  is large for large  $z$ , so Liouville's Theorem tells us nothing interesting about  $f$ . But consider the function

$$g(z) = \frac{1}{f(z)}.$$

If  $f(z)$  were never 0, then  $g$  would be entire; since  $f(z)$  becomes large for large  $z$ , the function  $g$  would also be bounded, contradicting Liouville's Theorem. Thus  $f(z) = 0$  for some  $z$ , and we have proved the Fundamental Theorem of Algebra.

## PROBLEMS

- Decide whether each of the following series converges, and whether it converges absolutely.

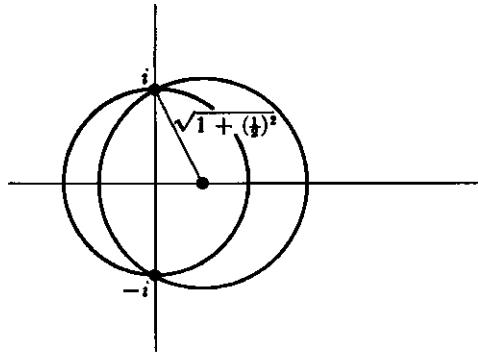


FIGURE 8

(i)  $\sum_{n=1}^{\infty} \frac{(1+i)^n}{n!}.$

(ii)  $\sum_{n=1}^{\infty} \frac{1+2i}{2^n}.$

(iii)  $\sum_{n=1}^{\infty} \frac{i^n}{n}.$

(iv)  $\sum_{n=1}^{\infty} \left(\frac{1}{2} + \frac{1}{2}i\right)^n.$

(v)  $\sum_{n=2}^{\infty} \frac{\log n}{n} + i^n \frac{\log n}{n}.$

2. Use the ratio test to show that the radius of convergence of each of the following power series is 1. (In each case the ratios of successive terms will approach a limit  $< 1$  if  $|z| < 1$ , but for  $|z| > 1$  the ratios will tend to  $\infty$  or to a limit  $> 1$ .)

(i)  $\sum_{n=1}^{\infty} \frac{z^n}{n^2}.$

(ii)  $\sum_{n=1}^{\infty} \frac{z^n}{n}.$

(iii)  $\sum_{n=1}^{\infty} z^n.$

(iv)  $\sum_{n=1}^{\infty} (n + 2^{-n}) z^n.$

(v)  $\sum_{n=1}^{\infty} 2^n z^{n!}.$

3. Use the root test (Problem 23-7) to find the radius of convergence of each of the following power series.

(i)  $\frac{z}{2} + \frac{z^2}{3} + \frac{z^3}{2^2} + \frac{z^4}{3^2} + \frac{z^5}{2^3} + \frac{z^6}{3^3} + \dots$

(ii)  $\sum_{n=1}^{\infty} \frac{n! z^n}{n^n}.$

(iii)  $\sum_{n=1}^{\infty} \frac{n}{2^n} z^n.$

(iv)  $\sum_{n=1}^{\infty} \frac{n^2}{2^n} z^n.$

(v)  $\sum_{n=1}^{\infty} 2^n z^{n!}.$

4. The root test can always be used, in theory at least, to find the radius of convergence of a power series; in fact, a close analysis of the situation leads to a formula for the radius of convergence, known as the “Cauchy-Hadamard formula.” Suppose first that the set of numbers  $\sqrt[n]{|a_n|}$  is bounded.
- Use Problem 23-7 to show that if  $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{|a_n|} |z| < 1$ , then  $\sum_{n=0}^{\infty} a_n z^n$  converges.
  - Also show that if  $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{|a_n|} |z| > 1$ , then  $\sum_{n=0}^{\infty} a_n z^n$  has unbounded terms.
  - Parts (a) and (b) show that the radius of convergence of  $\sum_{n=0}^{\infty} a_n z^n$  is  $1/\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{|a_n|}$  (where “1/0” means “ $\infty$ ”). To complete the formula, define  $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{|a_n|} = \infty$  if the set of all  $\sqrt[n]{|a_n|}$  is unbounded. Prove that in this case,  $\sum_{n=0}^{\infty} a_n z^n$  diverges for  $z \neq 0$ , so that the radius of convergence is 0 (which may be considered as “1/ $\infty$ ”).
5. Consider the following three series from Problem 2:
- $$\sum_{n=1}^{\infty} \frac{z^n}{n^2}, \quad \sum_{n=1}^{\infty} \frac{z^n}{n}, \quad \sum_{n=1}^{\infty} z^n.$$
- Prove that the first series converges everywhere on the unit circle; that the third series converges nowhere on the unit circle; and that the second series converges for at least one point on the unit circle and diverges for at least one point on the unit circle.
- (a) Prove that  $e^z \cdot e^w = e^{z+w}$  for all complex numbers  $z$  and  $w$  by showing that the infinite series for  $e^{z+w}$  is the Cauchy product of the series for  $e^z$  and  $e^w$ .  
(b) Show that  $\sin(z + w) = \sin z \cos w + \cos z \sin w$  and  $\cos(z + w) = \cos z \cos w - \sin z \sin w$  for all complex  $z$  and  $w$ .
  - (a) Prove that every complex number of absolute value 1 can be written  $e^{iy}$  for some real number  $y$ .  
(b) Prove that  $|e^{x+iy}| = e^x$  for real  $x$  and  $y$ .
  - (a) Prove that  $\exp$  takes on every complex value except 0.  
(b) Prove that  $\sin$  takes on every complex value.
  - For each of the following functions, compute the first three nonzero terms of the Taylor series centered at 0 by manipulating power series.
    - $f(z) = \tan z$ .
    - $f(z) = z(1-z)^{-1/2}$ .

- (iii)  $f(z) = \frac{e^{\sin z} - 1}{z}$ .
- (iv)  $f(z) = \log(1 - z^2)$ .
- (v)  $f(z) = \frac{\sin^2 z}{z^2}$ .
- (vi)  $f(z) = \frac{\sin(z^2)}{z \cos^2 z}$ .
- (vii)  $f(z) = \frac{1}{z^4 - 2z^2 + 3}$ .
- (viii)  $f(z) = \frac{1}{z} [e^{(\sqrt{1+z}-1)} - 1]$ .

10. (a) Suppose that we write a differentiable complex function  $f$  as  $f = u + iv$ , where  $u$  and  $v$  are real-valued. Let  $\bar{u}$  and  $\bar{v}$  denote the restrictions of  $u$  and  $v$  to the real numbers. In other words,  $\bar{u}(x) = u(x)$  for real numbers  $x$  (but  $\bar{u}$  is not defined for other  $x$ ). Using Problem 26-9, show that for real  $x$  we have

$$f'(x) = \bar{u}'(x) + i\bar{v}'(x),$$

where  $f'$  denotes the complex derivative, while  $\bar{u}'$  and  $\bar{v}'$  denote the ordinary derivatives of these real-valued functions on  $\mathbf{R}$ .

- (b) Show, more generally, that

$$f^{(k)}(x) = \bar{u}^{(k)}(x) + i\bar{v}^{(k)}(x).$$

- (c) Suppose that  $f$  satisfies the equation

$$(*) \quad f^{(n)} + a_{n-1}f^{(n-1)} + \cdots + a_0f = 0,$$

where the  $a_i$  are real numbers, and where the  $f^{(k)}$  denote higher-order complex derivatives. Show that  $\bar{u}$  and  $\bar{v}$  satisfy the same equation, where  $\bar{u}^{(k)}$  and  $\bar{v}^{(k)}$  now denote higher-order derivatives of real-valued functions on  $\mathbf{R}$ .

- (d) Show that if  $a = b+ci$  is a complex root of the equation  $z^n + a_{n-1}z^{n-1} + \cdots + a_0 = 0$ , then  $f(x) = e^{bx} \sin cx$  and  $f(x) = e^{bx} \cos cx$  are both solutions of  $(*)$ .
11. (a) Show that  $\exp$  is *not* one-one on  $\mathbf{C}$ .
- (b) Given  $w \neq 0$ , show that  $e^z = w$  if and only if  $z = x+iy$  with  $x = \log|w|$  (here  $\log$  denotes the real logarithm function), and  $y$  an argument of  $w$ .
- \*(c) Show that there does not exist a continuous function  $\log$  defined for nonzero complex numbers, such that  $\exp(\log(z)) = z$  for all  $z \neq 0$ . (Show that  $\log$  cannot even be defined continuously for  $|z| = 1$ .)

Since there is no way to define a continuous logarithm function we cannot speak of *the* logarithm of a complex number, but only of “a logarithm for  $w$ ,” meaning one of the infinitely many numbers  $z$  with  $e^z = w$ . And

for complex numbers  $a$  and  $b$  we define  $a^b$  to be a *set* of complex numbers, namely the set of all numbers  $e^{b \log a}$  or, more precisely, the set of all numbers  $e^{bz}$  where  $z$  is a logarithm for  $a$ .

- (d) If  $m$  is an integer, then  $a^m$  consists of only one number, the one given by the usual elementary definition of  $a^m$ .
- (e) If  $m$  and  $n$  are integers, then the set  $a^{m/n}$  coincides with the set of values given by the usual elementary definition, namely the set of all  $b^m$  where  $b$  is an  $n$ th root of  $a$ .
- (f) If  $a$  and  $b$  are real and  $b$  is irrational, then  $a^b$  contains infinitely many members, even for  $a > 0$ .
- (g) Find all logarithms of  $i$ , and find all values of  $i^i$ .
- (h) By  $(a^b)^c$  we mean the set of all numbers of the form  $z^c$  for some number  $z$  in the set  $a^b$ . Show that  $(1^i)^i$  has infinitely many values, while  $1^{i^i}$  has only one.
- (i) Show that all values of  $a^{b \cdot c}$  are also values of  $(a^b)^c$ . Is  $a^{b \cdot c} = (a^b)^c \cap (a^c)^b$ ?

12. (a) For real  $x$  show that we can choose  $\log(x+i)$  and  $\log(x-i)$  to be

$$\log(x+i) = \log(1+x^2) + i\left(\frac{\pi}{2} - \arctan x\right),$$

$$\log(x-i) = \log(1+x^2) - i\left(\frac{\pi}{2} - \arctan x\right).$$

(It will help to note that  $\pi/2 - \arctan x = \arctan 1/x$  for  $x \neq 0$ .)

- (b) The expression

$$\frac{1}{1+x^2} = \frac{1}{2i}\left(\frac{1}{x-i} - \frac{1}{x+i}\right)$$

yields, formally,

$$\int \frac{dx}{1+x^2} = \frac{1}{2i} [\log(x-i) - \log(x+i)].$$

Use part (a) to check that this answer agrees with the usual one.

13. (a) A sequence  $\{a_n\}$  of complex numbers is called a **Cauchy sequence** if  $\lim_{m,n \rightarrow \infty} |a_m - a_n| = 0$ . Suppose that  $a_n = b_n + ic_n$ , where  $b_n$  and  $c_n$  are real. Prove that  $\{a_n\}$  is a Cauchy sequence if and only if  $\{b_n\}$  and  $\{c_n\}$  are Cauchy sequences.
- (b) Prove that every Cauchy sequence of complex numbers converges.
- (c) Give direct proofs, without using theorems about real series, that an absolutely convergent series is convergent and that any rearrangement has the same sum. (It is permitted, and in fact advisable, to use the *proofs* of the corresponding theorems for real series.)

14. (a) Prove that

$$\sum_{k=1}^n e^{ikx} = e^{ix} \frac{1 - e^{inx}}{1 - e^{ix}} = \frac{\sin\left(\frac{n}{2}x\right)}{\sin\frac{x}{2}} e^{i(n+1)x/2}.$$

- (b) Deduce the formulas for  $\sum_{k=1}^n \cos kx$  and  $\sum_{k=1}^n \sin kx$  that are given in Problem 15-33.
15. Let  $\{a_n\}$  be the Fibonacci sequence,  $a_1 = a_2 = 1$ ,  $a_{n+2} = a_n + a_{n+1}$ .
- If  $r_n = a_{n+1}/a_n$ , show that  $r_{n+1} = 1 + 1/r_n$ .
  - Show that  $r = \lim_{n \rightarrow \infty} r_n$  exists, and  $r = 1 + 1/r$ . Conclude that  $r = (1 + \sqrt{5})/2$ .
  - Show that  $\sum_{n=1}^{\infty} a_n z^n$  has radius of convergence  $2/(1 + \sqrt{5})$ . (Using the unproved theorems in this chapter and the fact that  $\sum_{n=1}^{\infty} a_n z^n = -1/(z^2 + z - 1)$  from Problem 24-15 we could have predicted that the radius of convergence is the smallest absolute value of the roots of  $z^2 + z - 1 = 0$ ; since the roots are  $(-1 \pm \sqrt{5})/2$ , the radius of convergence should be  $(-1 + \sqrt{5})/2$ . Notice that this number is indeed equal to  $2/(1 + \sqrt{5})$ .)
16. Since  $(e^z - 1)/z$  can be written as the power series  $1 + z/2! + z^2/3! + \dots$  which is nonzero at 0, it follows that there is a power series

$$\frac{z}{e^z - 1} = \sum_{n=0}^{\infty} \frac{b_n}{n!} z^n$$

with nonzero radius of convergence. Using the unproved theorems in this chapter, we can even predict the radius of convergence; it is  $2\pi$ , since this is the smallest absolute value of the numbers  $z = 2k\pi i$  for which  $e^z - 1 = 0$ . The numbers  $b_n$  appearing here are called the **Bernoulli numbers**.\*

- (a) Clearly  $b_0 = 1$ . Now show that

$$\begin{aligned} \frac{z}{e^z - 1} &= -\frac{z}{2} + \frac{z}{2} \cdot \frac{e^z + 1}{e^z - 1}, \\ \frac{e^{-z} + 1}{e^{-z} - 1} &= -\frac{e^z + 1}{e^z - 1}, \end{aligned}$$

and deduce that

$$b_1 = -\frac{1}{2}, \quad b_n = 0 \quad \text{if } n \text{ is odd and } n > 1.$$

- (b) By finding the coefficient of  $z^n$  in the right side of the equation

$$z = \left( \sum_{k=0}^{\infty} \frac{b_k}{k!} \right) \left( z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots \right),$$

\* Sometimes the numbers  $B_n = (-1)^{n-1} b_{2n}$  are called the Bernoulli numbers, because  $b_n = 0$  if  $n$  is odd and  $> 1$  (see part (a)) and because the numbers  $b_{2n}$  alternate in sign, although we will not prove this. Other modifications of this nomenclature are also in use.

show that

$$\sum_{i=0}^{n-1} \binom{n}{i} b_i = 0 \quad \text{for } n > 1.$$

This formula allows us to compute any  $b_k$  in terms of previous ones, and shows that each is rational. Calculate two or three of the following:

$$b_2 = \frac{1}{6}, \quad b_4 = -\frac{1}{30}, \quad b_6 = \frac{1}{42}, \quad b_8 = -\frac{1}{30}.$$

\*(c) Part (a) shows that

$$\sum_{n=0}^{\infty} \frac{b_{2n}}{(2n)!} z^{2n} = \frac{z}{2} \cdot \frac{e^z + 1}{e^z - 1} = \frac{z}{2} \cdot \frac{e^{z/2} + e^{-z/2}}{e^{z/2} - e^{-z/2}}.$$

Replace  $z$  by  $2iz$  and show that

$$z \cot z = \sum_{n=0}^{\infty} \frac{b_{2n}}{(2n)!} (-1)^n 2^{2n} z^{2n}.$$

\*(d) Show that

$$\tan z = \cot z - 2 \cot 2z.$$

\*(e) Show that

$$\tan z = \sum_{n=1}^{\infty} \frac{b_{2n}}{(2n)!} (-1)^{n-1} 2^{2n} (2^{2n} - 1) z^{2n-1}.$$

(This series converges for  $|z| < \pi/2$ .)

17. The Bernoulli numbers play an important role in a theorem which is best introduced by some notational nonsense. Let us use  $D$  to denote the “differentiation operator,” so that  $Df$  denotes  $f'$ . Then  $D^k f$  will mean  $f^{(k)}$  and  $e^D f$  will mean  $\sum_{n=0}^{\infty} f^{(n)} / n!$  (of course this series makes no sense in general, but it will make sense if  $f$  is a polynomial function, for example). Finally, let  $\Delta$  denote the “difference operator” for which  $\Delta f(x) = f(x+1) - f(x)$ . Now Taylor’s Theorem implies, disregarding questions of convergence, that

$$f(x+1) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!},$$

or

$$(*) \quad f(x+1) - f(x) = \sum_{n=1}^{\infty} \frac{f^{(n)}(x)}{n!};$$

we may write this symbolically as  $\Delta f = (e^D - 1)f$ , where 1 stands for the “identity operator.” Even more symbolically this can be written  $\Delta = e^D - 1$ , which suggests that

$$D = \frac{D}{e^D - 1} \Delta.$$

Thus we obviously ought to have

$$D = \sum_{k=0}^{\infty} \frac{b_k}{k!} D^k \Delta,$$

i.e.,

$$(**) \quad f'(x) = \sum_{k=0}^{\infty} \frac{b_k}{k!} [f^{(k)}(x+1) - f^{(k)}(x)].$$

The beautiful thing about all this nonsense is that it works!

- (a) Prove that  $(**)$  is literally true if  $f$  is a polynomial function (in which case the infinite sum is really a finite sum). Hint: By applying  $(*)$  to  $f^{(k)}$ , find a formula for  $f^{(k)}(x+1) - f^{(k)}(x)$ ; then use the formula in Problem 16(b) to find the coefficient of  $f^{(j)}(x)$  in the right side of  $(**)$ .
- (b) Deduce from  $(**)$  that

$$f'(0) + \cdots + f'(n) = \sum_{k=0}^{\infty} \frac{b_k}{k!} [f^{(k)}(n+1) - f^{(k)}(0)].$$

- (c) Show that for any polynomial function  $g$  we have

$$g(0) + \cdots + g(n) = \int_0^{n+1} g(t) dt + \sum_{k=1}^{\infty} \frac{b_k}{k!} [g^{(k-1)}(n+1) - g^{(k-1)}(0)].$$

- (d) Apply this to  $g(x) = x^p$  to show that

$$\sum_{k=1}^n k^p = \frac{n^{p+1}}{p+1} + \sum_{k=1}^p \frac{b_k}{k} \binom{p}{k-1} n^{p-k+1}.$$

Using the fact that  $b_1 = -\frac{1}{2}$ , show that

$$\sum_{k=1}^n k^p = \frac{n^{p+1}}{p+1} + \frac{n^p}{2} + \sum_{k=2}^p \frac{b_k}{k} \binom{p}{k-1} n^{p-k+1}.$$

The first ten instances of this formula were written out in Problem 2-7, which offered as a challenge the discovery of the general pattern. This may now seem to be a preposterous suggestion, but the Bernoulli numbers were actually discovered in precisely this way! After writing out these 10 formulas, Bernoulli claims (in his posthumously printed work *Ars Conjectandi*, 1713): “Whoever will examine the series as to their regularity may be able to continue the table.” He then writes down the above formula, offering no proof at all, merely noting that the coefficients  $b_k$  (which he denoted simply by  $A, B, C, \dots$ ) satisfy the equation in Problem 16(b). The relation between these numbers and the coefficients in the power series for  $z/(e^z - 1)$  was discovered by Euler.

- \*18.** The formula in Problem 17(c) can be generalized to the case where  $g$  is not a polynomial function; the infinite sum must be replaced by a finite sum plus a remainder term. In order to find an expression for the remainder, it is useful to introduce some new functions.

(a) The *Bernoulli polynomials*  $\varphi_n$  are defined by

$$\varphi_n(x) = \sum_{k=0}^n \binom{n}{k} b_{n-k} x^k.$$

The first three are

$$\begin{aligned}\varphi_1(x) &= x - \frac{1}{2}, \\ \varphi_2(x) &= x^2 - x + \frac{1}{6}, \\ \varphi_3(x) &= x^3 - \frac{3x^2}{2} + \frac{x}{2}.\end{aligned}$$

Show that

$$\begin{aligned}\varphi_n(0) &= b_n, \\ \varphi_n(1) &= b_n \quad \text{if } n > 1, \\ \varphi_n'(x) &= n\varphi_{n-1}(x), \\ \varphi_n(x) &= (-1)^n \varphi_n(1-x) \quad \text{for } n > 1.\end{aligned}$$

Hint: Prove the last equation by induction on  $n$ , starting with  $n = 2$ .

- (b) Let  $R_N{}^k(x)$  be the remainder term in Taylor's Theorem for  $f^{(k)}$ , on the interval  $[x, x+1]$ , so that

$$(*) \quad f^{(k)}(x+1) - f^{(k)}(x) = \sum_{n=1}^N \frac{f^{(k+n)}(x)}{n!} + R_N{}^k(x).$$

Prove that

$$f'(x) = \sum_{k=0}^N \frac{b_k}{k!} [f^{(k)}(x+1) - f^{(k)}(x)] - \sum_{k=0}^N \frac{b_k}{k!} R_{N-k}{}^k(x).$$

Hint: Imitate Problem 17(a). Notice the subscript  $N - k$  on  $R$ .

- (c) Use the integral form of the remainder to show that

$$\sum_{k=0}^N \frac{b_k}{k!} R_{N-k}{}^k(x) = \int_x^{x+1} \frac{\varphi_n(x+1-t)}{N!} f^{(N+1)}(t) dt.$$

- (d) Deduce the “Euler-Maclaurin Summation Formula”:

$$\begin{aligned}g(x) + g(x+1) + \cdots + g(x+n) \\ = \int_x^{x+n+1} g(t) dt + \sum_{k=1}^N \frac{b_k}{k!} [g^{(k-1)}(x+n+1) - g^{(k-1)}(x)] + S_n(x, n),\end{aligned}$$

where

$$S_N(x, n) = - \sum_{j=0}^n \int_{x+j}^{x+j+1} \frac{\varphi_N(x+j+1-t)}{N!} g^{(N)}(t) dt.$$

- (e) Let  $\psi_n$  be the periodic function, with period 1, which satisfies  $\psi_n(t) = \varphi_n(t)$  for  $0 \leq t < 1$ . (Part (a) implies that if  $n > 1$ , then  $\psi_n$  is continuous, since  $\varphi_n(1) = \varphi_n(0)$ , and also that  $\psi_n$  is even if  $n$  is even and odd if  $n$  is odd.) Show that

$$\begin{aligned} S_N(x, n) &= - \int_x^{x+n+1} \frac{\psi_N(x-t)}{N!} g^{(N)}(t) dt \\ &\left( = (-1)^{N+1} \int_x^{x+n+1} \frac{\psi_N(t)}{N!} g^{(N)}(t) dt \quad \text{if } x \text{ is an integer} \right). \end{aligned}$$

Unlike the remainder in Taylor's Theorem, the remainder  $S_n(x, n)$  usually does not satisfy  $\lim_{N \rightarrow \infty} S_N(x, n) = 0$ , because the Bernoulli numbers and functions become large very rapidly (although the first few examples do not suggest this). Nevertheless, important information can often be obtained from the summation formula. The general situation is best discussed within the context of a specialized study ("asymptotic series"), but the next problem shows one particularly important example.

- \*\*19.** (a) Use the Euler-Maclaurin Formula, with  $N = 2$ , to show that

$$\begin{aligned} \log 1 + \cdots + \log(n-1) &= \int_1^n \log t dt - \frac{1}{2} \log n + \frac{1}{12} \left( \frac{1}{n} - 1 \right) + \int_1^n \frac{\psi_2(t)}{2t^2} dt. \end{aligned}$$

- (b) Show that

$$\log \left( \frac{n!}{n^{n+1/2} e^{-n+1/12n}} \right) = \frac{11}{12} + \int_1^n \frac{\psi_2(t)}{2t^2} dt.$$

- (c) Explain why the improper integral  $\beta = \int_1^\infty \psi_2(t)/2t^2 dt$  exists, and show that if  $\alpha = \exp(\beta + 11/12)$ , then

$$\log \left( \frac{n!}{\alpha n^{n+1/2} e^{-n+1/12n}} \right) = - \int_n^\infty \frac{\psi_2(t)}{2t^2} dt.$$

- (d) Problem 19-40(d) shows that

$$\sqrt{\pi} = \lim_{n \rightarrow \infty} \frac{(n!)^2 2^{2n}}{(2n)! \sqrt{n}}.$$

Use part (c) to show that

$$\sqrt{\pi} = \lim_{n \rightarrow \infty} \frac{\alpha^2 n^{2n+1} e^{-2n} 2^{2n}}{\alpha (2n)^{2n+1/2} e^{-2n} \sqrt{n}},$$

and conclude that  $\alpha = \sqrt{2\pi}$ .

(e) Show that

$$\int_0^{1/2} \varphi_2(t) dt = \int_0^1 \varphi_2(t) dt = 0.$$

(You can do the computations explicitly, but the result also follows immediately from Problem 18(a).) Now what can be said about the graphs of  $\bar{\psi}(x) = \int_0^x \psi_2(t) dt$  and  $\tilde{\psi}(x) = \int_0^x \bar{\psi}(t) dt$ ? Use this information and integration by parts to show that

$$\int_n^\infty \frac{\psi_2(t)}{2t^2} dt > 0.$$

(f) Show that the maximum value of  $|\varphi_2(x)|$  for  $x$  in  $[0, 1]$  is  $\frac{1}{6}$ , and conclude that

$$\left| \int_n^\infty \frac{\psi_2(t)}{2t^2} dt \right| < \frac{1}{12n}.$$

(g) Finally, conclude that

$$\sqrt{2\pi} n^{n+1/2} e^{-n} < n! < \sqrt{2\pi} n^{n+1/2} e^{-n+1/12n}.$$

The final result of Problem 19, a strong form of Stirling's Formula, shows that  $n!$  is approximately  $\sqrt{2\pi} n^{n+1/2} e^{-n}$ , in the sense that this expression differs from  $n!$  by an amount which is small compared to  $n$  when  $n$  is large. For example, for  $n = 10$  we obtain 3598696 instead of 3628800, with an error  $< 1\%$ .

A more general form of Stirling's Formula illustrates the "asymptotic" nature of the summation formula. The same argument which was used in Problem 19 can now be used to show that for  $N \geq 2$  we have

$$\log \left( \frac{n!}{\sqrt{2\pi} n^{n+1/2} e^{-n}} \right) = \sum_{k=2}^N \frac{b_k}{k(k-1)n^{k-1}} \pm \int_n^\infty \frac{\psi_N(t)}{Nt^N} dt.$$

Since  $\psi_N$  is bounded, we can obtain estimates of the form

$$\left| \int_n^\infty \frac{\psi_N(t)}{Nt^N} dt \right| \leq \frac{M_N}{n^{N-1}}.$$

If  $N$  is large, the constant  $M_N$  will also be large; but for very large  $n$  the factor  $n^{1-N}$  will make the product very small. Thus, the expression

$$\sqrt{2\pi} n^{n+1/2} e^{-n} \cdot \exp \left( \sum_{k=2}^N \frac{b_k}{k(k-1)n^{k-1}} \right)$$

may be a very bad approximation for  $n!$  when  $n$  is small, but for large  $n$  (*how* large depends on  $N$ ) it will be an extremely good one (*how* good depends on  $N$ ).

PART 5

EPILOGUE

*There was a most ingenious Architect  
who had contrived a new Method  
for building Houses,  
by beginning at the Roof, and working  
downwards to the Foundation.*

JONATHAN SWIFT

# CHAPTER 28 FIELDS

Throughout this book a conscientious attempt has been made to define all important concepts, even terms like “function,” for which an intuitive definition is often considered sufficient. But **Q** and **R**, the two main protagonists of this story, have only been named, never defined. What has never been defined can never be analyzed thoroughly, and “properties” P1–P13 must be considered assumptions, not theorems, about numbers. Nevertheless, the term “axiom” has been purposely avoided, and in this chapter the logical status of P1–P13 will be scrutinized more carefully.

Like **Q** and **R**, the sets **N** and **Z** have also remained undefined. True, some talk about all four was inserted in Chapter 2, but those rough descriptions are far from a definition. To say, for example, that **N** consists of 1, 2, 3, etc., merely names some elements of **N** without identifying them (and the “etc.” is useless). The natural numbers *can* be defined, but the procedure is involved and not quite pertinent to the rest of the book. The Suggested Reading list contains references to this problem, as well as to the other steps that are required if one wishes to develop calculus from its basic logical starting point. The further development of this program would proceed with the definition of **Z**, in terms of **N**, and the definition of **Q** in terms of **Z**. This program results in a certain well-defined set **Q**, certain explicitly defined operations + and ·, and properties P1–P12 as *theorems*. The final step in this program is the construction of **R**, in terms of **Q**. It is this last construction which concerns us. Assuming that **Q** has been defined, and that P1–P12 have been proved for **Q**, we shall ultimately *define R* and *prove* all of P1–P13 for **R**.

Our intention of proving P1–P13 means that we must define not only real numbers, but also addition and multiplication of real numbers. Indeed, the real numbers are of interest only as a set together with these operations: how the real numbers behave with respect to addition and multiplication is crucial; what the real numbers may actually be is quite irrelevant. This assertion can be expressed in a meaningful mathematical way, by using the concept of a “field,” which includes as special cases the three important number systems of this book. This extraordinarily important abstraction of modern mathematics incorporates the properties P1–P9 common to **Q**, **R**, and **C**. A **field** is a set  $F$  (of objects of any sort whatsoever), together with two “binary operations”  $+$  and  $\cdot$  defined on  $F$  (that is, two rules which associate to elements  $a$  and  $b$  in  $F$ , other elements  $a+b$  and  $a\cdot b$  in  $F$ ) for which the following conditions are satisfied:

- (1)  $(a+b)+c = a+(b+c)$  for all  $a$ ,  $b$ , and  $c$  in  $F$ .
- (2) There is some element **0** in  $F$  such that
  - (i)  $a+0=a$  for all  $a$  in  $F$ ,
  - (ii) for every  $a$  in  $F$ , there is some element  $b$  in  $F$  such that  $a+b=0$ .

- (3)  $a + b = b + a$  for all  $a$  and  $b$  in  $F$ .
- (4)  $(a \cdot b) \cdot c = a \cdot (b \cdot c)$  for all  $a$ ,  $b$ , and  $c$  in  $F$ .
- (5) There is some element  $\mathbf{1}$  in  $F$  such that  $\mathbf{1} \neq \mathbf{0}$  and
  - (i)  $a \cdot \mathbf{1} = a$  for all  $a$  in  $F$ ,
  - (ii) For every  $a$  in  $F$  with  $a \neq \mathbf{0}$ , there is some element  $b$  in  $F$  such that  $a \cdot b = \mathbf{1}$ .
- (6)  $a \cdot b = b \cdot a$  for all  $a$  and  $b$  in  $F$ .
- (7)  $a \cdot (b + c) = a \cdot b + a \cdot c$  for all  $a$ ,  $b$ , and  $c$  in  $F$ .

The familiar examples of fields are, as already indicated, **Q**, **R**, and **C**, with  $+$  and  $\cdot$  being the familiar operations of  $+$  and  $\cdot$ . It is probably unnecessary to explain why these are fields, but the explanation is, at any rate, quite brief. When  $+$  and  $\cdot$  are understood to mean the ordinary  $+$  and  $\cdot$ , the rules (1), (3), (4), (6), (7) are simply restatements of P1, P4, P5, P8, P9; the elements which play the role of  $\mathbf{0}$  and  $\mathbf{1}$  are the numbers  $0$  and  $1$  (which accounts for the choice of the symbols  $\mathbf{0}$ ,  $\mathbf{1}$ ); and the number  $b$  in (2) or (5) is  $-a$  or  $a^{-1}$ , respectively. (For this reason, in an arbitrary field  $F$  we denote by  $-a$  the element such that  $a + (-a) = \mathbf{0}$ , and by  $a^{-1}$  the element such that  $a \cdot a^{-1} = 1$ , for  $a \neq \mathbf{0}$ .)

In addition to **Q**, **R**, and **C**, there are several other fields which can be described easily. One example is the collection  $F_1$  of all numbers  $a + b\sqrt{2}$  for  $a$ ,  $b$  in **Q**. The operations  $+$  and  $\cdot$  will, once again, be the usual  $+$  and  $\cdot$  for real numbers. It is necessary to point out that these operations really do produce new elements of  $F_1$ :

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2}, \quad \text{which is in } F_1;$$

$$(a + b\sqrt{2}) \cdot (c + d\sqrt{2}) = (ac + 2bd) + (bc + ad)\sqrt{2}, \quad \text{which is in } F_1.$$

Conditions (1), (3), (4), (6), (7) for a field are obvious for  $F_1$ : since these hold for all real numbers, they certainly hold for all real numbers of the form  $a + b\sqrt{2}$ . Condition (2) holds because the number  $\mathbf{0} = 0 + 0\sqrt{2}$  is in  $F_1$  and, for  $\alpha = a + b\sqrt{2}$  in  $F_1$  the number  $\beta = (-a) + (-b)\sqrt{2}$  in  $F_1$  satisfies  $\alpha + \beta = \mathbf{0}$ . Similarly,  $\mathbf{1} = 1 + 0\sqrt{2}$  is in  $F_1$ , so (5i) is satisfied. The verification of (5ii) is the only slightly difficult point. If  $a + b\sqrt{2} \neq 0$ , then

$$a + b\sqrt{2} \cdot \frac{1}{a + b\sqrt{2}} = 1;$$

it is therefore necessary to show that  $1/(a + b\sqrt{2})$  is in  $F_1$ . This is true because

$$\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{(a - b\sqrt{2})(a + b\sqrt{2})} = \frac{a}{a^2 - 2b^2} + \frac{(-b)}{a^2 - 2b^2}\sqrt{2}.$$

(The division by  $a - b\sqrt{2}$  is valid because the relation  $a - b\sqrt{2} = 0$  could be true only if  $a = b = 0$  (since  $\sqrt{2}$  is irrational) which is ruled out by the hypothesis  $a + b\sqrt{2} \neq 0$ .)

The next example of a field,  $F_2$ , is considerably simpler in one respect: it contains only two elements, which we might as well denote by  $\mathbf{0}$  and  $\mathbf{1}$ . The operations

$+$  and  $\cdot$  are described by the following tables.

$+$	0	1	$\cdot$	0	1
0	0	1	0	0	0
1	1	0	1	0	1

The verification of conditions (1)–(7) are straightforward, case-by-case checks. For example, condition (1) may be proved by checking the 8 equations obtained by setting  $a, b, c = 0$  or  $1$ . Notice that in this field  $1 + 1 = 0$ ; this equation may also be written  $1 = -1$ .

Our final example of a field is rather silly:  $F_3$  consists of all pairs  $(a, a)$  for  $a$  in  $\mathbf{R}$ , and  $+$  and  $\cdot$  are defined by

$$(a, a) + (b, b) = (a + b, a + b), \\ (a, a) \cdot (b, b) = (a \cdot b, a \cdot b).$$

(The  $+$  and  $\cdot$  appearing on the right side are ordinary addition and multiplication for  $\mathbf{R}$ .) The verification that  $F_3$  is a field is left to you as a simple exercise.

A detailed investigation of the properties of fields is a study in itself, but for our purposes, fields provide an ideal framework in which to discuss the properties of numbers in the most economical way. For example, the consequences of P1–P9 which were derived for “numbers” in Chapter 1 actually hold for any field; in particular, they are true for the fields  $\mathbf{Q}$ ,  $\mathbf{R}$ , and  $\mathbf{C}$ .

Notice that certain common properties of  $\mathbf{Q}$ ,  $\mathbf{R}$ , and  $\mathbf{C}$  do not hold for all fields. For example, it is possible for the equation  $1 + 1 = 0$  to hold in some fields, and consequently  $a - b = b - a$  does not necessarily imply that  $a = b$ . For the field  $\mathbf{C}$  the assertion  $1 + 1 \neq 0$  was derived from the explicit description of  $\mathbf{C}$ ; for the fields  $\mathbf{Q}$  and  $\mathbf{R}$ , however, this assertion was derived from further properties which do not have analogues in the conditions for a field. There is a related concept which does use these properties. An **ordered field** is a field  $F$  (with operations  $+$  and  $\cdot$ ) together with a certain subset  $\mathbf{P}$  of  $F$  (the “positive” elements) with the following properties:

(8) For all  $a$  in  $F$ , one and only one of the following is true:

- (i)  $a = 0$ ,
- (ii)  $a$  is in  $\mathbf{P}$ ,
- (iii)  $-a$  is in  $\mathbf{P}$ .

(9) If  $a$  and  $b$  are in  $\mathbf{P}$ , then  $a + b$  is in  $\mathbf{P}$ .

(10) If  $a$  and  $b$  are in  $\mathbf{P}$ , then  $a \cdot b$  is in  $\mathbf{P}$ .

We have already seen that the field  $\mathbf{C}$  cannot be made into an ordered field. The field  $F_2$ , with only two elements, likewise cannot be made into an ordered field: in fact, condition (8), applied to  $1 = -1$ , shows that  $1$  must be in  $\mathbf{P}$ ; then (9) implies that  $1 + 1 = 0$  is in  $\mathbf{P}$ , contradicting (8). On the other hand, the field  $F_1$ ,

consisting of all numbers  $a + b\sqrt{2}$  with  $a, b$  in  $\mathbf{Q}$ , certainly can be made into an ordered field: let  $\mathbf{P}$  be the set of all  $a + b\sqrt{2}$  which are positive real numbers (in the ordinary sense). The field  $F_3$  can also be made into an ordered field; the description of  $\mathbf{P}$  is left to you.

It is natural to introduce notation for an arbitrary ordered field which corresponds to that used for  $\mathbf{Q}$  and  $\mathbf{R}$ : we define

$$\begin{aligned} a > b &\quad \text{if } a - b \text{ is in } \mathbf{P}, \\ a < b &\quad \text{if } b > a, \\ a \leq b &\quad \text{if } a < b \text{ or } a = b, \\ a \geq b &\quad \text{if } a > b \text{ or } a = b. \end{aligned}$$

Using these definitions we can reproduce, for an arbitrary ordered field  $F$ , the definitions of Chapter 7:

A set  $A$  of elements of  $F$  is **bounded above** if there is some  $x$  in  $F$  such that  $x \geq a$  for all  $a$  in  $A$ . Any such  $x$  is called an **upper bound** for  $A$ . An element  $x$  of  $F$  is a **least upper bound** for  $A$  if  $x$  is an upper bound for  $A$  and  $x \leq y$  for every  $y$  in  $F$  which is an upper bound for  $A$ .

Finally, it is possible to state an analogue of property P13 for  $\mathbf{R}$ ; this leads to the last abstraction of this chapter:

**A complete ordered field** is an ordered field in which every nonempty set which is bounded above has a least upper bound.

The consideration of fields may seem to have taken us far from the goal of constructing the real numbers. However, we are now provided with an intelligible means of formulating this goal. There are two questions which will be answered in the remaining two chapters:

1. Is there a complete ordered field?
2. Is there only one complete ordered field?

Our starting point for these considerations will be  $\mathbf{Q}$ , assumed to be an ordered field, containing  $\mathbf{N}$  and  $\mathbf{Z}$  as certain subsets. At one crucial point it will be necessary to assume another fact about  $\mathbf{Q}$ :

Let  $x$  be an element of  $\mathbf{Q}$  with  $x > 0$ . Then for any  $y$  in  $\mathbf{Q}$  there is some  $n$  in  $\mathbf{N}$  such that  $nx > y$ .

This assumption, which asserts that the rational numbers have the Archimedean property of the real numbers, does not follow from the other properties of an ordered field (for the example that demonstrates this conclusively see reference [17] of the Suggested Reading). The important point for us is that when  $\mathbf{Q}$  is explicitly constructed, properties P1–P12 appear as theorems, and so does this additional

assumption; if we really began from the beginning, no assumptions about  $\mathbf{Q}$  would be necessary.

### PROBLEMS

- Let  $F$  be the set  $\{0, 1, 2\}$  and define operations  $+$  and  $\cdot$  on  $F$  by the following tables. (The rule for constructing these tables is as follows: add or multiply in the usual way, and then subtract the highest possible multiple of 3; thus  $2 \cdot 2 = 4 = 3 + 1$ , so  $2 \cdot 2 = 1$ .)

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Show that  $F$  is a field, and prove that it cannot be made into an ordered field.

- Suppose now that we try to construct a field  $F$  having elements 0, 1, 2, 3 with operations  $+$  and  $\cdot$  defined as in the previous example, by adding or multiplying in the usual way, and then subtracting the highest possible multiple of 4. Show that  $F$  will *not* be a field.
- Let  $F = \{0, 1, \alpha, \beta\}$  and define operations  $+$  and  $\cdot$  on  $F$  by the following tables.

$+$	0	1	$\alpha$	$\beta$
0	0	1	$\alpha$	$\beta$
1	1	0	$\beta$	$\alpha$
$\alpha$	$\alpha$	$\beta$	0	1
$\beta$	$\beta$	$\alpha$	1	0

$\cdot$	0	1	$\alpha$	$\beta$
0	0	0	0	0
1	0	1	$\alpha$	$\beta$
$\alpha$	0	$\alpha$	$\beta$	1
$\beta$	0	$\beta$	1	$\alpha$

Show that  $F$  is a field.

- (a) Let  $F$  be a field in which  $1 + 1 = 0$ . Show that  $a + a = 0$  for all  $a$  (this can also be written  $a = -a$ ).
- (b) Suppose that  $a + a = 0$  for some  $a \neq 0$ . Show that  $1 + 1 = 0$  (and consequently  $b + b = 0$  for all  $b$ ).

5. (a) Show that in any field we have

$$\underbrace{(1 + \cdots + 1)}_{m \text{ times}} \cdot \underbrace{(1 + \cdots + 1)}_{n \text{ times}} = \underbrace{1 + \cdots + 1}_{mn \text{ times}}$$

for all natural numbers  $m$  and  $n$ .

- (b) Suppose that in the field  $F$  we have

$$\underbrace{1 + \cdots + 1}_{m \text{ times}} = 0$$

for some natural number  $m$ . Show that the smallest  $m$  with this property must be a prime number (this prime number is called the **characteristic** of  $F$ ).

6. Let  $F$  be any field with only finitely many elements.

- (a) Show that there must be distinct natural numbers  $m$  and  $n$  with

$$\underbrace{1 + \cdots + 1}_{m \text{ times}} = \underbrace{1 + \cdots + 1}_{n \text{ times}}.$$

- (b) Conclude that there is some natural number  $k$  with

$$\underbrace{1 + \cdots + 1}_{k \text{ times}} = 0.$$

7. Let  $a, b, c$ , and  $d$  be elements of a field  $F$  with  $a \cdot d - b \cdot c \neq 0$ . Show that the equations

$$\begin{aligned} a \cdot x + b \cdot y &= \alpha, \\ c \cdot x + d \cdot y &= \beta, \end{aligned}$$

can be solved for  $x$  and  $y$ .

8. Let  $a$  be an element of a field  $F$ . A “square root” of  $a$  is an element  $b$  of  $F$  with  $b^2 = b \cdot b = a$ .

- (a) How many square roots does  $0$  have?

- (b) Suppose  $a \neq 0$ . Show that if  $a$  has a square root, then it has two square roots, unless  $1 + 1 = 0$ , in which case  $a$  has only one.

9. (a) Consider an equation  $x^2 + b \cdot x + c = 0$ , where  $b$  and  $c$  are elements of a field  $F$ . Suppose that  $b^2 - 4 \cdot c$  has a square root  $r$  in  $F$ . Show that  $(-b + r)/2$  is a solution of this equation.

- (b) In the field  $F_2$  of the text, both elements clearly have a square root. On the other hand, it is easy to check that neither element satisfies the equation  $x^2 + x + 1 = 0$ . Thus some detail in part (a) must be incorrect. What is it?

10. Let  $F$  be a field and  $a$  an element of  $F$  which does *not* have a square root. This problem shows how to construct a bigger field  $F'$ , containing  $F$ , in which  $a$  does have a square root. (This construction has already been carried

through in a special case, namely,  $F = \mathbf{R}$  and  $a = -1$ ; this special case should guide you through this example.)

Let  $F'$  consist of all pairs  $(x, y)$  with  $x$  and  $y$  in  $F$ . If the operations on  $F$  are  $\oplus$  and  $\cdot$ , define operations  $\oplus$  and  $\odot$  on  $F'$  as follows:

$$(x, y) \oplus (z, w) = (x + z, y + w), \\ (x, y) \odot (z, w) = (x \cdot z + a \cdot y \cdot w, y \cdot z + x \cdot w).$$

- (a) Prove that  $F'$ , with the operations  $\oplus$  and  $\odot$ , is a field.
- (b) Prove that

$$(x, 0) \oplus (y, 0) = (x + y, 0), \\ (x, 0) \odot (y, 0) = (x \cdot y, 0),$$

so that we may agree to abbreviate  $(x, 0)$  by  $x$ .

- (c) Find a square root of  $a = (a, 0)$  in  $F'$ .

11. Let  $F$  be the set of all four-tuples  $(w, x, y, z)$  of real numbers. Define  $\oplus$  and  $\cdot$  by

$$(s, t, u, v) \oplus (w, x, y, z) = (s + w, t + x, u + y, v + z), \\ (s, t, u, v) \cdot (w, x, y, z) = (sw - tx - uy - vz, sx + tw + uz - vy, \\ sy + uw + vx - tz, sz + vw + ty - ux).$$

- (a) Show that  $F$  satisfies all conditions for a field, except (6). At times the algebra will become quite ornate, but the existence of multiplicative inverses is the only point requiring any thought.
- (b) It is customary to denote

$$(0, 1, 0, 0) \text{ by } i, \\ (0, 0, 1, 0) \text{ by } j, \\ (0, 0, 0, 1) \text{ by } k.$$

Find all 9 products of pairs  $i$ ,  $j$ , and  $k$ . The results will show in particular that condition (6) is definitely false. This “skew field”  $F$  is known as the **quaternions**.

The mass of drudgery which this chapter necessarily contains is relieved by one truly first-rate idea. In order to prove that a complete ordered field exists we will have to explicitly describe one in detail; verifying conditions (1)–(10) for an ordered field will be a straightforward ordeal, but the description of the field itself, of the elements in it, is ingenious indeed.

At our disposal is the set of rational numbers, and from this raw material it is necessary to produce the field which will ultimately be called the real numbers. To the uninitiated this must seem utterly hopeless—if only the rational numbers are known, where are the others to come from? By now we have had enough experience to realize that the situation may not be quite so hopeless as that casual consideration suggests. The strategy to be adopted in our construction has already been used effectively for defining functions and complex numbers. Instead of trying to determine the “real nature” of these concepts, we settled for a definition that described enough about them to determine their mathematical properties completely.

A similar proposal for defining real numbers requires a description of real numbers in terms of rational numbers. The observation, that a real number ought to be determined completely by the set of rational numbers less than it, suggests a strikingly simple and quite attractive possibility: a real number might (and in fact eventually will) be described as a collection of rational numbers. In order to make this proposal effective, however, some means must be found for describing “the set of rational numbers less than a real number” without mentioning real numbers, which are still nothing more than heuristic figments of our mathematical imagination.

If  $A$  is to be regarded as the set of rational numbers which are less than the real number  $\alpha$ , then  $A$  ought to have the following property: If  $x$  is in  $A$  and  $y$  is a rational number satisfying  $y < x$ , then  $y$  is in  $A$ . In addition to this property, the set  $A$  should have a few others. Since there should be some rational number  $x < \alpha$ , the set  $A$  should not be empty. Likewise, since there should be some rational number  $x > \alpha$ , the set  $A$  should not be all of  $\mathbb{Q}$ . Finally, if  $x < \alpha$ , then there should be another rational number  $y$  with  $x < y < \alpha$ , so  $A$  should not contain a greatest member.

If we temporarily regard the real numbers as known, then it is not hard to check (Problem 8-17) that a set  $A$  with these properties is indeed the set of rational numbers less than some real number  $\alpha$ . Since the real numbers are presently in limbo, your proof, if you supply one, must be regarded only as an unofficial comment on these proceedings. It will serve to convince you, however, that we have not failed to notice any crucial property of the set  $A$ . There appears to be no reason for hesitating any longer.

**DEFINITION**

A **real number** is a set  $\alpha$ , of rational numbers, with the following four properties:

- (1) If  $x$  is in  $\alpha$  and  $y$  is a rational number with  $y < x$ , then  $y$  is also in  $\alpha$ .
- (2)  $\alpha \neq \emptyset$ .
- (3)  $\alpha \neq \mathbf{Q}$ .
- (4) There is no greatest element in  $\alpha$ ; in other words, if  $x$  is in  $\alpha$ , then there is some  $y$  in  $\alpha$  with  $y > x$ .

The set of all real numbers is denoted by  $\mathbf{R}$ .

Just to remind you of the philosophy behind our definition, here is an explicit example of a real number:

$$\alpha = \{x \text{ in } \mathbf{Q} : x < 0 \text{ or } x^2 < 2\}.$$

It should be clear that  $\alpha$  is the real number which will eventually be known as  $\sqrt{2}$ , but it is not an entirely trivial exercise to show that  $\alpha$  actually is a real number. The whole point of such an exercise is to prove this using only facts about  $\mathbf{Q}$ ; the hard part will be checking condition (4), but this has already appeared as a problem in a previous chapter (finding out which one is up to you). Notice that condition (4), although quite bothersome here, is really essential in order to avoid ambiguity; without it both

$$\{x \text{ in } \mathbf{Q} : x < 1\}$$

and

$$\{x \text{ in } \mathbf{Q} : x \leq 1\}$$

would be candidates for the “real number 1.”

The shift from  $A$  to  $\alpha$  in our definition indicates both a conceptual and a notational concern. Henceforth, a real number *is*, by definition, a set of rational numbers. This means, in particular, that a rational number (a member of  $\mathbf{Q}$ ) is *not* a real number; instead every rational number  $x$  has a natural counterpart which is a real number, namely,  $\{y \text{ in } \mathbf{Q} : y < x\}$ . After completing the construction of the real numbers, we can mentally throw away the elements of  $\mathbf{Q}$  and agree that  $\mathbf{Q}$  will henceforth denote these special sets. For the moment, however, it will be necessary to work at the same time with rational numbers, real numbers (sets of rational numbers) and even sets of real numbers (sets of sets of rational numbers). Some confusion is perhaps inevitable, but proper notation should keep this to a minimum. Rational numbers will be denoted by lower case Roman letters ( $x, y, z, a, b, c$ ) and real numbers by lower case Greek letters ( $\alpha, \beta, \gamma$ ); capital Roman letters ( $A, B, C$ ) will be used to denote sets of real numbers.

The remainder of this chapter is devoted to the definition of  $+$ ,  $\cdot$ , and  $\mathbf{P}$  for  $\mathbf{R}$ , and a proof that with these structures  $\mathbf{R}$  is indeed a complete ordered field.

We shall actually begin with the definition of  $\mathbf{P}$ , and even here we shall work backwards. We first define  $\alpha \ll \beta$ ; later, when  $+$ ,  $\cdot$ , and  $\mathbf{0}$  are available, we shall define  $\mathbf{P}$  as the set of all  $\alpha$  with  $\mathbf{0} \ll \alpha$ , and prove the necessary properties for  $\mathbf{P}$ .

The reason for beginning with the definition of  $\ll$  is the simplicity of this concept in our present setup:

*Definition.* If  $\alpha$  and  $\beta$  are real numbers, then  $\alpha \ll \beta$  means that  $\alpha$  is contained in  $\beta$  (that is, every element of  $\alpha$  is also an element of  $\beta$ ), but  $\alpha \neq \beta$ .

A repetition of the definitions of  $\leq$ ,  $>$ ,  $\geq$  would be stultifying, but it is interesting to note that  $\leq$  can now be expressed more simply than  $\ll$ ; if  $\alpha$  and  $\beta$  are real numbers, then  $\alpha \leq \beta$  if and only if  $\alpha$  is contained in  $\beta$ .

If  $A$  is a bounded collection of real numbers, it is almost obvious that  $A$  should have a least upper bound. Each  $\alpha$  in  $A$  is a collection of rational numbers; if these rational numbers are all put in one collection  $\beta$ , then  $\beta$  is presumably  $\sup A$ . In the proof of the following theorem we check all the little details which have not been mentioned, not least of which is the assertion that  $\beta$  is a real number. (We will not bother numbering theorems in this chapter, since they all add up to one big Theorem: There is a complete ordered field.)

**THEOREM** If  $A$  is a set of real numbers and  $A \neq \emptyset$  and  $A$  is bounded above, then  $A$  has a least upper bound.

**PROOF** Let  $\beta = \{x : x \text{ is in some } \alpha \text{ in } A\}$ . Then  $\beta$  is certainly a collection of rational numbers; the proof that  $\beta$  is a real number requires checking four facts.

- (1) Suppose that  $x$  is in  $\beta$  and  $y < x$ . The first condition means that  $x$  is in  $\alpha$  for some  $\alpha$  in  $A$ . Since  $\alpha$  is a real number, the assumption  $y < x$  implies that  $y$  is in  $\alpha$ . Therefore it is certainly true that  $y$  is in  $\beta$ .
- (2) Since  $A \neq \emptyset$ , there is some  $\alpha$  in  $A$ . Since  $\alpha$  is a real number, there is some  $x$  in  $\alpha$ . This means that  $x$  is in  $\beta$ , so  $\beta \neq \emptyset$ .
- (3) Since  $A$  is bounded above, there is some real number  $\gamma$  such that  $\alpha \ll \gamma$  for every  $\alpha$  in  $A$ . Since  $\gamma$  is a real number, there is some rational number  $x$  which is not in  $\gamma$ . Now  $\alpha \ll \gamma$  means that  $\alpha$  is contained in  $\gamma$ , so it is also true that  $x$  is not in  $\alpha$  for any  $\alpha$  in  $A$ . This means that  $x$  is not in  $\beta$ ; so  $\beta \neq \mathbb{Q}$ .
- (4) Suppose that  $x$  is in  $\beta$ . Then  $x$  is in  $\alpha$  for some  $\alpha$  in  $A$ . Since  $\alpha$  does not have a greatest member, there is some rational number  $y$  with  $x < y$  and  $y$  in  $\alpha$ . But this means that  $y$  is in  $\beta$ ; thus  $\beta$  does not have a greatest member.

These four observations prove that  $\beta$  is a real number. The proof that  $\beta$  is the least upper bound of  $A$  is easier. If  $\alpha$  is in  $A$ , then clearly  $\alpha$  is contained in  $\beta$ ; this means that  $\alpha \leq \beta$ , so  $\beta$  is an upper bound for  $A$ . On the other hand, if  $\gamma$  is an upper bound for  $A$ , then  $\alpha \leq \gamma$  for every  $\alpha$  in  $A$ ; this means that  $\alpha$  is contained in  $\gamma$ , for every  $\alpha$  in  $A$ , and this surely implies that  $\beta$  is contained in  $\gamma$ . This, in turn, means that  $\beta \leq \gamma$ ; thus  $\beta$  is the least upper bound of  $A$ . ■

The definition of  $\dot{+}$  is both obvious and easy, but is must be complemented with a proof that this “obvious” definition makes any sense at all.

*Definition.* If  $\alpha$  and  $\beta$  are real numbers, then

$$\alpha + \beta = \{x : x = y + z \text{ for some } y \text{ in } \alpha \text{ and some } z \text{ in } \beta\}.$$

**THEOREM** If  $\alpha$  and  $\beta$  are real numbers, then  $\alpha + \beta$  is a real number.

**PROOF** Once again four facts must be verified.

- (1) Suppose  $w < x$  for some  $x$  in  $\alpha + \beta$ . Then  $x = y + z$  for some  $y$  in  $\alpha$  and some  $z$  in  $\beta$ , which means that  $w < y + z$ , and consequently,  $w - y < z$ . This shows that  $w - y$  is in  $\beta$  (since  $z$  is in  $\beta$ , and  $\beta$  is a real number). Since  $w = y + (w - y)$ , it follows that  $w$  is in  $\alpha + \beta$ .
- (2) It is clear that  $\alpha + \beta \neq \emptyset$ , since  $\alpha \neq \emptyset$  and  $\beta \neq \emptyset$ .
- (3) Since  $\alpha \neq \mathbf{Q}$  and  $\beta \neq \mathbf{Q}$ , there are rational numbers  $a$  and  $b$  with  $a$  not in  $\alpha$  and  $b$  not in  $\beta$ . Any  $x$  in  $\alpha$  satisfies  $x < a$  (for if  $a < x$ , then condition (1) for a real number would imply that  $a$  is in  $\alpha$ ); similarly any  $y$  in  $\beta$  satisfies  $y < b$ . Thus  $x + y < a + b$  for any  $x$  in  $\alpha$  and  $y$  in  $\beta$ . This shows that  $a + b$  is not in  $\alpha + \beta$ , so  $\alpha + \beta \neq \mathbf{Q}$ .
- (4) If  $x$  is in  $\alpha + \beta$ , then  $x = y + z$  for  $y$  in  $\alpha$  and  $z$  in  $\beta$ . There are  $y'$  in  $\alpha$  and  $z'$  in  $\beta$  with  $y < y'$  and  $z < z'$ ; then  $x < y' + z'$  and  $y' + z'$  is in  $\alpha + \beta$ . Thus  $\alpha + \beta$  has no greatest member. ■

By now you can see how tiresome this whole procedure is going to be. Every time we mention a new real number, we must prove that it is a real number; this requires checking four conditions, and even when trivial they require concentration. There is really no help for this (except that it will be less boring if you check the four conditions for yourself). Fortunately, however, a few points of interest will arise now and then, and some of our theorems will be easy. In particular, two properties of  $+$  present no problems.

**THEOREM** If  $\alpha$ ,  $\beta$ , and  $\gamma$  are real numbers, then  $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ .

**PROOF** Since  $(x + y) + z = x + (y + z)$  for all rational numbers  $x$ ,  $y$ , and  $z$ , every member of  $(\alpha + \beta) + \gamma$  is also a member of  $\alpha + (\beta + \gamma)$ , and vice versa. ■

**THEOREM** If  $\alpha$  and  $\beta$  are real numbers, then  $\alpha + \beta = \beta + \alpha$ .

**PROOF** Left to you (even easier). ■

To prove the other properties of  $+$  we first define **0**.

*Definition.* **0** = { $x$  in  $\mathbf{Q} : x < 0$ }.

It is, thank goodness, obvious that **0** is a real number, and the following theorem is also simple.

**THEOREM** If  $\alpha$  is a real number, then  $\alpha + \mathbf{0} = \alpha$ .

**PROOF** If  $x$  is in  $\alpha$  and  $y$  is in  $\mathbf{0}$ , then  $y < 0$ , so  $x + y < x$ . This implies that  $x + y$  is in  $\alpha$ . Thus every member of  $\alpha + \mathbf{0}$  is also a member of  $\alpha$ .

On the other hand, if  $x$  is in  $\alpha$ , then there is a rational number  $y$  in  $\alpha$  such that  $y > x$ . Since  $x = y + (x - y)$ , where  $y$  is in  $\alpha$ , and  $x - y < 0$  (so that  $x - y$  is in  $\mathbf{0}$ ), this shows that  $x$  is in  $\alpha + \mathbf{0}$ . Thus every member of  $\alpha$  is also a member of  $\alpha + \mathbf{0}$ . ■

The reasonable candidate for  $-\alpha$  would seem to be the set

$$\{x \text{ in } \mathbf{Q} : -x \text{ is not in } \alpha\}$$

(since  $-x$  not in  $\alpha$  means, intuitively, that  $-x > \alpha$ , so that  $x < -\alpha$ ). But in certain cases this set will not even be a real number. Although a real number  $\alpha$  does not have a greatest member, the set

$$\mathbf{Q} - \alpha = \{x \text{ in } \mathbf{Q} : x \text{ is not in } \alpha\}$$

may have a *least* element  $x_0$ ; when  $\alpha$  is a real number of this kind, the set  $\{x : -x \text{ is not in } \alpha\}$  will have a greatest element  $-x_0$ . It is therefore necessary to introduce a slight modification into the definition of  $-\alpha$ , which comes equipped with a theorem.

*Definition.* If  $\alpha$  is a real number, then

$$-\alpha = \{x \text{ in } \mathbf{Q} : -x \text{ is not in } \alpha, \text{ but } -x \text{ is not the least element of } \mathbf{Q} - \alpha\}.$$

**THEOREM** If  $\alpha$  is a real number, then  $-\alpha$  is a real number.

**PROOF**

- (1) Suppose that  $x$  is in  $-\alpha$  and  $y < x$ . Then  $-y > -x$ . Since  $-x$  is not in  $\alpha$ , it is also true that  $-y$  is not in  $\alpha$ . Moreover, it is clear that  $-y$  is not the smallest element of  $\mathbf{Q} - \alpha$ , since  $-x$  is a smaller element. This shows that  $y$  is in  $-\alpha$ .
- (2) Since  $\alpha \neq \mathbf{Q}$ , there is some rational number  $y$  which is not in  $\alpha$ . We can assume that  $y$  is not the smallest rational number in  $\mathbf{Q} - \alpha$  (since  $y$  can always be replaced by any  $y' > y$ ). Then  $-y$  is in  $-\alpha$ . Thus  $-\alpha \neq \emptyset$ .
- (3) Since  $\alpha \neq \emptyset$ , there is some  $x$  in  $\alpha$ . Then  $-x$  cannot possibly be in  $-\alpha$ , so  $-\alpha \neq \mathbf{Q}$ .
- (4) If  $x$  is in  $-\alpha$ , then  $-x$  is not in  $\alpha$ , and there is a rational number  $y < -x$  which is also not in  $\alpha$ . Let  $z$  be a rational number with  $y < z < -x$ . Then  $z$  is also not in  $\alpha$ , and  $z$  is clearly not the smallest element of  $\mathbf{Q} - \alpha$ . So  $-z$  is in  $-\alpha$ . Since  $-z > x$ , this shows that  $-\alpha$  does not have a greatest element. ■

The proof that  $\alpha + (-\alpha) = \mathbf{0}$  is not entirely straightforward. The difficulties are not caused, as you might presume, by the finicky details in the definition

of  $-\alpha$ . Rather, at this point we require the Archimedean property of  $\mathbf{Q}$  stated on page 574, which does not follow from P1–P12. This property is needed to prove the following lemma, which plays a crucial role in the next theorem.

**LEMMA** Let  $\alpha$  be a real number, and  $z$  a positive rational number. Then there are (Figure 1) rational numbers  $x$  in  $\alpha$ , and  $y$  not in  $\alpha$ , such that  $y - x = z$ . Moreover, we may assume that  $y$  is not the smallest element of  $\mathbf{Q} - \alpha$ .

**PROOF** Suppose first that  $z$  is in  $\alpha$ . If the numbers

$$z, 2z, 3z, \dots$$

were all in  $\alpha$ , then every rational number would be in  $\alpha$ , since every rational number  $w$  satisfies  $w < nz$  for some  $n$ , by the additional assumption on page 574. This contradicts the fact that  $\alpha$  is a real number, so there is some  $k$  such that  $x = kz$  is in  $\alpha$  and  $y = (k + 1)z$  is not in  $\alpha$ . Clearly  $y - x = z$ .

Moreover, if  $y$  happens to be the smallest element of  $\mathbf{Q} - \alpha$ , let  $x' > x$  be an element of  $\alpha$ , and replace  $x$  by  $x'$ , and  $y$  by  $y + (x' - x)$ .

If  $z$  is not in  $\alpha$ , there is a similar proof, based on the fact that the numbers  $(-n)z$  cannot all fail to be in  $\alpha$ . ■

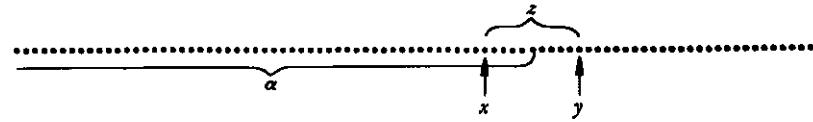


FIGURE 1

**THEOREM** If  $\alpha$  is a real number, then

$$\alpha + (-\alpha) = 0.$$

**PROOF** Suppose  $x$  is in  $\alpha$  and  $y$  is in  $-\alpha$ . Then  $-y$  is not in  $\alpha$ , so  $-y > x$ . Hence  $x + y < 0$ , so  $x + y$  is in  $0$ . Thus every member of  $\alpha + (-\alpha)$  is in  $0$ .

It is a little more difficult to go in the other direction. If  $z$  is in  $0$ , then  $-z > 0$ . According to the lemma, there is some  $x$  in  $\alpha$ , and some  $y$  not in  $\alpha$ , with  $y$  not the smallest element of  $\mathbf{Q} - \alpha$ , such that  $y - x = -z$ . This equation can be written  $x + (-y) = z$ . Since  $x$  is in  $\alpha$ , and  $-y$  is in  $-\alpha$ , this proves that  $z$  is in  $\alpha + (-\alpha)$ . ■

Before proceeding with multiplication, we define the “positive elements” and prove a basic property:

*Definition.*  $\mathbf{P} = \{\alpha \text{ in } \mathbf{R} : \alpha > 0\}$ .

Notice that  $\alpha + \beta$  is clearly in  $\mathbf{P}$  if  $\alpha$  and  $\beta$  are.

**THEOREM** If  $\alpha$  is a real number, then one and only one of the following conditions holds:

- (i)  $\alpha = 0$ ,
- (ii)  $\alpha$  is in  $\mathbf{P}$ ,
- (iii)  $-\alpha$  is in  $\mathbf{P}$ .

**PROOF** If  $\alpha$  contains any positive rational number, then  $\alpha$  certainly contains all negative rational numbers, so  $\alpha$  contains  $0$  and  $\alpha \neq 0$ , i.e.,  $\alpha$  is in  $\mathbf{P}$ . If  $\alpha$  contains no positive rational numbers, then one of two possibilities must hold:

- (1)  $\alpha$  contains all negative rational numbers; then  $\alpha = 0$ .
- (2) there is some negative rational number  $x$  which is not in  $\alpha$ ; it can be assumed that  $x$  is not the least element of  $\mathbf{Q} - \alpha$  (since  $x$  could be replaced by  $x/2 > x$ ); then  $-\alpha$  contains the positive rational number  $-x$ , so, as we have just proved,  $-\alpha$  is in  $\mathbf{P}$ .

This shows that *at least one* of (i)–(iii) must hold. If  $\alpha = 0$ , it is clearly impossible for condition (ii) or (iii) to hold. Moreover, it is impossible that  $\alpha > 0$  and  $-\alpha > 0$  both hold, since this would imply that  $0 = \alpha + (-\alpha) > 0$ . ■

Recall that  $\alpha > \beta$  was defined to mean that  $\alpha$  contains  $\beta$ , but is unequal to  $\beta$ . This definition was fine for proving completeness, but now we have to show that it is equivalent to the definition which would be made in terms of  $\mathbf{P}$ . Thus, we must show that  $\alpha - \beta > 0$  is equivalent to  $\alpha > \beta$ . This is clearly a consequence of the next theorem.

**THEOREM** If  $\alpha$ ,  $\beta$ , and  $\gamma$  are real numbers and  $\alpha > \beta$ , then  $\alpha + \gamma > \beta + \gamma$ .

**PROOF** The hypothesis  $\alpha > \beta$  implies that  $\beta$  is contained in  $\alpha$ ; it follows immediately from the definition of  $+$  that  $\beta + \gamma$  is contained in  $\alpha + \gamma$ . This shows that  $\alpha + \gamma \geq \beta + \gamma$ . We can easily rule out the possibility of equality, for if

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

then

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

which is false. Thus  $\alpha + \gamma > \beta + \gamma$ . ■

Multiplication presents difficulties of its own. If  $\alpha, \beta > 0$ , then  $\alpha \cdot \beta$  can be defined as follows.

*Definition.* If  $\alpha$  and  $\beta$  are real numbers and  $\alpha, \beta > 0$ , then

$$\alpha \cdot \beta = \{z : z \leq 0 \text{ or } z = x \cdot y \text{ for some } x \text{ in } \alpha \text{ and } y \text{ in } \beta \text{ with } x, y > 0\}.$$

**THEOREM** If  $\alpha$  and  $\beta$  are real numbers with  $\alpha, \beta > 0$ , then  $\alpha \cdot \beta$  is a real number.

**PROOF** As usual, we must check four conditions.

- (1) Suppose  $w < z$ , where  $z$  is in  $\alpha \cdot \beta$ . If  $w \leq 0$ , then  $w$  is automatically in  $\alpha \cdot \beta$ . Suppose that  $w > 0$ . Then  $z > 0$ , so  $z = x \cdot y$  for some positive  $x$  in  $\alpha$  and positive  $y$  in  $\beta$ . Now

$$w = \frac{wz}{z} = \frac{wxy}{z} = \left( \frac{w}{z} \cdot x \right) \cdot y.$$

Since  $0 < w < z$ , we have  $w/z < 1$ , so  $(w/z) \cdot x$  is in  $\alpha$ . Thus  $w$  is in  $\alpha \cdot \beta$ .

- (2) Clearly  $\alpha \cdot \beta \neq \emptyset$ .

- (3) If  $x$  is not in  $\alpha$ , and  $y$  is not in  $\beta$ , then  $x > x'$  for all  $x'$  in  $\alpha$ , and  $y > y'$  for all  $y'$  in  $\beta$ . Hence  $xy > x'y'$  for all such positive  $x'$  and  $y'$ . So  $xy$  is not in  $\alpha \cdot \beta$ ; thus  $\alpha \cdot \beta \neq \mathbb{Q}$ .

- (4) Suppose  $w$  is in  $\alpha \cdot \beta$ , and  $w \leq 0$ . There is some  $x$  in  $\alpha$  with  $x > 0$  and some  $y$  in  $\beta$  with  $y > 0$ . Then  $z = xy$  is in  $\alpha \cdot \beta$  and  $z > w$ . Now suppose  $w > 0$ . Then  $w = xy$  for some positive  $x$  in  $\alpha$  and some positive  $y$  in  $\beta$ . Moreover,  $\alpha$  contains some  $x' > x$ ; if  $z = x'y$ , then  $z > xy = w$ , and  $z$  is in  $\alpha \cdot \beta$ . Thus  $\alpha \cdot \beta$  does not have a greatest element. ■

Notice that  $\alpha \cdot \beta$  is clearly in  $\mathbf{P}$  if  $\alpha$  and  $\beta$  are. This completes the verification of all properties of  $\mathbf{P}$ . To complete the definition of  $\cdot$  we first define  $|\alpha|$ .

*Definition.* If  $\alpha$  is a real number, then

$$|\alpha| = \begin{cases} \alpha, & \text{if } \alpha \geq 0 \\ -\alpha, & \text{if } \alpha \leq 0. \end{cases}$$

*Definition.* If  $\alpha$  and  $\beta$  are real numbers, then

$$\alpha \cdot \beta = \begin{cases} 0, & \text{if } \alpha = 0 \text{ or } \beta = 0 \\ |\alpha| \cdot |\beta|, & \text{if } \alpha > 0, \beta > 0 \text{ or } \alpha < 0, \beta < 0 \\ -(|\alpha| \cdot |\beta|), & \text{if } \alpha > 0, \beta < 0 \text{ or } \alpha < 0, \beta > 0. \end{cases}$$

As one might suspect, the proofs of the properties of multiplication usually involve reduction to the case of positive numbers.

**THEOREM** If  $\alpha$ ,  $\beta$ , and  $\gamma$  are real numbers, then  $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma$ .

**PROOF** This is clear if  $\alpha$ ,  $\beta$ ,  $\gamma > 0$ . The proof for the general case requires considering separate cases (and is simplified slightly if one uses the following theorem). ■

**THEOREM** If  $\alpha$  and  $\beta$  are real numbers, then  $\alpha \cdot \beta = \beta \cdot \alpha$ .

**PROOF** This is clear if  $\alpha, \beta > 0$ , and the other cases are easily checked. ■

*Definition.*  $1 = \{x \text{ in } \mathbf{Q} : x < 1\}$ .

(It is clear that  $1$  is a real number.)

**THEOREM** If  $\alpha$  is a real number, then  $\alpha \cdot 1 = \alpha$ .

**PROOF** Let  $\alpha > 0$ . It is easy to see that every member of  $\alpha \cdot 1$  is also a member of  $\alpha$ . On the other hand, suppose  $x$  is in  $\alpha$ . If  $x \leq 0$ , then  $x$  is automatically in  $\alpha \cdot 1$ . If  $x > 0$ , then there is some rational number  $y$  in  $\alpha$  such that  $x < y$ . Then  $x = y \cdot (x/y)$ , and  $x/y$  is in  $1$ , so  $x$  is in  $\alpha \cdot 1$ . This proves that  $\alpha \cdot 1 = \alpha$  if  $\alpha > 0$ .

If  $\alpha < 0$ , then, applying the result just proved, we have

$$\alpha \cdot 1 = -(|\alpha| \cdot |1|) = -(|\alpha|) = \alpha.$$

Finally, the theorem is obvious when  $\alpha = 0$ . ■

*Definition.* If  $\alpha$  is a real number and  $\alpha > 0$ , then

$\alpha^{-1} = \{x \text{ in } \mathbf{Q} : x \leq 0, \text{ or } x > 0 \text{ and } 1/x \text{ is not in } \alpha, \text{ but } 1/x \text{ is not the smallest member of } \mathbf{Q} - \alpha\}$ ;

if  $\alpha < 0$ , then  $\alpha^{-1} = -(|\alpha|)^{-1}$ .

**THEOREM** If  $\alpha$  is a real number unequal to  $0$ , then  $\alpha^{-1}$  is a real number.

**PROOF** Clearly it suffices to consider only  $\alpha > 0$ . Four conditions must be checked.

- (1) Suppose  $y < x$ , and  $x$  is in  $\alpha^{-1}$ . If  $y \leq 0$ , then  $y$  is in  $\alpha^{-1}$ . If  $y > 0$ , then  $x > 0$ , so  $1/x$  is not in  $\alpha$ . Since  $1/y > 1/x$ , it follows that  $1/y$  is not in  $\alpha$ , and  $1/y$  is clearly not the smallest element of  $\mathbf{Q} - \alpha$ , so  $y$  is in  $\alpha^{-1}$ .
- (2) Clearly  $\alpha^{-1} \neq \emptyset$ .
- (3) Since  $\alpha > 0$ , there is some positive rational number  $x$  in  $\alpha$ . Then  $1/x$  is not in  $\alpha^{-1}$ , so  $\alpha^{-1} \neq \mathbf{Q}$ .
- (4) Suppose  $x$  is in  $\alpha^{-1}$ . If  $x \leq 0$ , there is clearly some  $y$  in  $\alpha^{-1}$  with  $y > x$  because  $\alpha^{-1}$  contains some positive rationals. If  $x > 0$ , then  $1/x$  is not in  $\alpha$ . Since  $1/x$  is not the smallest member of  $\mathbf{Q} - \alpha$ , there is a rational number  $y$  not in  $\alpha$ , with  $y < 1/x$ . Choose a rational number  $z$  with  $y < z < 1/x$ . Then  $1/z$  is in  $\alpha$ , and  $1/z > x$ . Thus  $\alpha^{-1}$  does not contain a largest member. ■

In order to prove that  $\alpha^{-1}$  is really the multiplicative inverse of  $\alpha$ , it helps to have another lemma, which is the multiplicative analogue of our first lemma.

**LEMMA** Let  $\alpha$  be a real number with  $\alpha > 0$ , and  $z$  a rational number with  $z > 1$ . Then there are rational numbers  $x$  in  $\alpha$ , and  $y$  not in  $\alpha$ , such that  $y/x = z$ . Moreover, we can assume that  $y$  is not the least element of  $\mathbf{Q} - \alpha$ .

**PROOF** Suppose first that  $z$  is in  $\alpha$ . Since  $z - 1 > 0$  and

$$z^n = (1 + (z - 1))^n \geq 1 + n(z - 1),$$

it follows that the numbers

$$z, z^2, z^3, \dots$$

cannot all be in  $\alpha$ . So there is some  $k$  such that  $x = z^k$  is in  $\alpha$ , and  $y = z^{k+1}$  is not in  $\alpha$ . Clearly  $y/x = z$ . Moreover, if  $y$  happens to be the least element of  $\mathbf{Q} - \alpha$ , let  $x' > x$  be an element of  $\alpha$ , and replace  $x$  by  $x'$  and  $y$  by  $yx'/x$ .

If  $z$  is not in  $\alpha$ , there is a similar proof, based on the fact that the numbers  $1/z^k$  cannot all fail to be in  $\alpha$ . ■

**THEOREM** If  $\alpha$  is a real number and  $\alpha \neq 0$ , then  $\alpha \cdot \alpha^{-1} = 1$ .

**PROOF** It obviously suffices to consider only  $\alpha > 0$ , in which case  $\alpha^{-1} > 0$ . Suppose that  $x$  is a positive rational number in  $\alpha$ , and  $y$  is a positive rational number in  $\alpha^{-1}$ . Then  $1/y$  is not in  $\alpha$ , so  $1/y > x$ ; consequently  $xy < 1$ , which means that  $xy$  is in 1. Since all rational numbers  $x \leq 0$  are also in 1, this shows that every member of  $\alpha \cdot \alpha^{-1}$  is in 1.

To prove the converse assertion, let  $z$  be in 1. If  $z \leq 0$ , then clearly  $z$  is in  $\alpha \cdot \alpha^{-1}$ . Suppose  $0 < z < 1$ . According to the lemma, there are positive rational numbers  $x$  in  $\alpha$ , and  $y$  not in  $\alpha$ , such that  $y/x = 1/z$ ; and we can assume that  $y$  is not the smallest element of  $\mathbf{Q} - \alpha$ . But this means that  $z = x \cdot (1/y)$ , where  $x$  is in  $\alpha$ , and  $1/y$  is in  $\alpha^{-1}$ . Consequently,  $z$  is in  $\alpha \cdot \alpha^{-1}$ . ■

We are almost done! Only the proof of the distributive law remains. Once again we must consider many cases, but do not despair. The case when all numbers are positive contains an interesting point, and the other cases can all be taken care of very neatly.

**THEOREM** If  $\alpha$ ,  $\beta$ , and  $\gamma$  are real numbers, then  $\alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$ .

**PROOF** Assume first that  $\alpha, \beta, \gamma > 0$ . Then both numbers in the equation contain all rational numbers  $\leq 0$ . A positive rational number in  $\alpha \cdot (\beta + \gamma)$  is of the form  $x \cdot (y + z)$  for positive  $x$  in  $\alpha$ ,  $y$  in  $\beta$ , and  $z$  in  $\gamma$ . Since  $x \cdot (y + z) = x \cdot y + x \cdot z$ , where  $x \cdot y$  is a positive element of  $\alpha \cdot \beta$ , and  $x \cdot z$  is a positive element of  $\alpha \cdot \gamma$ , this number is also in  $\alpha \cdot \beta + \alpha \cdot \gamma$ . Thus, every element of  $\alpha \cdot (\beta + \gamma)$  is also in  $\alpha \cdot \beta + \alpha \cdot \gamma$ .

On the other hand, a positive rational number in  $\alpha \cdot \beta + \alpha \cdot \gamma$  is of the form  $x_1 \cdot y + x_2 \cdot z$  for positive  $x_1, x_2$  in  $\alpha$ ,  $y$  in  $\beta$ , and  $z$  in  $\gamma$ . If  $x_1 \leq x_2$ , then  $(x_1/x_2) \cdot y \leq y$ , so  $(x_1/x_2) \cdot y$  is in  $\beta$ . Thus

$$x_1 \cdot y + x_2 \cdot z = x_2[(x_1/x_2)y + z]$$

is in  $\alpha \cdot (\beta + \gamma)$ . Of course, the same trick works if  $x_2 \leq x_1$ .

To complete the proof it is necessary to consider the cases when  $\alpha$ ,  $\beta$ , and  $\gamma$  are not all  $> 0$ . If any one of the three equals 0, the proof is easy and the cases

involving  $\alpha < 0$  can be derived immediately once all the possibilities for  $\beta$  and  $\gamma$  have been accounted for. Thus we assume  $\alpha > 0$  and consider three cases:  $\beta, \gamma < 0$ , and  $\beta < 0, \gamma > 0$ , and  $\beta > 0, \gamma < 0$ . The first follows immediately from the case already proved, and the third follows from the second by interchanging  $\beta$  and  $\gamma$ . Therefore we concentrate on the case  $\beta < 0, \gamma > 0$ . There are then two possibilities:

(1)  $\beta + \gamma \geq 0$ . Then

$$\alpha \cdot \gamma = \alpha \cdot ([\beta + \gamma] + |\beta|) = \alpha \cdot (\beta + \gamma) + \alpha \cdot |\beta|,$$

so

$$\begin{aligned}\alpha \cdot (\beta + \gamma) &= -(\alpha \cdot |\beta|) + \alpha \cdot \gamma \\ &= \alpha \cdot \beta + \alpha \cdot \gamma.\end{aligned}$$

(2)  $\beta + \gamma \leq 0$ . Then

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

so

$$\alpha \cdot (\beta + \gamma) = -(\alpha \cdot |\beta + \gamma|) = -(\alpha \cdot |\beta|) + \alpha \cdot \gamma = \alpha \cdot \beta + \alpha \cdot \gamma. \blacksquare$$

This proof completes the work of the chapter. Although long and frequently tedious, this chapter contains results sufficiently important to be read in detail at least once (and preferably not more than once!). For the first time we know that we have not been operating in a vacuum—there is indeed a complete ordered field, the theorems of this book are not based on assumptions which can never be realized. One interesting and horrid possibility remains: there may be several complete ordered fields. If this is true, then the theorems of calculus are unexpectedly rich in content, but the properties P1–P13 are disappointingly incomplete. The last chapter disposes of this possibility; properties P1–P13 completely characterize the real numbers—anything that can be proved about real numbers can be proved on the basis of these properties alone.

### PROBLEMS

There are only two problems in this set, but each asks for an entirely different construction of the real numbers! The detailed examination of another construction is recommended only for masochists, but the main idea behind these other constructions is worth knowing. The real numbers constructed in this chapter might be called “the algebraist’s real numbers,” since they were purposely defined so as to guarantee the least upper bound property, which involves the ordering  $<$ , an algebraic notion. The real number system constructed in the next problem might be called “the analyst’s real numbers,” since they are devised so that Cauchy sequences will always converge.

1. Since every real number ought to be the limit of some Cauchy sequence of rational numbers, we might try to *define* a real number to be a Cauchy sequence of rational numbers. Since two Cauchy sequences might converge to the same real number, however, this proposal requires some modifications.

- (a) Define two Cauchy sequences of rational numbers  $\{a_n\}$  and  $\{b_n\}$  to be *equivalent* (denoted by  $\{a_n\} \sim \{b_n\}$ ) if  $\lim_{n \rightarrow \infty} (a_n - b_n) = 0$ . Prove that  $\{a_n\} \sim \{a_n\}$ , that  $\{b_n\} \sim \{a_n\}$  if  $\{a_n\} \sim \{b_n\}$ , and that  $\{a_n\} \sim \{c_n\}$  if  $\{a_n\} \sim \{b_n\}$  and  $\{b_n\} \sim \{c_n\}$ .
- (b) Suppose that  $\alpha$  is the set of all sequences equivalent to  $\{a_n\}$ , and  $\beta$  is the set of all sequences equivalent to  $\{b_n\}$ . Prove that either  $\alpha \cap \beta = \emptyset$  or  $\alpha = \beta$ . (If  $\alpha \cap \beta \neq \emptyset$ , then there is some  $\{c_n\}$  in both  $\alpha$  and  $\beta$ . Show that in this case  $\alpha$  and  $\beta$  both consist precisely of those sequences equivalent to  $\{c_n\}$ .)

Part (b) shows that the collection of all Cauchy sequences can be split up into disjoint sets, each set consisting of all sequences equivalent to some fixed sequence. We define a real number to be such a collection, and denote the set of all real numbers by  $\mathbf{R}$ .

- (c) If  $\alpha$  and  $\beta$  are real numbers, let  $\{a_n\}$  be a sequence in  $\alpha$ , and  $\{b_n\}$  a sequence in  $\beta$ . Define  $\alpha + \beta$  to be the collection of all sequences equivalent to the sequence  $\{a_n + b_n\}$ . Show that  $\{a_n + b_n\}$  is a Cauchy sequence and also show that this definition does not depend on the particular sequences  $\{a_n\}$  and  $\{b_n\}$  chosen for  $\alpha$  and  $\beta$ . Check also that the analogous definition of multiplication is well defined.
- (d) Show that  $\mathbf{R}$  is a field with these operations; existence of a multiplicative inverse is the only interesting point to check.
- (e) Define the positive real numbers  $P$  so that  $\mathbf{R}$  will be an ordered field.
- (f) Prove that every Cauchy sequence of real numbers converges. Remember that if  $\{\alpha_n\}$  is a sequence of real numbers, then each  $\alpha_n$  is itself a collection of Cauchy sequences of rational numbers.

2. This problem outlines a construction of “the high-school student’s real numbers.” We define a real number to be a pair  $(a, \{b_n\})$ , where  $a$  is an integer and  $\{b_n\}$  is a sequence of natural numbers from 0 to 9, with the proviso that the sequence is not eventually 9; intuitively, this pair represents  $a + \sum_{n=1}^{\infty} b_n 10^{-n}$ . With this definition, a real number is a very concrete object, but the difficulties involved in defining addition and multiplication are formidable (how do you add infinite decimals without worrying about carrying digits infinitely far out?). A reasonable approach is outlined below; the trick is to use least upper bounds right from the start.

- (a) Define  $(a, \{b_n\}) \ll (c, \{d_n\})$  if  $a < c$ , or if  $a = c$  and for some  $n$  we have  $b_n < d_n$  but  $b_j = d_j$  for  $1 \leq j < n$ . Using this definition, prove the least upper bound property.
- (b) Given  $\alpha = (a, \{b_n\})$ , define  $\alpha_k = a + \sum_{n=1}^k b_n 10^{-n}$ ; intuitively,  $\alpha_k$  is the rational number obtained by changing all decimal places after the  $k$ th

to 0. Conversely, given a rational number  $r$  of the form  $a + \sum_{n=1}^k b_n 10^{-n}$ , let  $r'$  denote the real number  $(a, \{b_n'\})$ , where  $b_n' = b_n$  for  $1 \leq n \leq k$  and  $b_n' = 0$  for  $n > k$ . Now for  $\alpha = (a, \{b_n\})$  and  $\beta = (c, \{d_n\})$  define

$$\alpha + \beta = \sup\{(\alpha_k + \beta_k)': k \text{ a natural number}\}$$

(the least upper bound exists by part (a)). If multiplication is defined similarly, then the verification of all conditions for a field is a straightforward task, not highly recommended. Once more, however, existence of multiplicative inverses will be the hardest.

# CHAPTER 30

# UNIQUENESS OF THE REAL NUMBERS

We shall now revert to the usual notation for real numbers, reserving boldface symbols for other fields which may turn up. Moreover, we will regard integers and rational numbers as special kinds of real numbers, and forget about the specific way in which real numbers were defined. In this chapter we are interested in only one question: are there any complete ordered fields other than  $\mathbf{R}$ ? The answer to this question, if taken literally, is “yes.” For example, the field  $F_3$  introduced in Chapter 28 is a complete ordered field, and it is certainly not  $\mathbf{R}$ . This field is a “silly” example because the pair  $(a, a)$  can be regarded as just another name for the real number  $a$ ; the operations

$$(a, a) \dot{+} (b, b) = (a + b, a + b), \\ (a, a) \cdot (b, b) = (a \cdot b, a \cdot b),$$

are consistent with this renaming. This sort of example shows that any intelligent consideration of the question requires some mathematical means of discussing such renaming procedures.

If the elements of a field  $F$  are going to be used to rename elements of  $\mathbf{R}$ , then for each  $a$  in  $\mathbf{R}$  there should correspond a “name”  $f(a)$  in  $F$ . The notation  $f(a)$  suggests that renaming can be formulated in terms of functions. In order to do this we will need a concept of function much more general than any which has occurred until now; in fact, we will require the most general notion of “function” used in mathematics. A function, in this general sense, is simply a rule which assigns to some things, other things. To be formal, a **function** is a collection of ordered pairs (of objects of any sort) which does not contain two distinct pairs with the same first element. The **domain** of a function  $f$  is the set  $A$  of all objects  $a$  such that  $(a, b)$  is in  $f$  for some  $b$ ; this (unique)  $b$  is denoted by  $f(a)$ . If  $f(a)$  is in the set  $B$  for all  $a$  in  $A$ , then  $f$  is called a function **from  $A$  to  $B$** . For example,

if  $f(x) = \sin x$  for all  $x$  in  $\mathbf{R}$  (and  $f$  is defined only for  $x$  in  $\mathbf{R}$ ), then  $f$  is a function from  $\mathbf{R}$  to  $\mathbf{R}$ ; it is also a function from  $\mathbf{R}$  to  $[-1, 1]$ ;

if  $f(z) = \sin z$  for all  $z$  in  $\mathbf{C}$ , then  $f$  is a function from  $\mathbf{C}$  to  $\mathbf{C}$ ;

if  $f(z) = e^z$  for all  $z$  in  $\mathbf{C}$ , then  $f$  is a function from  $\mathbf{C}$  to  $\mathbf{C}$ ; it is also a function from  $\mathbf{C}$  to  $\{z \in \mathbf{C} : z \neq 0\}$ ;

$\theta$  is a function from  $\{z \in \mathbf{C} : z \neq 0\}$  to  $\{x \in \mathbf{R} : 0 \leq x < 2\pi\}$ ;

if  $f$  is the collection of all pairs  $(a, (a, a))$  for  $a$  in  $\mathbf{R}$ , then  $f$  is a function from  $\mathbf{R}$  to  $F_3$ .

Suppose that  $F_1$  and  $F_2$  are two fields; we will denote the operations in  $F_1$  by  $\oplus$ ,  $\odot$ , etc., and the operations in  $F_2$  by  $\dot{+}$ ,  $\cdot$ , etc. If  $F_2$  is going to be considered as a collection of new names for elements of  $F_1$ , then there should be a function from  $F_1$  to  $F_2$  with the following properties:

- (1) The function  $f$  should be one-one, that is, if  $x \neq y$ , then we should have  $f(x) \neq f(y)$ ; this means that no two elements of  $F_1$  have the same name.
- (2) The function  $f$  should be “onto,” that is, for every element  $z$  in  $F_2$  there should be some  $x$  in  $F_1$  such that  $z = f(x)$ ; this means that every element of  $F_2$  is used to name some element of  $F_1$ .
- (3) For all  $x$  and  $y$  in  $F_1$  we should have

$$\begin{aligned}f(x \oplus y) &= f(x) \dot{+} f(y), \\f(x \odot y) &= f(x) \cdot f(y);\end{aligned}$$

this means that the renaming procedure is consistent with the operations of the field.

If we are also considering  $F_1$  and  $F_2$  as ordered fields, we add one more requirement:

- (4) If  $x \otimes y$ , then  $f(x) \prec f(y)$ .

A function with these properties is called an *isomorphism* from  $F_1$  to  $F_2$ . This definition is so important that we restate it formally.

**DEFINITION**

If  $F_1$  and  $F_2$  are two fields, an **isomorphism** from  $F_1$  to  $F_2$  is a function  $f$  from  $F_1$  to  $F_2$  with the following properties:

- (1) If  $x \neq y$ , then  $f(x) \neq f(y)$ .
- (2) If  $z$  is in  $F_2$ , then  $z = f(x)$  for some  $x$  in  $F_1$ .
- (3) If  $x$  and  $y$  are in  $F_1$ , then

$$\begin{aligned}f(x \oplus y) &= f(x) \dot{+} f(y), \\f(x \odot y) &= f(x) \cdot f(y).\end{aligned}$$

If  $F_1$  and  $F_2$  are ordered fields we also require:

- (4) If  $x \otimes y$ , then  $f(x) \prec f(y)$ .

The fields  $F_1$  and  $F_2$  are called **isomorphic** if there is an isomorphism between them. Isomorphic fields may be regarded as essentially the same—any important property of one will automatically hold for the other. Therefore, we can, and should, reformulate the question asked at the beginning of the chapter; if  $F$  is a complete ordered field it is silly to expect  $F$  to equal  $\mathbf{R}$ —rather, we would like to know if  $F$  is isomorphic to  $\mathbf{R}$ . In the following theorem,  $F$  will be a field, with operations  $\dot{+}$  and  $\cdot$ , and “positive elements”  $\mathbf{P}$ ; we write  $a \ll b$  to mean that  $b - a$  is in  $\mathbf{P}$ , and so forth.

**THEOREM** If  $F$  is a complete ordered field, then  $F$  is isomorphic to  $\mathbf{R}$ .

**PROOF** Since two fields are defined to be isomorphic if there is an isomorphism between them, we must actually construct a function  $f$  from  $\mathbf{R}$  to  $F$  which is an isomorphism. We begin by defining  $f$  on the integers as follows:

$$\begin{aligned} f(0) &= \mathbf{0}, \\ f(n) &= \underbrace{\mathbf{1} + \dots + \mathbf{1}}_{n \text{ times}} \quad \text{for } n > 0, \\ f(n) &= -\underbrace{(\mathbf{1} + \dots + \mathbf{1})}_{|n| \text{ times}} \quad \text{for } n < 0. \end{aligned}$$

It is easy to check that

$$\begin{aligned} f(m+n) &= f(m) + f(n), \\ f(m \cdot n) &= f(m) \cdot f(n), \end{aligned}$$

for all integers  $m$  and  $n$ , and it is convenient to denote  $f(n)$  by  $n$ . We then define  $f$  on the rational numbers by

$$f(m/n) = m/n = m \cdot n^{-1}$$

(notice that the  $n$ -fold sum  $\mathbf{1} + \dots + \mathbf{1} \neq \mathbf{0}$  if  $n > 0$ , since  $F$  is an ordered field). This definition makes sense because if  $m/n = k/l$ , then  $ml = nk$ , so  $m \cdot l = k \cdot n$ , so  $m \cdot n^{-1} = k \cdot l^{-1}$ . It is easy to check that

$$\begin{aligned} f(r_1 + r_2) &= f(r_1) + f(r_2), \\ f(r_1 \cdot r_2) &= f(r_1) \cdot f(r_2), \end{aligned}$$

for all rational numbers  $r_1$  and  $r_2$ , and that  $f(r_1) < f(r_2)$  if  $r_1 < r_2$ .

The definition of  $f(x)$  for arbitrary  $x$  is based on the now familiar idea that any real number is determined by the rational numbers less than it. For any  $x$  in  $\mathbf{R}$ , let  $A_x$  be the subset of  $F$  consisting of all  $f(r)$ , for all rational numbers  $r < x$ . The set  $A_x$  is certainly not empty, and it is also bounded above, for if  $r_0$  is a rational number with  $r_0 > x$ , then  $f(r_0) > f(r)$  for all  $f(r)$  in  $A_x$ . Since  $F$  is a complete ordered field, the set  $A_x$  has a least upper bound; we define  $f(x)$  as  $\sup A_x$ .

We now have  $f(x)$  defined in two different ways, first for rational  $x$ , and then for any  $x$ . Before proceeding further, it is necessary to show that these two definitions agree for rational  $x$ . In other words, if  $x$  is a rational number, we want to show that

$$\sup A_x = f(x),$$

where  $f(x)$  here denotes  $m/n$ , for  $x = m/n$ . This is not automatic, but depends on the completeness of  $F$ ; a slight digression is thus required.

Since  $F$  is complete, the elements

$$\underbrace{\mathbf{1} + \dots + \mathbf{1}}_{n \text{ times}} \quad \text{for natural numbers } n$$

form a set which is not bounded above; the proof is exactly the same as the proof for  $\mathbf{R}$  (Theorem 8-2). The consequences of this fact for  $\mathbf{R}$  have exact analogues in  $F$ : in particular, if  $a$  and  $b$  are elements of  $F$  with  $a \lessdot b$ , then there is a rational number  $r$  such that

$$a \lessdot f(r) \lessdot b.$$

Having made this observation, we return to the proof that the two definitions of  $f(x)$  agree for rational  $x$ . If  $y$  is a rational number with  $y < x$ , then we have already seen that  $f(y) \lessdot f(x)$ . Thus every element of  $A_x$  is  $\lessdot f(x)$ . Consequently,

$$\sup A_x \leq f(x).$$

On the other hand, suppose that we had

$$\sup A_x \lessdot f(x).$$

Then there would be a rational number  $r$  such that

$$\sup A_x \lessdot f(r) \lessdot f(x).$$

But the condition  $f(r) \lessdot f(x)$  means that  $r < x$ , which means that  $f(r)$  is in the set  $A_x$ ; this clearly contradicts the condition  $\sup A_x \lessdot f(r)$ . This shows that the original assumption is false, so

$$\sup A_x = f(x).$$

We thus have a certain well-defined function  $f$  from  $\mathbf{R}$  to  $F$ . In order to show that  $f$  is an isomorphism we must verify conditions (1)–(4) of the definition. We will begin with (4).

If  $x$  and  $y$  are real numbers with  $x < y$ , then clearly  $A_x$  is contained in  $A_y$ . Thus

$$f(x) = \sup A_x \leq \sup A_y = f(y).$$

To rule out the possibility of equality, notice that there are rational numbers  $r$  and  $s$  with

$$x < r < s < y.$$

We know that  $f(r) \lessdot f(s)$ . It follows that

$$f(x) \leq f(r) \lessdot f(s) \leq f(y).$$

This proves (4).

Condition (1) follows immediately from (4): If  $x \neq y$ , then either  $x < y$  or  $y < x$ ; in the first case  $f(x) \lessdot f(y)$ , and in the second case  $f(y) \lessdot f(x)$ ; in either case  $f(x) \neq f(y)$ .

To prove (2), let  $a$  be an element of  $F$ , and let  $B$  be the set of all rational numbers  $r$  with  $f(r) \lessdot a$ . The set  $B$  is not empty, and it is also bounded above, because there is a rational number  $s$  with  $f(s) \gg a$ , so that  $f(s) \gg f(r)$  for  $r$  in  $B$ , which implies that  $s > r$ . Let  $x$  be the least upper bound of  $B$ ; we claim that  $f(x) = a$ . In order to prove this it suffices to eliminate the alternatives

$$\begin{aligned} f(x) &\lessdot a, \\ a &\lessdot f(x). \end{aligned}$$

In the first case there would be a rational number  $r$  with

$$f(x) < f(r) < a.$$

But this means that  $x < r$  and that  $r$  is in  $B$ , which contradicts the fact that  $x = \sup B$ . In the second case there would be a rational number  $r$  with

$$a < f(r) < f(x).$$

This implies that  $r < x$ . Since  $x = \sup B$ , this means that  $r < s$  for some  $s$  in  $B$ . Hence

$$f(r) < f(s) < a,$$

again a contradiction. Thus  $f(x) = a$ , proving (2).

To check (3), let  $x$  and  $y$  be real numbers and suppose that  $f(x + y) \neq f(x) + f(y)$ . Then either

$$f(x + y) < f(x) + f(y) \quad \text{or} \quad f(x) + f(y) < f(x + y).$$

In the first case there would be a rational number  $r$  such that

$$f(x + y) < f(r) < f(x) + f(y).$$

But this would mean that

$$x + y < r.$$

Therefore  $r$  could be written as the sum of two rational numbers

$$r = r_1 + r_2, \quad \text{where } x < r_1 \text{ and } y < r_2.$$

Then, using the facts checked about  $f$  for *rational* numbers, it would follow that

$$f(r) = f(r_1 + r_2) = f(r_1) + f(r_2) > f(x) + f(y),$$

a contradiction. The other case is handled similarly.

Finally, if  $x$  and  $y$  are positive real numbers, the same sort of reasoning shows that

$$f(x \cdot y) = f(x) \cdot f(y);$$

the general case is then a simple consequence. ■

This theorem brings to an end our investigation of the real numbers, and resolves any doubts about them: There *is* a complete ordered field and, up to isomorphism, only one complete ordered field. It is an important part of a mathematical education to follow a construction of the real numbers in detail, but it is not necessary to refer ever again to this particular construction. It is utterly irrelevant that a real number happens to be a collection of rational numbers, and such a fact should never enter the proof of any important theorem about the real numbers. Reasonable proofs should use only the fact that the real numbers are a complete ordered field, because this property of the real numbers characterizes them up to isomorphism, and any significant mathematical property of the real numbers will be true for all isomorphic fields. To be candid I should admit that this last assertion is just a prejudice of the author, but it is one shared by almost all other mathematicians.

## PROBLEMS

1. Let  $f$  be an isomorphism from  $F_1$  to  $F_2$ .
  - (a) Show that  $f(\mathbf{0}) = \mathbf{0}$  and  $f(\mathbf{1}) = \mathbf{1}$ . (Here  $\mathbf{0}$  and  $\mathbf{1}$  on the left denote elements in  $F_1$ , while  $\mathbf{0}$  and  $\mathbf{1}$  on the right denote elements of  $F_2$ .)
  - (b) Show that  $f(-a) = -f(a)$  and  $f(a^{-1}) = f(a)^{-1}$ , for  $a \neq \mathbf{0}$ .
2. Here is an opportunity to convince yourself that any significant property of a field is shared by any field isomorphic to it. The point of this problem is to write out very formal proofs until you are certain that all statements of this sort are obvious.  $F_1$  and  $F_2$  will be two fields which are isomorphic; for simplicity we will denote the operations in both by  $\mathbf{+}$  and  $\mathbf{\cdot}$ . Show that:
  - (a) If the equation  $x^2 \mathbf{+} 1 = \mathbf{0}$  has a solution in  $F_1$ , then it has a solution in  $F_2$ .
  - (b) If every polynomial equation  $x^n + a_{n-1} \cdot x^{n-1} + \cdots + a_0 = \mathbf{0}$  with  $a_0, \dots, a_{n-1}$  in  $F_1$ , has a root in  $F_1$ , then every polynomial equation  $x^n + b_{n-1} \cdot x^{n-1} + \cdots + b_0 = \mathbf{0}$  with  $b_0, \dots, b_{n-1}$  in  $F_2$  has a root in  $F_2$ .
  - (c) If  $1 + \cdots + 1$  (summed  $m$  times)  $= \mathbf{0}$  in  $F_1$ , then the same is true in  $F_2$ .
  - (d) If  $F_1$  and  $F_2$  are ordered fields (and the isomorphism  $f$  satisfies  $f(x) \lessdot f(y)$  for  $x \lessdot y$ ) and  $F_1$  is complete, then  $F_2$  is complete.
3. Let  $f$  be an isomorphism from  $F_1$  to  $F_2$  and  $g$  an isomorphism from  $F_2$  to  $F_3$ . Define the function  $g \circ f$  from  $F_1$  to  $F_3$  by  $(g \circ f)(x) = g(f(x))$ . Show that  $g \circ f$  is an isomorphism.
4. Suppose that  $F$  is a complete ordered field, so that there is an isomorphism  $f$  from  $\mathbf{R}$  to  $F$ . Show that there is actually only *one* isomorphism from  $\mathbf{R}$  to  $F$ . Hint: In case  $F = \mathbf{R}$ , this is Problem 3-17. Now if  $f$  and  $g$  are two isomorphisms from  $\mathbf{R}$  to  $F$  consider  $g^{-1} \circ f$ .
5. Find an isomorphism from  $\mathbf{C}$  to  $\mathbf{C}$  other than the identity function.

# SUGGESTED READING

*A man ought to read  
just as inclination leads him;  
for what he reads as a task  
will do him little good.*

**SAMUEL JOHNSON**



One purpose of this bibliography is to guide the reader to other sources, but the most important function it can serve is to indicate the variety of mathematical reading available. Consequently, there is an attempt to achieve diversity, but no pretense of being complete. The present plethora of mathematics books would make such an undertaking almost hopeless in any case, and since I have tried to encourage independent reading, the more standard a text, the less likely it is to appear here. In some cases, this philosophy may seem to have been carried to extremes, as some entries in the list cannot be read by a student just finishing a first course of calculus until several years have elapsed. Nevertheless, there are many selections which can be read now, and I can't believe that it hurts to have some idea of what lies ahead.

Many of these books have gone through numerous editions and printings, which will be reflected in more recent publication dates. Many of the books with older publications dates are out of print, though that generally doesn't apply to books from the redoubtable Dover Publications, or from the Mathematical Association of America. Those that are no longer in print can still often be found in well-stocked academic libraries.

One of the most elementary unproved theorems mentioned in this book is the fact that every natural number can be written as a product of primes in only one way. A proof of this basic theorem will be found near the beginning of almost any book on elementary number theory. Few books have won so enthusiastic an audience as

- [1] *An Introduction to the Theory of Numbers* (fifth edition), by G. H. Hardy and E. M. Wright; Oxford University Press, 1980.

The Pergamon Press published a series, Popular Lectures in Mathematics, with several titles worth investigating, among them

- [2] *A Selection of Problems in the Theory of Numbers*, by W. Sierpinski; Macmillan (Pergamon), 1964.

Finally, I will mention an intriguing little book, now out of print I fear,

- [3] *Three Pearls of Number Theory*, by A. Khinchin; Graylock Press, 1952.

The subject of irrational numbers straddles the fields of number theory and analysis. An excellent introduction will be found in

- [4] *Irrational Numbers*, by I. M. Niven; Mathematical Association of America, 1956.

Together with many historical notes, there are references to some fairly elementary articles in journals. There is also a proof that  $\pi$  is transcendental (see also [51]) and, finally, a proof of the “Gelfond-Schneider theorem”: If  $a$  and  $b$  are algebraic, with  $a \neq 0$  or 1, and  $b$  is irrational, then  $a^b$  is transcendental.

All the books listed so far begin with natural numbers, but whenever necessary take for granted the irrational numbers, not to mention the integers and rational

numbers. Several books present a construction of the rational numbers from the natural numbers, but one of the most lucid treatments is still to be found in

- [5] *Foundations of Analysis* (second edition), by E. Landau; Chelsea, 1960.

Incidentally, the original German edition,

- [6] *Grundlagen der Analysis* (fourth edition), by E. Landau; Chelsea, 1965.

has been printed in paper back, together with a complete German-English dictionary (of about 300 words) for the whole book—an excellent way to begin reading mathematical German. The basic idea for constructing the real numbers is derived from Dedekind, whose contributions can be found in

- [7] *Essays on the Theory of Numbers*, by R. Dedekind; Dover, 1963.

While many mathematicians are content to accept the natural numbers as a natural starting point, numbers can be defined in terms of sets, the most basic starting point of all. A charming exposition of set theory can be found in a sophisticated little book called

- [8] *Naive Set Theory*, by P. R. Halmos; Springer-Verlag, 1991.

Another very good introduction is

- [9] *Theory of Sets*, by E. Kamke; Dover, 1950.

Perhaps it is necessary to assure some victims of the “new math” that set theory does have some mathematical content (in fact, some very deep theorems). Using these deep results, Kamke proves that there is a discontinuous function  $f$  such that  $f(x + y) = f(x) + f(y)$  for all  $x$  and  $y$ . For those who enjoy reading the classics, the most important notions of set theory were first introduced by Cantor, whose work is reproduced in

- [10] *Contributions to the Founding of the Theory of Transfinite Numbers*, by G. Cantor; Dover, 1952.

Inequalities, which were treated as an elementary topic in Chapters 1 and 2, actually form a specialized field. A good elementary introduction is provided by

- [11] *Analytic Inequalities*, by N. Kazarinoff; Mathematical Association of America, 1961.

Twelve different proofs that the geometric mean is less than or equal to the arithmetic mean, each based on a different principle, can be found in the beginning of the more advanced book

- [12] *An Introduction to Inequalities*, by E. Beckenbach and R. Bellman; Mathematical Association of America, 1961.

The classic work on inequalities is

- [13] *Inequalities* (second edition), by G. H. Hardy, J. E. Littlewood, and G. Polya; Cambridge University Press, 1988.

Each of the authors of this triple collaboration has provided his own contribution to the sparse literature about the nature of mathematical thinking, written from a mathematician's point of view. My favorite is

- [14] *A Mathematician's Apology*, by G. H. Hardy; Cambridge University Press, 1992.

Littlewood's anecdotal selections are entitled

- [15] *A Mathematician's Miscellany*, by J. E. Littlewood; Methuen, 1953.

Polya's contribution is pedagogy at the highest level:

- [16] *Mathematics and Plausible Reasoning* (Vol. I: *Induction and Analogy in Mathematics*; Vol. II: *Patterns of Plausible Inference*), by G. Polya; Princeton University Press, 1990.

Geometry is the other main field which can be considered as background for calculus. Euclid's *Elements* is still a masterful mathematical work, but should perhaps be postponed until some preparation has been made, with a modern work on "classical geometry," like

- [17] *Elementary Geometry from an Advanced Standpoint* (second edition), by E. Moise; Addison-Wesley, 1974.

This beautiful book provides excellent historical perspectives and contains a thorough discussion of the role of the "Archimedean axiom" in geometry; in addition, Chapter 28 describes an ordered field in which the Archimedean axiom does not hold. Speaking of beautiful geometry books, all sorts of fascinating things can be found in

- [18] *Introduction to Geometry* (second edition), by H. S. Coxeter; Wiley, 1989.

Almost all treatments of geometry at least mention convexity, which forms another specialized topic. I cannot imagine a better introduction to convexity, or a better mathematical experience in general, than reading and working through

- [19] *Convex Figures*, by I. M. Yaglom and W. G. Boltyanskii; Holt, Rinehart and Winston, 1961.

This book contains a carefully arranged sequence of definitions and *statements* of theorems, whose proofs are to be supplied by the reader (worked-out proofs are supplied in the back of the book). Another geometry book has been modeled on the same principle:

- [20] *Combinatorial Geometry in the Plane*, by H. Hadwiger and H. Debrunner; Holt, Rinehart and Winston, 1964.

Along with these two out-of-the-ordinary books, I might mention an extremely valuable little book, also of a specialized sort,

- [21] *Counterexamples in Analysis*, by B. Gelbaum and J. Olmsted; Holden-Day, 1964.

Many of the example in this book come from more advanced topics in analysis, but quite a few can be appreciated by someone who knows calculus.

Of calculus books I will mention only two, each something of a classic:

- [22] *A Course of Pure Mathematics* (tenth edition), by G. H. Hardy; Cambridge University Press, 1952.
- [23] *Differential and Integral Calculus* (two volumes), by R. Courant; Wiley (Interscience), 1988.

Courant is especially strong on applications to physics. Speaking of such applications, an elegant exposition of the material in Chapter 17, together with much further discussion, can be found in the article

- [24] *On the Geometry of the Kepler Problem*, by John Milnor; in *The American Mathematical Monthly*, Volume 90 (1983), pp. 353–365.

(In this paper the curve  $c'$  of Chapter 17 is denoted by  $\mathbf{v}$ , and the derivative of the important composition  $\mathbf{v} \circ \theta^{-1}$  (page 331) is introduced quite off-handedly as  $d\mathbf{v}/d\theta$ .) A “straight-forward” derivation of Kepler’s laws, together with numerous references, can be found in another article in this same journal,

- [25] *The Mathematical Relationship Between Kepler’s Laws and Newton’s Laws*, by Andrew T. Hyman; in *The American Mathematical Monthly*, Volume 100 (1993), pp. 932–936.

The latter parts of Volume I of Courant contain material usually found in advanced calculus, including differential equations and Fourier series. An introduction to Fourier series (requiring a little advanced calculus) will also be found in

- [26] *An Introduction to Fourier Series and Integrals*, by R. Seeley; W. A. Benjamin, 1966.

The second volume of Courant (advanced calculus in earnest) contains additional material on differential equations, as well as an introduction to the calculus of variations. A widely admired, though somewhat more advanced, book on differential equations is

- [27] *Lectures on Ordinary Differential Equations*, by W. Hurewicz; Dover, 1990.

I will bypass the more or less standard advanced calculus books (which can easily be found by the reader) since nowadays there is a movement to revise the whole presentation of advanced calculus, basing it upon linear algebra. One of the first, and still one of the nicest, treatments of advanced calculus using linear algebra is

- [28] *Calculus of Vector Functions*, by R. H. Crowell and R. E. Williamson; Prentice-Hall, 1962.

Several recent books on advanced calculus attempt to acquaint undergraduates with very large areas of modern mathematics. My favorite, of course, is

- [29] *Calculus on Manifolds*, by M. Spivak; W. A. Benjamin, 1965.

There are three other topics which are somewhat out of place in this bibliography because they are rapidly becoming established as part of a standard undergraduate curriculum. The purposeful study of fields and related systems is the domain of "algebra." One of the favorite texts is

- [30] *Topics in Algebra* (second edition), by I. N. Herstein; Wiley, 1975.

A more advanced book is the great classic:

- [31] *Algebra*, by B. L. van der Waerden; Springer-Verlag, 1990.

By the way, this book contains a proof of the partial fraction decomposition of a rational function.

There are now several introductions to complex analysis, as well as many elementary books on topology. Although the latter subject has not been mentioned before, it has really been in the background of many discussions, since it is the natural generalization of the ideas about limits and continuity which play such a prominent role in Part II of this book.

The next few topics, ranging from elementary to very difficult, are included in this bibliography because they have been alluded to in the text. The proof that a nondecreasing function is differentiable at almost all points (and an explanation of just what this means) receives a beautiful exposition in

- [32] *Functional Analysis*, by F. Riesz and B. Sz.-Nagy; Ungar, 1955.

(After this elementary beginning, the book moves on to quite advanced material.) The gamma function has an elegant little book devoted entirely to its properties, most of them proved by using the theorem of Bohr and Mollerup which was mentioned in Problem 19-39:

- [33] *The Gamma Function*, by E. Artin; Holt, Rinehart and Winston, 1964.

The gamma function is only one of several important improper integrals in mathematics. In particular, the calculation of  $\int_0^\infty e^{-x^2} dx$  (see Problem 19-41) is important in probability theory, where the "normal distribution function"

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}y^2} dy$$

plays a fundamental role. A classic book on probability theory is

- [34] *An Introduction to Probability Theory and Its Applications* (third edition), by W. Feller; Wiley, 1968.

The impossibility of integrating certain functions in elementary terms (among them  $f(x) = e^{-x^2}$ ) is one of the most esoteric subjects in mathematics. An interesting discussion of the possibilities of integrating in elementary terms, with an

outline of the impossibility proofs, and references to the original papers of Liouville, will be found in

- [35] *The Integration of Functions of a Single Variable* (second edition), by G. H. Hardy; Cambridge University Press, 1958.

A complete presentation of the impossibility proofs will be found in

- [36] *Integration in Finite Terms*, by J. Ritt; Columbia University Press, 1948.

Oddly enough, a related but seemingly more difficult problem has a much neater solution. There are simple differential equations ( $y'' + xy = 0$  is a specific example) whose solutions cannot be expressed even in terms of indefinite integrals of elementary functions. This fact is proved on page 43 of the (60-page) book:

- [37] *An Introduction to Differential Algebra*, by I. Kaplansky; Hermann, 1957.

To read this book you will need to know quite a bit of algebra, however.

A few words should also be said in defense of the process of integrating in elementary terms, which many mathematicians look upon as an art (unlike differentiation, which is merely a skill). You are probably already aware that the process of integration can be expedited by tables of indefinite integrals. For those who enjoy pursuing tables there is a really beautiful collection, that includes indefinite integrals, definite improper integrals, and a great deal more besides (if you should ever happen to need the value of the thirty-fourth Bernoulli number, this is the place to look):

- [38] *Tables of Integrals, Series, and Products*, by I. S. Gradshteyn et al.; Academic Press, 1980.

For the thrifty, there is a paperback table of integrals:

- [39] *Tables of Indefinite Integrals*, by G. Petit Bois; Dover, 1961.

The remaining references are of a somewhat different sort. They fall into three categories, of which the first is historical. The letter of H. A. Schwarz referred to in Problem 11-65 will be found in

- [40] *Ways of Thought of Great Mathematicians*, by H. Meschkowski; Holden-Day, 1964.

Some historical remarks, and an attempt to incorporate them into the teaching of calculus, will be found in

- [41] *The Calculus: A Genetic Approach*, by O. Toeplitz; University of Chicago Press, 1981.

An admirable textbook on the history of mathematics is

- [42] *An Introduction to the History of Mathematics* (sixth edition), by H. Eves; Saunders College Publishing, 1990.

Three good scholarly works are

- [43] *History of Analytic Geometry*, by C. Boyer; Scholar's Bookshelf, 1988.
- [44] *A History of the Calculus, and Its Conceptual Development*, by C. Boyer; Dover, 1959.
- [45] *The Mathematics of Great Amateurs*, by J. Coolidge; Oxford University Press, 1990

and extracts from original sources will be found in

- [46] *A Source Book in Mathematics* (2 vols.), by D. Smith; Dover, 1959.

Despite the impression that might be given by the large number of books listed here, it is often hard to find specific concrete information about the origins of calculus. For example, it is almost impossible to find out who first proved the Mean Value Theorem (according to the *Encyklopädie der Mathematischen Wissenschaften*, Volume II, it was O. Bonnet, whose name is familiar to students of differential geometry from the “Gauss-Bonnet Theorem”). Similarly, though many history books tell us that Wallis proved Wallis’ formula by a “complicated method of interpolation,” most never bother to mention what it was, even though it inspired Euler’s investigations of the gamma function (a description is given in the answer book, along with the solution to Problem 19-40).

The second category in this final group of books might be described as “popularizations.” There are a surprisingly large number of first-rate ones by real mathematicians:

- [47] *What is Mathematics?* (fourth edition), by R. Courant and H. Robbins; Oxford University Press, 1979.
- [48] *Geometry and the Imagination*, by D. Hilbert and S. Cohn-Vossen; Chelsea, 1952.
- [49] *The Enjoyment of Mathematics*, by H. Rademacher and O. Toeplitz; Dover, 1990.
- [50] *Famous Problems of Mathematics* (second edition), by H. Tietze; Graylock Press, 1965.

One of the most renowned “popularizations” is especially concerned with the teaching of mathematics:

- [51] *Elementary Mathematics from an Advanced Standpoint*, by F. Klein (vol. 1: *Arithmetic, Algebra, Analysis*; vol. 2: *Geometry*); Dover, 1948.

Volume 1 contains a proof of the transcendence of  $\pi$  which, although not so elementary as the one in [4], is a direct analogue of the proof that  $e$  is transcendental, replacing integrals with complex line integrals. It can be read as soon as the basic facts about complex analysis are known.

The third category is the very opposite extreme—original papers. The difficulties encountered here are formidable, and I have only had the courage to list one such paper, the source of the quotation for Part IV. It is not even in English,

although you do have a choice of foreign languages. The article in the original French is in

- [52] *Oeuvres Complètes d'Abel*; Christiania. Johnson Reprint Corporation, New York, 1965.

It first appeared in a German translation in the *Journal für die reine und angewandte Mathematik*, Volume 1, 1826. To compound the difficulties, these references will usually be available only in university libraries. Yet the study of this paper will probably be as valuable as any other reading mentioned here. The reason is suggested by a remark of Abel himself, who attributed his profound knowledge of mathematics to the fact that he read the masters, rather than the pupils.

ANSWERS  
TO SELECTED  
PROBLEMS



**CHAPTER 1**

1. (i)  $1 = a^{-1}a = a^{-1}(ax) = (a^{-1}a)x = 1 \cdot x = x$ .  
 (iii) If  $x^2 = y^2$ , then  $0 = x^2 - y^2 = (x - y)(x + y)$ , so either  $x - y = 0$  or  $x + y = 0$ , that is, either  $x = -y$  or  $x = y$ .  
 (vi) Replace  $y$  by  $-y$  in (iv).
2. One step requires dividing by  $x - y = 0$ .
3. (i)  $a/b = ab^{-1} = (ac)(b^{-1}c^{-1}) = (ac)(bc)^{-1}$  (by (iii)) =  $ac/bc$ .  
 (ii)  $(ad + bc)/(bd) = (ad + bc)(bd)^{-1} = (ad + bc)(b^{-1}d^{-1})$  (by (iii)) =  $ab^{-1} + cd^{-1} = a/b + c/d$ .  
 (iii)  $ab(a^{-1}b^{-1}) = (a \cdot a^{-1})(b \cdot b^{-1}) = 1$ , so  $a^{-1} \cdot b^{-1} = (ab)^{-1}$ .  
 (v)  $(a/b)/(c/d) = (a/b)(c/d)^{-1} = (a \cdot b^{-1})(c \cdot d^{-1})^{-1} = (a \cdot b^{-1})(c^{-1} \cdot d) = ad(b^{-1} \cdot c^{-1}) = ad(bc)^{-1} = (ad)/(bc)$ .
4. (i)  $x < -1$ .  
 (iii)  $x > \sqrt{7}$  or  $x < -\sqrt{7}$ .  
 (v) All  $x$ , since  $x^2 - 2x + 2 = (x - 1)^2 + 1$ .  
 (vii)  $x > 3$  or  $x < -2$ , since 3 and -2 are the roots of  $x^2 - x - 6 = 0$ .  
 (ix)  $x > \pi$  or  $-5 < x < 3$ .  
 (xi)  $x < 3$ .  
 (xiii)  $x > 1$  or  $0 < x < 1$ .
5. (i)  $b - a$  and  $d - c$  are in  $P$ , so  $(b - a) + (d - c) = (b + d) - (a + c)$  is in  $P$ . Thus,  $b + d > a + c$ .  
 (iii) Using (ii),  $-c < -d$ ; then (i) implies that  $a + (-c) < b + (-d)$ .  
 (v)  $(b - a)$  and  $-c$  are in  $P$ , so  $-c(b - a) = ac - bc$  is in  $P$ , that is,  $ac > bc$ .  
 (vii) Using (iv),  $a > 0$  and  $a < 1$ , so  $a^2 < a$ .  
 (ix) Substitute  $a$  for  $c$  and  $b$  for  $d$  in (vii).
9. (i)  $\sqrt{2} + \sqrt{3} - \sqrt{5} + \sqrt{7}$ .  
 (iii)  $|a + b| + |c| - |a + b + c|$ .  
 (v)  $\sqrt{2} + \sqrt{3} + \sqrt{5} - \sqrt{7}$ .
10. (i)  $a$  if  $a \geq -b$  and  $b \geq 0$ ;  
 $-a$  if  $a \leq -b$  and  $b \leq 0$ ;  
 $a + 2b$  if  $a \geq -b$  and  $b \leq 0$ ;  
 $-a - 2b$  if  $a \leq -b$  and  $b \geq 0$ .  
 (iii)  $x - x^2$  if  $x \geq 0$ ;  
 $-x - x^2$  if  $x \leq 0$ .
11. (i)  $x = 11, -5$ .  
 (iii)  $-6 < x < -2$ .  
 (v) No  $x$  (the distance from  $x$  to 1 plus the distance from  $x$  to -1 is at least 2).  
 (vii)  $x = 1, -1$ .
12. (i)  $(|xy|)^2 = (xy)^2 = x^2y^2 = |x|^2|y|^2 = (|x| \cdot |y|)^2$ ; since  $|xy|$  and  $|x| \cdot |y|$  are both  $\geq 0$ , this proves that  $|xy| = |x| \cdot |y|$ .  
 (iii)  $|x|/|y| = |x| \cdot |y|^{-1} = |x| \cdot |y^{-1}|$  by (ii) =  $|xy^{-1}|$  (by (i)) =  $|x/y|$ .  
 (v) It follows from (iv) that  $|x| = |y - (y - x)| \leq |y| + |y - x|$ , so  $|x| - |y| \leq |x - y|$ .  
 (vii)  $|x + y + z| \leq |x + y| + |z| \leq |x| + |y| + |z|$ . If equality holds, then  $|x + y| = |x| + |y|$ , so  $x$  and  $y$  have the same sign. Moreover,  $z$  must

have the same sign as  $x + y$ , so  $x$ ,  $y$ , and  $z$  must all have the same sign (unless one is 0).

## CHAPTER 2

1. (i) Since  $1^2 = 1 \cdot (2) \cdot (2 \cdot 1 + 1)/6$ , the formula is true for  $n = 1$ . Suppose that the formula is true for  $k$ . Then

$$\begin{aligned} 1^2 + \cdots + k^2 + (k+1)^2 &= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\ &= \frac{(k+1)}{6}[k(2k+1) + 6(k+1)] \\ &= \frac{(k+1)}{6}[(k+2)(2k+3)] \\ &= \frac{(k+1)(k+2)(2[k+1]+1)}{6}, \end{aligned}$$

so the formula is true for  $k + 1$ .

2. (i)

$$\begin{aligned} \sum_{i=1}^n (2i-1) &= 1 + 3 + 5 + \cdots + (2n-1) \\ &= 1 + 2 + 3 + \cdots + 2n - 2(1 + \cdots + n) \\ &= \frac{(2n)(2n+1)}{2} - n(n+1) \\ &= n^2. \end{aligned}$$

5. (a) Since

$$1+r=\frac{1-r^2}{1-r},$$

the formula is true for  $n = 1$ . Suppose that

$$1+r+\cdots+r^n=\frac{1-r^{n+1}}{1-r}.$$

Then

$$\begin{aligned} 1+r+\cdots+r^n+r^{n+1} &= \frac{1-r^{n+1}}{1-r} + r^{n+1} \\ &= \frac{1-r^{n+1}+r^{n+1}(1-r)}{1-r} \\ &= \frac{1-r^{n+2}}{1-r}. \end{aligned}$$

(b)

$$\begin{aligned} S &= 1+r+\cdots+r^n \\ rS &= \quad\quad r+\cdots+r^n+r^{n+1}. \end{aligned}$$

Thus

$$S(1-r)=S-rS=1-r^{n+1},$$

so

$$S = \frac{1 - r^{n+1}}{1 - r}.$$

6. (i) From

$$(k+1)^4 - k^4 = 4k^3 + 6k^2 + 4k + 1, \quad k = 1, \dots, n$$

we obtain

$$(n+1)^4 - 1 = 4 \sum_{k=1}^n k^3 + 6 \sum_{k=1}^n k^2 + 4 \sum_{k=1}^n k + n,$$

so

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

- (iii) From

$$\frac{1}{k} - \frac{1}{k+1} = \frac{1}{k(k+1)}, \quad k = 1, \dots, n$$

we obtain

$$1 - \frac{1}{n+1} = \sum_{k=1}^n \frac{1}{k(k+1)}.$$

8. 1 is either even or odd, in fact it is odd. Suppose  $n$  is either even or odd; then  $n$  can be written either as  $2k$  or  $2k+1$ . In the first case  $n+1 = 2k+1$  is odd; in the second case  $n+1 = 2k+1+1 = 2(k+1)$  is even. In either case,  $n+1$  is either even or odd. (Admittedly, this looks fishy, but it is really correct.)
9. Let  $B$  be the set of all natural numbers  $l$  such that  $n_0 - 1 + l$  is in  $A$ . Then 1 is in  $B$ , and  $l+1$  is in  $B$  if  $l$  is in  $B$ , so  $B$  contains all natural numbers, which means that  $A$  contains all natural numbers  $\geq n_0$ .
12. (a) Yes, for if  $a+b$  were rational, then  $b = (a+b)-a$  would be rational. If  $a$  and  $b$  are irrational, then  $a+b$  could be rational, for  $b$  could be  $r-a$  for some rational number  $a$ .  
(b) If  $a = 0$ , then  $ab$  is rational. But if  $a \neq 0$ , then  $ab$  could not be rational, for then  $b = (ab) \cdot a^{-1}$  would be rational.  
(c) Yes; for example,  $\sqrt[4]{2}$ .  
(d) Yes; for example,  $\sqrt{2}$  and  $-\sqrt{2}$ .
13. (a) Since

$$(3n+1)^2 = 9n^2 + 6n + 1 = 3(3n^2 + 2n) + 1,$$

$$(3n+2)^2 = 9n^2 + 12n + 4 = 3(3n^2 + 4n + 1) + 1,$$

it follows that if  $k^2$  is divisible by 3, then  $k$  must also be divisible by 3.

Now suppose that  $\sqrt{3}$  were rational, and let  $\sqrt{3} = p/q$  where  $p$  and

$q$  have no common factor. Then  $p^2 = 3q^2$ , so  $p^2$  is divisible by 3, so  $p$  must be. Thus,  $p = 3p'$  for some natural number  $p'$ , and consequently  $(3p')^2 = 3q^2$ , or  $3(p')^2 = q^2$ . Thus,  $q$  is also divisible by 3, a contradiction.

The same proofs work for  $\sqrt{5}$  and  $\sqrt{6}$ , because the equations

$$\begin{aligned}(5n+1)^2 &= 25n^2 + 10n + 1 = 5(5n^2 + 2n) + 1, \\ (5n+2)^2 &= 25n^2 + 20n + 4 = 5(5n^2 + 4n) + 4, \\ (5n+3)^2 &= 25n^2 + 30n + 9 = 5(5n^2 + 6n + 1) + 4, \\ (5n+4)^2 &= 25n^2 + 40n + 16 = 5(5n^2 + 8n + 3) + 1,\end{aligned}$$

and the corresponding equations for numbers of the form  $6n+m$ , show that if  $k^2$  is divisible by 5 or 6, then  $k$  must be. The proof fails for  $\sqrt{4}$ , because  $(4n+2)^2$  is divisible by 4. (For precisely this reason this proof cannot be used to show that in general  $\sqrt{a}$  is irrational if  $a$  is not a perfect square—we have no guarantee that  $(an+m)^2$  might not be a multiple of  $a$  for some  $m < a$ . Actually, this assertion is true, but the proof requires the information in Problem 17.)

(b) Since

$$(2n+1)^3 = 8n^3 + 12n^2 + 6n + 1 = 2(4n^3 + 6n^2 + 3n) + 1,$$

it follows that if  $k^3$  is even, then  $k$  is even. If  $\sqrt[3]{2} = p/q$  where  $p$  and  $q$  have no common factors, then  $p^3 = 2q^3$ , so  $p^3$  is divisible by 2, so  $p$  must be. Thus,  $p = 2p'$  for some natural number  $p'$ , and consequently  $(2p')^3 = 2q^3$ , or  $4(p')^3 = q^3$ . Thus,  $q$  is also even, a contradiction.

The proof for  $\sqrt[3]{3}$  is similar, using the equations

$$\begin{aligned}(3n+1)^3 &= 27n^3 + 27n^2 + 9n + 1 = 3(9n^3 + 9n^2 + 3n) + 1, \\ (3n+2)^3 &= 27n^3 + 54n^2 + 36n + 8 = 3(9n^3 + 18n^2 + 12n + 2) + 2.\end{aligned}$$

19. If  $n = 1$ , then  $(1+h)^n = 1+nh$ . Suppose that  $(1+h)^n \geq 1+nh$ . Then

$$\begin{aligned}(1+h)^{n+1} &= (1+h)(1+h)^n \geq (1+h)(1+nh), \quad \text{since } 1+h > 0 \\ &= 1+(n+1)h+nh^2 \geq 1+(n+1)h.\end{aligned}$$

For  $h > 0$ , the inequality follows directly from the binomial theorem, since all the other terms appearing in the expansion of  $(1+h)^n$  are positive.

- CHAPTER 3**
1. (i)  $(x+1)/(x+2)$ ; the expression  $f(f(x))$  makes sense only when  $x \neq -1$  and  $x \neq -2$ .  
 (iii)  $1/(1+cx)$  (for  $x \neq -1/c$  if  $c \neq 0$ ).  
 (v)  $(x+y+2)/(x+1)(y+1)$  (for  $x, y \neq -1$ ).  
 (vii) Only  $c = 1$ , since  $f(x) = f(cx)$  implies that  $x = cx$ , and this must be true for at least one  $x \neq 0$ .
  2. (i)  $y \geq 0$  and rational, or  $y \geq 1$ .  
 (iii) 0.  
 (v)  $-1, 0, 1$ .

3. (i)  $\{x : -1 \leq x \leq 1\}$ .  
 (iii)  $\{x : x \neq 1 \text{ and } x \neq 2\}$ .  
 (v)  $\emptyset$ .
4. (i)  $2^{2y}$ .  
 (iii)  $2^{2\sin t} + \sin(2t)$ .
5. (i)  $P \circ s$ .  
 (iii)  $s \circ S$ .  
 (v)  $P \circ P$ .  
 (vii)  $s \circ s \circ s \circ P \circ P \circ P \circ s$ .
11. (a)  $y$ .  
 (b)  $H(y)$ .  
 (c)  $H(y)$ .
12. (a)

	even	odd
even	even	neither
odd	neither	odd

(b)

	even	odd
even	even	odd
odd	odd	even

(c)

	$f$ even	$f$ odd
$g$ even	even	even
$g$ odd	even	odd

21. (d) Let  $g(x) = f(x)$  for  $x \geq 0$  and define  $g$  arbitrarily for  $x < 0$ .  
 (i) Let  $g(x) = h(x) = 1$  and let  $f$  be a function for which  $f(2) \neq f(1) + f(1)$ . Then  $f \circ (g+h) \neq f \circ g + f \circ h$ .  
 (ii)  $[(g+h) \circ f](x) = (g+h)(f(x)) = g(f(x)) + h(f(x)) = (g \circ f)(x) + (h \circ f)(x) = [(g \circ f) + (h \circ f)](x)$ .  
 (iii)  $\frac{1}{f \circ g}(x) = \frac{1}{f(g(x))} = \frac{1}{f}(g(x)) = \left(\frac{1}{f} \circ g\right)(x)$ .  
 (iv) Let  $g(x) = 2$  and let  $f$  be a function for which  $f(\frac{1}{2}) \neq 1/f(2)$ . Then  $1/(f \circ g) \neq f \circ (1/g)$ .

**CHAPTER 4**

1. (i)  $(2, 4)$ .  
 (iii)  $[2, 4]$ .  
 (v)  $(-2, 2)$ .  
 (vii)  $(-\infty, 1] \cup [1, \infty)$ .
3. (i) All points below the graph of  $f(x) = x$ .  
 (iii) All points below the graph of  $f(x) = x^2$ .  
 (v) All points between the graphs of  $f(x) = x + 1$  and  $f(x) = x - 1$ .  
 (vii) A collection of straight lines parallel to the graph of  $f(x) = -x$ , intersecting the horizontal axis at the points  $(n, 0)$  for integers  $n$ .  
 (ix) All points inside the circle of radius 1 and around  $(1, 2)$ .
4. (i) A square with vertices  $(1, 0)$ ,  $(0, 1)$ ,  $(-1, 0)$ , and  $(0, -1)$ .  
 (iii) The union of the graph of  $f(x) = x$  and of  $f(x) = 2 - x$ .  
 (v) The point  $(0, 0)$ .  
 (vii) The circle of radius  $\sqrt{5}$  around  $(1, 0)$ , since  $x^2 - 2x + y^2 = (x - 1)^2 + y^2 - 1$ .
6. (a) Simply observe that the graph of  $f(x) = m(x - a) + b = mx + (b - ma)$  is a straight line with slope  $m$ , which goes through the point  $(a, b)$ . (The important point about this exercise is simply to remember the point slope form.)  
 (b) The straight line through  $(a, b)$  and  $(c, d)$  has slope  $(d - b)/(c - a)$ , so the equation follows from part (a).  
 (c) When  $m = m'$  and  $b \neq b'$ . In that case, there is clearly no number  $x$  with  $f(x) = g(x)$ , while such a number  $x$  always exists if  $m \neq m'$ , namely,  $x = (b' - b)/(m - m')$ .
7. (a) If  $B = 0$  and  $A \neq 0$ , then the set is the vertical straight line formed by all points  $(x, y)$  with  $x = -C/A$ . If  $B \neq 0$ , the set is the graph of  $f(x) = (-A/B)x + (-C/A)$ .  
 (b) The points  $(x, y)$  on the vertical line with  $x = a$  are precisely the ones which satisfy  $1 \cdot x + 0 \cdot y + (-a) = 0$ . The points  $(x, y)$  on the graph of  $f(x) = mx + b$  are precisely the ones which satisfy  $(-m)x + 1 \cdot y + (-b) = 0$ .
11. (i) The graph of  $f$  is symmetric with respect to the vertical axis.  
 (ii) The graph of  $f$  is symmetric with respect to the origin. Equivalently, the part of the graph to the left of the vertical axis is obtained by reflecting first through the vertical axis, and then through the horizontal axis.  
 (iii) The graph of  $f$  lies above or on the horizontal axis.  
 (iv) The graph of  $f$  repeats the part between 0 and  $a$  over and over.
21. (a) The square of the distance from  $(x, x^2)$  to  $(0, \frac{1}{4})$  is

$$\begin{aligned} x^2 + \left(x^2 - \frac{1}{4}\right)^2 &= x^2 + x^4 - \frac{x^2}{2} + \frac{1}{16} \\ &= x^4 + \frac{x^2}{2} + \frac{1}{16} \\ &= (x^2 + \frac{1}{4})^2, \end{aligned}$$

which is the square of the distance from  $(x, x^2)$  to the graph of  $g$ .

- (b) The point  $(x, y)$  satisfies this condition if and only if

$$(x - \alpha)^2 + (y - \beta)^2 = (y - \gamma)^2,$$

or

$$x^2 - 2\alpha x + \alpha^2 + y^2 - 2\beta y + \beta^2 = y^2 - 2\gamma y + \gamma^2,$$

or

$$y = \left( \frac{1}{2\beta - 2\gamma} \right) x^2 + \left( \frac{\alpha}{\gamma - \beta} \right) x + \left( \frac{\alpha^2 + \beta^2 - \gamma^2}{2\beta - 2\gamma} \right).$$

(This solution works only for  $\beta \neq \gamma$ , which is just the condition that  $P$  is not on  $L$ . If  $P$  is on  $L$ , then the solution is the vertical line through  $P$ .)

**CHAPTER 5**

1. (ii)

$$\lim_{x \rightarrow 2} \frac{x^3 - 8}{x - 2} = \lim_{x \rightarrow 2} (x^2 + 2x + 4) = 12.$$

- (iv)

$$\begin{aligned} \lim_{x \rightarrow y} \frac{x^n - y^n}{x - y} &= \lim_{x \rightarrow y} x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1} \\ &= y^{n-1} + y^{n-1} + \cdots + y^{n-1} = ny^{n-1}. \end{aligned}$$

- (vi)

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\sqrt{a+h} - \sqrt{a}}{h} &= \lim_{h \rightarrow 0} \frac{(\sqrt{a+h} - \sqrt{a})(\sqrt{a+h} + \sqrt{a})}{h(\sqrt{a+h} + \sqrt{a})} \\ &= \lim_{h \rightarrow 0} \frac{1}{\sqrt{a+h} + \sqrt{a}} \\ &= \frac{1}{2\sqrt{a}}. \end{aligned}$$

3. (i) It is possible to find  $\delta$  by beginning with the equation

$$x^4 - a^4 = (x - a)(x^3 + ax^2 + a^2x + a^3).$$

If  $|x - a| < 1$ , then  $|x| < 1 + |a|$ , so

$$\begin{aligned} |x^3 + ax^2 + a^2x + a^3| &\leq |x|^3 + |a| \cdot |x|^2 + |a|^2 \cdot |x| + |a|^3 \\ &< (1 + |a|)^3 + |a|(1 + |a|)^2 + |a|^2(1 + |a|) + |a|^3; \end{aligned}$$

therefore we can choose

$$\delta = \min \left( 1, \frac{\varepsilon}{(1 + |a|)^3 + |a|(1 + |a|)^2 + |a|^2(1 + |a|) + |a|^3} \right).$$

It is instructive, and probably easier, to use part (2) of the lemma. This shows that  $|x^4 - a^4| < \varepsilon$  when

$$|x^2 - a^2| < \min \left( 1, \frac{\varepsilon}{2(|a|^2 + 1)} \right),$$

which is true when

$$\begin{aligned}|x - a| &< \min\left(1, \frac{\min\left(1, \frac{\varepsilon}{2(|a|^2 + 1)}\right)}{2(|a| + 1)}\right) \\&= \min\left(1, \frac{\varepsilon}{4(|a|^2 + 1)(|a| + 1)}\right) = \delta.\end{aligned}$$

(ii) By part (3) of the lemma,  $|1/x - 1| < \varepsilon$  when

$$|x - 1| < \min\left(\frac{1}{2}, \frac{\varepsilon}{2}\right) = \delta.$$

(iii) By part (1) of the lemma,  $|(x^4 + 1/x) - 2| < \varepsilon$  when  $|1/x - 1| < \varepsilon/2$  and  $|x^4 - 1| < \varepsilon/2$ . According to parts (i) and (ii) of this problem, this happens when

$$|x - 1| < \min\left(\frac{1}{2}, \frac{\varepsilon}{4}, 1, \frac{\varepsilon}{8 \cdot 2 \cdot 2}\right) = \min\left(\frac{1}{2}, \frac{\varepsilon}{32}\right) = \delta.$$

(v) Let  $\delta = \varepsilon^2$ , since  $0 < |x| < \varepsilon^2$  implies that  $\sqrt{|x|} < \varepsilon$ .

6. (i) We need  $|f(x) - 2| < \varepsilon/2$  and  $|g(x) - 4| < \varepsilon/2$ , so we need

$$0 < |x - 2| < \min\left(\sin^2\left(\frac{\varepsilon^2}{36}\right) + \frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right) = \delta.$$

(iii) We need

$$|g(x) - 4| < \min\left(\frac{|4|}{2}, \frac{\varepsilon|4|^2}{2}\right),$$

so we need

$$0 < |x - 2| < [\min(2, 8\varepsilon)]^2 = \delta.$$

9. Let  $l = \lim_{x \rightarrow a} f(x)$  and define  $g(h) = f(a + h)$ . Then for every  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ . Now, if  $0 < |h| < \delta$ , then  $0 < |(a + h) - a| < \delta$ , so  $|f(a + h) - l| < \varepsilon$ . This inequality can be written  $|g(h) - l| < \varepsilon$ . Thus,  $\lim_{h \rightarrow 0} g(h) = l$ , which can also be written  $\lim_{h \rightarrow 0} f(a + h) = l$ . The same sort of argument shows that if  $\lim_{h \rightarrow 0} f(a + h) = m$ , then  $\lim_{x \rightarrow a} f(x) = m$ . So either limit exists if the other does, and in this case they are equal.
10. (a) Intuitively, we can get  $f(x)$  as close to  $l$  as we like if and only if we can get  $f(x) - l$  as close to 0 as we like. The formal proof is so trivial that it takes a bit of work to make it look like a proof at all. To be very precise, suppose  $\lim_{x \rightarrow a} f(x) = l$  and let  $g(x) = f(x) - l$ . Then for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ . This last inequality can be written  $|g(x) - 0| < \varepsilon$ , so  $\lim_{x \rightarrow a} g(x) = 0$ . The argument in the other direction is similarly uninteresting.

- (b) Intuitively, making  $x$  close to  $a$  is the same as making  $x - a$  close to 0. Formally: Suppose that  $\lim_{x \rightarrow a} f(x) = l$ , and let  $g(x) = f(x - a)$ . Then for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ . Now, if  $0 < |y| < \delta$ , then  $0 < |(y + a) - a| < \delta$ , so  $|f(y + a) - l| < \varepsilon$ . But this last inequality can be written  $|g(y) - l| < \varepsilon$ . So  $\lim_{y \rightarrow 0} g(y) = l$ . The argument in the reverse direction is similar.
- (c) Intuitively,  $x$  is close to 0 if and only if  $x^3$  is. Formally: Let  $\lim_{x \rightarrow 0} f(x) = l$ . For every  $\varepsilon > 0$  there is a  $\delta > 0$  such that if  $0 < |x| < \delta$ , then  $|f(x) - l| < \varepsilon$ . Then if  $0 < |x| < \min(1, \delta)$ , we have  $0 < |x^3| < \delta$ , so  $|f(x^3) - l| < \varepsilon$ . Thus,  $\lim_{x \rightarrow 0} f(x^3) = l$ . On the other hand, if we assume that  $\lim_{x \rightarrow 0} f(x^3)$  exists, say  $\lim_{x \rightarrow 0} f(x^3) = m$ , then for all  $\varepsilon > 0$  there is a  $\delta$  such that if  $0 < |x| < \delta$ , then  $|f(x^3) - m| < \varepsilon$ . Then if  $0 < |x| < \delta^3$ , we have  $0 < |\sqrt[3]{x}| < \delta$ , so  $|f([\sqrt[3]{x}]^3) - m| < \varepsilon$ , or  $|f(x) - m| < \varepsilon$ . Thus  $\lim_{x \rightarrow 0} f(x) = m$ .
- (d) Let  $f(x) = 1$  for  $x \geq 0$ , and  $f(x) = -1$  for  $x < 0$ . Then  $\lim_{x \rightarrow 0} f(x^2) = 1$ , but  $\lim_{x \rightarrow 0} f(x)$  does not exist.
17. (a) The function  $f(x) = 1/x$  cannot approach a limit at 0, since it becomes arbitrarily large near 0. In fact, no matter what  $\delta > 0$  may be, there is some  $x$  satisfying  $0 < |x| < \delta$ , but  $1/x > |l| + \varepsilon$ , namely,  $x = \min(\delta, 1/(|l| + \varepsilon))$ . This  $x$  does not satisfy  $|1/(x - l)| < \varepsilon$ .
- (b) No matter what  $\delta > 0$  may be, there is some  $x$  satisfying  $0 < |x - 1| < \delta$ , but  $1/(x - 1) > |l| + \varepsilon$ , namely,  $x = \min(1 + \delta, 1 + 1/(|l| + \varepsilon))$ . This  $x$  does not satisfy  $|1/(x - 1) - l| < \varepsilon$ . (It is also possible to apply Problem 10(b):  $\lim_{x \rightarrow 0} 1/x = \lim_{x \rightarrow 1} 1/(x - 1)$  if the latter exists, so this limit does not exist, because of part (a).)
25. (i) This is the usual definition, simply calling the numbers  $\delta$  and  $\varepsilon$ , instead of  $\varepsilon$  and  $\delta$ .
- (ii) This is a minor modification of (i): if the condition is true for *all*  $\delta > 0$ , then it applies to  $\delta/2$ , so there is an  $\varepsilon > 0$  such that if  $0 < |x - a| < \varepsilon$ , then  $|f(x) - l| \leq \delta/2 < \delta$ .
- (iii) This is a similar modification: apply it to  $\delta/5$  to obtain (i).
- (iv) This is also a modification: it says the same thing as (i), since  $\varepsilon/10 > 0$ , and it is only the existence of *some*  $\varepsilon > 0$  that is in question.
29. If  $\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^-} f(x) = l$ , then for every  $\varepsilon > 0$  there are  $\delta_1, \delta_2 > 0$  such that, for all  $x$ ,

$$\begin{aligned} \text{if } a < x < a + \delta_1, \text{ then } |f(x) - l| < \varepsilon, \\ \text{if } a - \delta_2 < x < a, \text{ then } |f(x) - l| < \varepsilon. \end{aligned}$$

Let  $\delta = \min(\delta_1, \delta_2)$ . If  $0 < |x - a| < \delta$ , then either  $a - \delta_2 < a - \delta < x < a$  or else  $a < x < a + \delta < a + \delta_1$ , so  $|f(x) - l| < \varepsilon$ .

30. (i) If  $l = \lim_{x \rightarrow 0^+} f(x)$ , then for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that  $|f(x) - l| < \varepsilon$  for  $0 < x < \delta$ . If  $-\delta < x < 0$ , then  $0 < -x < \delta$ , so  $|f(-x) - l| < \varepsilon$ . Thus  $\lim_{x \rightarrow 0^-} f(-x) = l$ . Similarly, if  $\lim_{x \rightarrow 0^-} f(x)$  exists, then  $\lim_{x \rightarrow 0^+} f(x)$  exists and has the same value. (Intuitively,  $x$  is close to 0 and positive if and only if  $-x$  is close to 0 and negative.)
- (ii) If  $l = \lim_{x \rightarrow 0^+} f(x)$ , then for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that  $|f(x) - l| < \varepsilon$  for  $0 < x < \delta$ . So if  $0 < |x| < \delta$ , then  $|f(|x|) - l| < \varepsilon$ . Thus  $\lim_{x \rightarrow 0} f(|x|) = l$ . The reverse direction is similar. (Intuitively, if  $x$  is close to 0, then  $|x|$  is close to 0 and positive.)
- (iii) If  $l = \lim_{x \rightarrow 0^+} f(x)$ , then for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that  $|f(x) - l| < \varepsilon$  for  $0 < x < \delta$ . If  $0 < |x| < \sqrt{\delta}$ , then  $0 < x^2 < \delta$ , so  $|f(x^2) - l| < \varepsilon$ . Thus  $\lim_{x \rightarrow 0} f(x^2) = l$ . The reverse direction is similar. (Intuitively, if  $x$  is close to 0, then  $x^2$  is close to 0 and positive.)
34. If  $l = \lim_{x \rightarrow \infty} f(x)$ , then for every  $\varepsilon > 0$  there is some  $N$  such that  $|f(x) - l| < \varepsilon$  for  $x > N$ , and we can clearly assume that  $N > 0$ . Now, if  $0 < x < 1/N$ , then  $1/x > N$ , so  $|f(1/x) - l| < \varepsilon$ . Thus  $\lim_{x \rightarrow 0^+} f(1/x) = l$ . The reverse direction is similar.

**CHAPTER 6**

1. (i)  $F(x) = x + 2$  for all  $x$ .  
 (iii)  $F(x) = 0$  for all  $x$ .

**CHAPTER 7**

1. (i) Bounded above and below; minimum value 0; no maximum value.  
 (iii) Bounded below but not above; minimum value 0.  
 (v) Bounded above and below. It is understood that  $a > -1$  (so that  $-a - 1 < a + 1$ ). If  $-1 < a \leq -\frac{1}{2}$ , then  $a \leq -a - 1$ , so  $f(x) = a + 2$  for all  $x$  in  $(-a - 1, a + 1)$ , so  $a + 2$  is the maximum and minimum value. If  $-\frac{1}{2} < a \leq 0$ , then  $f$  has the minimum value  $a^2$ , and if  $a \geq 0$ , then  $f$  has the minimum value 0. Since  $a + 2 > (a + 1)^2$  only for  $[-1 - \sqrt{5}]/2 < a < [1 + \sqrt{5}]/2$ , when  $a \geq -\frac{1}{2}$  the function  $f$  has a maximum value only for  $a \leq [1 + \sqrt{5}]/2$  (the maximum value being  $a + 2$ ).  
 (vii) Bounded above and below; maximum value 1; minimum value 0.  
 (ix) Bounded above and below; maximum value 1; minimum value -1.  
 (xi)  $f$  has a maximum and minimum value, since  $f$  is continuous.
2. (i)  $n = -2$ , since  $f(-2) < 0 < f(-1)$ .  
 (iii)  $n = -1$ , since  $f(-1) = -1 < 0 < f(0)$ .
3. (i) If  $f(x) = x^{179} + 163/(1 + x^2 + \sin^2 x)$ , then  $f$  is continuous on  $\mathbb{R}$  and  $f(1) > 0$ , while  $f(-2) < 0$ , so  $f(x) = 0$  for some  $x$  in  $(-2, 1)$ .
5.  $f$  is constant, for if  $f$  took on two different values, then  $f$  would take on all values in between, which would include irrational values.
7. (1)  $f(x) = x$ ;  
 (2)  $f(x) = -x$ ;

- (3)  $f(x) = |x|$ ;  
 (4)  $f(x) = -|x|$ .
10. Apply Theorem 1 to  $f - g$ .
11. If  $f(0) = 0$  or  $f(1) = 1$ , choose  $x = 0$  or 1. If  $f(0) > 0 = I(0)$  and  $f(1) < 1 = I(1)$ , then Problem 10 applied to  $f$  and  $I$  implies that  $f(x) = x$  for some  $x$ .
- CHAPTER 8**
1. (i) 1 is the greatest element, and the greatest lower bound is 0, which is not in the set.  
 (iii) 1 is the greatest element, and 0 is the least element.  
 (v) Since  $\{x : x^2 + x + 1 \geq 0\} = \mathbf{R}$ , there is no least upper bound or greatest lower bound.  
 (vii) Since  $\{x : x < 0 \text{ and } x^2 + x - 1 < 0\} = (-1 - \sqrt{5})/2, 0)$ , the greatest lower bound is  $(-1 - \sqrt{5})/2$ , and the least upper bound is 0; neither belongs to the set.
  2. (a) Since  $A \neq \emptyset$ , there is some  $x$  in  $A$ . Then  $-x$  is in  $-A$ , so  $-A \neq \emptyset$ . Since  $A$  is bounded above, there is some  $y$  such that  $y \geq x$  for all  $x$  in  $A$ . Then  $-y \leq -x$  for all  $x$  in  $A$ , so  $-y \leq z$  for all  $z$  in  $-A$ , so  $-A$  is bounded below. Let  $\alpha = \sup(-A)$ . Then  $\alpha$  is an upper bound for  $-A$ , so, reversing the argument just given,  $-\alpha$  is a lower bound for  $A$ . Moreover, if  $\beta$  is any lower bound for  $A$ , then  $-\beta$  is an upper bound for  $-A$ , so  $-\beta \geq \alpha$ , so  $\beta \leq -\alpha$ . Thus  $-\alpha$  is the greatest lower bound for  $A$ .
  5. (a) If  $l$  is the largest integer with  $l \leq x$ , then  $l+1 > x$ , but  $l+1 \leq x+1 < y$ . So we can let  $k = l+1$ . (Proof that a largest such integer  $l$  exists: Since  $\mathbf{N}$  is not bounded above, there is some natural number  $n$  with  $-n < x < n$ . There are consequently only a finite number of integers  $l$  with  $-n \leq l \leq x$ . Pick the largest.)  
 (b) Since  $y - x > 0$ , there is some natural number  $n$  with  $1/n < y - x$ . Since  $ny - nx > 1$ , there is, by part (a), an integer  $k$  with  $nx < k < ny$ , which means that  $x < k/n < y$ .  
 (c) Choose  $r + \sqrt{2}(s - r)/2$ .  
 (d) By part (b), there is a rational number  $r$  with  $x < r < y$ , and therefore a rational number  $s$  with  $x < r < s < y$ . Apply part (c) to  $r < s$ .
  10. Let  $k$  be the largest integer  $\leq x/\alpha$  (the solution to Problem 5 shows that such a  $k$  exists), and let  $x' = x - k\alpha \geq 0$ . If  $x - k\alpha = x' \geq \alpha$ , then  $x \geq (k+1)\alpha$ , so  $k+1 \leq x/\alpha$ , contradicting the choice of  $k$ . So  $0 \leq x' < \alpha$ .
  12. (a) Since any  $y$  in  $B$  satisfies  $y \geq x$  for all  $x$  in  $A$ , any  $y$  in  $B$  is an upper bound for  $A$ , so  $y \geq \sup A$ .  
 (b) Part (a) shows that  $\sup A$  is a lower bound for  $B$ , so  $\sup A \leq \inf B$ .
  13. Since  $x \leq \sup A$  and  $y \leq \sup B$  for every  $x$  in  $A$ , and  $y$  in  $B$ , it follows that  $x + y \leq \sup A + \sup B$ . Thus,  $\sup A + \sup B$  is an upper bound for  $A + B$ , so  $\sup(A + B) \leq \sup A + \sup B$ . If  $x$  and  $y$  are chosen in  $A$  and  $B$ , respectively, so that  $\sup A - x < \varepsilon/2$  and  $\sup B - y < \varepsilon/2$ , then  $\sup A + \sup B - (x + y) < \varepsilon$ . Hence,

$$\sup(A + B) \geq x + y > \sup A + \sup B - \varepsilon.$$

## CHAPTER 9

1. (a)

$$\begin{aligned}f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \frac{\frac{1}{a+h} - \frac{1}{a}}{h} \\&= \lim_{h \rightarrow 0} \frac{-1}{a(a+h)} = -\frac{1}{a^2}.\end{aligned}$$

(b) The tangent line through  $(a, 1/a)$  is the graph of

$$\begin{aligned}g(x) &= \frac{-1}{a^2}(x-a) + \frac{1}{a} \\&= \frac{-x}{a^2} + \frac{2}{a}.\end{aligned}$$

If  $f(x) = g(x)$ , then

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

or

$$x^2 - 2ax + a^2 = 0,$$

so  $x = a$ .

2. (a)

$$\begin{aligned}f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \frac{\frac{1}{(a+h)^2} - \frac{1}{a^2}}{h} \\&= \lim_{h \rightarrow 0} \frac{(-2ah - h^2)}{ha^2(a+h)^2} = -\frac{2}{a^3}.\end{aligned}$$

(b) The tangent line through  $(a, 1/a^2)$  is the graph of

$$\begin{aligned}g(x) &= -\frac{2}{a^3}(x-a) + \frac{1}{a^2} \\&= -\frac{2x}{a^3} + \frac{3}{a^2}.\end{aligned}$$

If  $f(x) = g(x)$ , then

$$\frac{1}{x^2} = -\frac{2x}{a^3} + \frac{3}{a^2},$$

or

$$2x^3 - 3ax^2 + a^3 = 0,$$

or

$$0 = (x-a)(2x^2 - ax - a^2) = (x-a)(2x+a)(x-a).$$

So  $x = a$  or  $x = -a/2$ ; the point  $(-a/2, 4/a^2)$  lies on the opposite side of the vertical axis from  $(a, 1/a^2)$ .

3.

$$\begin{aligned}
 f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \frac{\sqrt{a+h} - \sqrt{a}}{h} \\
 &= \lim_{h \rightarrow 0} \frac{(\sqrt{a+h} - \sqrt{a})(\sqrt{a+h} + \sqrt{a})}{h(\sqrt{a+h} + \sqrt{a})} = \lim_{h \rightarrow 0} \frac{h}{h(\sqrt{a+h} + \sqrt{a})} \\
 &= \frac{1}{2\sqrt{a}}.
 \end{aligned}$$

4. Conjecture:  $S_n'(x) = nx^{n-1}$ . Proof:

$$\begin{aligned}
 S_n'(x) &= \lim_{h \rightarrow 0} \frac{S_n(x+h) - S_n(x)}{h} = \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\sum_{j=0}^n \binom{n}{j} x^{n-j} h^j - x^n}{h} \\
 &= \lim_{h \rightarrow 0} \sum_{j=1}^n \binom{n}{j} x^{n-j} h^{j-1} \\
 &= \binom{n}{1} x^{n-1} = nx^{n-1}, \quad \text{since } \lim_{h \rightarrow 0} h^{j-1} = 0 \text{ for } j > 1.
 \end{aligned}$$

5.  $f'(x) = 0$  for  $x$  not an integer, and  $f'(x)$  is not defined if  $x$  is an integer.

6. (a)

$$\begin{aligned}
 g'(x) &= \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h} = \lim_{h \rightarrow 0} \frac{[f(x+h) + c] - [f(x) + c]}{h} \\
 &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = f'(x).
 \end{aligned}$$

(b)

$$\begin{aligned}
 g'(x) &= \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h} = \lim_{h \rightarrow 0} \frac{cf(x+h) - cf(x)}{h} \\
 &= c \cdot \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = cf'(x).
 \end{aligned}$$

7. (a)  $f'(9) = 3 \cdot 9^2$ ;  $f'(25) = 3 \cdot (25)^2$ ;  $f'(36) = 3 \cdot (36)^2$ .(b)  $f'(3^2) = f'(9) = 3 \cdot 9^2$ ;  $f'(5^2) = f'(25) = 3 \cdot (25)^2$ ;  $f'(6^2) = f'(36) = 3 \cdot (36)^2$ .(c)  $f'(a^2) = 3(a^2)^2 = 3a^4$ ;  $f'(x^2) = 3(x^2)^2 = 3x^4$ .(d)  $f'(x^2) = 3x^4$ ; but  $g(x) = x^6$ , so  $g'(x) = 6x^5$ .

8. (a)

$$\begin{aligned}
 g'(x) &= \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h} = \lim_{h \rightarrow 0} \frac{f(x+h+c) - f(x+c)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{f([x+c]+h) - f(x+c)}{h} = f'(x+c).
 \end{aligned}$$

(b)

$$\begin{aligned}
 g'(x) &= \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h} = \lim_{h \rightarrow 0} \frac{f(cx+ch) - f(cx)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{c[f(cx+ch) - f(cx)]}{ch} = \lim_{k \rightarrow 0} \frac{c[f(cx+k) - f(cx)]}{k} \\
 &= c \cdot \lim_{k \rightarrow 0} \frac{f(cx+k) - f(cx)}{k} = c \cdot f'(cx).
 \end{aligned}$$

(Compare the manipulations in this calculation with Problem 5-14.)

- (c) If  $g(x) = f(x+a)$ , then  $g'(x) = f'(x+a)$ , by part (a). But  $g = f$ , so  $f'(x) = g'(x) = f'(x+a)$  for all  $x$ , which means that  $f'$  is periodic, with period  $a$ .
9. (i) If  $g(x) = x^5$ , then  $g'(x) = 5x^4$ . Now  $f(x) = g(x+3)$ , so by Problem 8(a),  $f'(x) = g'(x+3) = 5(x+3)^4$ . And  $f'(x+3) = 5(x+6)^4$ .  
(ii)  $f(x) = (x-3)^5$ , so  $f'(x) = 5(x-3)^4$ , as in part (i). And  $f'(x+3) = 5 \cdot 0^4 = 0$ .  
(iii)  $f(x) = (x+2)^7$ , so  $f'(x) = 7(x+2)^6$ , as in part (i). And  $f'(x+3) = 7(x+5)^6$ .
10. If  $f(x) = g(t+x)$ , then  $f'(x) = g'(t+x)$ , by Problem 8(a). If  $f(t) = g(t+x)$ , then  $f'(t) = g'(t+x)$ , by Problem 8(a), so  $f'(x) = g'(2x)$ .
11. (a) If  $s(t) = ct^2$ , then  $s'(t) = 2ct$ , and there is no number  $k$  such that  $s'(t) = ks(t)$  [that is,  $2ct = kct^2$ ] for all  $t$ .  
(By the way, at this point we do not know any nonzero function  $f$  for which  $f'$  is proportional to  $f$ . After Chapter 18 it might be amusing to determine what the world would be like if Galileo were correct.)  
(b) (i) If  $s(t) = (a/2)t^2$ , then  $s'(t) = at$ , so  $s''(t) = a$ .  
(ii)  $[s'(t)]^2 = (at)^2 = 2a \cdot (a/2)t^2 = 2as(t)$ .  
(c) The chandelier falls  $s(t) = 16t^2$  feet in  $t$  seconds, so it falls 400 feet in  $t$  seconds, if  $400 = 16t^2$ , or  $t = 5$ . After 5 seconds the velocity will be  $s'(5) = 5a = 5 \cdot 32 = 160$  feet per second. The speed was half this amount when  $80 = s'(t) = 32t$ , or  $t = \frac{5}{2}$ .
21. (a) This is another way of writing the definition (see Problem 5-9).  
(b) This follows from Problem 5-11, applied to the functions  $\alpha(h) = [f(a+h) - f(a)]/h$  and  $\beta(h) = [g(a+h) - g(a)]/h$ .
26. (i)  $f''(x) = 6x$ .  
(iii)  $f''(x) = 4x^3$ .
30. (i) means that  $f'(a) = na^{n-1}$  if  $f(x) = x^n$ .  
(iii) means that  $g'(a) = f'(a)$  if  $g(x) = f(x) + c$ .  
(v) means the same as (iii).  
(vii) means that  $g'(b) = f'(b+a)$  if  $g(x) = f(x+a)$ .  
(ix) means that  $g'(b) = cf'(cb)$  if  $g(x) = f(cx)$ .

1. (i)  $(1+2x) \cdot \cos(x+x^2)$ .  
(iii)  $(-\sin x) \cdot \cos(\cos x)$ .

(v)  $\cos\left(\frac{\cos x}{x}\right) \cdot \frac{-x \sin x - \cos x}{x^2}.$

(vii)  $(\cos(x + \sin x)) \cdot (1 + \cos x).$

2. (i)  $(\cos((x+1)^2(x+2))) \cdot [2(x+1)(x+2) + (x+1)^2].$

(iii)  $[2 \sin((x+\sin x)^2) \cos((x+\sin x)^2)] \cdot 2(x+\sin x)(1+\cos x).$

(v)  $(\cos(x \sin x)) \cdot (\sin x + x \cos x) + (\cos(\sin x^2)(\cos x^2)) \cdot 2x.$

(vii)  $(2 \sin x \cos x \sin x^2 \sin^2 x^2) + (2x \cos x^2 \sin^2 x \sin^2 x^2)$   
 $+ (4x \sin x^2 \cos x^2 \sin^2 x \sin x^2).$

(ix)  $6(x + \sin^5 x)^5(1 + 5 \sin^4 x \cos x).$

(xi)  $\cos(\sin^7 x^7 + 1)^7 \cdot 7(\sin^7 x^7 + 1)^6 \cdot (7 \sin^6 x^7 \cdot \cos x^7 \cdot 7x^6).$

(xiii)  $\cos(x^2 + \sin(x^2 + \sin x^2)) \cdot [(2x + \cos(x^2 + \sin x^2) \cdot (2x + 2x \cos x^2)).]$

(xv)  $\frac{(1 + \sin x)(2x \cos x^2 \cdot \sin^2 x + \sin x^2 \cdot 2 \sin x \cos x) - \cos x \sin x^2 \sin^2 x}{(1 + \sin x)^2}.$

(xvii)  $\cos\left(\frac{x^3}{\sin\left(\frac{x^3}{\sin x}\right)}\right).$

$$\frac{3x^2 \sin\left(\frac{x^3}{\sin x}\right) - x^3 \cos\left(\frac{x^3}{\sin x}\right) \cdot \left(\frac{3x^2 \sin x - x^3 \cos x}{\sin^2 x}\right)}{\sin^2\left(\frac{x^3}{\sin x}\right)}.$$

4. (i)  $-\frac{(x+1)^2}{(x+2)^2}.$

(iii)  $2x^2.$

5. (i)  $-x^2.$

(iii)  $17.$

6. (i)  $f'(x) = g'(x + g(a)).$

(iii)  $f'(x) = g'(x + g(x)) \cdot (1 + g'(x)).$

(v)  $f'(x) = g(a).$

7. (a)  $A'(t) = 2\pi r(t)r'(t).$  Since  $r'(t) = 4$  for that  $t$  with  $r(t) = 6,$  it follows that  $A'(t) = 2\pi \cdot 6 \cdot 4 = 48\pi$  when  $r(t) = 6.$

(b) If  $V(t)$  is the volume at time  $t,$  then  $V(t) = 4\pi r(t)^3/3,$  so  $V'(t) = 4\pi r(t)^2 r'(t) = 4\pi \cdot 6^2 \cdot 4 = 576\pi$  when  $r(t) = 6.$

(c) First method: Since  $A'(t) = 2\pi r(t)r'(t),$  and  $A'(t) = 5$  for  $r(t) = 3,$  it follows that

$$r'(t) = \frac{A'(t)}{2\pi r(t)} = \frac{5}{6\pi} \quad \text{when } r(t) = 3.$$

Thus

$$\begin{aligned} V'(t) &= 4\pi r(t)^2 r'(t) \\ &= 4\pi \cdot 9 \cdot \frac{5}{6\pi} \\ &= 30 \quad \text{when } r(t) = 3. \end{aligned}$$

To apply the second method, we first note that if

$$f(t) = A(t)^{3/2} = \sqrt{A(t)^3},$$

then, using Problem 9-3 and the Chain Rule,

$$\begin{aligned} f'(t) &= \frac{1}{2\sqrt{A(t)^3}} \cdot 3A(t)^2 A'(t) \\ &= \frac{1}{2A(t)^{3/2}} \cdot 3A(t)^2 A'(t) \\ &= \frac{3}{2} A(t)^{1/2} A'(t) \quad (\text{just as we might have guessed}). \end{aligned}$$

Now

$$\begin{aligned} V(t) &= \frac{4\pi r(t)^3}{3} = \frac{4\pi [r(t)^2]^{3/2}}{3} \\ &= \frac{4[\pi r(t)^2]^{3/2}}{3\pi^{1/2}} \\ &= \frac{4A(t)^{3/2}}{3\pi^{1/2}}. \end{aligned}$$

So

$$\begin{aligned} V'(t) &= \frac{4}{3\pi^{1/2}} \cdot \frac{3}{2} \sqrt{A(t)} A'(t) \\ &= \frac{2}{\pi^{1/2}} \cdot \pi^{1/2} r(t) A'(t) \\ &= 2 \cdot 3 \cdot 5 = 30. \end{aligned}$$

10. (i)  $(f \circ h)'(0) = f'(h(0)) \cdot h'(0) = f'(3) \cdot \sin^2(\sin 1) = [6 \sin \frac{1}{3} - \cos \frac{1}{3}] \sin^2(\sin 1).$   
(ii)  $\alpha'(x^2) = h'(x^4) \cdot 2x^2 = \sin^2(\sin(x^4 + 1)) \cdot 2x^2.$

12. The Chain Rule implies that

$$\begin{aligned} \left(\frac{1}{g}\right)'(x) &= (f \circ g)'(x) = f'(g(x)) \cdot g'(x) \\ &= -\frac{1}{g(x)^2} \cdot g'(x). \end{aligned}$$

33. (i)  $\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx} = (\cos y) \cdot (1 + 2x) = (\cos(x + x^2)) \cdot (1 + 2x).$

$$(iii) \frac{dz}{dx} = \frac{dz}{du} \cdot \frac{du}{dx} = (-\sin u) \cdot (\cos x) = (-\cos(\sin x)) \cdot (\cos x).$$

- CHAPTER 11**
1. (i)  $0 = f'(x) = 3x^2 - 2x - 8$  for  $x = 2$  and  $x = -\frac{4}{3}$ , both of which are in  $[-2, 2]$ ;  
 $f(-2) = 5$ ,  $f(2) = -11$ ,  $f(-\frac{4}{3}) = \frac{203}{27}$ ;  
maximum  $= \frac{203}{27}$ , minimum  $= -11$ .
  - (iii)  $0 = f'(x) = 12x^3 - 24x^2 + 12x = 12x(x^2 - 2x + 1)$  for  $x = 0$  and  $x = 1$ , of which only 0 is in  $[-\frac{1}{2}, \frac{1}{2}]$ ;  
 $f(-\frac{1}{2}) = \frac{43}{16}$ ,  $f(\frac{1}{2}) = \frac{11}{16}$ ,  $f(0) = 0$ ;  
maximum  $= \frac{43}{16}$ , minimum  $= 0$ .
  - (v)  $0 = f'(x) =$ 

$$\frac{x^2 + 1 - (x + 1)2x}{(x^2 + 1)^2} = \frac{1 - 2x - x^2}{(x^2 + 1)^2}$$
for  $x = -1 + \sqrt{2}$  and  $x = -1 - \sqrt{2}$ , of which only  $-1 + \sqrt{2}$  is in  $[-1, \frac{1}{2}]$ ;  
 $f(-1) = 0$ ,  $f(\frac{1}{2}) = \frac{6}{5}$ ,  $f(-1 + \sqrt{2}) = (1 + \sqrt{2})/2$ ;  
maximum  $= (1 + \sqrt{2})/2$ , minimum  $= 0$ .
  2. (i)  $-\frac{4}{3}$  is a local maximum point, and 2 is a local minimum point.
  - (iii) 0 is a local minimum point, and there are no local maximum points.
  - (v)  $-1 + \sqrt{2}$  is a local maximum point, and  $-1 - \sqrt{2}$  is a local minimum point.
  4. (a) Notice that  $f$  actually has a minimum value, since  $f$  is a polynomial function of even degree. The minimum occurs at a point  $x$  with

$$0 = f'(x) = 2 \sum_{i=1}^n (x - a_i),$$

so  $x = (a_1 + \dots + a_n)/n$ .

5. (i) 3 and 7 are local maximum points, and 5 and 9 are local minimum points.
- (iii) All irrational  $x > 0$  are local minimum points, and all irrational  $x < 0$  are local maximum points.
- (v)  $x$  is a local maximum (minimum) point if the decimal expansion contains (does not contain) a 5.
8. If  $f(x)$  is the total length of the path, then

$$f(x) = \sqrt{x^2 + a^2} + \sqrt{(1-x)^2 + b^2}.$$

The positive function  $f$  clearly has a minimum, since  $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(x) = \infty$ , and  $f$  is differentiable everywhere, so the minimum occurs at a point  $x$  with  $f'(x) = 0$ . Now,  $f'(x) = 0$  when

$$\frac{x}{\sqrt{x^2 + a^2}} - \frac{(1-x)}{\sqrt{(1-x)^2 + b^2}} = 0.$$

This equation says that  $\cos \alpha = \cos \beta$ .

It is also possible to notice that  $f(x)$  is equal to the sum of the lengths of the dashed line segment and the line segment from  $(x, 0)$  to  $(1, b)$ . This is shortest when the two line segments lie along a line (because of Problem 4-9(b), if a rigorous reason is required); a little plane geometry shows that this happens when  $\alpha = \beta$ .

9. If  $x$  is the length of one side of a rectangle of perimeter  $P$ , then the length of the other side is  $(P - 2x)/2$ , so the area is

$$A(x) = \frac{x(P - 2x)}{2}.$$

So the rectangle with greatest area occurs when  $x$  is the maximum point for  $f$  on  $(0, P/2)$ . Since  $A$  is continuous on  $[0, P/2]$ , and  $A(0) = A(P/2) = 0$ , and  $A(x) > 0$  for  $x$  in  $(0, P/2)$ , the maximum exists. Since  $A$  is differentiable on  $(0, P/2)$ , the minimum point  $x$  satisfies

$$\begin{aligned} 0 = A'(x) &= \frac{P - 2x}{2} - x \\ &= \frac{P - 4x}{2}, \end{aligned}$$

so  $x = P/4$ .

10. Let  $S(r)$  be the surface area of the right circular cylinder of volume  $V$  with radius  $r$ . Since

$$V = \pi r^2 h \quad \text{where } h \text{ is the height,}$$

we have  $h = V/\pi r^2$ , so

$$\begin{aligned} S(r) &= 2\pi r^2 + 2\pi r h \\ &= 2\pi r^2 + \frac{2V}{r}. \end{aligned}$$

We want the minimum point of  $S$  on  $(0, \infty)$ ; this exists, since  $\lim_{r \rightarrow 0} S(r) = \lim_{r \rightarrow \infty} S(r) = \infty$ . Since  $S$  is differentiable on  $(0, \infty)$ , the minimum point  $r$  satisfies

$$\begin{aligned} 0 = S'(r) &= 4\pi r - \frac{2V}{r^2} \\ &= \frac{4\pi r^3 - 2V}{r^2}, \end{aligned}$$

or

$$r = \sqrt[3]{\frac{V}{2\pi}}.$$

19. 1 is a local maximum point, and 3 is a local minimum point.  
25. (a) We have

$$\begin{aligned} \frac{f(b) - f(a)}{b - a} &= f'(x) \quad \text{for some } x \text{ in } (a, b) \\ &\geq M, \end{aligned}$$

so  $f(b) - f(a) \geq M(b - a)$ .

(b) We have

$$\begin{aligned}\frac{f(b) - f(a)}{b - a} &= f'(x) \quad \text{for some } x \text{ in } (a, b) \\ &\leq m,\end{aligned}$$

so  $f(b) - f(a) \leq m(b - a)$ .

(c) If  $|f'(x)| \leq M$  for all  $x$  in  $[a, b]$ , then  $-M \leq f(x) \leq M$ , so

$$f(a) - M(b - a) \leq f(b) \leq f(a) + M(b - a),$$

or

$$|f(b) - f(a)| \leq M(b - a).$$

28. (a)  $f(x) = -\cos x + a$  for some number  $a$  (because  $f(x) = -\cos x$  is one such function, and any two such functions differ by a constant function).  
 (b)  $f'(x) = x^4/4 + a$  for some number  $a$ , so  $f(x) = x^5/20 + ax + b$  for some numbers  $a$  and  $b$ .  
 (c)  $f''(x) = x^2 + x^3/3 + a$  for some  $a$ , so  $f'(x) = x^3/6 + x^4/12 + ax + b$  for some  $a$  and  $b$ , so  $f(x) = x^4/24 + x^5/60 + ax^2/2 + bx + c$  for some numbers  $a$ ,  $b$ , and  $c$ . Equivalently, and more simply,  $f(x) = x^4/24 + x^5/60 + ax^2 + bx + c$  for some numbers  $a$ ,  $b$ , and  $c$ .
29. (a) Since  $s''(t) = -32$ , we have  $s'(t) = -32t + \alpha$  for some  $\alpha$ , so  $s(t) = -16t^2 + \alpha t + \beta$  for some  $\alpha$  and  $\beta$ .  
 (b) Clearly,  $s(0) = 0 + 0 + \beta$  and  $s'(0) = 0 + \alpha$ . Thus,  $\alpha = v_0$  and  $\beta = s_0$ .  
 (c) In this case,  $s_0 = 0$  and  $v_0 = v$ , so  $s(t) = -16t^2 + vt$ . The maximum value of  $s$  occurs when  $0 = s'(t) = -32t + v$ , or  $t = v/32$ , so the maximum value is

$$\begin{aligned}s\left(\frac{v}{32}\right) &= -16\left(\frac{v}{32}\right)^2 + v \cdot \left(\frac{v}{32}\right) \\ &= \frac{-v^2}{64} + \frac{v^2}{32} \\ &= \frac{v^2}{64}.\end{aligned}$$

At that moment the velocity is clearly 0, but the acceleration is  $-32$  (as at any time). The weight hits the ground at time  $t > 0$  when

$$0 = s(t) = -16t^2 + vt,$$

or  $t = v/16$  (it takes as long to fall back down as it took to reach the top). The velocity is then

$$\begin{aligned}s'(v/16) &= -32\left(\frac{v}{16}\right) + v \\ &= -v\end{aligned}$$

(the same velocity with which it was initially moving upward).

44. Apply the Mean Value Theorem to  $f(x) = \sqrt{x}$  on  $[64, 66]$ :

$$\frac{\sqrt{66} - \sqrt{64}}{66 - 64} = f'(x) = \frac{1}{2\sqrt{x}} \quad \text{for some } x \text{ in } [64, 66].$$

Since  $64 < x < 66$ , we have  $8 < \sqrt{x} < 9$ , so

$$\frac{1}{2 \cdot 9} < \frac{\sqrt{66} - 8}{2} < \frac{1}{2 \cdot 8}.$$

48. l'Hôpital's Rule does not lead to the equation

$$\lim_{x \rightarrow 1} \frac{3x^2 + 1}{2x - 3} = \lim_{x \rightarrow 1} \frac{6x}{2}$$

because  $\lim_{x \rightarrow 1} 3x^2 + 1 \neq 0$ .

49. (i)

$$\lim_{x \rightarrow 0} \frac{x}{\tan x} = \lim_{x \rightarrow 0} \frac{1}{\sec^2 x} = \lim_{x \rightarrow 0} \cos^2 x = 1.$$

- (ii)

$$\lim_{x \rightarrow 0} \frac{\cos^2 x - 1}{x^2} = \lim_{x \rightarrow 0} \frac{-2 \sin x \cos x}{2x} = -1.$$

#### CHAPTER 12

1. (i)  $f^{-1}(x) = (x - 1)^{1/3}$ . (If  $y = f^{-1}(x)$ , then  $x = f(y) = y^3 + 1$ , so  $y = (x - 1)^{1/3}$ .)  
 (iii)  $f^{-1} = f$ . (If  $y = f^{-1}(x)$ , then

$$x = f(y) = \begin{cases} y, & y \text{ rational} \\ -y, & y \text{ irrational}; \end{cases}$$

since  $\pm y$  is rational or irrational if and only if  $y$  is, we have  $y = x$  if  $x$  is rational and  $y = -x$  if  $x$  is irrational, so  $y = f(x)$ .)

- (v)

$$f^{-1}(x) = \begin{cases} x, & x \neq a_1, \dots, a_n \\ a_{i-1}, & x = a_i, \quad i = 2, \dots, n \\ a_n, & x = a_1. \end{cases}$$

- (vii)  $f^{-1} = f$ .

2. (i)  $f^{-1}$  is increasing and  $f^{-1}(x)$  is not defined for  $x \leq 0$ .  
 (iii)  $f^{-1}$  is decreasing and  $f^{-1}(x)$  is not defined for  $x \leq 0$ .  
 3. Suppose  $f$  is increasing. Let  $a < b$ . Then  $f^{-1}(a) \neq f^{-1}(b)$ , since  $f^{-1}$  is one-one. So either  $f^{-1}(a) < f^{-1}(b)$  or  $f^{-1}(a) > f^{-1}(b)$ . But if  $f^{-1}(a) > f^{-1}(b)$ , then

$$b = f(f^{-1}(b)) < f(f^{-1}(a)) = a,$$

a contradiction. The proof is similar for decreasing  $f$ , or one can consider  $-f$  instead.

4. Clearly,  $f + g$  is increasing, for if  $f(a) < f(b)$  and  $g(a) < g(b)$ , then  $(f + g)(a) = f(a) + g(a) < f(b) + g(b) = (f + g)(b)$ .  
 $f \cdot g$  is not necessarily increasing; for example, if  $f(x) = g(x) = x$ . (But  $f \cdot g$

is increasing if  $f(x) \geq 0$  for all  $x$ .)

$f \circ g$  is increasing, for if  $a < b$ , then  $g(a) < g(b)$ , so  $f(g(a)) < f(g(b))$ .

5. (a) If  $(f \circ g)(x) = (f \circ g)(y)$ , so that  $f(g(x)) = f(g(y))$ , then  $g(x) = g(y)$ , since  $f$  is one-one, so  $x = y$ , since  $g$  is one-one.  
 $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$ : for if  $y = (f \circ g)^{-1}(x)$ , then  $x = (f \circ g)(y) = f(g(y))$ , so  $g(y) = f^{-1}(x)$ , so  $y = g^{-1}(f^{-1}(x))$ .

6. If  $f(x) = f(y)$ , then

$$\frac{ax+b}{cx+d} = \frac{ay+b}{cy+d},$$

so

$$acxy + bcy + adx + bd = acxy + ady + bcx + bd,$$

or

$$ad(x - y) = bc(x - y).$$

If  $ad \neq bc$ , this implies that  $x - y = 0$ . (But if  $ad = bc$ , then  $f(x) = f(y)$  for all  $x$  and  $y$  in the domain of  $f$ .)

If  $y = f^{-1}(x)$ , then  $x = f(y)$ , so

$$x = \frac{ay+b}{cy+d}$$

so

$$f^{-1}(x) = y = \frac{-dx + b}{cx - a} \quad \text{for } x \neq a/c.$$

7. (i) Those intervals  $[a, b]$  which are contained in  $(-\infty, 0]$  or  $[0, 2]$  or  $[2, \infty)$ , since  $f$  is increasing on  $(-\infty, 0]$  and  $[2, \infty)$ , and decreasing on  $[0, 2]$ .  
(ii) Those intervals  $[a, b]$  which are contained in  $(-\infty, 0]$  or  $[0, \infty)$ , since  $f$  is increasing on  $(-\infty, 0]$  and decreasing on  $[0, \infty)$ .  
17. The formula for the derivative reads:

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}.$$

(In this formula, it is understood that  $dx/dy$  means  $(f^{-1})'(y)$ , while  $dy/dx$  is an “expression involving  $x$ ,” and in the final answer  $x$  must be replaced by  $y$ , by means of the equation  $y = f(x)$ .)

The computation in Problem 17, when completed, shows that

$$\begin{aligned} \frac{dx^{1/n}}{dx} &= \frac{1}{n(x^{1/n})^{n-1}} = \frac{1}{nx^{1-(1/n)}} \\ &= \frac{1}{n}x^{(1/n)-1}. \end{aligned}$$

18.

$$\begin{aligned} G'(x) &= x(f^{-1})'(x) + f^{-1}(x) - F'(f^{-1}(x)) \cdot (f^{-1})'(x) \\ &= x(f^{-1})'(x) + f^{-1}(x) - f(f^{-1}(x)) \cdot (f^{-1})'(x) \\ &= x(f^{-1})'(x) + f^{-1}(x) - x(f^{-1})'(x) \\ &= f^{-1}(x). \end{aligned}$$

19. (i)

$$(h^{-1})'(3) = \frac{1}{h'(h^{-1}(3))} = \frac{1}{h'(0)} = \frac{1}{\sin^2(\sin 1)}.$$

20. Since

$$(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))},$$

we have

$$\begin{aligned} (f^{-1})''(x) &= \frac{-f''(f^{-1}(x)) \cdot (f^{-1})'(x)}{[f'(f^{-1}(x))]^2} \\ &= \frac{-f''(f^{-1}(x))}{[f'(f^{-1}(x))]^3}. \end{aligned}$$

CHAPTER 13 1. If  $P_n = \{t_0, \dots, t_n\}$  is the partition with  $t_i = ib/n$ , then

$$\begin{aligned} L(f, P_n) &= \sum_{i=1}^n (t_i - t_{i-1})^3 \cdot (t_i - t_{i-1}) \\ &= \sum_{i=1}^n (i-1)^3 \cdot \frac{b_3}{n^3} \cdot \frac{b}{n} \\ &= \frac{b^4}{n^4} \sum_{j=0}^{n-1} j^3 \\ &= \frac{b^4}{n^4} \left[ \frac{(n-1)^4}{4} + \frac{(n-1)^3}{2} + \frac{(n-1)^2}{4} \right], \end{aligned}$$

and similarly

$$\begin{aligned} U(f, P_n) &= \frac{b^4}{n^4} \sum_{j=1}^n j^3 \\ &= \frac{b^4}{n^4} \left[ \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4} \right]. \end{aligned}$$

Clearly  $L(f, P_n)$  and  $U(f, P_n)$  can be made as close to  $b^4/4$  as desired by choosing  $n$  sufficiently large, so  $U(f, P_n) - L(f, P_n)$  can be made as small as desired, by choosing  $n$  large enough. This shows that  $f$  is integrable. Moreover, there is only one number  $a$  with  $L(f, P_n) \leq a \leq U(f, P_n)$  for all  $n$ ; since  $\int_0^b x^3 dx$  has this property, the proof that  $\int_0^b x^3 dx = b^4/4$  will be complete once we show that  $L(f, P_n) \leq b^4/4 \leq U(f, P_n)$  for all  $n$ . This

can be done by a straightforward computation, but it actually follows from the fact that  $L(f, P_n)$  and  $U(f, P_n)$  can be made as close to  $b^4/4$  as desired by choosing  $n$  sufficiently large. In fact, if it were true that  $b^4/4 < \int_0^b x^3 dx$ , then it would not be possible to make  $U(f, P_n)$  as close as desired to  $b^4/4$  by choosing  $n$  large enough, since each  $U(f, P_n) \geq \int_0^b x^3 dx$ , and similarly we cannot have  $b^4/4 > \int_0^b x^3 dx$ .

2. We have

$$L(f, P_n) = \frac{b^5}{n^5} \left[ \frac{(n-1)^5}{5} + \frac{(n-1)^4}{2} + \frac{(n-1)^3}{3} - \frac{(n-1)}{30} \right],$$

$$U(f, P_n) = \frac{b^5}{n^5} \left[ \frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30} \right].$$

Clearly  $L(f, P_n)$  and  $U(f, P_n)$  can be made as close to  $b^5/5$  as desired by choosing  $n$  large enough. As in Problem 1, this implies that  $\int_0^b x^4 dx = b^5/5$ .

7. (i)  $\int_0^2 f = 0$ .

(iii)  $\int_0^2 f = 3$ .

(v)  $f$  is not integrable.

(vii)  $\int_0^2 f = 1$ .

(For a rigorous proof that the functions in (i), (iii), and (vii) are integrable, see Problem 20. The values of the integrals, which are clear from the geometric picture, can also be deduced rigorously by using the ideas in the proof of Problem 20, together with known integrals.)

8. (i)

$$\int_{-2}^2 \left[ \left( \frac{x^2}{2} + 2 \right) - x^2 \right] dx = \frac{16}{3}.$$

(iii)

$$\int_{-\sqrt{2}/2}^{\sqrt{2}/2} [(1-x^2) - x^2] dx = \frac{2\sqrt{2}}{3}.$$

(v)

$$\int_0^2 [(x^2 - 2x + 4) - x^2] dx = 4.$$

- 9.

$$\begin{aligned} \int_a^b \left( \int_c^d f(x)g(y) dy \right) dx &= \int_a^b \left( f(x) \int_c^d g(y) dy \right) dx \quad (\text{here } f(x) \text{ is the constant}) \\ &= \int_c^d g(y) dy \cdot \int_a^b f(x) dx \\ &\quad (\text{here } \int_c^d g(y) dy \text{ is the constant}). \end{aligned}$$

13. (a) Clearly  $L(f, P) \geq 0$  for every partition  $P$ .

(b) Apply part (a) to  $f - g$ , and use the fact that

$$\int_a^b (f - g) = \int_a^b f - \int_a^b g.$$

23. (a) Clearly

$$m(b-a) \leq L(f, P) \leq U(f, P) \leq M(b-a)$$

for all partitions  $P$  of  $[a, b]$ . Consequently,

$$m(b-a) \leq \int_a^b f(x) dx \leq M(b-a).$$

Thus

$$\mu = \frac{\int_a^b f(x) dx}{b-a}$$

satisfies  $m \leq \mu \leq M$ .

(b) Let  $m$  and  $M$  be the minimum and maximum values of  $f$  on  $[a, b]$ . Since  $f$  is continuous, it takes on the values  $m$  and  $M$ , and consequently the number  $\mu$  of part (a).

33. (a) 0.

(b)  $\frac{1}{2}$ .

37. Since

$$-|f| \leq f \leq |f|,$$

we have

$$-\int_a^b |f| \leq \int_a^b f \leq \int_a^b |f|,$$

so

$$\left| \int_a^b f \right| \leq \int_a^b |f|.$$

(Problem 36 implies that  $\int_a^b |f|$  makes sense.)

**CHAPTER 14**

1. (i)  $(\sin^3 x^3) \cdot 3x^2$ .

(iii)  $\int_8^x \frac{1}{1+t^2 + \sin^2 t} dt$ .

(v)  $\int_a^b \frac{1}{1+t^2 + \sin^2 t} dt$ .

(vii)  $(F^{-1})'(x) = \frac{1}{F'(F^{-1}(x))} = F^{-1}(x)$ .

2. (i) All  $x \neq 1$ .

(iii) All  $x \neq 1$ .

(v) All  $x$ .

(vii) All  $x \neq 0$ . ( $F$  is not differentiable at 0 because  $F(x) = 0$  for  $x \leq 0$ , but there are  $x > 0$  arbitrarily close to 0 with  $\frac{F(x)}{x} = \frac{1}{2}$ .)

5. (i)

$$\begin{aligned}(f^{-1})'(0) &= \frac{1}{f'(f^{-1}(0))} = \frac{1}{1 + \sin(\sin(f^{-1}(0)))} \\ &= \frac{1}{1 + \sin(\sin 0)} = 1.\end{aligned}$$

11.  $F(x) = x \int_0^x f(t) dt$ , so

$$F'(x) = xf(x) + \int_0^x f(t) dt.$$

14.

$$f(x) = \int_0^x \left( \int_0^y \left( \int_0^z \frac{1}{\sqrt{1 + \sin^2 t}} dt \right) dz \right) dy.$$

16. We can choose

$$f(x) = \frac{x^{(1/n)+1}}{\frac{1}{n} + 1}.$$

Then

$$\int_0^b \sqrt[n]{x} dx = f(b) - f(0) = \frac{b^{(1/n)+1}}{\frac{1}{n} + 1}.$$

CHAPTER 15

1. (i)

$$\frac{1}{1 + \arctan^2(\arctan x)} \cdot \frac{1}{1 + \arctan^2 x} \cdot \frac{1}{1 + x^2}.$$

(iii)

$$\frac{1}{1 + (\tan x \arctan x)^2} \cdot \left( \sec^2 x \arctan x + \frac{\tan x}{1 + x^2} \right).$$

2. (i) 0.

(iii) 0.

(v) 0.

7. (a)

$$\sin 2x = \sin(x + x) = \sin x \cos x + \cos x \sin x = 2 \sin x \cos x.$$

$$\cos 2x = \cos^2 x - \sin^2 x = 2 \cos^2 x - 1 = 1 - 2 \sin^2 x.$$

$$\sin 3x = \sin(2x + x) = \sin 2x \cos x + \cos 2x \sin x$$

$$= 2 \sin x \cos^2 x + (\cos^2 x - \sin^2 x) \sin x$$

$$= 3 \sin x \cos^2 x - \sin^3 x.$$

$$\cos 3x = \cos(2x + x) = \cos 2x \cos x - \sin 2x \sin x$$

$$= (\cos^2 x - \sin^2 x) \cos x = 2 \sin^2 x \cos x$$

$$= \cos^3 x - \sin^2 x \cos x - 2 \sin^2 x \cos x$$

$$= \cos^3 x - 3 \sin^2 x \cos x$$

$$= 4 \cos^3 x - 3 \cos x.$$

(b) Since  $\cos \pi/4 > 0$  and

$$0 = \cos \frac{\pi}{2} = \cos 2 \cdot \frac{\pi}{4} = 2 \cos^2 \frac{\pi}{4} - 1,$$

we have  $\cos \pi/4 = \sqrt{2}/2$ . It follows, since  $\sin \pi/4 > 0$  and  $\sin^2 + \cos^2 = 1$ , that  $\sin \pi/4 = \sqrt{2}/2$ , and consequently  $\tan \pi/4 = 1$ . Similarly, since  $\cos \pi/6 > 0$  and

$$0 = \cos \frac{\pi}{2} = \cos 3 \cdot \frac{\pi}{6} = 4 \cos^3 \frac{\pi}{6} - 3 \cos \frac{\pi}{6},$$

we have  $\cos \pi/6 = \sqrt{3}/2$ . It follows, since  $\sin \pi/6 > 0$ , that  $\sin \pi/6 = \sqrt{1 - (\sqrt{3}/2)^2} = \frac{1}{2}$ .

9. (a)

$$\begin{aligned}\tan(x+y) &= \frac{\sin(x+y)}{\cos(x+y)} \\ &= \frac{\sin x \cos y + \cos x \sin y}{\cos x \cos y - \sin x \sin y} \\ &= \frac{\frac{\sin x \cos y}{\cos x \cos y} + \frac{\cos x \sin y}{\cos x \cos y}}{\frac{\cos x \cos y}{\cos x \cos y} - \frac{\sin x \sin y}{\cos x \cos y}} \\ &= \frac{\tan x + \tan y}{1 - \tan x \tan y}.\end{aligned}$$

(b) From part (a) we have

$$\begin{aligned}\tan(\arctan x + \arctan y) &= \frac{\tan(\arctan x) + \tan(\arctan y)}{1 - \tan(\arctan x) \tan(\arctan y)} \\ &= \frac{x + y}{1 - xy},\end{aligned}$$

provided that  $\arctan x$ ,  $\arctan y$ , and  $\arctan x + \arctan y \neq k\pi + \pi/2$ . Since  $-\pi/2 < \arctan x$ ,  $\arctan y < \pi/2$ , this is always the case except when  $\arctan x + \arctan y = \pm\pi/2$ , which is equivalent to  $xy = 1$ . From this equation we can conclude that

$$\arctan x + \arctan y = \arctan \left( \frac{x + y}{1 - xy} \right)$$

provided that  $\arctan x + \arctan y$  lies in  $(-\pi/2, \pi/2)$ , which is true whenever  $xy < 1$ . (If  $x, y > 0$  and  $xy > 1$ , so that  $\arctan x + \arctan y > \pi/2$ , then we must add  $\pi$  to the right side, and if  $x, y < 0$  and  $xy > 1$ , so that  $\arctan x + \arctan y < -\pi/2$ , then we must subtract  $\pi$ .)

11. The first formula is derived by subtracting the second of the following two equations from the first:

$$\begin{aligned}\cos(m-n)x &= \cos(mx-nx) = \cos mx \cos(-nx) - \sin mx \sin(-nx) \\ &= \cos mx \cos nx + \sin mx \sin nx, \\ \cos(m+n)x &= \cos mx \cos nx - \sin mx \sin nx.\end{aligned}$$

The other formulas are derived similarly.

12. It follows from Problem 11 that if  $m \neq n$ , then

$$\begin{aligned}\int_{-\pi}^{\pi} \sin mx \sin nx dx &= \frac{1}{2} \int_{-\pi}^{\pi} [\cos(m-n)x - \cos(m+n)x] dx \\ &= \frac{1}{2} \left\{ \left[ \frac{\sin(m-n)\pi}{m-n} - \frac{\sin(m+n)\pi}{m+n} \right] \right. \\ &\quad \left. - \left[ \frac{\sin(m-n)\pi}{m-n} - \frac{\sin(m+n)\pi}{m+n} \right] \right\} \\ &= 0.\end{aligned}$$

But if  $m = n$ , then

$$\begin{aligned}\int_{-\pi}^{\pi} \sin mx \sin nx dx &= \frac{1}{2} \int_{-\pi}^{\pi} 1 - \cos(m+n)x dx \\ &= \frac{1}{2} \{ [\pi - \cos(m+n)\pi] - [-\pi - \cos(m+n)\pi] \} \\ &= \pi.\end{aligned}$$

The other formulas are proved similarly.

15. (a) We have

$$\begin{aligned}\cos 2x &= \cos^2 x - \sin^2 x \\ &= 1 - 2 \sin^2 x \\ &= 2 \cos^2 x - 1.\end{aligned}$$

So

$$\begin{aligned}\sin^2 x &= \frac{1 - \cos 2x}{2}, \\ \cos^2 x &= \frac{1 + \cos 2x}{2}.\end{aligned}$$

- (b) These formulas follow from part (a), because  $\cos x/2 \geq 0$  and  $\sin x/2 \geq 0$  (since  $0 \leq x \leq \pi/2$ ).

(c)

$$\begin{aligned}\int_a^b \sin^2 x dx &= \int_a^b \frac{1 - \cos 2x}{2} dx = \frac{1}{2}(b-a) - \frac{1}{4}(\sin 2b - \sin 2a). \\ \int_a^b \cos^2 x dx &= \int_a^b \frac{1 + \cos 2x}{2} dx = \frac{1}{2}(b-a) + \frac{1}{4}(\sin 2b - \sin 2a).\end{aligned}$$

19. (a)  $\arctan 1 - \arctan 0 = \pi/4$ .

(b)  $\lim_{x \rightarrow \infty} \arctan x - \arctan 0 = \pi/2.$

20.  $\lim_{x \rightarrow \infty} x \sin \frac{1}{x} = \lim_{x \rightarrow 0^+} \frac{1}{x} \sin x = 1.$

21. (a)

$$(\sin^\circ)'(x) = \frac{\pi}{180} \cos\left(\frac{\pi x}{180}\right) = \frac{\pi}{180} \cos^\circ(x).$$

$$(\cos^\circ)'(x) = \frac{\pi}{180} \cdot -\sin\left(\frac{\pi x}{180}\right) = \frac{-\pi}{180} \sin^\circ(x).$$

(b)  $\lim_{x \rightarrow 0} \frac{\sin^\circ x}{x} = \lim_{x \rightarrow 0} \frac{\sin(\pi x/180)}{x} = \lim_{x \rightarrow 0} \frac{\pi}{180} \cdot \frac{\sin(\pi x/180)}{\pi x/180} = \frac{\pi}{180}.$

$$\lim_{x \rightarrow \infty} x \sin^\circ \frac{1}{x} = \lim_{x \rightarrow 0^+} \frac{\sin^\circ x}{x} = \frac{\pi}{180}.$$

**CHAPTER 18**

1. (i)  $e^{e^{e^x}} \cdot e^{e^x} \cdot e^x \cdot e^x.$

(iii)  $(\sin x)^{\sin(\sin x)} [(\log(\sin x)) \cdot \cos(\sin x) \cdot \cos x + (\cos x / \sin x) \cdot \sin(\sin x)]$

(v)  $\sin x^{\sin x^{\sin x}} [(\log(\sin x)) \cdot \sin x^{\sin x} \cdot \{(\log(\sin x)) \cdot \cos x + (\cos x / \sin x) \cdot \sin x\} + (\cos x / \sin x) \cdot \sin x^{\sin x}].$

(vii)  $\left[ \arcsin\left(\frac{x}{\sin x}\right) \right]^{\log(\sin e^x)} \left[ (\log(\arcsin(\frac{x}{\sin x}))) \cdot \frac{(\cos e^x)e^x}{\sin e^x} + \log(\sin e^x) \cdot \frac{\sin x - x \cos x}{\arcsin\left(\frac{x}{\sin x}\right) \sqrt{1 - \left(\frac{x}{\sin x}\right)^2 \cdot \sin^2 x}} \right].$

(ix)  $(\log x)^{\log x} \cdot \left[ \log(\log x) \cdot \frac{1}{x} + \log x \cdot \frac{1}{\log x} \cdot \frac{1}{x} \right].$

5. (i) 0.

(iii)  $\frac{1}{6}.$

(v)  $\frac{1}{3}.$

7. (a)  $\cosh^2 x - \sinh^2 x = \left(\frac{e^x + e^{-x}}{2}\right)^2 - \left(\frac{e^x - e^{-x}}{2}\right)^2$

$$= \left[\frac{e^{2x}}{4} + \frac{1}{2} + \frac{e^{-2x}}{4}\right] - \left[\frac{e^{2x}}{4} - \frac{1}{2} + \frac{e^{-2x}}{4}\right]$$

$$= 1.$$

(c)

$$\begin{aligned}\sinh x \cosh y + \cosh x \sinh y &= \left(\frac{e^x - e^{-x}}{2}\right)\left(\frac{e^y + e^{-y}}{2}\right) + \left(\frac{e^x + e^{-x}}{2}\right)\left(\frac{e^y - e^{-y}}{2}\right) \\ &= \left[\frac{e^{x+y}}{4} + \frac{e^{-x-y}}{4} - \frac{e^{-x+y}}{4} + \frac{e^{x-y}}{4}\right] + \left[\frac{e^{x+y}}{4} + \frac{e^{-x-y}}{4} + \frac{e^{-x+y}}{4} - \frac{e^{-x-y}}{4}\right] \\ &= \frac{e^{x+y} + e^{-(x+y)}}{2} = \sinh(x+y).\end{aligned}$$

(e) Since

$$\sinh x = \frac{e^x - e^{-x}}{2},$$

we have

$$\sinh'(x) = \frac{e^x - e^{-x}}{2} = \cosh x.$$

(g) Since

$$\tanh x = \frac{\sinh x}{\cosh x},$$

we have

$$\begin{aligned}\tanh'(x) &= \frac{(\cosh x)^2 - (\sinh x)^2}{\cosh^2 x} \\ &= \frac{1}{\cosh^2 x} \quad \text{by part (a).}\end{aligned}$$

8. (a) If  $y = \arg \cosh x$ , then  $x \geq 0$  and

$$x = \cosh y = \sqrt{1 + \sinh^2 y} \quad \text{by part (a).}$$

So

$$\sinh(\arg \cosh x) = \sinh y = \sqrt{x^2 - 1} \quad \text{since } \sinh y \geq 0 \text{ for } y \geq 0.$$

(c)

$$\begin{aligned}(\arg \sinh)'(x) &= \frac{1}{\sinh'(\arg \sinh(x))} \\ &= \frac{1}{\cosh(\arg \sinh(x))} \\ &= \frac{1}{\sqrt{1+x^2}} \quad \text{by part (b).}\end{aligned}$$

(e)

$$\begin{aligned}(\arg \tanh)'(x) &= \frac{1}{\tanh'(\arg \tanh(x))}, \\ &= \cosh^2(\arg \tanh(x)).\end{aligned}$$

Now,

$$\tanh^2 y + \frac{1}{\cosh^2 y} = 1 \quad \text{by Problem 7(b),}$$

so

$$\tanh^2(\arg \tanh(x)) + \frac{1}{\cosh^2(\arg \tanh(x))} = 1,$$

or

$$\cosh^2(\arg \tanh(x)) = \frac{1}{1-x^2}.$$

9. (a) If  $y = \arg \sinh x$ , then

$$x = \sinh y = \frac{e^y - e^{-y}}{2}$$

so

$$\begin{aligned} e^y - e^{-y} &= 2x, \\ e^{2y} - 2xe^y - 1 &= 0, \\ e^y &= \frac{2x \pm \sqrt{4x^2 + 4}}{2} \end{aligned}$$

so

$$e^y = x + \sqrt{1+x^2} \quad \text{since } e^y > 0$$

or

$$y = \arg \sinh x = \log(x + \sqrt{1+x^2}).$$

Similarly,

$$\begin{aligned} \arg \cosh x &= \log(x + \sqrt{x^2 - 1}), \\ \arg \tanh x &= \frac{1}{2} \log(1+x) - \frac{1}{2} \log(1-x). \end{aligned}$$

(b)

$$\int_a^b \frac{1}{\sqrt{1+x^2}} dx = \arg \sinh b - \arg \sinh a \quad \text{by Problem 8(c)}$$

$$= \log(b + \sqrt{1+b^2}) - \log(a + \sqrt{1+a^2}).$$

$$\int_a^b \frac{1}{\sqrt{x^2 - 1}} dx = \log(b + \sqrt{b^2 - 1}) - \log(a + \sqrt{a^2 - 1}).$$

$$\int_a^b \frac{1}{1-x^2} dx = \frac{1}{2} [\log(1+b) - \log(1-b) - \log(1+a) + \log(1-a)].$$

12. (a)  $\lim_{x \rightarrow \infty} a^x = \lim_{x \rightarrow \infty} e^{x \log a}$ . Since  $\log a < 0$ , we have  $\lim_{x \rightarrow \infty} x \log a = -\infty$ , so  $\lim_{x \rightarrow \infty} e^{x \log a} = 0$ .

$$(c) \lim_{x \rightarrow \infty} \frac{(\log x)^n}{x} = \lim_{y \rightarrow \infty} \frac{y^n}{e^y} = 0.$$

$$(e) \lim_{x \rightarrow 0^+} x^x = \lim_{x \rightarrow 0^+} e^{x \log x}. \text{ Now, } \lim_{x \rightarrow 0^+} x \log x = 0 \text{ by part (d), so } \lim_{x \rightarrow 0^+} x^x = 1.$$

16. (a)  $\lim_{y \rightarrow 0} \log(1+y)/y = \log'(1) = 1$ .

(b)  $\lim_{x \rightarrow \infty} x \log(1 + 1/x) = \lim_{y \rightarrow 0^+} \log(1 + y)/y = 1.$

(c)

$$\begin{aligned} e &= \exp(1) = \exp(\lim_{x \rightarrow \infty} x \log(1 + 1/x)) \\ (*) &= \lim_{x \rightarrow \infty} \exp(x \log(1 + 1/x)) \\ &= \lim_{x \rightarrow \infty} (1 + 1/x)^x. \end{aligned}$$

(The starred equality depends on the continuity of  $\exp$  at 1, and can be justified as follows. For every  $\varepsilon > 0$  there is some  $\delta > 0$  such that  $|e - \exp y| < \varepsilon$  for  $|y - 1| < \delta$ . Moreover, there is some  $N$  such that  $|x \log(1 + 1/x) - 1| < \delta$  for  $x > N$ . So  $|e - \exp(x \log(1 + 1/x))| < \varepsilon$  for  $x > N$ .

(d)

$$\begin{aligned} e^a &= [\lim_{x \rightarrow \infty} (1 + 1/x)^x]^a = \lim_{x \rightarrow \infty} (1 + 1/x)^{ax} \\ &= \lim_{ax \rightarrow \infty} (1 + 1/x)^{ax} \\ &= \lim_{y \rightarrow \infty} (1 + a/y)^y. \end{aligned}$$

18. After one year the number of dollars yielded by an initial investment of one dollar will be

$$\lim_{x \rightarrow \infty} (1 + a/100x)^x = e^{a/100}.$$

19. (a) Clearly  $f'(x) = 1/x$  for  $x > 0$ . If  $x < 0$ , then  $f(x) = \log(-x)$ , so  $f'(x) = (-1) \cdot 1/(-x) = 1/x$ .  
(b) We can write  $\log|f|$  as  $g \circ f$  where  $g(x) = \log|x|$  is the function of part (a). So  $(\log|f|)' = (g' \circ f) \cdot f' = 1/f \cdot f'$ .  
20. (c) Let  $g(x) = f(x)/e^{cx}$ . Then

$$g'(x) = \frac{e^{cx} f'(x) - f(x)c e^{cx}}{e^{2cx}} = 0,$$

so there is some number  $k$  such that  $g(x) = k$  for all  $x$ .

21. (a) According to Problem 20, there is some  $k$  such that  $A(t) = ke^{ct}$ . Then  $k = ke^{0 \cdot t} = A_0$ . So  $A(t) = A_0 e^{ct}$ .  
(b) If  $A(t + \tau) = A(t)/2$ , then

$$A_0 e^{ct+ct} = \frac{A_0 e^{ct}}{2},$$

so  $e^{ct} e^{c\tau} = e^{ct}/2$  or  $e^{c\tau} = \frac{1}{2}$ , so  $\tau = -(\log 2)/c$ . It is easy to check that this  $\tau$  does work.

22. Newton's law states that, for a certain (positive) number  $c$ ,

$$T'(t) = c(T - M),$$

which can be written

$$(T - M)' = c(T - M).$$

So by Problem 20 there is some number  $k$  such that

$$T(t) - M = ke^{ct},$$

and  $k = ke^{0t} = T(0) = T_0$ . So  $T(t) = M + T_0 e^{ct}$ .

## CHAPTER 19

1. (i)  $(\sqrt[5]{x^3} + \sqrt[6]{x})/\sqrt{x} = x^{1/10} + x^{-1/3}$ .  
 (ii)  $\frac{1}{\sqrt{x-1} + \sqrt{x+1}} = \frac{\sqrt{x-1} - \sqrt{x+1}}{-2}$ .  
 (iii)  $(e^x + e^{2x} + e^{3x})/e^{4x} = e^{-3x} + e^{-2x} + e^{-x}$ .  
 (iv)  $a^x/b^x = (a/b)^x = e^{x \log(a/b)}$ .  
 (v)  $\tan^2 x = \sec^2 x - 1$ .  
 (vi)  $\frac{1}{a^2 + x^2} = \frac{1/a^2}{1 + (\frac{x}{a})^2}$ .  
 (vii)  $\frac{1}{\sqrt{a^2 - x^2}} = \frac{1/a}{\sqrt{1 - (x/a)^2}}$ .  
 (viii)  $\frac{1}{1 + \sin x} = \frac{1 - \sin x}{1 - \sin^2 x} = \frac{1 - \sin x}{\cos^2 x} = \sec^2 x - \sec x \tan x$ .  
 (ix)  $\frac{8x^2 + 6x + 4}{x + 1} = 8x - 2 + \frac{6}{x + 1}$ .  
 (x)  $\frac{1}{\sqrt{2x - x^2}} = \frac{1}{\sqrt{1 - (x-1)^2}}$ .
2. (i)  $-\cos e^x$ . (Let  $u = e^x$ .)  
 (iii)  $(\log x)^2/2$ . (Let  $u = \log x$ .)  
 (v)  $e^{e^x}$ . (Let  $u = e^{e^x}$ .)  
 (vii)  $2e^{\sqrt{x}}$ . (Let  $u = \sqrt{x}$ .)  
 (ix)  $-(\log(\cos x))^2/2$ . (Let  $u = \log(\cos x)$ .)
3. (i)  $\int x^2 e^x dx = x^2 e^x - \int 2x e^x dx = x^2 e^x - \left[ 2x e^x - \int e^x dx \right] = x^2 e^x - 2x e^x + 2e^x$ .

(iii) We have

$$\begin{aligned} \int e^{ax} \sin bx dx &= \frac{e^{ax} \sin bx}{a} - \frac{b}{a} \int e^{ax} \cos bx dx \\ &= \frac{e^{ax} \sin bx}{a} - \frac{b}{a} \left[ \frac{e^{ax} \cos bx}{a} - \frac{b}{a} \int e^{ax} (-\sin bx) dx \right], \end{aligned}$$

so

$$\int e^{ax} \sin bx dx = \frac{a}{a^2 + b^2} e^{ax} \sin bx - \frac{b}{a^2 + b^2} e^{ax} \cos bx.$$

- (v) Using the result  $\int (\log x)^2 dx = x(\log x)^2 - 2x(\log x) + 2x$  from the text, we have

$$\begin{aligned} \int (\log x)^3 dx &= [x(\log x)^2 - 2x(\log x) + 2x] \log x \\ &\quad - \int \frac{1}{x} [x(\log x)^2 - 2x(\log x) + 2x] dx \\ &= x(\log x)^3 - 2x(\log x)^2 + 2x \log x \\ &\quad - \int (\log x)^2 dx + 2[x \log x - x] - 2x \\ &= x(\log x)^3 - 2x(\log x)^2 + 2x \log x \\ &\quad - [x(\log x)^2 - 2x(\log x) + 2x] + 2[x \log x - x] - 2x \\ &= x(\log x)^3 - 3x(\log x)^2 + 6x \log x - 6x. \end{aligned}$$

(vii)

$$\begin{aligned} \int \sec^3 x dx &= \int (\sec^2 x)(\sec x) dx = \tan x \sec x - \int (\tan x)(\sec x \tan x) dx \\ &= \tan x \sec x - \int \sec x (\sec^2 x - 1) dx \\ &= \tan x \sec x - \int \sec^3 x dx + \int \sec x dx, \end{aligned}$$

so

$$\int \sec^3 x dx = \frac{1}{2} [\tan x \sec x + \log(\sec x + \tan x)].$$

(ix)

$$\begin{aligned} \int \sqrt{x} \log x dx &= \frac{2x^{3/2}}{3} \log x - \frac{2}{3} \int x^{3/2} \cdot \frac{1}{x} dx \\ &= \frac{2x^{3/2}}{3} \log x - \frac{2}{3} \int x^{1/2} dx \\ &= \frac{2x^{3/2}}{3} \log x - \frac{4}{9} x^{3/2}. \end{aligned}$$

4. (i) Let  $x = \sin u$ ,  $dx = \cos u du$ . The integral becomes

$$\int \frac{\cos u du}{\sqrt{1 - \sin^2 u}} = \int 1 du = u = \arcsin x.$$

- (iii) Let  $x = \sec u$ ,  $dx = \sec u \tan u du$ . The integral becomes

$$\begin{aligned} \int \frac{\sec u \tan u du}{\sqrt{\sec^2 u - 1}} &= \int \sec u du = \log(\sec u + \tan u) \\ &= \log(x + \sqrt{x^2 - 1}). \end{aligned}$$

(v) Let  $x = \sin u$ ,  $dx = \cos u du$ . The integral becomes

$$\begin{aligned}\int \frac{\cos u du}{\sin u \sqrt{1 - \sin^2 u}} &= \int \csc u du = -\log(\csc u + \cot u) \\ &= -\log\left(\frac{1}{x} + \frac{\sqrt{1-x^2}}{x}\right).\end{aligned}$$

(vii) Let  $x = \sin u$ ,  $dx = \cos u du$ . The integral becomes

$$\begin{aligned}\int (\sin^3 u \cos u) \cos u du &= \int \sin^3 u \cos^2 u du = \int (\sin u)(1 - \cos^2 u) \cos^2 u du \\ &= \int (\sin u)(\cos^2 u - \cos^4 u) du = -\frac{\cos^3 u}{3} + \frac{\cos^5 u}{5} \\ &= -\frac{(1-x^2)^{3/2}}{3} + \frac{(1-x^2)^{5/2}}{5}.\end{aligned}$$

(ix) Let  $x = \tan u$ ,  $dx = \sec^2 u du$ . The integral becomes

$$\begin{aligned}\int \sec u \sec^2 u du &= \int \sec^3 u du \\ &= \frac{1}{2}[\tan u \sec u + \log(\sec u + \tan u)] \quad \text{by Problem 3(vii)} \\ &= \frac{1}{2}[x\sqrt{1+x^2} + \log(x + \sqrt{1+x^2})].\end{aligned}$$

5. (i) Let  $u = \sqrt{x+1}$ ,  $x = u^2 - 1$ ,  $dx = 2u du$ . The integral becomes

$$\begin{aligned}\int \frac{2u du}{1+u} &= \int \left(2 + \frac{-2}{1+u}\right) du \\ &= 2u - 2\log(1+u) = 2\sqrt{x+1} - 2\log(1+\sqrt{x+1}).\end{aligned}$$

(iii) Let  $u = x^{1/6}$ ,  $x = u^6$ ,  $dx = 6u^5 du$ . The integral becomes

$$\begin{aligned}\int \frac{6u^5 du}{u^3 + u^2} &= 6 \int \left(u^2 - u + 1 - \frac{1}{u+1}\right) du = 2u^3 - 3u^2 + 6u - 6\log(u+1) \\ &= 2\sqrt{x} - 3\sqrt[3]{x} + 6\sqrt[6]{x} - 6\log(\sqrt[6]{x}+1).\end{aligned}$$

(v) Let  $u = \tan x$ ,  $x = \arctan u$ ,  $dx = du/(1+u^2)$ . The integral becomes

$$\begin{aligned}\int \frac{du}{(1+u^2)(2+u)} &= \frac{1}{5} \int \left(\frac{1}{2+u} - \frac{u-2}{1+u^2}\right) du \\ &= \frac{1}{5} \int \frac{du}{2+u} - \frac{1}{10} \int \frac{2u}{1+u^2} du + \frac{2}{5} \int \frac{du}{1+u^2} \\ &= \frac{1}{5} \log(2+u) - \frac{1}{10} \log(1+u^2) + \frac{2}{5} \arctan u \\ &= \frac{1}{5} \log(2+\tan x) - \frac{1}{10} \log(1+\tan^2 x) + \frac{2}{5}x.\end{aligned}$$

(vii) Let  $u = 2^x$ ,  $x = (\log u)/(\log 2)$ ,  $dx = du/(u \log 2)$ . The integral becomes

$$\begin{aligned}\frac{1}{\log 2} \int \frac{u^2 + 1}{(u+1)u} du &= \frac{1}{\log 2} \int \left(1 + \frac{1-u}{u(u+1)}\right) du \\ &= \frac{1}{\log 2} \int \left(1 + \frac{1}{u} - \frac{2}{u+1}\right) du \\ &= \frac{1}{\log 2} [u + \log u - 2 \log(u+1)] \\ &= \frac{1}{\log 2} [2^x + x \log 2 - 2 \log(2^x + 1)].\end{aligned}$$

(ix) Let  $u = \sqrt{x}$ ,  $x = u^2$ ,  $dx = 2u$ . The integral becomes

$$\int \frac{\sqrt{1-u^2} \cdot 2u \, du}{1-u}.$$

Now let  $u = \sin y$ ,  $du = \cos y \, dy$ . The integral becomes

$$\begin{aligned}\int \frac{2 \cos y \sin y \cos y}{1-\sin y} \, dy &= 2 \int \frac{(1-\sin^2 y) \sin y}{1-\sin y} \, dy \\ &= 2 \int (1+\sin y) \sin y \, dy \\ &= 2 \int \sin y \, dy + \int 1-\cos 2y \, dy \\ &= -2 \cos y + y - \frac{\sin 2y}{2} = -2 \cos y + y - \sin y \cos y \\ &= -2\sqrt{1-u^2} + \arcsin u - u\sqrt{1-u^2} \\ &= -2\sqrt{1-x} + \arcsin \sqrt{x} - \sqrt{x}\sqrt{1-x}.\end{aligned}$$

The substitution  $u = \sqrt{1-x}$ ,  $x = 1-u^2$ ,  $dx = -2u \, du$  leads to

$$\int \frac{-2u^2 \, du}{1-\sqrt{1-u^2}}$$

and the substitution  $u = \sin y$  then leads to

$$\begin{aligned}\int \frac{-2 \sin^2 y \cos y \, dy}{1-\cos y} &= -2 \sin y - y - \sin y \cos y \\ &= -2u - \arcsin u - u\sqrt{1-u^2} \\ &= -2\sqrt{1-x} - \arcsin \sqrt{1-x} - \sqrt{1-x}\sqrt{x}.\end{aligned}$$

These answers agree, since

$$\arcsin \sqrt{x} = \frac{\pi}{2} - \arcsin \sqrt{1-x}$$

(check this by comparing their derivatives and their values for  $x = 0$ ).

6. In these problems  $I$  will denote the original integral.

(i)

$$\begin{aligned} I &= \int \frac{2}{x-1} dx + \int \frac{3}{(x+1)^2} dx \\ &= 2 \log(x-1) - \frac{3}{x+1}. \end{aligned}$$

(iii)

$$\begin{aligned} I &= \int \frac{1}{(x-1)^2} dx + \int \frac{4}{(x+1)^3} dx \\ &= -\frac{1}{(x-1)} - \frac{2}{(x+1)^2}. \end{aligned}$$

(v)

$$\begin{aligned} I &= \frac{1}{2} \int \frac{2x}{x^2+1} dx + \int \frac{4}{x^2+1} dx \\ &= \frac{1}{2} \log(x^2+1) + 4 \arctan x. \end{aligned}$$

(vii)

$$\begin{aligned} I &= \int \frac{1}{(x+1)} dx + \int \frac{2x}{(x^2+x+1)} dx \\ &= \int \frac{1}{x+1} dx + \int \frac{2x+1}{x^2+x+1} dx - \int \frac{1}{x^2+x+1} dx. \end{aligned}$$

Now

$$\begin{aligned} \int \frac{1}{x^2+x+1} dx &= \int \frac{1}{(x+\frac{1}{2})^2 + \frac{3}{4}} dx \\ &= \frac{4}{3} \int \frac{1}{\left[\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right]^2 + 1} dx \\ &= \frac{4}{3} \cdot \frac{\sqrt{3}}{2} \arctan\left(\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right) \\ &= \frac{2\sqrt{3}}{3} \arctan\left(\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right), \end{aligned}$$

so

$$I = \log(x+1) + \log(x^2+x+1) - \frac{2\sqrt{3}}{3} \arctan\left(\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right).$$

(ix)

$$\begin{aligned} I &= \int \frac{2x+1}{(x^2+x+1)^2} dx - \int \frac{1}{(x^2+x+1)^2} dx \\ &= \int \frac{2x+1}{(x^2+x+1)^2} dx - \frac{16}{9} \int \frac{1}{\left(\left[\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right]^2 + 1\right)^2} dx. \end{aligned}$$

Now the substitution

$$u = \frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right), \quad dx = \frac{\sqrt{3}}{2} du$$

changes the second integral to

$$-\frac{16}{9} \cdot \frac{\sqrt{3}}{2} \int \frac{du}{(u^2 + 1)^2}.$$

Using the reduction formula, this can be written

$$-\frac{8\sqrt{3}}{9} \left[ \frac{u}{2(u^2 + 1)} + \frac{1}{2} \int \frac{du}{u^2 + 1} \right] = -\frac{8\sqrt{3}}{9} \left[ \frac{\log(u^2 + 1)}{4} + \frac{1}{2} \arctan u \right],$$

so

$$I = -\frac{1}{x^2 + x + 1} - \frac{2\sqrt{3}}{9} \log \left( \frac{4}{3}(x^2 + x + 1) \right) - \frac{4\sqrt{3}}{9} \arctan \left( \frac{2}{\sqrt{3}} \left( x + \frac{1}{2} \right) \right).$$

13. The equation  $\int e^x \sin x \, dx = e^x \sin x - e^x \cos x - \int e^x \sin x \, dx$  means that any function  $F$  with  $F'(x) = e^x \sin x$  can be written  $F(x) = e^x \sin x - e^x \cos x - G(x)$  where  $G$  is another function with  $G'(x) = e^x \sin x$ . Of course,  $G = F + c$  for some number  $c$ , but it is not necessarily true that  $F = G$ .
15. (a)

$$\begin{aligned} \int \arcsin x \, dx &= \int 1 \cdot \arcsin x \, dx = x \arcsin x - \int \frac{x}{\sqrt{1-x^2}} \, dx \\ &= x \arcsin x + \sqrt{1-x^2}. \end{aligned}$$

16. (a)

$$\begin{aligned} \int \sin^4 x \, dx &= -\frac{\sin^3 x \cos x}{4} + \frac{3}{4} \int \sin^2 x \, dx \\ &= -\frac{\sin^3 x \cos x}{4} + \frac{3}{4} \left[ -\frac{\sin x \cos x}{2} + \frac{1}{2} \int 1 \, dx \right] \\ &= -\frac{\sin^3 x \cos x}{4} - \frac{3 \sin x \cos x}{8} + \frac{3}{8}x. \end{aligned}$$

$$\begin{aligned} \int \sin^4 x \, dx &= \int \left( \frac{1 - \cos 2x}{2} \right)^2 \, dx = \int \left( \frac{1}{4} - \frac{\cos 2x}{2} + \frac{\cos^2 2x}{4} \right) \, dx \\ &= \frac{x}{4} - \frac{\sin 2x}{4} + \frac{1}{4} \int \frac{1 + \cos 4x}{2} \, dx \\ &= \frac{x}{4} - \frac{\sin 2x}{4} + \frac{1}{4} \left[ \frac{x}{2} + \frac{\sin 4x}{8} \right] \\ &= \frac{3x}{8} - \frac{\sin 2x}{4} + \frac{\sin 4x}{32}. \end{aligned}$$

- (b) It follows that these two answers are the same, since they have the same value for  $x = 0$ .

20. (a)

$$\begin{aligned}\sin^n x \, dx &= \int (\sin x)(\sin^{n-1} x) \, dx \\ &= -\cos x \sin^{n-1} x + (n-1) \int \cos x (\sin^{n-2} x) \cos x \, dx \\ &= -\cos x \sin^{n-1} x + (n-1) \int (\sin^{n-2} x - \sin^n x) \, dx,\end{aligned}$$

so

$$\int \sin^n x \, dx = -\frac{1}{n} \cos x \sin^{n-1} x + \frac{n-1}{n} \int \sin^{n-2} x \, dx.$$

(b)

$$\begin{aligned}\int \cos^n x \, dx &= \int (\cos x)(\cos^{n-1} x) \, dx \\ &= \sin x \cos^{n-1} x + (n-1) \int \sin x (\cos^{n-2} x) \sin x \, dx \\ &= \sin x \cos^{n-1} x + (n-1) \int (\cos^{n-2} x - \cos^n x) \, dx,\end{aligned}$$

so

$$\int \cos^n x \, dx = \frac{1}{n} \sin x \cos^{n-1} x + \frac{n-1}{n} \int \cos^{n-2} x \, dx.$$

(c)

$$\begin{aligned}\int \frac{dx}{(1+x^2)^n} &= \int \frac{dx}{(1+x^2)^{n-1}} - \int \frac{x^2 \, dx}{(1+x^2)^n} \\ &= \int \frac{dx}{(1+x^2)^{n-1}} - \int x \cdot \frac{x}{(1+x^2)^n} \, dx \\ &= \int \frac{dx}{(1+x^2)^{n-1}} - \left[ \frac{x}{2(1-n)(1+x^2)^{n-1}} \right. \\ &\quad \left. - \int \frac{dx}{2(1-n)(1+x^2)^{n-1}} \right]\end{aligned}$$

so

$$\int \frac{dx}{(1+x^2)^n} = \frac{1}{2(n-1)} \frac{x}{(x^2+1)^{n-1}} - \frac{(2n-3)}{2(n-1)} \int \frac{1}{(x^2+1)^{n-1}} \, dx.$$

We can also use the substitution  $x = \tan u$ ,  $dx = \sec^2 u du$ , which changes the integral to

$$\begin{aligned}\int \frac{\sec^2 u du}{\sec^{2n} u} &= \int \cos^{2n-2} u du \\ &= \frac{1}{2n-2} \cos^{2n-3} u \sin u + \frac{2n-3}{2n-2} \int \cos^{2n-4} u du \\ &= \frac{1}{2n-2} \cdot \frac{1}{(\sqrt{1+x^2})^{2n-3}} \cdot \frac{x}{\sqrt{1+x^2}} + \frac{2n-3}{2n-2} \int \frac{dx}{(1+x^2)^{n-1}} \\ &= \frac{1}{2(n-1)} \frac{x}{(1+x^2)^{n-1}} + \frac{2n-3}{2n-2} \int \frac{dx}{(1+x^2)^{n-1}}.\end{aligned}$$

**CHAPTER 20**

1. (i)  $P_{3,0}(x) = e + ex + ex^2 + (5e/3!)x^3.$   
 (iii)  $P_{2n,\pi/2}(x) = 1 - \frac{(x-\pi/2)^2}{2!} + \frac{(x-\pi/2)^4}{4!} - \dots + \frac{(-1)^n(x-\pi/2)^{2n}}{(2n)!}.$   
 (v)  $P_{n,1}(x) = e + e(x-1) + \frac{e(x-1)^2}{2!} + \dots + \frac{e(x-1)^n}{n!}.$   
 (vii)  $P_{4,0}(x) = x + x^3.$   
 (ix)  $P_{2n+1,0}(x) = 1 - x^2 + x^4 - \dots + (-1)^n x^{2n}.$
2. If  $f$  is a polynomial function of degree  $n$ , then  $f^{(n+1)} = 0$ . It follows from Taylor's Theorem that  $R_{n,a}(x) = 0$ , so  $f(x) = P_{n,a}(x)$ .
  - (i)  $-12 + 2(x-3) + (x-3)^2.$   
 (iii)  $243 + 405(x-3) + 270(x-3)^2 + 90(x-3)^3 + 15(x-3)^4 + (x-3)^5.$
  3. (i)  $\sum_{i=0}^9 \frac{(-1)^i}{(2i+1)!} \left( \text{since } \frac{1}{(2n+2)!} < 10^{-17} \text{ for } 2n+2 \geq 19, \text{ or } n \geq 9 \right).$   
 (iii)  $\sum_{i=0}^8 \frac{(-1)^i}{2^i(2i+1)!} \left( \text{since } \frac{1}{2^{2n+2}(2n+2)!} < 10^{-20} \text{ for } 2n+2 \geq 18,$   
 $\text{or } n \geq 8 \right).$   
 (v)  $\sum_{i=0}^{11} \frac{2^i}{i!} \left( \text{since } \frac{3^2 \cdot 2^{n+1}}{(n+1)!} < 10^{-5} \text{ for } n+1 \geq 12, \text{ or } n \geq 11 \right).$
  8. (i)  $c_i = a_i + b_i.$   
 (iii)  $c_i = (i+1)a_i.$   
 (v)  $c_0 = \int_0^a f(t) dt; c_i = a_{i-1}/i \text{ for } i > 0.$

**CHAPTER 22**

1. (i)  $1 - n/(n+1) = 1/(n+1) < \epsilon$  for  $n+1 > 1/\epsilon$ .  
 (iii)  $\lim_{n \rightarrow \infty} \sqrt[8]{n^2+1} - \sqrt[4]{n+1} = \lim_{n \rightarrow \infty} (\sqrt[8]{n^2+1} - \sqrt[8]{n^2}) + \lim_{n \rightarrow \infty} (\sqrt[4]{n} - \sqrt[4]{n+1})$   
 $= 0 + 0 = 0.$  (Each of these two limits can be proved in the same way that  $\lim_{n \rightarrow \infty} (\sqrt{n+1} - \sqrt{n}) = 0$  was proved in the text.)

- (v) Clearly  $\lim_{n \rightarrow \infty} (\log a)/n = 0$ . So  $\lim_{n \rightarrow \infty} \sqrt[n]{a} = \lim_{n \rightarrow \infty} e^{(\log a)/n} = e^0$  (by Theorem 1) = 1.
- (vii)  $\sqrt[n^2]{n^2} \leq \sqrt[n^2+n]{n^2+n} \leq \sqrt[n^2]{2n^2}$ , so  $(\sqrt[n]{n})^2 \leq \sqrt[n^2+n]{n^2+n} \leq \sqrt[2]{(\sqrt[n]{n})^2}$ , and  $\lim_{n \rightarrow \infty} (\sqrt[n]{n})^2 = \lim_{n \rightarrow \infty} \sqrt[2]{(\sqrt[n]{n})^2} = 1$  by parts (v) and (vi).
- (ix) Clearly  $\alpha(n) \leq \log_2 n$ , and  $\lim_{n \rightarrow \infty} (\log_2 n)/n = 0$ .
5. (a) If  $0 < a < 2$ , then  $a^2 < 2a < 4$ , so  $a < \sqrt{2a} < 2$ .
- (b) Part (a) shows that

$$\sqrt{2} < \sqrt{2\sqrt{2}} < \sqrt{2\sqrt{2\sqrt{2}}} < \dots < 2,$$

so the sequence converges by Theorem 2.

- (c) If this sequence is denoted by  $\{a_n\}$ , then the sequence  $\{\sqrt{2a_n}\}$  is the same as  $\{a_{n+1}\}$ . So the hint shows that  $l = \sqrt{2l}$ , or  $l = 2$ .
8. If  $x$  is rational, then  $n! \pi x$  is a multiple of  $\pi$  for sufficiently large  $n$ , so  $(\cos n! \pi x)^{2k} = 1$  for all such  $n$ , so  $\lim_{n \rightarrow \infty} (\lim_{k \rightarrow \infty} (\cos n! \pi x)^{2k}) = 1$ . If  $x$  is irrational, then  $n! \pi x$  is not a multiple of  $\pi$  for any  $n$ , so  $|\cos n! \pi x| < 1$ , so  $\lim_{k \rightarrow \infty} (\cos n! \pi x)^{2k} = 0$ , so  $f(x) = 0$ .
9. (i)  $\int_0^1 e^x dx = e - 1$ . (Use partitions of  $[0, 1]$  into  $n$  equal parts.)
- (iii)  $\int_0^1 \frac{1}{1+x} dx = \log 2$ .
- (v)  $\int_0^1 \frac{1}{(1+x)^2} dx = \frac{1}{2}$ .

## CHAPTER 23

1. (i) (Absolutely) convergent, since  $|(\sin n\theta)/n^2| \leq 1/n^2$ .
- (iii) Divergent, since the first  $2n$  terms have sum  $\frac{1}{2} + \dots + 1/n$ . (Leibniz's Theorem does not apply since the terms are not decreasing in absolute value.)
- (v) Divergent, since

$$\frac{1}{\sqrt[3]{n^2 - 1}} \geq \frac{1}{2n^{2/3}}$$

for sufficiently large  $n$ .

- (vii) Convergent, since

$$\lim_{n \rightarrow \infty} \frac{(n+1)^2/(n+1)!}{n^2/n!} = \lim_{n \rightarrow \infty} \left(\frac{n+1}{n}\right)^2 \cdot \frac{1}{n+1} = 0.$$

- (ix) Divergent, since  $1/(\log n) > 1/n$ .

- (xi) Convergent, since  $1/(\log n)^n < \frac{1}{2^n}$  for  $n > 9$ .

- (xiii) Divergent, since

$$\frac{n^2}{n^3 + 1} > \frac{1}{2n}$$

for large enough  $n$ .

(xv) Divergent, since

$$\int_2^N \frac{1}{x \log x} dx = \log(\log N) - \log(\log 2) \rightarrow \infty \text{ as } N \rightarrow \infty.$$

(Notice that  $f(x) = 1/(x \log x)$  is decreasing on  $[2, \infty)$ , since

$$f'(x) = \frac{-[1 + \log x]}{(x \log x)^2} < 0 \quad \text{for } x > 1.$$

(xvii) Convergent, since  $1/n^2(\log n) < 1/n^2$  for  $n > 2$ .

(xix) Convergent, since

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{2^{n+1}(n+1)!/(n+1)^{n+1}}{2^n n!/n^n} &= \lim_{n \rightarrow \infty} \frac{2(n+1)n^n}{(n+1)^{n+1}} \\ &= \lim_{n \rightarrow \infty} \frac{2}{\left(1 + \frac{1}{n}\right)^n} = \frac{2}{e}, \end{aligned}$$

by Problem 18-16.

5. (a) For each  $N$  we clearly have

$$0 \leq \sum_{n=1}^N a_n 10^{-n} < 9 \sum_{n=1}^{\infty} 10^{-n} = 1,$$

so  $\sum_{n=1}^{\infty} a_n 10^{-n}$  converges by the boundedness criterion, and lies between 0 and 1. (Actually, this number is denoted by  $0.a_1a_2a_3a_4\dots$  only when the sequence  $\{a_n\}$  is not eventually 0.)

17. The area of the shaded region is  $\frac{1}{2}$ . The integral is

$$\begin{aligned} \frac{1}{2}([1 - \frac{1}{2}] + [\frac{1}{4} - \frac{1}{8}] + [\frac{1}{16} - \frac{1}{32}] + \dots) - \frac{1}{2}([\frac{1}{2} - \frac{1}{4}] + [\frac{1}{8} - \frac{1}{16}] + \dots) \\ = \frac{1}{2}(\frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \dots) - \frac{1}{2}(\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \dots) \\ = \frac{1}{4}(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \dots) - \frac{1}{8}(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \dots) \\ = \frac{1}{8} \left(1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \dots\right) \\ = \frac{1}{8} \cdot \frac{1}{1 - \frac{1}{4}} \\ = \frac{1}{6}. \end{aligned}$$

#### CHAPTER 24

1. (i)

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) = \begin{cases} 0, & x = 0 \\ 1, & 0 < x \leq 1. \end{cases}$$

$\{f_n\}$  does not converge uniformly to  $f$ .

(iii)  $f(x) = \lim_{n \rightarrow \infty} f_n(x) = 0$  (since  $\lim_{n \rightarrow \infty} x^n = \infty$  for  $x > 1$ ). The sequence  $\{f_n\}$  does not converge uniformly to  $f$ ; in fact, for any  $n$  we have  $f_n(x)$  large for sufficiently large  $x$ .

(v)  $f(x) = \lim_{n \rightarrow \infty} f_n(x) = 0$ , and  $\{f_n\}$  converges uniformly to  $f$ , since  $|f_n(x)| \leq 1/n$  for all  $x$ .

3. (i)  $-\frac{1}{a} - \frac{x}{a^2} - \frac{x^2}{a^3} - \dots$

(iii)  $\sum_{k=0}^{\infty} (-1)^k \binom{-\frac{1}{2}}{k} x^k$ .

(v)  $\sum_{k=0}^{\infty} \frac{(-1)^k \binom{-\frac{1}{2}}{k}}{2k+1} x^{2k+1}$ .

4. (i)  $e^{-x}$ .

(iii) If

$$f(x) = \frac{x^2}{2} - \frac{x^3}{3 \cdot 2} + \frac{x^4}{4 \cdot 3} - \dots, \quad |x| \leq 1$$

then

$$\begin{aligned} f'(x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \dots \\ &= \log(1+x) \quad |x| < 1, \end{aligned}$$

so for  $|x| < 1$  we have  $f(x) = (1+x)\log(1+x) - (1+x) + c$  for some number  $c$ . Since  $f(0) = 0$ , we have  $c = 1$ , so  $f(x) = (1+x) \cdot \log(1+x) - x$  for  $|x| < 1$ .

6. Since

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

we have

$$f(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!}$$

(notice that the right side is 1 for  $x = 0$ ). So

$$f^{(k)}(0) = \begin{cases} \frac{(-1)^n}{(2n+1)!}, & k = 2n \\ 0, & k \text{ odd.} \end{cases}$$

#### CHAPTER 25

1. (i)  $|3+4i| = 5$ ;  $\theta = \arctan \frac{4}{3}$ .

(iii)  $|(1+i)^5| = (|1+i|)^5 = (\sqrt{2})^5$ ; since  $\pi/4 = \arctan 1/1$  is an argument for  $1+i$ , an argument for  $(1+i)^5$  is  $5\pi/4$ .

(v)  $|(|3+4i|)| = |5| = 5$ ;  $\theta = 0$ .

2. (i)

$$\begin{aligned}x &= \frac{-i \pm \sqrt{-1 - 4}}{2} \\&= \frac{-i \pm \sqrt{5}i}{2} \\&= \frac{(-1 + \sqrt{5})i}{2} \quad \text{or} \quad \frac{(-1 - \sqrt{5})i}{2}.\end{aligned}$$

- (iii)  $x^2 + 2ix - 1 = (x + i)^2$ , so the only solution is  $x = -i$ .  
 (v)  $x^3 - x^2 - x - 2 = (x - 2)(x^2 + x + 1)$ . The solutions are

$$2, \quad -\frac{1}{2} + \frac{\sqrt{3}}{2}i, \quad -\frac{1}{2} - \frac{\sqrt{3}}{2}i.$$

3. (i) All  $z = iy$  with  $y$  real.

(iii) All  $z$  on the perpendicular bisector of the line segment between  $a$  and  $b$ .  
 (v) For  $z = x + iy$ , we clearly need  $1 - x = 1 - \text{real part of } z \geq 0$ , or  $x \leq 1$ . For such  $x$ , our condition  $\sqrt{x^2 + y^2} < 1 - x$  is equivalent to  $x^2 + y^2 < (1 - x)^2$ , or  $x < (1 - y^2)/2$ . The set of points with  $x = (1 - y^2)/2$  is the parabola pointing along the second axis, with the point  $(0, 1/2)$  closest to the origin, and intersecting the line  $x = 1$  at  $(1, 1)$  and  $(1, -1)$ . The area bounded by this parabola and the line  $x = 1$  is the desired set of points.

4.  $|x + iy|^2 = x^2 + y^2 = x^2 + (-y)^2 = |x - iy|^2$ .

$$(z + \bar{z})/2 = [(x + iy) + (x - iy)]/2 = x.$$

$$(z - \bar{z})/2 = [(x + iy) - (x - iy)]/2i = y.$$

5.  $|z + w|^2 + |z - w|^2 = (z + w)(\bar{z} + \bar{w}) + (z - w)(\bar{z} - \bar{w}) = 2z\bar{z} + 2w\bar{w} = 2(|z|^2 + |w|^2)$ . Geometrically, this says that the sum of the squares of the diagonals of a parallelogram equal the sum of the squares of the sides.

#### CHAPTER 27

1. (i) Converges absolutely, since  $|(1+i)^n/n!| \leq (\sqrt{2})^n/n!$ , and  $\sum_{n=1}^{\infty} (\sqrt{2})^n/n!$  converges.  
 (iii) Converges, but not absolutely, since the real terms form the series

$$-\frac{1}{2} + \frac{1}{4} - \frac{1}{6} + \frac{1}{8} - \dots$$

and the imaginary terms form the series

$$i\left(\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots\right).$$

(v) Diverges, since the real terms form the series

$$\frac{\log 3}{3} + 2\frac{\log 4}{4} + \frac{\log 5}{5} + \frac{\log 7}{7} + 2\frac{\log 8}{8} + \frac{\log 9}{9} + \dots$$

2. (i) The limit

$$\lim_{n \rightarrow \infty} \frac{|z|^{n+1}/(n+1)^2}{|z|^n/n^2} = \lim_{n \rightarrow \infty} \left(\frac{n+1}{n}\right)^2 |z| = |z|$$

is  $< 1$  for  $|z| < 1$ , but  $> 1$  for  $|z| > 1$ .

(iii) The limit

$$\lim_{n \rightarrow \infty} \frac{|z|^{n+1}}{|z|^n} = |z|$$

is  $< 1$  for  $|z| < 1$  but  $> 1$  for  $|z| > 1$ .

(v) The limit

$$\lim_{n \rightarrow \infty} \frac{2^{n+1} |z|^{(n+1)!}}{2^n |z|^n} = \lim_{n \rightarrow \infty} 2 |z|^{(n+1)! - n!}$$

is 0 for  $|z| < 1$ , but  $\infty$  for  $|z| > 1$ .

3. (i) The limits

$$\lim_{n \rightarrow \infty} \sqrt[2n]{\frac{|z|^{2n}}{3^n}} = \frac{|z|}{\sqrt{3}} \quad \text{and} \quad \lim_{n \rightarrow \infty} \sqrt[2n+1]{\frac{|z|^{2n+1}}{2^{n+1}}} = \frac{|z|}{\sqrt{2}}$$

are  $< 1$  for  $|z| < \sqrt{2}$ , so the series converges absolutely for  $|z| < \sqrt{2}$ . But the series does not converge absolutely for  $|z| > \sqrt{2}$ , so the radius of convergence is  $\sqrt{2}$ .

(iii) Since

$$\lim_{n \rightarrow \infty} \sqrt[n]{\frac{n! |z|^n}{n^n}} = \lim_{n \rightarrow \infty} \frac{|z| \sqrt[n]{n!}}{n} \leq \lim_{n \rightarrow \infty} \frac{|z|}{n} = 0,$$

the series converges absolutely for all  $z$ , so the radius of convergence is  $\infty$ .

(v) The limit

$$\lim_{n \rightarrow \infty} \sqrt[n]{2^n z^n} = 2 \lim_{n \rightarrow \infty} z^{(n-1)!}$$

is 0 for  $|z| < 1$ , but  $\infty$  for  $|z| > 1$ , so the radius of convergence is 1.

# GLOSSARY OF SYMBOLS



$P$	9	$(-\infty, a)$	57	$e$	241, 328
$ a $	11	$(-\infty, a]$	57	$c + d$	242
$\sqrt{x}$	12	$(-\infty, \infty)$	57	$a \cdot c$	242
$\max(x, y)$	16	$[x]$	72	$c \cdot d$	243
$\min(x, y)$	16	$\{x\}$	72	$\det(c, d)$	243
$\epsilon$ ("epsilon")	18	$v + w$	75	$c'$	243
$N$	21	$v \cdot w$	78	$R(f, a, b)$	250
$\emptyset$	23	$\ v\ $	78	$L(f, P)$	251
$n!$	23	$\det(v, w)$	79	$U(f, P)$	251
$\sum_{i=1}^n a_i$		$\delta$ ("delta")	96	$\int_a^b f$	
		$\lim_{x \rightarrow a} f(x)$	99	$\int_a^b f(x) dx$	
$Z$	25	$\lim_a f$	99	$\ell(f, P)$	274
$Q$	25	$\lim_{x \rightarrow a^+} f(x)$	104	$\mathcal{L}(x)$	275
$R$	25, 579	$\lim_{x \downarrow a} f(x)$	104	$\mathbf{L} \int_a^b f$	292
$\binom{n}{k}$	27	$\lim_{x \rightarrow a^-} f(x)$	104	$\mathbf{U} \int_a^b f$	292
$f(x)$	40, 47, 591	$\lim_{x \uparrow a} f(x)$	104	$\int_a^\infty f$	298
$I$	43	$\lim_{x \rightarrow \infty} f(x)$	105	$\int_a^\infty f(x) dx$	298
$f + g$	43, 242	$\lim_{x \rightarrow -\infty} f(x)$	111	$\int_{-\infty}^a f$	298
$A \cap B$	43	$\lim_{x \rightarrow \infty} f(x) = \infty$	111	$\int_{-\infty}^\infty f$	298
$f \cdot g$	43	$\lim_{x \rightarrow \infty} f(x) = \infty$	111	$\sin^\circ$	301
$f/g$	43	$\sup A$	132	$\sin'$	301
$c \cdot g$	43	$\text{lub } A$	132	$\pi$	302
$\{x : \dots\}$	43	$\inf A$	132	$A(x)$	303
$\{a, \dots, z\}$	44	$\text{glb } A$	132	$\cos$	303, 305, 554
$f + g + h$	44	$\overline{\lim} A$	141	$\sin$	303, 305, 554
$f \cdot g \cdot h$	44	$\limsup A$	141	$\sec$	397
$f \circ g$	44	$\underline{\lim} A$	141	$\tan$	307
$f \circ g \circ h$	45	$f'(a)$	149	$\csc$	307
$x \rightarrow f(x)$	45	$f'$	149	$\cot$	307
$\prod_{i=1}^n a_i$		$\frac{df(x)}{dx}$	152	$\arcsin$	307
		$\frac{d}{dx} \Big _{x=a}$	153	$\arccos$	308
$a^{bc}$	49	$f''$	159	$\arctan$	308
$C_A$	50	$f'''$	159	$e$	328
$A \cup B$	50	$f^{(k)}$	159	$\log$	338
$R - A$	50	$\frac{d^2 f(x)}{dx^2}$	160	$\exp$	340, 554
$ f $	51	$f^{-1}$	228	$e$	340
$\max(f, g)$	51				
$\min(f, g)$	51				
$f < g$	53				
the pair $(a, b)$	54				
the open interval $(a, b)$	56				
$[a, b]$	57				
$(a, \infty)$	57				
$[a, \infty)$	57				

$e^x$	341	$\lim_{n \rightarrow \infty} a_n = \infty$	449	$\sin$	554
$a^x$	342	$\gamma$	456	$\cos$	554
$\log_a$	343	$\overline{\lim}_{n \rightarrow \infty} x_n$	460	$\exp$	554
$\sinh$	349	$\limsup_{n \rightarrow \infty} x_n$	460	$b_n$	563
$\cosh$	349	$\lim_{n \rightarrow \infty} x_n$	461	$B_n$	563
$\tanh$	349	$\liminf_{n \rightarrow \infty} x_n$	461	$D$	564
$\arg \cosh$	350	$N(n; a, b)$	462	$D^k$	564
$\arg \sinh$	350	$\sum_{i=1}^{\infty} a_n$	465	$e^D$	564
$\arg \tanh$	350	$i$	517, 523	$\Delta$	564
$\text{Nap log}$	354	$\mathbf{C}$	522	$\varphi_n$	566
$F(x) \Big _a^b$	359	$\bar{z}$	525	$\psi_n$	567
$\int f$	360	$ z $	525	$+$	571, 581
$\int f(x) dx$	360	$\text{Re}$	532	$\cdot$	571, 584
$\Gamma(x)$	390	$\text{Im}$	532	$0$	571, 581
$P_{n,a}$	406	$\theta$	533	$1$	572, 586
$P_{n,a,f}$	406	$\lim_{z \rightarrow a} f(z)$	533	$-a$	572, 582
$R_{n,a}$	415	$f'(a)$	542	$a^{-1}$	572, 586
$\binom{a}{n}$	429			$\mathbf{P}$	573
$\{a_n\}$	445			$\triangleright$	574, 580
$\lim_{n \rightarrow \infty} a_n$	446			$\triangleleft$	574, 580
				$\geq$	574, 580
				$\leq$	574, 580
				$ \alpha $	585

# INDEX



AalbmndoE, 273  
 Abel, Niels Henrik, 404, 513  
 Abel summable, 514  
 Abel's formula for summation by parts, 388  
 Abel's Lemma, 389  
 Abel's test, 488  
 Abel's Theorem, 513  
 Absolute value, 11  
     of a complex number, 525  
 Absolutely convergent, 473, 547  
 Absolutely summable, 473  
 Acceleration, 159  
*Acta Eruditorum*, 146  
 Addition, 3  
     associative law for, 9  
     commutative law for, 9  
     of complex numbers, 522  
         geometric interpretation of, 526  
     of vector-valued functions, 242  
     of vectors, 75  
 Addition formula  
     for arcsin, 314  
     for arctan, 314  
     for cos, 311  
     for sin, 310, 311  
     for tan, 314  
 Additive identity  
     existence of, 9  
     for vectors, 76  
 Additive inverses  
     existence of, 9  
 Algebra, Fundamental Theorem of, 373, 529, 539, 558  
 Algebraic functions, 359  
 Algebraic number, 435  
 Algebraist's real numbers, 588  
 Almost lower bound, 140  
 Almost upper bound, 140  
 Analyst's real numbers, 588  
 Angle, 300  
     directed, 300  
 Antidiagonal, 239  
 Arabic numerals, multiplication of, 8  
 Arc length, 275, 281  
 Arccos, 308  
     derivative of, 308  
 Archimedes, 136, 139, 260  
 Archimedean property  
     for the rational numbers, 574  
     for the real numbers, 136  
 Archimedean spiral, 85, 246  
 Arcsec, 317, 379  
 Arcsin, 307  
     addition formula for, 314  
     derivative of, 308  
     Taylor series for, 509  
 Arctan, 308  
     addition formula for, 314  
     derivative of, 308  
     Taylor polynomial for, 407, 414  
     remainder term for, 422  
 Area, 250, 255  
 Arg cosh, 350  
 Arg sinh, 350  
 Arg tanh, 350  
 Argument, 527  
 Argument function, 533  
     discontinuities of, 537  
 Argument of the hyperbolic functions, 350  
 Arithmetic mean, 33  
 Arrow, 75, 76  
     “ $x \rightarrow \sin(x^2)$ ”, 45  
 Associative law  
     for addition, 9  
     of vectors, 76  
     for multiplication, 9  
 Average velocity, 150  
 Axis  
     horizontal, 57  
     imaginary, 525  
     real, 524  
     vertical, 57  
 Bacon, Francis, vi  
 Basic properties of numbers, 3  
 “Bent graphs”, 147  
 Bernoulli, 146, 565  
 Bernoulli numbers, 563  
 Bernoulli polynomials, 566  
 Bernoulli's inequality, 32  
 Big game hunting, mathematical theory of, 543  
 Binary operation, 571

- Binomial coefficient, 27, 429
- Binomial series, 487, 510
- Binomial theorem, 28
- Bisection argument, 140, 543
- Bohr, Harold, 390
- Bolzano-Weierstrass Theorem, 451, 461, 543
- Bound
  - almost lower, 140
  - almost upper, 140
  - greatest lower, 132
  - least upper, 131, 574
  - lower, 132
  - upper, 131
- Bounded above, 120, 131, 450, 574
- Bounded below, 132, 450
- Bourbaki, Nicholas, 146
  
- Cardioid, 89, 247
- Cartesian coordinates, 84
- Cauchy, 278
- Cauchy Condensation Theorem, 488
- Cauchy criterion, 466
- Cauchy form of the remainder, 417, 419
- Cauchy Mean Value Theorem, 201
- Cauchy product, 486, 505
- Cauchy sequence, 452, 562
  - equivalence of, 589
- Cauchy-Hadamard formula, 560
- Cauchy-Schwarz inequality, 278
- Cavalieri, 272
- Cesaro summable, 486
- Chain Rule, 172 ff.
  - proof of, 176
- Change, rate of, 150
- Characteristic (of a field), 576
- Circle, 65
  - “ $f$  circle  $g$ ”, 44
  - unit, 66
- Circle of convergence, 550
- Classical notation
  - for derivatives, 152–154, 160, 165, 184, 238
  - for integrals, 262
- Cleio, 183
- Closed interval, 57
  
- Closed rectangle, 538
- Closure under addition, 9
- Closure under multiplication, 9
- Commutative law
  - for addition, 9
  - of vectors, 76
  - for multiplication, 9
- Comparison test, 467, 468
- Comparison Theorem, Sturm, 320
- Complete induction, 23
- Complete ordered field, 574, 593
- Completing the square, 17, 375
- Complex analysis, 556
- Complex function
  - continuous, 536
  - differentiable, 541
  - graph of, 533
  - limit of, 533
  - nondifferentiable, 542
  - Taylor series for, 554
- Complex  $n$ th root, 527
- Complex numbers, 517, 522
  - absolute value of, 525
  - addition of, 522
    - geometric interpretation of, 526
    - geometric interpretation of, 525
  - imaginary part of, 522
  - infinite sequence of, 546
  - infinite series of, 546–548
  - logarithm of, 561
  - modulus of, 525
  - multiplication of, 522
    - geometric interpretation of, 526–527
  - real part of, 522
- Complex plane, 524
- Complex power series, 548
  - circle of convergence of, 550
  - radius of convergence of, 550
- Complex-valued functions, 532
- Composition of functions, 44
- Concave function, 217
- Conditionally convergent series, 474
- Cone, 80
  - generating line of, 80
  - surface area of, 399
- Conic sections, 80; *see also* Ellipse, Hyperbola, Parabola
- Conjugate, 525, 530

- Conjugate function, 532
- Constant function, 43
- Construction of the real numbers, 578 ff.
- Continued fraction, 455
- Continuous, uniformly, 142
- Continuous at  $a$ , 113, 536
- Continuous function, 113, 116, 537
  - nowhere differentiable, 157, 501
- Continuous on  $(a, b)$ , 116
- Continuous on  $[a, b]$ , 116
- Contraction, 459
- Contraction lemma, 459
- Converge
  - pointwise, 494
  - uniformly, 494, 498
- Convergent sequence, 446, 546
- Convergent series, 465, 547
  - absolutely, 473, 547
  - conditionally, 474
- Convex function, 216
  - strictly, 226
  - weakly, 226
- Convex subset of the plane, 226, 544
- Cooling, Newton's law of, 352
- Coordinate
  - first, 57
  - second, 57
- Coordinate system, 57
  - cartesian, 84
  - origin of, 57
- Coordinates
  - polar, 84
- "Corner", 60
- Cos, 300, 303, 318–319, 554
  - addition formula for, 311
  - derivative of, 170, 304
  - inverse of, *see* Arccos
  - Taylor polynomials for, 407
  - remainder term for, 420
- Cosh, 349
- Cosine, hyperbolic, 349
- Cot, 307
  - derivative of, 307
- Countable, 442
- Counting numbers, 21
- Critical point, 187
- Critical value, 187
- Csc, 307
- derivative of, 307
- Cubic equation, general solution, 519–520
- Curve
  - parameterized, tangent line of, 243
  - parametric representation of, 241
  - reparameterization of, 244
- Cycloid, 247
- Darboux's Theorem, 211
- De Moivre's Theorem, 527
- Decimal expansion, 73, 485
- Decreasing function, 192
- Decreasing sequence, 450
- Dedekind, Richard, 38
- Defined implicitly, 238
- Definite integral, 361
- DEFINITION**, 47
- Definition, recursive, 23
- Degree (of a polynomial), 42
- Degree measurement, 63, 301–302
- Delicate ratio test, 486
- Delicate root test, 486
- Dense, 138
- Derivative, 147 ff., 149
  - classical notation for, 152–154, 160, 165, 184, 238
  - higher-order, 159
  - "infinite", 156
  - left-hand, 154
  - Leibnizian notation for, *see* Derivative, classical notation for
  - logarithmic, 348
  - "negative infinity", 156
  - of  $f$ , 149
  - of  $f$  at  $a$ , 149
  - of vector-valued function, 243
  - right-hand, 154
  - Schwarzian, 182
  - second, 159
  - Schwarz, 431
- Derivative of quotient, incantation for, 169
- Descartes, René, 84
- Determinant, 79
  - of vector-valued functions, 243
- Diagonal, 230

- Difference operator, 564
- Differentiable, 149, 541
- Differential equation, 289, 297, 318, 320, 352, 357, 432
  - initial conditions for, 433
- Differentiation, 166 ff.
  - implicit, 238
  - logarithmic, 348
- Differentiation operator, 564
- Dini's Theorem, 515
- Directed angle, 300
- Dirichlet's test, 488
- Disc method, 397
- Discontinuities of a nondecreasing function, 443
- Discontinuity, removable, 119
- Disraeli, Benjamin, 2
- Distance, 58, 525
  - shortest between two points, 275
- Distributive law, 9
- Diverge, 446, 547
- Division, 6
- Division by zero, 6
- Domain, 40, 41, 47, 591
- Dot product
  - of vectors, 78
  - of vector-valued functions, 243
- Double intersection, 163
- Double root, 183
- Durège, 38
  
- $e$ , 340
  - irrationality of, 425
  - relation with  $\pi$ , 441, 555
  - transcendentality of, 437
  - value of, 341, 422
- Eccentricity of ellipse, 87
- Elementary function, 359
- Ellipse, 66, 82
  - axes of, 87
  - eccentricity of, 87
  - equation in polar coordinates, 86
  - focus point of, 66, 86
  - major axis of, 87
  - minor axis of, 87
- Ellipsoid of revolution, 400
- Empty collection, 23
  
- Entire function, 558
- Epsilon, 18
- Equal up to order  $n$ , 412
- Equality, order of, 412
- Equations, differential, *see* Differential equations
- Equivalent Cauchy sequences, 589
- Etymology lesson, 82
- Euler, 565
- Euler's number, 456
- Euler-Maclaurin Summation Formula, 566
- Even function, 51, 196
- Even number, 25
- Eventually inside, 546
- Exhaustion, method of, 139
- Exp, 340 ff., 554
  - classical approach to, 354
  - elementary definition of, 461
  - Taylor polynomials for, 407
  - remainder term for, 422
- Expansion, decimal, 73, 485
- Extension of a function, 113–114
  
- Factorial, 23
- Factorials, table of, 428
- Factorization into primes, 31
- Fibonacci, 32
- Fibonacci Association, 32
- Fibonacci Quarterly, The*, 32
- Fibonacci sequence, 32, 512, 563
- Field, 571
  - characteristic of, 576
  - complete ordered, 574, 593
  - ordered, 573
- First coordinate, 57
- First Fundamental Theorem of Calculus, 282
- Fixed point of a function, 458
- Focus point, 66, 86
- Force, as vector, 76
- Four leaf clover, 88
- Fourier series, 315, 317, 320
- Fraction, continued, 455
- Function, 39, 47
  - absolute value, 532

Function (*continued*)

argument, 533  
 discontinuities of, 537  
 complex valued, 532  
 composition of, 44  
 concave, 217  
 conjugate, 532  
 constant, 43  
 continuous, 113 ff.  
 convex, 216  
 critical point of, 187  
 critical value of, 187  
 decreasing, 192  
 derivative of, 147 ff.  
 differentiable, 149, 541  
 elementary, 359  
 entire, 558  
 even, 51, 196  
 exponential, 340–341  
 extension of, 113–114  
 fixed point of, 458  
 from  $A$  to  $B$ , 591  
 from real numbers to the plane, 241  
 graphs of, 57–65, 195, 533  
 hyperbolic, 349  
 identity, 43  
 imaginary part, 532  
 increasing, 192  
 integrable, 255  
 integral of, 255  
 inverse, 228 ff.  
 linear, 58  
 local maximum point of, 186  
 local minimum point of, 186  
 local strict maximum point of, 215  
 logarithm, 338, 343  
 maximum point of, 185  
 maximum value of, 185  
 minimum value of, 185  
 most general definition of, 591  
 negative part of, 51  
 nondecreasing, 240  
 nonincreasing, 240  
 nonnegative, 51  
 notation for, 40, 45  
 odd, 51, 196  
 periodic, 71, 162, 296

polynomial, 42  
 positive part of, 51  
 power, 60  
 product of, 43  
 quotient of, 43  
 rational, 42  
 real part, 532  
 real-valued, 532  
 “reasonable”, 68, 116, 178  
 regulated, 515  
 step, 275  
 strict maximum point of, 215  
 sum of, 43  
 trigonometric, 300 ff.  
 value at  $x$ , 40  
 vector-valued, 241

Fundamental Theorem of Algebra, 373,  
 529, 539, 558

## Fundamental Theorem of Calculus

First, 282  
 Second, 286

Gabriel, 402  
 Galileo, 146, 162  
 Gamma function, 390, 437  
 Generating line, of a cone, 80  
 Geometric mean, 33  
 Geometric series, 466  
 Global property, 121  
 Goes to, “ $x$  goes to  $\sin(x^2)$ ”, 45  
 Graph of polynomial function, 194  
 Graph sketching, 193–198  
 Graphs, 57–65, 85 ff., 90–91, 196, 533  
 Gravitation, 327  
 Greatest lower bound, 132  
 Grin and bear it, 381–382  
 Gronwall’s inequality, 353  
 Grow  
 at the same rate as, 358  
 faster than, 358

- Hadamard, 560
- Half-life (of radioactive substance), 352
- Hermite, 436
- High-school student's real numbers, 589
- Higher-order derivatives, 159
- Hilbert, 436
- Horizontal axis, 57
- Hyperbola, 67, 82
  - equation in polar coordinates, 88
- Hyperbolic cosine, 349
- Hyperbolic functions, 349
- Hyperbolic sine, 349
- Hyperbolic spiral, 313
- Identity
  - additive, 9
  - multiplicative, 9
- Identity function, 43
- Identity operator, 564
- Imaginary axis, 525
- Imaginary part function, 532
- Imaginary part of a complex number, 522
- Implicit differentiation, 238
- Implicitly defined, 238
- Improper integral, 298–299, 391–393
- Incantation for derivative of quotient, 169
- Increasing at  $a$ , 214
- Increasing function, 192
- Increasing sequence, 450
- Indefinite integral, 361
- Indefinite integrals, short table of, 361–362
- Induction, mathematical, 21
  - complete, 23
- Inductive set of real numbers, 34
- Inequalities, 9
  - in an ordered field, 574
- Inequality
  - Bernoulli's, 32
  - Cauchy-Schwarz, 278
  - geometric-arithmetic mean, 33
  - Gronwall's, 353
  - Jensen's, 225
- Liouville's, 441
- Schwarz, 17, 32, 278
- triangle, 71
- Young's, 273
- Infimum, 132
- “Infinite” derivative, 156
- Infinite intervals; 57
- Infinite product, 326, 391
- Infinite products, 489
- Infinite sequence, 445, 546
- Infinite series, 465
  - multiplication of, 479–481
- Infinite sum, 426, 464
- Infinite trumpet, 402
- Infinitely many primes, 32
- “Infinitely small”, 153, 261
- Infinity, 57
  - minus, 57
- Inflection point, 222
- Initial conditions for differential equations, 433
- Instantaneous speed, 150
- Instantaneous velocity, 150
- Integer, 25
- Integrable, 255
- Integral, 255
  - classical notation for, 262
  - definite, 361
  - improper, 298–299, 391–393
  - indefinite, 361
    - short table of, 361–362
  - Leibnizian notation for, *see* Integral, classical notation for
  - lower, 292
  - Mean Value Theorem for, 274
  - Second Mean Value Theorem for, 387
  - upper, 292
- Integral form of the remainder, 417, 418
- Integral sign, 255
- Integral test, 471
- Integration
  - by parts, 362 ff.
  - by substitution, 365 ff.
  - limits of, 255
  - of rational functions, 373 ff.
- Interest (finance), 351

- Intermediate Value Theorem, 122, 129, 133, 296
- Interpolation, Lagrange, 49
- Intersection of sets, 43
- Interval, 56
  - closed, 57
  - infinite, 57
  - open, 56; *see also* Nested Intervals Theorem
- Inverse
  - additive, 9
  - multiplicative, 9
- Inverse of a function, 228 ff.
- Inverse square law, 327
- Inverses of trigonometric functions, *see* Trigonometric functions
- Irrational numbers, 25
- Isomorphic fields, 592
- Isomorphism, 592
  
- Jensen's inequality, 225
- Johnson, Samuel, 597
- Jump, 60
  
- Kepler, 327
- Kepler's laws of planetary motion, 327
  
- Lagrange form of the remainder, 417, 418
- Lagrange interpolation formula, 49
- Large negative, 64
- Least upper bound, 131 ff., 574
- Least upper bound property, 133
- Lebesgue, *see* Riemann-Lebesgue Lemma
- Left-hand derivative, 154
- Leibniz, 153, 261
- Leibniz's formula, 182
- Leibniz's Theorem, 474
- Leibnizian notation for derivatives, 153–154, 165, 184, 238
- for higher order derivatives, 160
  
- Lemma, 100
- Lemniscate, 89
- Length, 275, 281
- L'Hôpital, Marquis de, 146
- L'Hôpital's Rule, 201, 210–211
- Limit, 90 ff., 96, 533
  - at infinity, 105
  - "does not exist", 99
  - from above, 104
  - from below, 104
  - of a sequence, 446
  - of vector-valued function, 243, 249
  - uniqueness of, 98
- Limit of integration, 255
- Limit point, 462, 543
- Limit superior, 141, 460
- Lindemann, 440
- Line, real, 56
- Line, tangent, *see* Tangent line
- Linear functions, 58
- Liouville, 441
- Liouville's inequality, 441
- Liouville's Theorem, 558
- Lipschitz of order  $\alpha$ , 207
- Local maximum point of function, 186
  - higher-order derivative test for, 411
  - second derivative test for, 198
- Local minimum point of function, 186; *see also* Local maximum point
- Local property, 107, 121, 164
- Local strict maximum point, 215
- Log, 338, 343
  - Taylor polynomials for, 407
  - remainder term for, 423
- Logarithm
  - classical approach to, 354
  - Napierian, 354
  - of a complex number, 561
  - to the base 10, 336
- Logarithmic derivative, 348
- Lower bound, 132
  - almost, 140
  - greatest, 132
- Lower integral, 292

- Lower limit of integration, 255  
 Lower sum, 251  
 Lowest terms, 73  
  
 Maclaurin, 566  
 Major axis of ellipse, 87  
 Mass, rate of change of, 150  
 Mathematical induction, 21  
 Maximum of two numbers, 16  
 Maximum point of a function, 185  
     local, 186; *see also* Local maximum point  
     local strict, 215  
     strict, 215  
 Maximum value of function, 185  
 Mean  
     arithmetic, 33  
     geometric, 33  
 Mean Value Theorem, 190, 191  
     Cauchy, 201  
     for integrals, 274  
     Second, 387  
 Method of exhaustion, 139  
 Minimum of function, 185  
 Minimum of two numbers, 16  
 Minimum point of a function, local,  
     186; *see also* Local minimum point  
 Minor axis of ellipse, 87  
 Minus infinity, 57  
*Mirifici logarithmonum canonis description,*  
     355  
 Modulus of a complex number, 525  
 Mollerup, Johannes, 390  
 Multiplication, 5  
     associative law for, 9  
     closure under, 9  
     commutative law for, 9  
     of arabic numerals, 8  
     of complex numbers, 522  
         geometric interpretation, 526–527  
     of function and vector-valued function, 242  
     of infinite series, 479–481  
     of number and vector, 77  
     of vectors, 77  
 Multiplicative identity, existence of, 9  
 Multiplicative inverses, existence of, 9  
 Multiplicity (of a root), 128  
  
 Napier, 355  
 Napierian logarithm, 354  
 Natural numbers, 21, 34  
 Negative, large, 64  
 “Negative infinity”, derivative, 156  
 Negative number, 9  
 Negative numbers, product of two, 7  
 Negative part of a function, 51  
 Nested Interval Theorem, 140  
 Newton, 153, 273, 327  
 Newton’s law of cooling, 352  
 Newton’s laws of motion, 159  
 Newton’s method, 457  
 Nondecreasing function, 240  
 Nondecreasing sequence, 450  
 Nondifferentiable complex functions, 542  
 Nonincreasing function, 240  
 Nonincreasing sequence, 450  
 Nonnegative function, 51  
 Nonnegative sequence, 467  
 Norm, 78, 249  
 Notational nonsense, 564  
 Nowhere differentiable continuous function, 501  
 $n^{\text{th}}$  root, 71, 527  
     existence of, 123, 527, 544  
     primitive, 531  
 Null set, 23  
 Number  
     algebraic, 435  
     complex, 517, 522  
     counting, 21  
     even, 25  
     imaginary, 517  
     irrational, 25  
     natural, 21, 34  
     odd, 25  
     prime, 31  
     rational, 25  
     real, 25, 525, 579  
     transcendental, 435  
 Numbers, basic properties of, 3

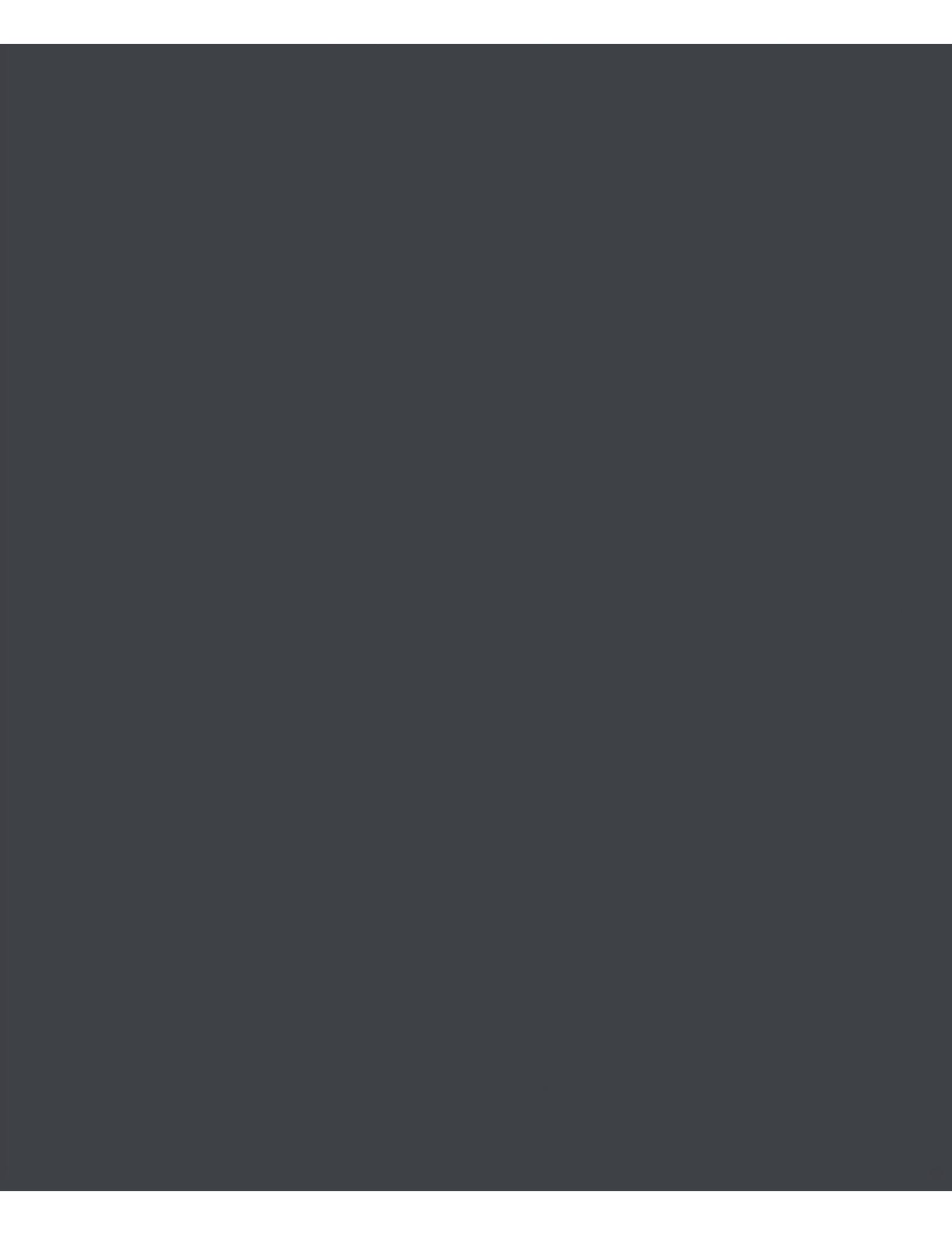
Odd function, 51, 196  
 Odd number, 25  
 One-one function, 227  
 Open interval, 56  
 "Or", 6  
 Order of equality, 412  
 Ordered field, 573  
     complete, 574  
 Ordered pair, 47 (footnote), 54  
 Origin (of a coordinate system), 57  
  
 Pair, 46  
     ordered, 47 (footnote), 54  
 Parabola, 60, 82  
     area under, 260  
     equation in polar coordinates, 88  
 Parallelogram, 76  
 Parameterized curve, tangent line of, 243  
 Parametric representation of a curve, 241  
 Partial fraction decomposition, 374  
 Partial sums, 464  
 Partition, 251  
 Parts  
     Abel's formula for summation by, 388  
     integration by, 362 ff.  
 Pascal's triangle, 27  
 "Peak", 61  
 Peak point, 451  
 Period of a function, 71, 162, 296  
 Periodic function, 71, 162, 296  
 Perpendicularity of lines, 70  
 Petard, H, 543  
 Pig, yellow, v, 371  
 Pigheaded, 183  
 Plane, 58  
     complex, 524  
 Planetary motion, Kepler's laws of, 327  
 Point, 56  
 Point of contact, 217  
 Point-slope form of equation of a line, 59, 70  
 Polar coordinates, 84 ff.  
 Polynomial function, 42  
     graph of, 61, 194  
     multiplicity of roots, 128, 183

Polynomials, Bernoulli, 566  
 Pope, Alexander, 327  
 Position, rate of change of, 150  
 Positive element of  $\mathbb{R}$ , 583  
 Positive elements of an ordered field, 573  
 Positive number, 9  
 Positive part of a function, 51  
 Power functions, 60  
 Power series, complex, 548  
 Power series centered at  $a$ , 502, 555  
 Power series representation, uniqueness of, 512  
 Powers of 2, table of, 428  
 Prime number, 31  
     characteristic of a field, 576  
     infinitely many of, 32  
     unique factorization into, 31  
 Primitive, 359  
 Primitive  $n$ th root, 531  
*Principia*, 273  
 Product, 5  
     Cauchy, 486, 505  
     infinite, 326, 391, 489  
     of function and vector-valued function, 242  
     of functions, 43  
     of number and vector, 77  
     of two negative numbers, 7  
     of vectors, 77  
 Pyramid  
     surface area of, 398  
     volume of, 402  
 Pythagorean theorem, 25, 58  
 $\pi$ , 302  
     Archimedes' approximation of, 139  
     irrationality of, 323  
     relation to  $e$ , 441, 555  
     transcendentality of, 440  
     value of, 429  
 Viète's product for  $2/\pi$ , 326  
 Wallis' product for  $\pi/2$ , 391  
  
 Quaternions, 577  
 Quotient of functions, 43

- Rabbits
  - growth of population, 32
- Radian measure, 63, 301–302
- Radioactive decay, 352
- Radius of convergence of complex power series, 550
- Rate of change of mass, 150
- Rate of change of position, 150
- Ratio test, 469
  - delicate, 486
- Rational functions, 42
  - integration of, 373 ff.
- Rational numbers, 25
- Real axis, 524
- Real line, 56
- Real number (formal definition), 579
- Real numbers, 25
  - algebraist's, 588
  - analyst's, 588
  - Archimedean property of, 136
  - construction of, 578 ff.
  - high-school student's, 589
  - inductive set of, 34
- Real part function, 532
- Real part of a complex number, 522
- Real-valued function, 532
- Rearrangement of a sequence, 476
- "Reasonable" function, 68, 116, 147, 178
- Rectangle, closed, 538
- Recursive definition, 23
- Reduction formulas, 373
- Regulated function, 515
- Remainder term for Taylor polynomials, 415
- Removable discontinuity, 119
- Reparameterization, 244
- Revolution
  - ellipsoid of, 400
  - solid of, 397
- Riemann sum, 279
- Riemann-Lebesgue Lemma, 317, 387
- Right-hand derivative, 154
- Rising Sun Lemma, 141
- Rolle, 183
- Rolle's Theorem, 190
- Root
  - multiplicity of, 128
- Root of a polynomial function, 50
- double, 183; *see also* *nth roots*
- Root test, 485
- delicate, 486
- Same sign, 12
- Scalar, 78
- Scalar product of vectors, 78
- Schwarz, H. A., 17, 215
- Schwarz inequality, 17, 32, 278
- Schwarz second derivative, 431
- Schwarzian derivative, 182
- Sec, 307
  - derivative of, 307
  - inverse of, *see* Arcsec
- Secant line, 148
- Second coordinate, 57
- Second derivative, 159
  - Schwarz, 431
- Second derivative test for maxima and minima, 198
- Second Fundamental Theorem of Calculus, 286
- Second Mean Value Theorem for Integrals, 387
- Sequence
  - absolutely summable, 473
  - Gaussian, 452
    - complex, 562
    - equivalence of, 589
  - complex numbers, 546
  - convergent, 446
    - pointwise, 494
    - uniformly, 494
  - decreasing, 450
  - divergent, 446
  - Fibonacci, 32, 512, 563
  - increasing, 450
  - infinite, 445
  - limit of, 446
  - nondecreasing, 450
  - nonincreasing, 450
  - nonnegative, 467
  - rearrangement of, 476
  - summable, 465
- Series
  - absolutely convergent, 473
  - conditionally convergent, 474

- Series (*continued*)**
- convergent, 465
  - Fourier, 315, 317, 320
  - geometric, 466
  - power, 502, 548
  - Taylor, 503
- Set, 22**
- empty, 23
- Sets**
- intersection of, 43
  - notation for, 43–44
- Shadow point, 141**
- Shell method, 398**
- Sigma, 24**
- Sign, 12**
- Simpson's rule, 396**
- Sin, 43, 300, 303, 318–319, 554**
- addition formula for, 310, 311
  - derivative of, 170, 304
  - inverse of, *see Arcsin*
  - Taylor polynomials for, 406
  - remainder term for, 420
- Sine, hyperbolic, 349**
- Sine function, 43**
- Sinh, 349**
- Sketching graphs, 193–198**
- Skew field, 577**
- Slope of a straight line, 58**
- Solid of revolution, 397**
- Speed, instantaneous, 150**
- Spiral**
- Archimedean, 85
  - hyperbolic, 313
- Square root, 12, 518**
- existence of, 122
- Square root function, 537–538**
- Square root in a field, 576**
- Squaring the circle, 440**
- Step function, 275**
- Stirling's Formula, 568**
- Straight line**
- analytic definition, 58
  - shortest distance between two points, 275
  - slope of, 58
- Strict maximum point, 215**
- Strictly convex, 226**
- Sturm Comparison Theorem, 320**
- Subsequence, 451**
- Substitution**
- integration by, 365 ff.
  - world's sneakiest, 382
- Substitution formula, 365**
- Subtraction, 5**
- Sum**
- finite, 3–4
  - infinite, 426, 464
  - lower, 251
  - of an infinite sequence, 465
  - of an infinite sequence of complex numbers, 546
  - of functions, 43
  - of vector-valued functions, 242
  - of vectors, 75
  - partial, 464
  - sigma notation for, 24
  - upper, 251
- Sum of squares, 543**
- Summable, 465, 547**
- Abel, 514
  - absolutely, 473
  - Cesaro, 486
  - uniformly, 498
- Summation by parts, Abel's formula for, 388**
- Supremum, 132**
- Surface area**
- of cone, 399
  - of pyramid, 398
  - of solid of revolution, 397
- Swift, Jonathan, 570**
- Symmetry in graphs, 196**
- Tan, 307**
- derivative of, 307
  - inverse of, *see Arctan*
  - Taylor series for, 564
- Tangent, hyperbolic, 349**
- Tangent line, 147, 149**
- of parameterized curve, 243
  - point of contact of, 217
- "Tangent line", vertical, 156**
- Tanh, 349**
- Taylor polynomial, 406 ff.**
- remainder term of, 415, 417, 418;
  - see also specific functions*

- Taylor series, 503, 554
- Taylor's Theorem, 417
- Torus, 400
- Transcendental number, 435
- Trapezoid rule, 394
- Triangle inequality, 71
- Trichotomy law, 9
- Trigonometric functions, 300, *see also*
  - cos, cot, csc, sec, sin, tan
  - integration of, 372–373
  - inverses of, 307; *see also* Arccos, arcsec, arcsin, arctan
- Trumpet
  - infinite, 402
- Two-time differentiable, 159
  
- Uniform limit, 494
- Uniformly continuous function, 142
- Uniformly convergent sequence, 494
- Uniformly convergent series, 498
- Uniformly distributed sequence, 462
- Uniformly summable, 498
- Uniqueness
  - of factorization into primes, 31
  - of limits, 98
  - of power series representations, 512
- Unit circle, 66
- Upper bound, 131, 574
  - almost, 140
  - least, 131
- Upper integral, 292
- Upper limit of integration, 255
- Upper sum, 251
  
- “Valley”, 61
- Value
  - absolute, *see* Absolute value
- Value of  $f$  at  $x$ , 40
  
- Vanishing condition, 466
- Vector-valued functions, 241
- Vector-valued functions
  - determinant of, 243
  - derivative of, 243
  - dot product of, 243
  - limit of, 243, 249
  - multiplication of function by, 242
  - sum of, 242
- Vectors, 75
  - addition of, 75
  - as forces, 76
  - dot product of, 78
  - multiplication by numbers, 77
  - multiplication of, 77
  - scalar product of, 78
- Velocity
  - average, 150
  - instantaneous, 150
- Vertical axis, 57
- Viète, François, 326
- Volume, 397–398
  
- Wallis' product, 391
- Weakly convex, 226
- Weierstrass, *see* Bolzano-Weierstrass Theorem
- Weierstrass  $M$ -test, 499
- Well-ordering principle, 23
- Wright, 383
  
- Young's inequality, 273



# Discrete Structures for Computer Science: Counting, Recursion, and Probability

Michiel Smid

*School of Computer Science  
Carleton University  
Ottawa, Ontario  
Canada*  
`michiel@scs.carleton.ca`

July 22, 2019





# Contents

<b>Preface</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Ramsey Theory . . . . .	1
1.2 Sperner's Theorem . . . . .	4
1.3 The Quick-Sort Algorithm . . . . .	5
<b>2 Mathematical Preliminaries</b>	<b>9</b>
2.1 Basic Concepts . . . . .	9
2.2 Proof Techniques . . . . .	11
2.2.1 Direct proofs . . . . .	13
2.2.2 Constructive proofs . . . . .	14
2.2.3 Nonconstructive proofs . . . . .	14
2.2.4 Proofs by contradiction . . . . .	15
2.2.5 Proofs by induction . . . . .	16
2.2.6 More examples of proofs . . . . .	18
2.3 Asymptotic Notation . . . . .	20
2.4 Logarithms . . . . .	22
2.5 Exercises . . . . .	23
<b>3 Counting</b>	<b>25</b>
3.1 The Product Rule . . . . .	25
3.1.1 Counting Bitstrings of Length $n$ . . . . .	26
3.1.2 Counting Functions . . . . .	26
3.1.3 Placing Books on Shelves . . . . .	29
3.2 The Bijection Rule . . . . .	31
3.3 The Complement Rule . . . . .	33
3.4 The Sum Rule . . . . .	34

3.5	The Principle of Inclusion and Exclusion . . . . .	35
3.6	Permutations and Binomial Coefficients . . . . .	37
3.6.1	Some Examples . . . . .	39
3.6.2	Newton's Binomial Theorem . . . . .	40
3.7	Combinatorial Proofs . . . . .	43
3.8	Pascal's Triangle . . . . .	46
3.9	More Counting Problems . . . . .	50
3.9.1	Reordering the Letters of a Word . . . . .	50
3.9.2	Counting Solutions of Linear Equations . . . . .	51
3.10	The Pigeonhole Principle . . . . .	55
3.10.1	India Pale Ale . . . . .	55
3.10.2	Sequences Containing Divisible Numbers . . . . .	56
3.10.3	Long Monotone Subsequences . . . . .	57
3.10.4	There are Infinitely Many Primes . . . . .	58
3.11	Exercises . . . . .	59
<b>4</b>	<b>Recursion</b>	<b>83</b>
4.1	Recursive Functions . . . . .	83
4.2	Fibonacci Numbers . . . . .	85
4.2.1	Counting 00-Free Bitstrings . . . . .	87
4.3	A Recursively Defined Set . . . . .	88
4.4	A Gossip Problem . . . . .	91
4.5	Euclid's Algorithm . . . . .	94
4.5.1	The Modulo Operation . . . . .	95
4.5.2	The Algorithm . . . . .	95
4.5.3	The Running Time . . . . .	97
4.6	The Merge-Sort Algorithm . . . . .	99
4.6.1	Correctness of Algorithm MERGESORT . . . . .	100
4.6.2	Running Time of Algorithm MERGESORT . . . . .	101
4.7	Computing the Closest Pair . . . . .	104
4.7.1	The Basic Approach . . . . .	105
4.7.2	The Recursive Algorithm . . . . .	111
4.8	Counting Regions when Cutting a Circle . . . . .	115
4.8.1	A Polynomial Upper Bound on $R_n$ . . . . .	115
4.8.2	A Recurrence Relation for $R_n$ . . . . .	118
4.8.3	Simplifying the Recurrence Relation . . . . .	123
4.8.4	Solving the Recurrence Relation . . . . .	124
4.9	Exercises . . . . .	125

<b>5 Discrete Probability</b>	<b>165</b>
5.1 Anonymous Broadcasting . . . . .	165
5.2 Probability Spaces . . . . .	170
5.2.1 Examples . . . . .	171
5.3 Basic Rules of Probability . . . . .	174
5.4 Uniform Probability Spaces . . . . .	179
5.4.1 The Probability of Getting a Full House . . . . .	180
5.5 The Birthday Paradox . . . . .	181
5.5.1 Throwing Balls into Boxes . . . . .	184
5.6 The Big Box Problem . . . . .	185
5.6.1 The Probability of Finding the Big Box . . . . .	187
5.7 The Monty Hall Problem . . . . .	189
5.8 Conditional Probability . . . . .	190
5.8.1 Anil's Children . . . . .	191
5.8.2 Rolling a Die . . . . .	192
5.8.3 Flip and Flip or Roll . . . . .	195
5.9 The Law of Total Probability . . . . .	198
5.9.1 Flipping a Coin and Rolling Dice . . . . .	200
5.10 Please Take a Seat . . . . .	202
5.10.1 Determining $p_{n,k}$ Using a Recurrence Relation . . . . .	203
5.10.2 Determining $p_{n,k}$ by Modifying the Algorithm . . . . .	206
5.11 Independent Events . . . . .	209
5.11.1 Rolling Two Dice . . . . .	209
5.11.2 A Basic Property of Independent Events . . . . .	211
5.11.3 Pairwise and Mutually Independent Events . . . . .	212
5.12 Describing Events by Logical Propositions . . . . .	213
5.12.1 Flipping a Coin and Rolling a Die . . . . .	214
5.12.2 Flipping Coins . . . . .	215
5.12.3 The Probability of a Circuit Failing . . . . .	215
5.13 Choosing a Random Element in a Linked List . . . . .	217
5.14 Long Runs in Random Bitstrings . . . . .	219
5.15 Infinite Probability Spaces . . . . .	224
5.15.1 Infinite Series . . . . .	225
5.15.2 Who Flips the First Heads . . . . .	227
5.15.3 Who Flips the Second Heads . . . . .	229
5.16 Exercises . . . . .	231

<b>6 Random Variables and Expectation</b>	<b>279</b>
6.1 Random Variables . . . . .	279
6.1.1 Flipping Three Coins . . . . .	280
6.1.2 Random Variables and Events . . . . .	281
6.2 Independent Random Variables . . . . .	283
6.3 Distribution Functions . . . . .	286
6.4 Expected Values . . . . .	287
6.4.1 Some Examples . . . . .	288
6.4.2 Comparing the Expected Values of Comparable Random Variables . . . . .	290
6.4.3 An Alternative Expression for the Expected Value . . . . .	291
6.5 Linearity of Expectation . . . . .	293
6.6 The Geometric Distribution . . . . .	296
6.6.1 Determining the Expected Value . . . . .	297
6.7 The Binomial Distribution . . . . .	299
6.7.1 Determining the Expected Value . . . . .	299
6.7.2 Using the Linearity of Expectation . . . . .	302
6.8 Indicator Random Variables . . . . .	303
6.8.1 Runs in Random Bitstrings . . . . .	304
6.8.2 Largest Elements in Prefixes of Random Permutations . . . . .	306
6.8.3 Estimating the Harmonic Number . . . . .	309
6.9 The Insertion-Sort Algorithm . . . . .	311
6.10 The Quick-Sort Algorithm . . . . .	313
6.11 Skip Lists . . . . .	316
6.11.1 Algorithm SEARCH . . . . .	318
6.11.2 Algorithms INSERT and DELETE . . . . .	319
6.11.3 Analysis of Skip Lists . . . . .	321
6.12 Exercises . . . . .	329
<b>7 The Probabilistic Method</b>	<b>369</b>
7.1 Large Bipartite Subgraphs . . . . .	369
7.2 Ramsey Theory . . . . .	371
7.3 Sperner's Theorem . . . . .	374
7.4 The Jaccard Distance between Finite Sets . . . . .	377
7.4.1 The First Proof . . . . .	378
7.4.2 The Second Proof . . . . .	380
7.5 Planar Graphs and the Crossing Lemma . . . . .	381
7.5.1 Planar Graphs . . . . .	382

**Contents****vii**

---

7.5.2	The Crossing Number of a Graph . . . . .	386
7.6	Exercises . . . . .	391



# Preface

This is a free textbook for an undergraduate course on Discrete Structures for Computer Science students, which I have been teaching at Carleton University since the fall term of 2013. The material is offered as the second-year course COMP 2804 (Discrete Structures II). Students are assumed to have taken COMP 1805 (Discrete Structures I), which covers mathematical reasoning, basic proof techniques, sets, functions, relations, basic graph theory, asymptotic notation, and countability.

During a 12-week term with three hours of classes per week, I cover most of the material in this book, except for Chapter 2, which has been included so that students can review the material from COMP 1805.

Chapter 2 is largely taken from the free textbook *Introduction to Theory of Computation* by Anil Maheshwari and Michiel Smid, which is available at <http://cg.scs.carleton.ca/~michiel/TheoryOfComputation/>

Please let me know if you find errors, typos, simpler proofs, comments, omissions, or if you think that some parts of the book “need improvement”.

**x**

---

# Chapter 1

## Introduction

In this chapter, we introduce some problems that will be solved later in this book. Along the way, we recall some notions from discrete mathematics that you are assumed to be familiar with. These notions are reviewed in more detail in Chapter 2.

### 1.1 Ramsey Theory

Ramsey Theory studies problems of the following form: How many elements of a given type must there be so that we can guarantee that some property holds? In this section, we consider the case when the elements are people and the property is “there is a large group of friends or there is a large group of strangers”.

**Theorem 1.1.1** *In any group of six people, there are*

- *three friends, i.e., three people who know each other,*
- *or three strangers, i.e., three people, none of which knows the other two.*

In order to prove this theorem, we denote the six people by  $P_1, P_2, \dots, P_6$ . Any two people  $P_i$  and  $P_j$  are either *friends* or *strangers*. We define the complete graph  $G = (V, E)$  with vertex set

$$V = \{P_i : 1 \leq i \leq 6\}$$

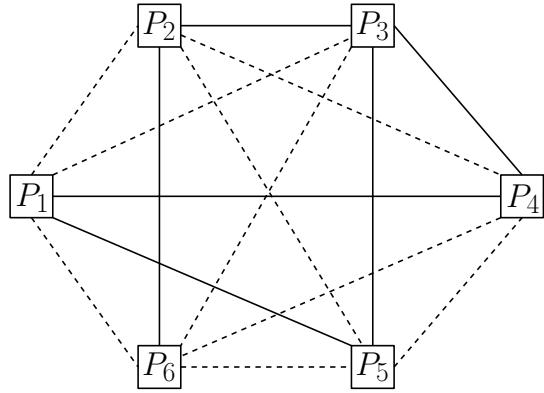
and edge set

$$E = \{P_iP_j : 1 \leq i < j \leq 6\}.$$

Observe that  $|V| = 6$  and  $|E| = 15$ . We draw each edge  $P_iP_j$  as a straight-line segment according to the following rule:

- If  $P_i$  and  $P_j$  are friends, then the edge  $P_iP_j$  is drawn as a *solid* segment.
- If  $P_i$  and  $P_j$  are strangers, then the edge  $P_iP_j$  is drawn as a *dashed* segment.

In the example below,  $P_3$  and  $P_5$  are friends, whereas  $P_1$  and  $P_3$  are strangers.



Observe that a group of three friends corresponds to a solid triangle in the graph  $G$ , whereas a group of three strangers corresponds to a dashed triangle. In the example above, there is no solid triangle and, thus, there is no group of three friends. Since the triangle  $P_2P_4P_5$  is dashed, there is a group of three strangers.

Using this terminology, Theorem 1.1.1 is equivalent to the following:

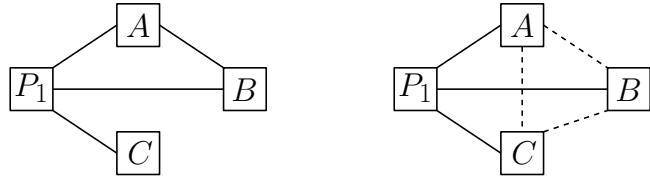
**Theorem 1.1.2** *Consider a complete graph on six vertices, in which each edge is either solid or dashed. Then there is a solid triangle or a dashed triangle.*

**Proof.** As before, we denote the vertices by  $P_1, \dots, P_6$ . Consider the five edges that are incident on  $P_1$ . Using a proof by contradiction, it can easily be shown that one of the following two claims must hold:

- At least three of these five edges are solid.
- At least three of these five edges are dashed.

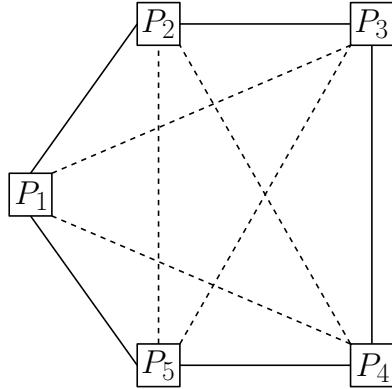
We may assume, without loss of generality, that the first claim holds. (Do you see why?) Consider three edges incident on  $P_1$  that are solid and denote them by  $P_1A$ ,  $P_1B$ , and  $P_1C$ .

If at least one of the edges  $AB$ ,  $AC$ , and  $BC$  is solid, then there is a solid triangle. In the left figure below,  $AB$  is solid and we obtain the solid triangle  $P_1AB$ .



Otherwise, all edges  $AB$ ,  $AC$ , and  $BC$  are dashed, in which case we obtain the dashed triangle  $ABC$ ; see the right figure above. ■

You should convince yourself that Theorem 1.1.2 also holds for complete graphs with more than six vertices. The example below shows an example of a complete graph with five vertices without any solid triangle and without any dashed triangle. Thus, Theorem 1.1.2 does not hold for complete graphs with five vertices. Equivalently, Theorem 1.1.1 does not hold for groups of five people.



What about larger groups of friends/strangers? Let  $k \geq 3$  be an integer. The following theorem states that even if we take  $\lfloor 2^{k/2} \rfloor$  people, we are not guaranteed that there is a group of  $k$  friends or a group of  $k$  strangers.

A  $k$ -clique in a graph is a set of  $k$  vertices, any two of which are connected by an edge. For example, a 3-clique is a triangle.

**Theorem 1.1.3** Let  $k \geq 3$  and  $n \geq 3$  be integers with  $n \leq \lfloor 2^{k/2} \rfloor$ . There exists a complete graph with  $n$  vertices, in which each edge is either solid or dashed, such that this graph does not contain a solid  $k$ -clique and does not contain a dashed  $k$ -clique.

We will prove this theorem in Section 7.2, using elementary counting techniques and probability theory. This probably sounds surprising to you, because Theorem 1.1.3 does not have anything to do with probability. In fact, in Section 7.2, we will prove the following claim: Take  $k = 20$  and  $n = 1024$ . If you go to the ByWard Market in downtown Ottawa and take a random group of  $n$  people, then with very high probability, this group does not contain a subgroup of  $k$  friends *and* does not contain a subgroup of  $k$  strangers. In other words, with very high probability, *every* subgroup of  $k$  people contains two friends *and* two strangers.

## 1.2 Sperner's Theorem

Consider a set  $S$  with five elements, say,  $S = \{1, 2, 3, 4, 5\}$ . Let  $S_1, S_2, \dots, S_m$  be a sequence of  $m$  subsets of  $S$ , such that for all  $i$  and  $j$  with  $i \neq j$ ,

$$S_i \not\subseteq S_j \text{ and } S_j \not\subseteq S_i,$$

i.e.,  $S_i$  is not a subset of  $S_j$  and  $S_j$  is not a subset of  $S_i$ . How large can  $m$  be? The following example shows that  $m$  can be as large as 10:

$$\begin{aligned} S_1 &= \{1, 2\}, & S_2 &= \{1, 3\}, & S_3 &= \{1, 4\}, & S_4 &= \{1, 5\}, & S_5 &= \{2, 3\}, \\ S_6 &= \{2, 4\}, & S_7 &= \{2, 5\}, & S_8 &= \{3, 4\}, & S_9 &= \{3, 5\}, & S_{10} &= \{4, 5\}. \end{aligned}$$

Observe that these are all subsets of  $S$  having size two. Can there be such a sequence of more than 10 subsets? The following theorem states that the answer is “no”.

**Theorem 1.2.1 (Sperner)** Let  $n \geq 1$  be an integer and let  $S$  be a set with  $n$  elements. Let  $S_1, S_2, \dots, S_m$  be a sequence of  $m$  subsets of  $S$ , such that for all  $i$  and  $j$  with  $i \neq j$ ,

$$S_i \not\subseteq S_j \text{ and } S_j \not\subseteq S_i.$$

Then

$$m \leq \binom{n}{\lfloor n/2 \rfloor}.$$

The right-hand side of the last line is a binomial coefficient, which we will define in Section 3.6. Its value is equal to the number of subsets of  $S$  having size  $\lfloor n/2 \rfloor$ . Observe that these subsets satisfy the property in Theorem 1.2.1.

We will prove Theorem 1.2.1 in Section 7.3, using elementary counting techniques and probability theory. Again, this probably sounds surprising to you, because Theorem 1.2.1 does not have anything to do with probability.

## 1.3 The Quick-Sort Algorithm

You are probably familiar with the `QUICKSORT` algorithm. This algorithm sorts any sequence  $S$  of  $n \geq 0$  pairwise distinct numbers in the following way:

- If  $n = 0$  or  $n = 1$ , then there is nothing to do.
- If  $n \geq 2$ , then the algorithm picks one of the numbers in  $S$ , let us call it  $p$  (which stands for *pivot*), scans the sequence  $S$ , and splits it into three subsequences: One subsequence  $S_1$  contains all elements in  $S$  that are less than  $p$ , one subsequence only consists of the element  $p$ , and the third subsequence  $S_2$  contains all elements in  $S$  that are larger than  $p$ ; see the figure below.

$< p$	$p$	$> p$
$S_1$		$S_2$

The algorithm then *recursively* runs `QUICKSORT` on the subsequence  $S_1$ . After this recursive call has terminated, the algorithm runs, again *recursively*, `QUICKSORT` on the subsequence  $S_2$ .

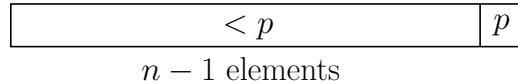
Running `QUICKSORT` recursively on the subsequence  $S_1$  means that we first check if  $S_1$  has size at most one; if this is the case, nothing needs to be done, because  $S_1$  is sorted already. If  $S_1$  has size at least two, then we choose a pivot  $p_1$  in  $S_1$ , use  $p_1$  to split  $S_1$  into three subsequences, recursively run `QUICKSORT` on the subsequence of  $S_1$  consisting of all elements that are less than  $p_1$ , and, finally, recursively run `QUICKSORT` on the subsequence of  $S_1$  consisting of all elements that are larger than  $p_1$ . (We will see recursive algorithms in more detail in Chapter 4.)

Algorithm `QUICKSORT` correctly sorts any sequence of numbers, no matter how we choose the pivot element. It turns out, however, that the running

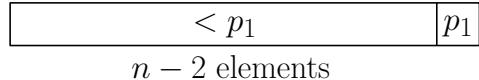
time of the algorithm heavily depends on the pivots that are chosen in the recursive calls.

For example, assume that in each (recursive) call to the algorithm, the pivot happens to be the largest element in the sequence. Then, in each call, the subsequence of elements that are larger than the pivot is empty. Let us see what happens in this case:

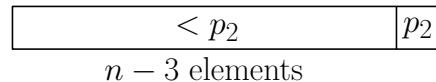
- We start with a sequence  $S$  of size  $n$ . The first pivot  $p$  is the largest element in  $S$ . Thus, using the notation given above, the subsequence  $S_1$  contains  $n - 1$  elements (all elements of  $S$  except for  $p$ ), whereas the subsequence  $S_2$  is empty. Computing these subsequences can be done in  $n$  “steps”, after which we are in the following situation:



- We now run **QUICKSORT** on a sequence of  $n - 1$  elements. Again, the pivot  $p_1$  is the largest element. In  $n - 1$  “steps”, we obtain a subsequence of  $n - 2$  elements that are less than  $p_1$ , and an empty subsequence of elements that are larger than  $p_1$ ; see the figure below.



- Next we run **QUICKSORT** on a sequence of  $n - 2$  elements. As before, the pivot  $p_2$  is the largest element. In  $n - 2$  “steps”, we obtain a subsequence of  $n - 3$  elements that are less than  $p_2$ , and an empty subsequence of elements that are larger than  $p_2$ ; see the figure below.



You probably see the pattern. The total running time of the algorithm, i.e., the total number of “steps”, is proportional to

$$n + (n - 1) + (n - 2) + (n - 3) + \cdots + 3 + 2 + 1,$$

which, by Theorem 2.2.10, is equal to

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

which, using the Big-Theta notation (see Section 2.3) is  $\Theta(n^2)$ , i.e., quadratic in  $n$ . It can be shown that this is, in fact, the worst-case running time of the QUICKSORT algorithm.

What would be a good choice for the pivot elements? Intuitively, a pivot is good if the sequences  $S_1$  and  $S_2$  have (roughly) the same size. Thus, after the first call, we are in the following situation:

$< p$	$p$	$> p$
$(n-1)/2$		$(n-1)/2$

In Section 4.6, we will prove that, if this happens in each recursive call, the running time of the QUICKSORT algorithm is only  $O(n \log n)$ . Obviously, it is not clear at all how we can guarantee that we always choose a good pivot. It turns out that there is a simple strategy: In each call, choose the pivot *randomly*! That is, among all elements involved in the recursive call, pick one uniformly at random; thus, each element has the same probability of being chosen. In Section 6.10, we will prove that this leads to an *expected* running time of  $O(n \log n)$ .



# Chapter 2

## Mathematical Preliminaries

### 2.1 Basic Concepts

Throughout this book, we will assume that you know the following mathematical concepts:

1. A *set* is a collection of well-defined objects. Examples are (i) the set of all Dutch Olympic Gold Medallists, (ii) the set of all pubs in Ottawa, and (iii) the set of all even natural numbers.
2. The set of *natural numbers* is  $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ .
3. The set of *integers* is  $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ .
4. The set of *rational numbers* is  $\mathbb{Q} = \{m/n : m \in \mathbb{Z}, n \in \mathbb{Z}, n \neq 0\}$ .
5. The set of *real numbers* is denoted by  $\mathbb{R}$ .
6. The *empty set* is the set that does not contain any element. This set is denoted by  $\emptyset$ .
7. If  $A$  is a finite set, then the *size* (or *cardinality*) of  $A$ , denoted by  $|A|$ , is the number of elements in  $A$ . Observe that  $|\emptyset| = 0$ .
8. If  $A$  and  $B$  are sets, then  $A$  is a *subset* of  $B$ , written as  $A \subseteq B$ , if every element of  $A$  is also an element of  $B$ . For example, the set of even natural numbers is a subset of the set of all natural numbers. Every set  $A$  is a subset of itself, i.e.,  $A \subseteq A$ . The empty set is a subset of

every set  $A$ , i.e.,  $\emptyset \subseteq A$ . We say that  $A$  is a *proper subset* of  $B$ , written as  $A \subset B$ , if  $A \subseteq B$  and  $A \neq B$ .

9. If  $B$  is a set, then the *power set*  $\mathcal{P}(B)$  of  $B$  is defined to be the set of all subsets of  $B$ :

$$\mathcal{P}(B) = \{A : A \subseteq B\}.$$

Observe that  $\emptyset \in \mathcal{P}(B)$  and  $B \in \mathcal{P}(B)$ .

10. If  $A$  and  $B$  are two sets, then

- (a) their *union* is defined as

$$A \cup B = \{x : x \in A \text{ or } x \in B\},$$

- (b) their *intersection* is defined as

$$A \cap B = \{x : x \in A \text{ and } x \in B\},$$

- (c) their *difference* is defined as

$$A \setminus B = \{x : x \in A \text{ and } x \notin B\},$$

- (d) the *Cartesian product* of  $A$  and  $B$  is defined as

$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\},$$

- (e) the *complement* of  $A$  is defined as

$$\overline{A} = \{x : x \notin A\}.$$

11. A *binary relation* on two sets  $A$  and  $B$  is a subset of  $A \times B$ .

12. A *function*  $f$  from  $A$  to  $B$ , denoted by  $f : A \rightarrow B$ , is a binary relation  $R$ , having the property that for each element  $a$  in  $A$ , there is exactly one ordered pair in  $R$ , whose first component is  $a$ . If this unique pair is  $(a, b)$ , then we will say that  $f(a) = b$ , or  $f$  maps  $a$  to  $b$ , or the image of  $a$  under  $f$  is  $b$ . The set  $A$  is called the *domain* of  $f$ , and the set

$$\{b \in B : \text{there is an } a \in A \text{ with } f(a) = b\}$$

is called the *range* of  $f$ .

13. A function  $f : A \rightarrow B$  is *one-to-one* (or *injective*), if for any two distinct elements  $a$  and  $a'$  in  $A$ , we have  $f(a) \neq f(a')$ . The function  $f$  is *onto* (or *surjective*), if for each element  $b$  in  $B$ , there exists an element  $a$  in  $A$ , such that  $f(a) = b$ ; in other words, the range of  $f$  is equal to the set  $B$ . A function  $f$  is a *bijection*, if  $f$  is both injective and surjective.
14. A set  $A$  is *countable*, if  $A$  is finite or there is a bijection  $f : \mathbb{N} \rightarrow A$ . The sets  $\mathbb{N}$ ,  $\mathbb{Z}$ , and  $\mathbb{Q}$  are countable, whereas  $\mathbb{R}$  is not.
15. A *graph*  $G = (V, E)$  is a pair consisting of a set  $V$ , whose elements are called *vertices*, and a set  $E$ , where each element of  $E$  is a pair of distinct vertices. The elements of  $E$  are called *edges*.
16. The *Boolean values* are 1 and 0, that represent *true* and *false*, respectively. The basic Boolean operations include
  - (a) negation (or *NOT*), represented by  $\neg$ ,
  - (b) conjunction (or *AND*), represented by  $\wedge$ ,
  - (c) disjunction (or *OR*), represented by  $\vee$ ,
  - (d) exclusive-or (or *XOR*), represented by  $\oplus$ ,
  - (e) equivalence, represented by  $\leftrightarrow$  or  $\Leftrightarrow$ ,
  - (f) implication, represented by  $\rightarrow$  or  $\Rightarrow$ .

The following table explains the meanings of these operations.

<i>NOT</i>	<i>AND</i>	<i>OR</i>	<i>XOR</i>	equivalence	implication
$\neg 0 = 1$	$0 \wedge 0 = 0$	$0 \vee 0 = 0$	$0 \oplus 0 = 0$	$0 \leftrightarrow 0 = 1$	$0 \rightarrow 0 = 1$
$\neg 1 = 0$	$0 \wedge 1 = 0$	$0 \vee 1 = 1$	$0 \oplus 1 = 1$	$0 \leftrightarrow 1 = 0$	$0 \rightarrow 1 = 1$
	$1 \wedge 0 = 0$	$1 \vee 0 = 1$	$1 \oplus 0 = 1$	$1 \leftrightarrow 0 = 0$	$1 \rightarrow 0 = 0$
	$1 \wedge 1 = 1$	$1 \vee 1 = 1$	$1 \oplus 1 = 0$	$1 \leftrightarrow 1 = 1$	$1 \rightarrow 1 = 1$

## 2.2 Proof Techniques

A proof is a proof. What kind of a proof? It's a proof. A proof is a proof. And when you have a good proof, it's because it's proven.

— Jean Chrétien, Prime Minister of Canada (1993–2003)

In mathematics, a theorem is a statement that is true. A proof is a sequence of mathematical statements that form an argument to show that a theorem is true. The statements in the proof of a theorem include axioms (assumptions about the underlying mathematical structures), hypotheses of the theorem to be proved, and previously proved theorems. The main question is “How do we go about proving theorems?” This question is similar to the question of how to solve a given problem. Of course, the answer is that finding proofs, or solving problems, is not easy; otherwise life would be dull! There is no specified way of coming up with a proof, but there are some generic strategies that could be of help. In this section, we review some of these strategies. The best way to get a feeling of how to come up with a proof is by solving a large number of problems. Here are some useful tips. (You may take a look at the book *How to Solve It*, by George Pólya).

1. Read and completely understand the statement of the theorem to be proved. Most often this is the hardest part.
2. Sometimes, theorems contain theorems inside them. For example, “Property  $A$  if and only if property  $B$ ”, requires showing two statements:
  - (a) If property  $A$  is true, then property  $B$  is true ( $A \Rightarrow B$ ).
  - (b) If property  $B$  is true, then property  $A$  is true ( $B \Rightarrow A$ ).

Another example is the theorem “Set  $A$  equals set  $B$ .” To prove this, we need to prove that  $A \subseteq B$  and  $B \subseteq A$ . That is, we need to show that each element of set  $A$  is in set  $B$ , and each element of set  $B$  is in set  $A$ .

3. Try to work out a few simple cases of the theorem just to get a grip on it (i.e., crack a few simple cases first).
4. Try to write down the proof once you think you have it. This is to ensure the correctness of your proof. Often, mistakes are found at the time of writing.
5. Finding proofs takes time, we do not come prewired to produce proofs. Be patient, think, express and write clearly, and try to be precise as much as possible.

In the next sections, we will go through some of the proof strategies.

### 2.2.1 Direct proofs

As the name suggests, in a direct proof of a theorem, we just approach the theorem directly.

**Theorem 2.2.1** *If  $n$  is an odd positive integer, then  $n^2$  is odd as well.*

**Proof.** An odd positive integer  $n$  can be written as  $n = 2k + 1$ , for some integer  $k \geq 0$ . Then

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

Since  $2(2k^2 + 2k)$  is even, and “even plus one is odd”, we can conclude that  $n^2$  is odd.  $\blacksquare$

For a graph  $G = (V, E)$ , the *degree* of a vertex  $v$ , denoted by  $\deg(v)$ , is defined to be the number of edges that are incident on  $v$ .

**Theorem 2.2.2** *Let  $G = (V, E)$  be a graph. Then the sum of the degrees of all vertices is an even integer, i.e.,*

$$\sum_{v \in V} \deg(v)$$

*is even.*

**Proof.** If you do not see the meaning of this statement, then first try it out for a few graphs. The reason why the statement holds is very simple: Each edge contributes 2 to the summation (because an edge is incident on exactly two distinct vertices).  $\blacksquare$

Actually, the proof above proves the following theorem.

**Theorem 2.2.3** *Let  $G = (V, E)$  be a graph. Then the sum of the degrees of all vertices is equal to twice the number of edges, i.e.,*

$$\sum_{v \in V} \deg(v) = 2|E|.$$

### 2.2.2 Constructive proofs

This technique not only shows the existence of a certain object, it actually gives a method of creating it:

**Theorem 2.2.4** *There exists an object with property  $\mathcal{P}$ .*

**Proof.** Here is the object: [...]

And here is the proof that the object satisfies property  $\mathcal{P}$ : [...] ■

A graph is called *3-regular*, if each vertex has degree three. We prove the following theorem using a constructive proof.

**Theorem 2.2.5** *For every even integer  $n \geq 4$ , there exists a 3-regular graph with  $n$  vertices.*

**Proof.** Let

$$V = \{0, 1, 2, \dots, n - 1\},$$

and

$$E = \{\{i, i+1\} : 0 \leq i \leq n-2\} \cup \{\{n-1, 0\}\} \cup \{\{i, i+n/2\} : 0 \leq i \leq n/2-1\}.$$

Then the graph  $G = (V, E)$  is 3-regular.

Convince yourself that this graph is indeed 3-regular. It may help to draw the graph for, say,  $n = 8$ . ■

### 2.2.3 Nonconstructive proofs

In a nonconstructive proof, we show that a certain object exists, without actually creating it. Here is an example of such a proof:

**Theorem 2.2.6** *There exist irrational numbers  $x$  and  $y$  such that  $x^y$  is rational.*

**Proof.** There are two possible cases.

**Case 1:**  $\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}$ .

In this case, we take  $x = y = \sqrt{2}$ . In Theorem 2.2.9 below, we will prove that  $\sqrt{2}$  is irrational.

**Case 2:**  $\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$ .

In this case, we take  $x = \sqrt{2}^{\sqrt{2}}$  and  $y = \sqrt{2}$ . Since

$$x^y = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^2 = 2,$$

the claim in the theorem follows. ■

Observe that this proof indeed proves the theorem, but it does not give an example of a pair of irrational numbers  $x$  and  $y$  such that  $x^y$  is rational.

#### 2.2.4 Proofs by contradiction

This is how a proof by contradiction looks like:

**Theorem 2.2.7** *Statement  $\mathcal{S}$  is true.*

**Proof.** Assume that statement  $\mathcal{S}$  is false. Then, derive a contradiction (such as  $1 + 1 = 3$ ).

In other words, we show that the statement “ $\neg\mathcal{S} \Rightarrow \text{false}$ ” is true. This is sufficient, because the contrapositive of the statement “ $\neg\mathcal{S} \Rightarrow \text{false}$ ” is the statement “ $\text{true} \Rightarrow \mathcal{S}$ ”. The latter logical formula is equivalent to  $\mathcal{S}$ , and that is what we wanted to show. ■

Below, we give two examples of proofs by contradiction.

**Theorem 2.2.8** *Let  $n$  be a positive integer. If  $n^2$  is even, then  $n$  is even.*

**Proof.** We will prove the theorem by contradiction. Thus, we assume that  $n^2$  is even, but  $n$  is odd. Since  $n$  is odd, we know from Theorem 2.2.1 that  $n^2$  is odd. This is a contradiction, because we assumed that  $n^2$  is even. ■

**Theorem 2.2.9**  *$\sqrt{2}$  is irrational, i.e.,  $\sqrt{2}$  cannot be written as a fraction of two integers.*

**Proof.** We will prove the theorem by contradiction. Thus, we assume that  $\sqrt{2}$  is rational. Then  $\sqrt{2}$  can be written as a fraction of two integers  $m \geq 1$  and  $n \geq 1$ , i.e.,  $\sqrt{2} = m/n$ . We may assume that  $m$  and  $n$  do not share

any common factors, i.e., the greatest common divisor of  $m$  and  $n$  is equal to one; if this is not the case, then we can get rid of the common factors. By squaring  $\sqrt{2} = m/n$ , we get  $2n^2 = m^2$ . This implies that  $m^2$  is even. Then, by Theorem 2.2.8,  $m$  is even, which means that we can write  $m$  as  $m = 2k$ , for some positive integer  $k$ . It follows that  $2n^2 = m^2 = 4k^2$ , which implies that  $n^2 = 2k^2$ . Hence,  $n^2$  is even. Again by Theorem 2.2.8, it follows that  $n$  is even.

We have shown that  $m$  and  $n$  are both even. But we know that  $m$  and  $n$  are *not* both even. Hence, we have a contradiction. Our assumption that  $\sqrt{2}$  is rational is wrong. Thus, we can conclude that  $\sqrt{2}$  is irrational. ■

There is a nice discussion of this proof in the book *My Brain is Open: The Mathematical Journeys of Paul Erdős* by Bruce Schechter.

### 2.2.5 Proofs by induction

This is a very powerful and important technique for proving theorems.

For each positive integer  $n$ , let  $P(n)$  be a mathematical statement that depends on  $n$ . Assume we wish to prove that  $P(n)$  is true for all positive integers  $n$ . A proof by induction of such a statement is carried out as follows:

**Base Case:** Prove that  $P(1)$  is true.

**Induction Step:** Prove that for all  $n \geq 1$ , the following holds: If  $P(n)$  is true, then  $P(n + 1)$  is also true.

In the induction step, we choose an arbitrary integer  $n \geq 1$  and assume that  $P(n)$  is true; this is called the *induction hypothesis*. We then prove that  $P(n + 1)$  is also true.

**Theorem 2.2.10** *For all positive integers  $n$ , we have*

$$1 + 2 + 3 + \cdots + n = \frac{n(n + 1)}{2}.$$

**Proof.** We start with the base case of the induction. If  $n = 1$ , then both the left-hand side and the right-hand side are equal to 1. Therefore, the theorem is true for  $n = 1$ .

For the induction step, let  $n \geq 1$  and assume that the theorem is true for  $n$ , i.e., assume that

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

We have to prove that the theorem is true for  $n + 1$ , i.e., we have to prove that

$$1 + 2 + 3 + \cdots + (n + 1) = \frac{(n + 1)(n + 2)}{2}.$$

Here is the proof:

$$\begin{aligned} 1 + 2 + 3 + \cdots + (n + 1) &= \underbrace{1 + 2 + 3 + \cdots + n}_{=\frac{n(n+1)}{2}} + (n + 1) \\ &= \frac{n(n+1)}{2} + (n + 1) \\ &= \frac{(n + 1)(n + 2)}{2}. \end{aligned}$$

■

By the way, here is an alternative proof of the theorem above: Let  $S = 1 + 2 + 3 + \cdots + n$ . Then,

$$\begin{array}{ccccccccccccccccc} S & = & 1 & + & 2 & + & 3 & + & \cdots & + & (n-2) & + & (n-1) & + & n \\ \hline \frac{S}{2} & = & (n+1) & + & (n+1) & + & (n+1) & + & \cdots & + & (n+1) & + & (n+1) & + & (n+1) \end{array}$$

Since there are  $n$  terms on the right-hand side, we have  $2S = n(n + 1)$ . This implies that  $S = n(n + 1)/2$ .

**Theorem 2.2.11** *For every positive integer  $n$ ,  $a - b$  is a factor of  $a^n - b^n$ .*

**Proof.** A direct proof can be given by providing a factorization of  $a^n - b^n$ :

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \cdots + ab^{n-2} + b^{n-1}).$$

We now prove the theorem by induction. For the base case, let  $n = 1$ . The claim in the theorem is “ $a - b$  is a factor of  $a - b$ ”, which is obviously true.

Let  $n \geq 1$  and assume that  $a - b$  is a factor of  $a^n - b^n$ . We have to prove that  $a - b$  is a factor of  $a^{n+1} - b^{n+1}$ . We have

$$a^{n+1} - b^{n+1} = a^{n+1} - a^n b + a^n b - b^{n+1} = a^n(a - b) + (a^n - b^n)b.$$

The first term on the right-hand side is divisible by  $a - b$ . By the induction hypothesis, the second term on the right-hand side is divisible by  $a - b$  as well. Therefore, the entire right-hand side is divisible by  $a - b$ . Since the right-hand side is equal to  $a^{n+1} - b^{n+1}$ , it follows that  $a - b$  is a factor of  $a^{n+1} - b^{n+1}$ .  $\blacksquare$

We now give an alternative proof of Theorem 2.2.3:

**Theorem 2.2.12** *Let  $G = (V, E)$  be a graph with  $m$  edges. Then the sum of the degrees of all vertices is equal to twice the number of edges, i.e.,*

$$\sum_{v \in V} \deg(v) = 2m.$$

**Proof.** The proof is by induction on the number  $m$  of edges. For the base case of the induction, assume that  $m = 0$ . Then the graph  $G$  does not contain any edges and, therefore,  $\sum_{v \in V} \deg(v) = 0$ . Thus, the theorem is true if  $m = 0$ .

Let  $m \geq 0$  and assume that the theorem is true for every graph with  $m$  edges. Let  $G$  be an arbitrary graph with  $m + 1$  edges. We have to prove that  $\sum_{v \in V} \deg(v) = 2(m + 1)$ .

Let  $\{a, b\}$  be an arbitrary edge in  $G$ , and let  $G'$  be the graph obtained from  $G$  by removing the edge  $\{a, b\}$ . Since  $G'$  has  $m$  edges, we know from the induction hypothesis that the sum of the degrees of all vertices in  $G'$  is equal to  $2m$ . Using this, we obtain

$$\sum_{v \in G} \deg(v) = \sum_{v \in G'} \deg(v) + 2 = 2m + 2 = 2(m + 1).$$

$\blacksquare$

## 2.2.6 More examples of proofs

Recall Theorem 2.2.5, which states that for every *even* integer  $n \geq 4$ , there exists a 3-regular graph with  $n$  vertices. The following theorem explains why we stated this theorem for even values of  $n$ .

**Theorem 2.2.13** *Let  $n \geq 5$  be an odd integer. There is no 3-regular graph with  $n$  vertices.*

**Proof.** The proof is by contradiction. Thus, we assume that there exists a graph  $G = (V, E)$  with  $n$  vertices that is 3-regular. Let  $m$  be the number of edges in  $G$ . Since  $\deg(v) = 3$  for every vertex  $v$ , we have

$$\sum_{v \in V} \deg(v) = 3n.$$

On the other hand, by Theorem 2.2.3, we have

$$\sum_{v \in V} \deg(v) = 2m.$$

It follows that

$$3n = 2m.$$

Since  $n$  is an odd integer, the left-hand side in this equation is an odd integer as well. The right-hand side, however, is an even integer. This is a contradiction.  $\blacksquare$

Let  $K_n$  be the *complete graph* on  $n$  vertices. This graph has a vertex set of size  $n$ , and every pair of distinct vertices is joined by an edge.

If  $G = (V, E)$  is a graph with  $n$  vertices, then the *complement*  $\overline{G}$  of  $G$  is the graph with vertex set  $V$  that consists of those edges of  $K_n$  that are not present in  $G$ .

**Theorem 2.2.14** *Let  $n \geq 2$  and let  $G$  be a graph on  $n$  vertices. Then  $G$  is connected or  $\overline{G}$  is connected.*

**Proof.** We prove the theorem by induction on the number  $n$  of vertices. For the base case, assume that  $n = 2$ . There are two possibilities for the graph  $G$ :

1.  $G$  contains one edge. In this case,  $G$  is connected.
2.  $G$  does not contain an edge. In this case, the complement  $\overline{G}$  contains one edge and, therefore,  $\overline{G}$  is connected.

Thus, for  $n = 2$ , the theorem is true.

Let  $n \geq 2$  and assume that the theorem is true for every graph with  $n$  vertices. Let  $G$  be graph with  $n + 1$  vertices. We have to prove that  $G$  is connected or  $\overline{G}$  is connected. We consider three cases.

**Case 1:** There is a vertex  $v$  whose degree in  $G$  is equal to  $n$ .

Since  $G$  has  $n+1$  vertices,  $v$  is connected by an edge to every other vertex of  $G$ . Therefore,  $G$  is connected.

**Case 2:** There is a vertex  $v$  whose degree in  $G$  is equal to 0.

In this case, the degree of  $v$  in the graph  $\overline{G}$  is equal to  $n$ . Since  $\overline{G}$  has  $n+1$  vertices,  $v$  is connected by an edge to every other vertex of  $\overline{G}$ . Therefore,  $\overline{G}$  is connected.

**Case 3:** For every vertex  $v$ , the degree of  $v$  in  $G$  is in  $\{1, 2, \dots, n-1\}$ .

Let  $v$  be an arbitrary vertex of  $G$ . Let  $G'$  be the graph obtained by deleting from  $G$  the vertex  $v$ , together with all edges that are incident on  $v$ . Since  $G'$  has  $n$  vertices, we know from the induction hypothesis that  $G'$  is connected or  $\overline{G'}$  is connected.

Let us first assume that  $G'$  is connected. Then the graph  $G$  is connected as well, because there is at least one edge in  $G$  between  $v$  and some vertex of  $G'$ .

If  $G'$  is not connected, then  $\overline{G'}$  must be connected. Since we are in Case 3, we know that the degree of  $v$  in  $G$  is in the set  $\{1, 2, \dots, n-1\}$ . It follows that the degree of  $v$  in the graph  $\overline{G}$  is in this set as well. Hence, there is at least one edge in  $\overline{G}$  between  $v$  and some vertex in  $\overline{G'}$ . This implies that  $\overline{G}$  is connected. ■

The previous theorem can be rephrased as follows:

**Theorem 2.2.15** *Let  $n \geq 2$  and consider the complete graph  $K_n$  on  $n$  vertices. Color each edge of this graph as either red or blue. Let  $R$  be the graph consisting of all the red edges, and let  $B$  be the graph consisting of all the blue edges. Then  $R$  is connected or  $B$  is connected.*

If you like surprising proofs of various mathematical results, you should read the book *Proofs from THE BOOK* by Martin Aigner and Günter Ziegler.

## 2.3 Asymptotic Notation

Let  $f : \mathbb{N} \rightarrow \mathbb{R}$  and  $g : \mathbb{N} \rightarrow \mathbb{R}$  be functions such that  $f(n) > 0$  and  $g(n) > 0$  for all  $n \in \mathbb{N}$ .

- We say that  $f(n) = O(g(n))$  if the following is true: There exist constants  $c > 0$  and  $k > 0$  such that for all  $n \geq k$ ,

$$f(n) \leq c \cdot g(n).$$

- We say that  $f(n) = \Omega(g(n))$  if the following is true: There exist constants  $c > 0$  and  $k > 0$  such that for all  $n \geq k$ ,

$$f(n) \geq c \cdot g(n).$$

- We say that  $f(n) = \Theta(g(n))$  if  $f(n) = O(g(n))$  and  $f(n) = \Omega(g(n))$ . Thus, there exist constants  $c > 0$ ,  $c' > 0$ , and  $k > 0$  such that for all  $n \geq k$ ,

$$c \cdot g(n) \leq f(n) \leq c' \cdot g(n).$$

For example, for all  $n \geq 1$ , we have

$$\begin{aligned} 13 + 7n - 5n^2 + 8n^3 &\leq 13 + 7n + 8n^3 \\ &\leq 13n^3 + 7n^3 + 8n^3 \\ &= 28n^3. \end{aligned}$$

Thus, by taking  $c = 28$  and  $k = 1$ , it follows that

$$13 + 7n - 5n^2 + 8n^3 = O(n^3). \quad (2.1)$$

We also have

$$13 + 7n - 5n^2 + 8n^3 \geq -5n^2 + 8n^3.$$

Since  $n^3 \geq 5n^2$  for all  $n \geq 5$ , it follows that, again for all  $n \geq 5$ ,

$$\begin{aligned} 13 + 7n - 5n^2 + 8n^3 &\geq -5n^2 + 8n^3 \\ &\geq -n^3 + 8n^3 \\ &= 7n^3. \end{aligned}$$

Hence, by taking  $c = 7$  and  $k = 5$ , we have shown that

$$13 + 7n - 5n^2 + 8n^3 = \Omega(n^3). \quad (2.2)$$

It follows from (2.1) and (2.2) that

$$13 + 7n - 5n^2 + 8n^3 = \Theta(n^3).$$

## 2.4 Logarithms

If  $b$  and  $x$  are real numbers with  $b > 1$  and  $x > 0$ , then  $\log_b x$  denotes the logarithm of  $x$  with base  $b$ . Note that

$$\log_b x = y \text{ if and only if } b^y = x.$$

If  $b = 2$ , then we write  $\log x$  instead of  $\log_2 x$ . We write  $\ln x$  to refer to the natural logarithm of  $x$  with base  $e$ .

**Lemma 2.4.1** *If  $b > 1$  and  $x > 0$ , then*

$$b^{\log_b x} = x.$$

**Proof.** We have seen above that  $y = \log_b x$  if and only if  $b^y = x$ . Thus, if we write  $y = \log_b x$ , then  $b^{\log_b x} = b^y = x$ . ■

For example, if  $x > 0$ , then

$$2^{\log x} = x.$$

**Lemma 2.4.2** *If  $b > 1$ ,  $x > 0$ , and  $a$  is a real number, then*

$$\log_b(x^a) = a \log_b x.$$

**Proof.** Let  $y = \log_b x$ . Then

$$a \log_b x = ay.$$

Since  $y = \log_b x$ , we have  $b^y = x$  and, thus,

$$x^a = (b^y)^a = b^{ay}.$$

Taking logarithms (with base  $b$ ) on both sides gives

$$\log_b(x^a) = \log_b(b^{ay}) = ay = a \log_b x.$$
■

For example, for  $x > 1$ , we get

$$2 \log \log x = \log(\log^2 x)$$

and

$$2^{2 \log \log x} = 2^{\log(\log^2 x)} = \log^2 x.$$

**Lemma 2.4.3** *If  $b > 1$ ,  $c > 1$ , and  $x > 0$ , then*

$$\log_b x = \frac{\log_c x}{\log_c b}.$$

**Proof.** Let  $y = \log_b x$ . Then  $b^y = x$ , and we get

$$\log_c x = \log_c (b^y) = y \log_c b = \log_b x \log_c b.$$

■

For example, if  $x > 0$ , then

$$\log x = \frac{\ln x}{\ln 2}.$$

## 2.5 Exercises

### Proofs that use a big hammer:

**Theorem:** For any integer  $n \geq 3$ ,  $\sqrt[n]{2}$  is irrational.

**Proof:** Assume  $\sqrt[n]{2}$  is rational. Then there exist positive integers  $a$  and  $b$  such that  $\sqrt[n]{2} = a/b$ . Thus, we have  $2 = (a/b)^n$ , which is equivalent to  $2 \cdot b^n = a^n$ , which is equivalent to

$$b^n + b^n = a^n.$$

This contradicts Fermat's Last Theorem. ■

**2.1** Prove that  $\sqrt{p}$  is irrational for every prime number  $p$ .

**2.2** Let  $n$  be a positive integer that is not a perfect square. Prove that  $\sqrt{n}$  is irrational.

**2.3** Use induction to prove that every integer  $n \geq 2$  can be written as a product of prime numbers.

**2.4** Prove by induction that  $n^4 - 4n^2$  is divisible by 3, for all integers  $n \geq 1$ .

**2.5** Prove that

$$\sum_{i=1}^n \frac{1}{i^2} < 2 - 1/n,$$

for all integers  $n \geq 2$ .

**2.6** Prove that 9 divides  $n^3 + (n+1)^3 + (n+2)^3$ , for all integers  $n \geq 0$ .

**2.7** The Fermat numbers  $F_0, F_1, F_2, \dots$  are defined by  $F_n = 2^{2^n} + 1$  for  $n \geq 0$ .

- Prove by induction that

$$F_0 F_1 F_2 \cdots F_{n-1} = F_n - 2$$

for all integers  $n \geq 1$ .

- Prove that for any two distinct integers  $n \geq 0$  and  $m \geq 0$ , the greatest common divisor of  $F_n$  and  $F_m$  is equal to 1.
- Conclude that there are infinitely many prime numbers.

**2.8** Prove by induction that  $n! > 2^{1+n/2}$  for all integers  $n \geq 3$ .

# Chapter 3

## Counting

There are three types of people, those who can count and those who cannot count.

Given a set of 23 elements, how many subsets of size 17 are there? How many solutions are there to the equation

$$x_1 + x_2 + \cdots + x_{12} = 873,$$

where  $x_1 \geq 0, x_2 \geq 0, \dots, x_{12} \geq 0$  are integers? In this chapter, we will introduce some general techniques that can be used to answer questions of these types.

### 3.1 The Product Rule

How many strings of two characters are there, if the first character must be an uppercase letter and the second character must be a digit? Examples of such strings are  $A0$ ,  $K7$ , and  $Z9$ . The answer is obviously  $26 \cdot 10 = 260$ , because there are 26 choices for the first character and, no matter which letter we choose for being the first character, there are 10 choices for the second character. We can look at this in the following way: Consider the “procedure” of writing a string of two characters, the first one being an uppercase letter, and the second one being a digit. Then our original question becomes “how many ways are there to perform this procedure?” Observe that the procedure consists of two “tasks”, the first one being writing the first character, and the

second one being writing the second character. Obviously, there are 26 ways to do the first task. Next, observe that, regardless of how we do the first task, there are 10 ways to do the second task. The Product Rule states that the total number of ways to perform the entire procedure is  $26 \cdot 10 = 260$ .

**Product Rule:** Assume a procedure consists of performing a sequence of  $m$  tasks in order. Furthermore, assume that for each  $i = 1, 2, \dots, m$ , there are  $N_i$  ways to do the  $i$ -th task, regardless of how the first  $i - 1$  tasks were done. Then, there are  $N_1 N_2 \cdots N_m$  ways to do the entire procedure.

In the example above, we have  $m = 2$ ,  $N_1 = 26$ , and  $N_2 = 10$ .

### 3.1.1 Counting Bitstrings of Length $n$

Let  $n \geq 1$  be an integer. A *bitstring* of length  $n$  is a sequence of 0's and 1's. How many bitstrings of length  $n$  are there? To apply the Product Rule, we have to specify the “procedure” and the “tasks”:

- The procedure is “write a bitstring of length  $n$ ”.
- For  $i = 1, 2, \dots, n$ , the  $i$ -th task is “write one bit”.

There are two ways to do the  $i$ -th task, regardless of how we did the first  $i - 1$  tasks. Therefore, we can apply the Product Rule with  $N_i = 2$  for  $i = 1, 2, \dots, n$ , and conclude that there are  $N_1 N_2 \cdots N_n = 2^n$  ways to do the entire procedure. As a result, the number of bitstrings of length  $n$  is equal to  $2^n$ .

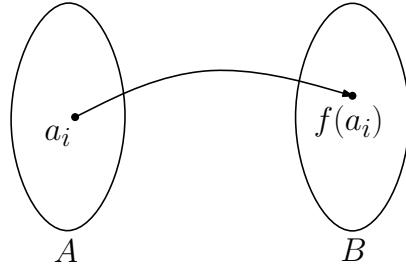
**Theorem 3.1.1** *For any integer  $n \geq 1$ , the number of bitstrings of length  $n$  is equal to  $2^n$ .*

### 3.1.2 Counting Functions

Let  $m \geq 1$  and  $n \geq 1$  be integers, let  $A$  be a set of size  $m$ , and let  $B$  be a set of size  $n$ . How many functions  $f : A \rightarrow B$  are there?

Write the set  $A$  as  $A = \{a_1, a_2, \dots, a_m\}$ . Any function  $f : A \rightarrow B$  is completely specified by the values  $f(a_1), f(a_2), \dots, f(a_m)$ , where each such value can be any element of  $B$ . Again, we are going to apply the Product Rule. Thus, we have to specify the “procedure” and the “tasks”:

- The procedure is “specify the values  $f(a_1), f(a_2), \dots, f(a_m)$ ”.
- For  $i = 1, 2, \dots, m$ , the  $i$ -th task is “specify the value  $f(a_i)$ ”.



For each  $i$ ,  $f(a_i)$  can be any of the  $n$  elements of  $B$ . As a result, there are  $N_i = n$  ways to do the  $i$ -th task, regardless of how we did the first  $i - 1$  tasks. By the Product Rule, there are  $N_1 N_2 \cdots N_m = n^m$  ways to do the entire procedure and, hence, this many functions  $f : A \rightarrow B$ . We have proved the following result:

**Theorem 3.1.2** *Let  $m \geq 1$  and  $n \geq 1$  be integers, let  $A$  be a set of size  $m$ , and let  $B$  be a set of size  $n$ . The number of functions  $f : A \rightarrow B$  is equal to  $n^m$ .*

Recall that a function  $f : A \rightarrow B$  is *one-to-one* if for any  $i$  and  $j$  with  $i \neq j$ , we have  $f(a_i) \neq f(a_j)$ . How many one-to-one functions  $f : A \rightarrow B$  are there?

If  $m > n$ , then there is no such function. (Do you see why?) Assume that  $m \leq n$ . To determine the number of one-to-one functions, we use the same procedure and tasks as before.

- Since  $f(a_1)$  can be any of the  $n$  elements of  $B$ , there are  $N_1 = n$  ways to do the first task.
- In the second task, we have to specify the value  $f(a_2)$ . Since the function  $f$  is one-to-one and since we have already specified  $f(a_1)$ , we can choose  $f(a_2)$  to be any of the  $n - 1$  elements in the set  $B \setminus \{f(a_1)\}$ . As a result, there are  $N_2 = n - 1$  ways to do the second task. Note that this is true, regardless of how we did the first task.

- In general, in the  $i$ -th task, we have to specify the value  $f(a_i)$ . Since we have already specified  $f(a_1), f(a_2), \dots, f(a_{i-1})$ , we can choose  $f(a_i)$  to be any of the  $n - i + 1$  elements in the set

$$B \setminus \{f(a_1), f(a_2), \dots, f(a_{i-1})\}.$$

As a result, there are  $N_i = n - i + 1$  ways to do the  $i$ -th task. Note that this is true, regardless of how we did the first  $i - 1$  tasks.

By the Product Rule, there are

$$N_1 N_2 \cdots N_m = n(n-1)(n-2) \cdots (n-m+1)$$

ways to do the entire procedure, which is also the number of one-to-one functions  $f : A \rightarrow B$ .

Recall the *factorial function*

$$k! = \begin{cases} 1 & \text{if } k = 0, \\ 1 \cdot 2 \cdot 3 \cdots k & \text{if } k \geq 1. \end{cases}$$

We can simplify the product

$$n(n-1)(n-2) \cdots (n-m+1)$$

by observing that it is “almost” a factorial:

$$\begin{aligned} & n(n-1)(n-2) \cdots (n-m+1) \\ &= n(n-1)(n-2) \cdots (n-m+1) \cdot \frac{(n-m)(n-m-1) \cdots 1}{(n-m)(n-m-1) \cdots 1} \\ &= \frac{n(n-1)(n-2) \cdots 1}{(n-m)(n-m-1) \cdots 1} \\ &= \frac{n!}{(n-m)!}. \end{aligned}$$

We have proved the following result:

**Theorem 3.1.3** *Let  $m \geq 1$  and  $n \geq 1$  be integers, let  $A$  be a set of size  $m$ , and let  $B$  be a set of size  $n$ .*

1. *If  $m > n$ , then there is no one-to-one function  $f : A \rightarrow B$ .*
2. *If  $m \leq n$ , then the number of one-to-one functions  $f : A \rightarrow B$  is equal to*

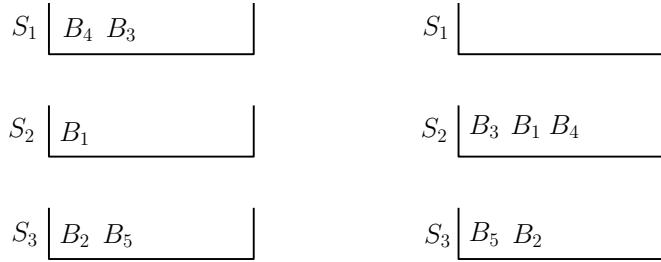
$$\frac{n!}{(n-m)!}.$$

### 3.1.3 Placing Books on Shelves

Let  $m \geq 1$  and  $n \geq 1$  be integers, and consider  $m$  books  $B_1, B_2, \dots, B_m$  and  $n$  bookshelves  $S_1, S_2, \dots, S_n$ . How many ways are there to place the books on the shelves? Placing the books on the shelves means that

- we specify for each book the shelf at which this book is placed, and
- we specify for each shelf the left-to-right order of the books that are placed on that shelf.

Some bookshelves may be empty. We assume that each shelf is large enough to fit all books. In the figure below, you see two different placements.



We are again going to use the Product Rule to determine the number of placements.

- The procedure is “place the  $m$  books on the  $n$  shelves”.
- For  $i = 1, 2, \dots, m$ , the  $i$ -th task is “place book  $B_i$  on the shelves”. When placing book  $B_i$ , we can place it on the far left or far right of any shelf or between any two of the books  $B_1, \dots, B_{i-1}$  that have already been placed.

Let us see how many ways there are to do each task.

- Just before we place book  $B_1$ , all shelves are empty. Therefore, there are  $N_1 = n$  ways to do the first task.
- In the second task, we have to place book  $B_2$ . Since  $B_1$  has already been placed, we have the following possibilities for placing  $B_2$ :
  - We place  $B_2$  on the far left of any of the  $n$  shelves.

- We place  $B_2$  immediately to the right of  $B_1$ .

As a result, there are  $N_2 = n + 1$  ways to do the second task. Note that this is true, regardless of how we did the first task.

- In general, in the  $i$ -th task, we have to place book  $B_i$ . Since the books  $B_1, B_2, \dots, B_{i-1}$  have already been placed, we have the following possibilities for placing  $B_i$ :

- We place  $B_i$  on the far left of any of the  $n$  shelves.
- We place  $B_i$  immediately to the right of one of  $B_1, B_2, \dots, B_{i-1}$ .

As a result, there are  $N_i = n + i - 1$  ways to do the  $i$ -th task. Note that this is true, regardless of how we did the first  $i - 1$  tasks.

Thus, by the Product Rule, there are

$$N_1 N_2 \cdots N_m = n(n+1)(n+2) \cdots (n+m-1)$$

ways to do the entire procedure, which is also the number of placements of the  $m$  books on the  $n$  shelves. As before, we use factorials to simplify this product:

$$\begin{aligned} & n(n+1)(n+2) \cdots (n+m-1) \\ &= \frac{1 \cdot 2 \cdot 3 \cdots (n-1)}{1 \cdot 2 \cdot 3 \cdots (n-1)} \cdot n(n+1)(n+2) \cdots (n+m-1) \\ &= \frac{(n+m-1)!}{(n-1)!}. \end{aligned}$$

We have proved the following result:

**Theorem 3.1.4** *Let  $m \geq 1$  and  $n \geq 1$  be integers. The number of ways to place  $m$  books on  $n$  bookshelves is equal to*

$$\frac{(n+m-1)!}{(n-1)!}.$$

## 3.2 The Bijection Rule

Let  $n \geq 0$  be an integer and let  $S$  be a set with  $n$  elements. How many subsets does  $S$  have? If  $n = 0$ , then  $S = \emptyset$  and there is only one subset of  $S$ , namely  $S$  itself. Assume from now on that  $n \geq 1$ . As we will see below, asking for the number of subsets of  $S$  is exactly the same as asking for the number of bitstrings of length  $n$ .

Let  $A$  and  $B$  be finite sets. Recall that a function  $f : A \rightarrow B$  is a *bijection* if

- $f$  is one-to-one, i.e., if  $a \neq a'$  then  $f(a) \neq f(a')$ , and
- $f$  is onto, i.e., for each  $b$  in  $B$ , there is an  $a$  in  $A$  such that  $f(a) = b$ .

This means that

- each element of  $A$  corresponds to a unique element of  $B$  and
- each element of  $B$  corresponds to a unique element of  $A$ .

It should be clear that this means that  $A$  and  $B$  contain the same number of elements.

**Bijection Rule:** Let  $A$  and  $B$  be finite sets. If there exists a bijection  $f : A \rightarrow B$ , then  $|A| = |B|$ , i.e.,  $A$  and  $B$  have the same size.

Let us see how we can apply this rule to the subset problem. We define the following two sets  $A$  and  $B$ :

- $A = \mathcal{P}(S)$ , i.e., the power set of  $S$ , which is the set of all subsets of  $S$ :

$$\mathcal{P}(S) = \{T : T \subseteq S\}.$$

- $B$  is the set of all bitstrings of length  $n$ .

We have seen in Theorem 3.1.1 that the set  $B$  has size  $2^n$ . Therefore, if we can show that there exists a bijection  $f : A \rightarrow B$ , then, according to the Bijection Rule, we have  $|A| = |B|$  and, thus, the number of subsets of  $S$  is equal to  $2^n$ .

Write the set  $S$  as  $S = \{s_1, s_2, \dots, s_n\}$ . We define the function  $f : A \rightarrow B$  in the following way:

- For any  $T \in A$  (i.e.,  $T \subseteq S$ ),  $f(T)$  is the bitstring  $b_1 b_2 \dots b_n$ , where

$$b_i = \begin{cases} 1 & \text{if } s_i \in T, \\ 0 & \text{if } s_i \notin T. \end{cases}$$

For example, assume that  $n = 5$ .

- If  $T = \{s_1, s_3, s_4\}$ , then  $f(T) = 10110$ .
- If  $T = S = \{s_1, s_2, s_3, s_4, s_5\}$ , then  $f(T) = 11111$ .
- If  $T = \emptyset$ , then  $f(T) = 00000$ .

It is not difficult to see that each subset of  $S$  corresponds to a unique bitstring of length  $n$ , and each bitstring of length  $n$  corresponds to a unique subset of  $S$ . Therefore, this function  $f$  is a bijection between  $A$  and  $B$ .

Thus, we have shown that there exists a bijection  $f : A \rightarrow B$ . This, together with Theorem 3.1.1 and the Bijection Rule, implies the following result:

**Theorem 3.2.1** *For any integer  $n \geq 0$ , the number of subsets of a set of size  $n$  is equal to  $2^n$ .*

You will probably have noticed that we could have proved this result directly using the Product Rule: The procedure “specify a subset of  $S = \{s_1, s_2, \dots, s_n\}$ ” can be carried out by specifying, for  $i = 1, 2, \dots, n$ , whether or not  $s_i$  is contained in the subset. For each  $i$ , there are two choices. As a result, there are  $2^n$  ways to do the procedure.

To conclude this section, we remark that we have already been using the Bijection Rule in Section 3.1!

**The Product Rule and the Bijection Rule:** In order to apply the Product Rule to a counting problem, we need the following:

1. Phrase the counting problem in terms of doing a procedure, consisting of a number of tasks.
2. There must be a one-to-one correspondence between the different ways to do the procedure and the objects we are counting. In other words:
  - (a) Each way to do the procedure must correspond to a unique object we are counting.
  - (b) Conversely, each object we are counting must correspond to a unique way to do the procedure.
3. Once we have this one-to-one correspondence, the Bijection Rule implies that the number of objects is equal to the number of ways to do the procedure.

### 3.3 The Complement Rule

Consider strings consisting of 8 characters, each character being a lowercase letter or a digit. Such a string is called a *valid password*, if it contains at least one digit. How many valid passwords are there? One way to answer this question is to first count the valid passwords with exactly one digit, then count the valid passwords with exactly two digits, then count the valid passwords with exactly three digits, etc. As we will see below, it is easier to first count the strings that do *not* contain any digit.

Recall that the *difference*  $U \setminus A$  of the two sets  $U$  and  $A$  is defined as

$$U \setminus A = \{x : x \in U \text{ and } x \notin A\}.$$

**Complement Rule:** Let  $U$  be a finite set and let  $A$  be a subset of  $U$ .

Then

$$|A| = |U| - |U \setminus A|.$$

This rule follows easily from the fact that  $|U| = |A| + |U \setminus A|$ , which holds because each element in  $U$  is either in  $A$  or in  $U \setminus A$ .

To apply the Complement Rule to the password problem, let  $U$  be the set of all strings consisting of 8 characters, each character being a lowercase letter or a digit, and let  $A$  be the set of all valid passwords, i.e., all strings in  $U$  that contain at least one digit. Note that  $U \setminus A$  is the set of all strings of 8 characters, each character being a lowercase letter or a digit, that do not contain any digit. In other words,  $U \setminus A$  is the set of all strings of 8 characters, where each character is a lowercase letter.

By the Product Rule, the set  $U$  has size  $36^8$ , because each string in  $U$  has 8 characters, and there are  $26 + 10 = 36$  choices for each character. Similarly, the set  $U \setminus A$  has size  $26^8$ , because there are 26 choices for each of the 8 characters. Then, by the Complement Rule, the number of valid passwords is equal to

$$|A| = |U| - |U \setminus A| = 36^8 - 26^8 = 2,612,282,842,880.$$

### 3.4 The Sum Rule

If  $A$  and  $B$  are two finite sets that are *disjoint*, i.e.,  $A \cap B = \emptyset$ , then it is obvious that the size of  $A \cup B$  is equal to the sum of the sizes of  $A$  and  $B$ .

**Sum Rule:** Let  $A_1, A_2, \dots, A_m$  be a sequence of finite and pairwise disjoint sets. Then

$$|A_1 \cup A_2 \cup \dots \cup A_m| = |A_1| + |A_2| + \dots + |A_m|.$$

Note that we already used this rule in Section 3.3 when we argued why the Complement Rule is correct!

To give an example, consider strings consisting of 6, 7, or 8 characters, each character being a lowercase letter or a digit. Such a string is called a *valid password*, if it contains at least one digit. Let  $A$  be the set of all valid passwords. What is the size of  $A$ ?

For  $i = 6, 7, 8$ , let  $A_i$  be the set of all valid passwords of length  $i$ . It is obvious that  $A = A_6 \cup A_7 \cup A_8$ . Since the three sets  $A_6$ ,  $A_7$ , and  $A_8$  are pairwise disjoint, we have, by the Sum Rule,

$$|A| = |A_6| + |A_7| + |A_8|.$$

We have seen in Section 3.3 that  $|A_8| = 36^8 - 26^8$ . By the same arguments, we have  $|A_6| = 36^6 - 26^6$  and  $|A_7| = 36^7 - 26^7$ . Thus, the number of valid passwords is equal to

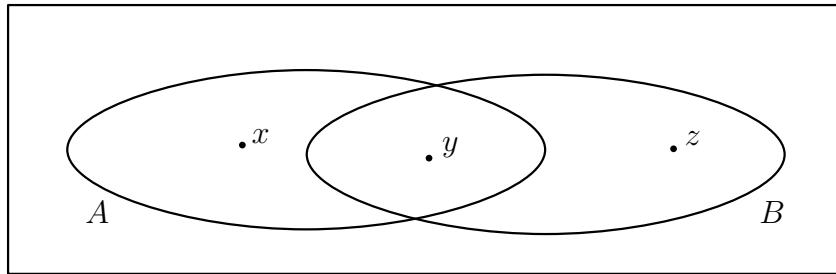
$$|A| = (36^6 - 26^6) + (36^7 - 26^7) + (36^8 - 26^8) = 2,684,483,063,360.$$

## 3.5 The Principle of Inclusion and Exclusion

The Sum Rule holds only for sets that are pairwise disjoint. Consider two finite sets  $A$  and  $B$  that are not necessarily disjoint. How can we determine the size of the union  $A \cup B$ ? We can start with the sum  $|A| + |B|$ , i.e., we *include* both  $A$  and  $B$ . In the Venn diagram below,

- $x$  is in  $A$  but not in  $B$ ; it is counted exactly once in  $|A| + |B|$ ,
- $z$  is in  $B$  but not in  $A$ ; it is counted exactly once in  $|A| + |B|$ ,
- $y$  is in  $A$  and in  $B$ ; it is counted exactly twice in  $|A| + |B|$ .

Based on this, if we subtract the size of the intersection  $A \cap B$ , i.e., we *exclude*  $A \cap B$ , then we have counted every element of  $A \cup B$  exactly once.



**Inclusion-Exclusion:** Let  $A$  and  $B$  be finite sets. Then

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

To give an example, let us count the bitstrings of length 17 that start with 010 or end with 11. Let  $S$  be the set of all such bitstrings. Define  $A$  to be the set of all bitstrings of length 17 that start with 010, and define  $B$  to be the set of all bitstrings of length 17 that end with 11. Then  $S = A \cup B$  and, thus, we have to determine the size of  $A \cup B$ .

- The size of  $A$  is equal to the number of bitstrings of length 14, because the first three bits of every string in  $A$  are fixed. Therefore, by the Product Rule, we have  $|A| = 2^{14}$ .
- The size of  $B$  is equal to the number of bitstrings of length 15, because the last two bits of every string in  $B$  are fixed. Therefore, by the Product Rule, we have  $|B| = 2^{15}$ .
- Each string in  $A \cap B$  starts with 010 *and* ends with 11. Thus, five bits are fixed for every string in  $A \cap B$ . It follows that the size of  $A \cap B$  is equal to the number of bitstrings of length 12. Therefore, by the Product Rule, we have  $|A \cap B| = 2^{12}$ .

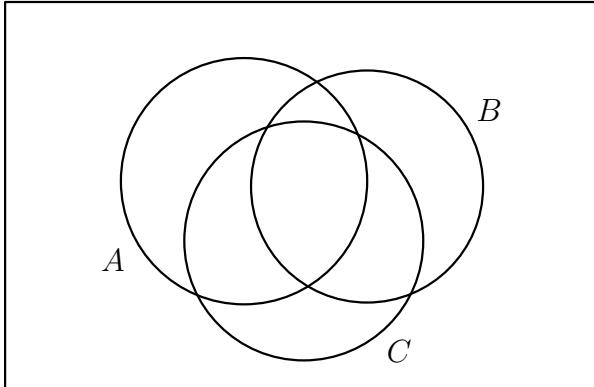
By applying the Inclusion-Exclusion formula, it follows that

$$|S| = |A \cup B| = |A| + |B| - |A \cap B| = 2^{14} + 2^{15} - 2^{12} = 45,056.$$

The Inclusion-Exclusion formula can be generalized to more than two sets. You are encouraged to verify, using the Venn diagram below, that the following formula is the correct one for three sets.

**Inclusion-Exclusion:** Let  $A$ ,  $B$ , and  $C$  be finite sets. Then

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$



To give an example, how many bitstrings of length 17 are there that start with 010, or end with 11, or have 10 at positions<sup>1</sup> 7 and 8? Let  $S$  be the set

---

<sup>1</sup>The positions are numbered 1, 2, ..., 17.

of all such bitstrings. Define  $A$  to be the set of all bitstrings of length 17 that start with 010, define  $B$  to be the set of all bitstrings of length 17 that end with 11, and define  $C$  to be the set of all bitstrings of length 17 that have 10 at positions 7 and 8. Then  $S = A \cup B \cup C$  and, thus, we have to determine the size of  $A \cup B \cup C$ .

- We have seen before that  $|A| = 2^{14}$ ,  $|B| = 2^{15}$ , and  $|A \cap B| = 2^{12}$ .
- We have  $|C| = 2^{15}$ , because the bits at positions 7 and 8 are fixed for every string in  $C$ .
- We have  $|A \cap C| = 2^{12}$ , because 5 bits are fixed for every string in  $A \cap C$ .
- We have  $|B \cap C| = 2^{13}$ , because 4 bits are fixed for every string in  $B \cap C$ .
- We have  $|A \cap B \cap C| = 2^{10}$ , because 7 bits are fixed for every string in  $A \cap B \cap C$ .

By applying the Inclusion-Exclusion formula, it follows that

$$\begin{aligned} |S| &= |A \cup B \cup C| \\ &= |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C| \\ &= 2^{14} + 2^{15} + 2^{15} - 2^{12} - 2^{12} - 2^{13} + 2^{10} \\ &= 66,560. \end{aligned}$$

## 3.6 Permutations and Binomial Coefficients

A *permutation* of a finite set  $S$  is an *ordered* sequence of the elements of  $S$ , in which each element occurs exactly once. For example, the set  $S = \{a, b, c\}$  has six permutations:

$$abc, acb, bac, bca, cab, cba$$

**Theorem 3.6.1** *Let  $n \geq 0$  be an integer and let  $S$  be a set with  $n$  elements. There are exactly  $n!$  many permutations of  $S$ .*

**Proof.** If  $n = 0$ , then  $S = \emptyset$  and the only permutation of  $S$  is the empty sequence. Since  $0! = 1$ , the claim holds for  $n = 0$ . Assume that  $n \geq 1$  and

denote the elements of  $S$  by  $s_1, s_2, \dots, s_n$ . Consider the procedure “write a permutation of  $S$ ” and, for  $i = 1, 2, \dots, n$ , the task “write the  $i$ -th element in the permutation”. When we do the  $i$ -th task, we have already written  $i - 1$  elements of the permutation; we cannot take any of these elements for the  $i$ -th task. Therefore, there are  $n - (i - 1) = n - i + 1$  ways to do the  $i$ -th task. By the Product Rule, there are

$$n \cdot (n - 1) \cdot (n - 2) \cdots 2 \cdot 1 = n!$$

ways to do the procedure. This number is equal to the number of permutations of  $S$ . ■

Note that we could also have used Theorem 3.1.3 to prove Theorem 3.6.1: A permutation of  $S$  can be regarded to be a one-to-one function  $f : S \rightarrow S$ . Therefore, by applying Theorem 3.1.3 with  $A = S$ ,  $B = S$  and, thus,  $m = n$ , we obtain Theorem 3.6.1.

Consider the set  $S = \{a, b, c, d, e\}$ . How many 3-element subsets does  $S$  have? Recall that in a set, the order of the elements does not matter. Here is a list of all 10 subsets of  $S$  having size 3:

$$\begin{aligned} &\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \\ &\{a, d, e\}, \{b, c, d\}, \{b, c, e\}, \{b, d, e\}, \{c, d, e\} \end{aligned}$$

**Definition 3.6.2** Let  $n \geq 0$  and  $k \geq 0$  be integers. The *binomial coefficient*  $\binom{n}{k}$  denotes the number of  $k$ -element subsets of an  $n$ -element set.

The symbol  $\binom{n}{k}$  is pronounced as “ $n$  choose  $k$ ”.

The example above shows that  $\binom{5}{3} = 10$ . Since the empty set has exactly one subset of size zero (the empty set itself), we have  $\binom{0}{0} = 1$ . Note that  $\binom{n}{k} = 0$  if  $k > n$ . Below, we derive a formula for the value of  $\binom{n}{k}$  if  $0 \leq k \leq n$ .

Let  $S$  be a set with  $n$  elements and let  $A$  be the set of all *ordered* sequences consisting of exactly  $k$  pairwise distinct elements of  $S$ . We are going to count the elements of  $A$  in two different ways.

The first way is by using the Product Rule. This gives

$$|A| = n(n - 1)(n - 2) \cdots (n - k + 1) = \frac{n!}{(n - k)!}. \quad (3.1)$$

Observe that (3.1) also follows from Theorem 3.1.3. (Do you see why?)

In the second way, we do the following:

- Write all  $\binom{n}{k}$  subsets of  $S$  having size  $k$ .
- For each of these subsets, write a list of all  $k!$  permutations of this subset.

If we put all these lists together, then we obtain a big list in which each ordered sequence of  $k$  pairwise distinct elements of  $S$  appears exactly once. In other words, the big list contains each element of  $A$  exactly once. Since the big list has size  $\binom{n}{k}k!$ , it follows that

$$|A| = \binom{n}{k}k!. \quad (3.2)$$

Since the right-hand sides of (3.1) and (3.2) are equal (because they are both equal to  $|A|$ ), we obtain the following result:

**Theorem 3.6.3** *Let  $n$  and  $k$  be integers with  $0 \leq k \leq n$ . Then*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

For example,

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2} = 10$$

and

$$\binom{0}{0} = \frac{0!}{0!0!} = \frac{1}{1 \cdot 1} = 1;$$

recall that we defined  $0!$  to be equal to 1.

### 3.6.1 Some Examples

**First Example:** Consider a standard deck of 52 cards. How many hands of 5 cards are there? Any such hand is a 5-element subset of the set of 52 cards and, therefore, the number of hands of 5 cards is equal to

$$\binom{52}{5} = \frac{52!}{5!47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2,598,960.$$

**Second Example:** Let  $n$  and  $k$  be integers with  $0 \leq k \leq n$ . How many bitstrings of length  $n$  have exactly  $k$  many 1s? We can answer this question using the Product Rule:

- The procedure is “write a bitstring of length  $n$  having exactly  $k$  many 1s”.
- Task 1: Consider the set  $\{1, 2, \dots, n\}$  of positions for the bits of the string. Choose a  $k$ -element subset of this set.
- Task 2: Write a 1 in each of the  $k$  positions of the chosen subset.
- Task 3: Write a 0 in each of the  $n - k$  remaining positions.

There are  $\binom{n}{k}$  ways to do the first task, there is one way to do the second task, and there is one way to do the third task. Thus, by the Product Rule, the number of ways to do the procedure and, therefore, the number of bitstrings of length  $n$  having exactly  $k$  many 1s, is equal to

$$\binom{n}{k} \cdot 1 \cdot 1 = \binom{n}{k}.$$

We can also use the Bijection Rule, by observing, in the same way as we did in Section 3.2, that there is a bijection between

- the set of all bitstrings of length  $n$  having exactly  $k$  many 1s, and
- the set of all  $k$ -element subsets of an  $n$ -element set.

Since the latter set has size  $\binom{n}{k}$ , the former set has size  $\binom{n}{k}$  as well.

**Theorem 3.6.4** *Let  $n$  and  $k$  be integers with  $0 \leq k \leq n$ . The number of bitstrings of length  $n$  having exactly  $k$  many 1s is equal to  $\binom{n}{k}$ .*

### 3.6.2 Newton’s Binomial Theorem

You have learned in high school that

$$(x + y)^2 = x^2 + 2xy + y^2.$$

You have probably also seen that

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

What is the expansion of  $(x + y)^5$ ? Observe that

$$(x + y)^5 = (x + y)(x + y)(x + y)(x + y)(x + y).$$

If we expand the expression on the right-hand side, we get terms

$$x^5, x^4y, x^3y^2, x^2y^3, xy^4, y^5,$$

each with some coefficient. What is the coefficient of  $x^2y^3$ ? We obtain a term  $x^2y^3$ , by

- choosing 3 of the 5 terms  $x + y$ ,
- taking  $y$  in each of the 3 chosen terms  $x + y$ , and
- taking  $x$  in each of the other 2 terms  $x + y$ .

Since there are  $\binom{5}{3}$  ways to do this, the coefficient of  $x^2y^3$  is equal to  $\binom{5}{3} = 10$ .

**Theorem 3.6.5 (Newton's Binomial Theorem)** *For any integer  $n \geq 0$ , we have*

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

**Proof.** The expression  $(x+y)^n$  is the product of  $n$  terms  $x+y$ . By expanding this product, we get a term  $x^{n-k}y^k$  for each  $k = 0, 1, \dots, n$ , each with some coefficient. We get a term  $x^{n-k}y^k$  by

- choosing  $k$  of the  $n$  terms  $x + y$ ,
- taking  $y$  in each of the  $k$  chosen terms  $x + y$ , and
- taking  $x$  in each of the other  $n - k$  terms  $x + y$ .

Since there are  $\binom{n}{k}$  ways to do this, the coefficient of  $x^{n-k}y^k$  is equal to  $\binom{n}{k}$ .

■

For example, we have

$$\begin{aligned} (x + y)^3 &= \binom{3}{0}x^3 + \binom{3}{1}x^2y + \binom{3}{2}xy^2 + \binom{3}{3}y^3 \\ &= x^3 + 3x^2y + 3xy^2 + y^3. \end{aligned}$$

To determine the coefficient of  $x^{12}y^{13}$  in  $(x+y)^{25}$ , we take  $n = 25$  and  $k = 13$  in Newton's Binomial Theorem, and get  $\binom{25}{13}$ .

What is the coefficient of  $x^{12}y^{13}$  in  $(2x-5y)^{25}$ ? Observe that

$$(2x-5y)^{25} = ((2x) + (-5y))^{25}.$$

By replacing  $x$  by  $2x$ , and  $y$  by  $-5y$  in Newton's Binomial Theorem, we get

$$(2x-5y)^{25} = \sum_{k=0}^{25} \binom{25}{k} (2x)^{25-k} (-5y)^k.$$

By taking  $k = 13$ , we obtain the coefficient of  $x^{12}y^{13}$ :

$$\binom{25}{13} \cdot 2^{25-13} \cdot (-5)^{13} = -\binom{25}{13} \cdot 2^{12} \cdot 5^{13}.$$

Newton's Binomial Theorem leads to identities for summations involving binomial coefficients:

**Theorem 3.6.6** *For any integer  $n \geq 0$ , we have*

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

**Proof.** Take  $x = y = 1$  in Newton's Binomial Theorem. ■

In Section 3.7, we will see a proof of Theorem 3.6.6 that does not use Newton's Binomial Theorem.

**Theorem 3.6.7** *For any integer  $n \geq 1$ , we have*

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

**Proof.** Take  $x = 1$  and  $y = -1$  in Newton's Binomial Theorem. ■

## 3.7 Combinatorial Proofs

In a combinatorial proof, we show the validity of an identity, such as the one in Theorem 3.6.6, by interpreting it as the answer to a counting problem. The identity is proved by solving this counting problem in two different ways. This gives two answers to the same counting problem. Obviously, these two answers must be equal. Observe that we have already used this approach in Section 3.6: When we determined the formula for  $\binom{n}{k}$ , we counted, in two different ways, the number of ordered sequences of  $k$  pairwise distinct elements from an  $n$ -element set. In this section, we will give several other examples of combinatorial proofs.

**Theorem 3.7.1** *For any integers  $n$  and  $k$  with  $0 \leq k \leq n$ , we have*

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

**Proof.** The claim can be proved using Theorem 3.6.3. To obtain a combinatorial proof, let  $S$  be a set with  $n$  elements. Recall that

- $\binom{n}{k}$  is the number of ways to choose  $k$  elements from the set  $S$ ,

which is the same as

- the number of ways to *not* choose  $n - k$  elements from the set  $S$ .

The latter number is equal to  $\binom{n}{n-k}$ .

We can also prove the claim using Theorem 3.6.4:

- The number of bitstrings of length  $n$  with exactly  $k$  many 1s is equal to  $\binom{n}{k}$ .
- The number of bitstrings of length  $n$  with exactly  $n - k$  many 0s is equal to  $\binom{n}{n-k}$ .

Since these two quantities are equal, the theorem follows. ■

**Theorem 3.7.2 (Pascal's Identity)** *For any integers  $n$  and  $k$  with  $1 \leq k \leq n$ , we have*

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

**Proof.** As in the previous theorem, the claim can be proved using Theorem 3.6.3. To obtain a combinatorial proof, let  $S$  be a set with  $n+1$  elements. We are going to count the  $k$ -element subsets of  $S$  in two different ways.

First, by definition, the number of  $k$ -element subsets of  $S$  is equal to

$$\binom{n+1}{k}. \quad (3.3)$$

For the second way, we choose an element  $x$  in  $S$  and consider the set  $T = S \setminus \{x\}$ , i.e., the set obtained by removing  $x$  from  $S$ . Any  $k$ -element subset of  $S$  is of exactly one of the following two types:

- The  $k$ -element subset of  $S$  does not contain  $x$ .
  - Any such subset is a  $k$ -element subset of  $T$ . Since  $T$  has size  $n$ , there are  $\binom{n}{k}$  many  $k$ -element subsets of  $S$  that do not contain  $x$ .
- The  $k$ -element subset of  $S$  contains  $x$ .
  - If  $A$  is any such subset, then  $B = A \setminus \{x\}$  is a  $(k-1)$ -element subset of  $T$ .
  - Conversely, for any  $(k-1)$ -element subset  $B$  of  $T$ , the set  $A = B \cup \{x\}$  is a  $k$ -element subset of  $S$  that contains  $x$ .
  - It follows that the number of  $k$ -element subsets of  $S$  containing  $x$  is equal to the number of  $(k-1)$ -element subsets of  $T$ . The latter number is equal to  $\binom{n}{k-1}$ .

Thus, the second way of counting shows that the number of  $k$ -element subsets of  $S$  is equal to

$$\binom{n}{k} + \binom{n}{k-1}. \quad (3.4)$$

Since the expressions in (3.3) and (3.4) count the same objects, they must be equal. Therefore, the proof is complete. ■

**Theorem 3.7.3** *For any integer  $n \geq 0$ , we have*

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

**Proof.** We have seen in Theorem 3.6.6 that this identity follows from Newton's Binomial Theorem. Below, we give a combinatorial proof.

Consider a set  $S$  with  $n$  elements. According to Theorem 3.2.1, this set has  $2^n$  many subsets. A different way to count the subsets of  $S$  is by dividing them into (pairwise disjoint) groups according to their sizes. For each  $k$  with  $0 \leq k \leq n$ , consider all  $k$ -element subsets of  $S$ . The number of such subsets is equal to  $\binom{n}{k}$ . If we take the sum of all these binomial coefficients, then we have counted each subset of  $S$  exactly once. Thus,

$$\sum_{k=0}^n \binom{n}{k}$$

is equal to the total number of subsets of  $S$ . ■

**Theorem 3.7.4 (Vandermonde's Identity)** *For any integers  $m \geq 0$ ,  $n \geq 0$ , and  $r \geq 0$  with  $r \leq m$  and  $r \leq n$ , we have*

$$\sum_{k=0}^r \binom{m}{k} \binom{n}{r-k} = \binom{m+n}{r}.$$

**Proof.** Consider a set  $S$  with  $m + n$  elements. We are going to count the  $r$ -element subsets of  $S$  in two different ways.

First, by using the definition of binomial coefficients, the number of  $r$ -element subsets of  $S$  is equal to  $\binom{m+n}{r}$ .

For the second way, we partition the set  $S$  into two subsets  $A$  and  $B$ , where  $A$  has size  $m$  and  $B$  has size  $n$ . Observe that any  $r$ -element subset of  $S$  contains

- some elements of  $A$ , say  $k$  many, and
- $r - k$  elements of  $B$ .

The value of  $k$  can be any integer in the set  $\{0, 1, 2, \dots, r\}$ .

Let  $k$  be any integer with  $0 \leq k \leq r$ , and let  $N_k$  be the number of  $r$ -element subsets of  $S$  that contain exactly  $k$  elements of  $A$  (and, thus,  $r - k$  elements of  $B$ ). Then,  $\sum_{k=0}^r N_k$  is equal to the total number of  $r$ -element subsets of  $S$  and, thus,

$$\sum_{k=0}^r N_k = \binom{m+n}{r}.$$

To determine  $N_k$ , we use the Product Rule: We obtain any subset that is counted in  $N_k$ , by

- choosing  $k$  elements in  $A$  (there are  $\binom{m}{k}$  ways to do this) and
- choosing  $r - k$  elements in  $B$  (there are  $\binom{n}{r-k}$  ways to do this).

It follows that

$$N_k = \binom{m}{k} \binom{n}{r-k}.$$

■

**Corollary 3.7.5** *For any integer  $n \geq 0$ , we have*

$$\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}.$$

**Proof.** By taking  $m = n = r$  in Vandermonde's Identity, we get

$$\sum_{k=0}^n \binom{n}{k} \binom{n}{n-k} = \binom{2n}{n}.$$

Using Theorem 3.7.1, we get

$$\binom{n}{k} \binom{n}{n-k} = \binom{n}{k} \binom{n}{k} = \binom{n}{k}^2.$$

■

## 3.8 Pascal's Triangle

The computational method at the heart of Pascal's work was actually discovered by a Chinese mathematician named Jia Xian around 1050, published by another Chinese mathematician, Zhu Shijie, in 1303, discussed in a work by Cardano in 1570, and plugged into the greater whole of probability theory by Pascal, who ended up getting most of the credit.

— Leonard Mlodinow, *The Drunkard's Walk*, 2008

We have seen that

- $\binom{n}{0} = 1$  for all integers  $n \geq 0$ ,
- $\binom{n}{n} = 1$  for all integers  $n \geq 0$ ,
- $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$  for all integers  $n \geq 2$  and  $k$  with  $1 \leq k \leq n-1$ ;  
see Theorem 3.7.2.

These relations lead to an algorithm for generating binomial coefficients:

**Algorithm GENERATEBINOMCOEFF:**

```

BCoeff(0, 0) = 1;
for n = 1, 2, 3, ...
do BCoeff(n, 0) = 1;
   for k = 1 to n - 1
      do BCoeff(n, k) = BCoeff(n - 1, k - 1) + BCoeff(n - 1, k)
      endfor;
      BCoeff(n, n) = 1
   endfor
endfor

```

The values  $BCoeff(n, k)$  that are computed by this (non-terminating) algorithm satisfy

$$BCoeff(n, k) = \binom{n}{k} \text{ for } 0 \leq k \leq n.$$

The triangle obtained by arranging these binomial coefficients, with the  $n$ -th row containing all values  $\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n}$ , is called *Pascal's Triangle*. The figure below shows rows 0, 1, ..., 6:

$$\begin{array}{ccccccccc}
& & \binom{0}{0} & & & & & & \\
& & \binom{1}{0} & \quad & \binom{1}{1} & & & & \\
& & \binom{2}{0} & \quad & \binom{2}{1} & \quad & \binom{2}{2} & & \\
& & \binom{3}{0} & \quad & \binom{3}{1} & \quad & \binom{3}{2} & \quad & \binom{3}{3} \\
& & \binom{4}{0} & \quad & \binom{4}{1} & \quad & \binom{4}{2} & \quad & \binom{4}{3} \\
& & \binom{5}{0} & \quad & \binom{5}{1} & \quad & \binom{5}{2} & \quad & \binom{5}{3} \\
& & \binom{6}{0} & \quad & \binom{6}{1} & \quad & \binom{6}{2} & \quad & \binom{6}{3} \\
& & \binom{6}{4} & \quad & \binom{6}{5} & \quad & \binom{6}{6} & \quad & \binom{6}{6}
\end{array}$$

We obtain the values for the binomial coefficients by using the following rules:

- Each value along the boundary is equal to 1.
  - Each value in the interior is equal to the sum of the two values above it.

In Figure 3.1, you see rows 0, 1, . . . , 12.

Below, we state some of our earlier results using Pascal's Triangle.

- The values in the  $n$ -th row are equal to the coefficients in Newton's Binomial Theorem (i.e., Theorem 3.6.5). For example, the coefficients in the expansion of  $(x + y)^5$  are given in the 5-th row:

$$(x+y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5.$$

- Theorem 3.6.6 states that the sum of all values in the  $n$ -th row is equal to  $2^n$ .
  - Theorem 3.7.1 states that reading the  $n$ -th row from left to right gives the same sequence as reading this row from right to left.
  - Corollary 3.7.5 states that the sum of the squares of all values in the  $n$ -th row is equal to the middle element in the  $2n$ -th row.

													1
													1      1
													1      2      1
													1      3      3      1
													1      4      6      4      1
													1      5      10     10     5      1
													1      6      15     20     15     6      1
													1      7      21     35     35     21     7      1
													1      8      28     56     70     56     28     8      1
													1      9      36     84     126    126    84     36     9      1
													1      10     45     120    210    252    210    120    45     10     1
													1      11     55     165    330    462    462    330    165    55     11     1
1	12	66	220	495	792	924	792	495	220	66	12	1	

Figure 3.1: Rows 0, 1, . . . , 12 of Pascal’s Triangle.

## 3.9 More Counting Problems

### 3.9.1 Reordering the Letters of a Word

How many different strings can be made by reordering the letters of the 7-letter word

SUCCESS.

It should be clear that the answer is not  $7!$ : If we swap, for example, the two occurrences of C, then we obtain the same string.

The correct answer can be obtained by applying the Product Rule. We start by counting the frequencies of each letter:

- The letter S occurs 3 times.
- The letter C occurs 2 times.
- The letter U occurs 1 time.
- The letter E occurs 1 time.

To apply the Product Rule, we have to specify the procedure and the tasks:

- The procedure is “write the letters occurring in the word SUCCESS”.
- The first task is “choose 3 positions out of 7, and write the letter S in each chosen position”.
- The second task is “choose 2 positions out of the remaining 4, and write the letter C in each chosen position”.
- The third task is “choose 1 position out of the remaining 2, and write the letter U in the chosen position”.
- The fourth task is “choose 1 position out of the remaining 1, and write the letter E in the chosen position”.

Since there are  $\binom{7}{3}$  ways to do the first task,  $\binom{4}{2}$  ways to do the second task,  $\binom{2}{1}$  ways to do the third task, and  $\binom{1}{1}$  way to do the fourth task, it follows that the total number of different strings that can be made by reordering the letters of the word SUCCESS is equal to

$$\binom{7}{3} \binom{4}{2} \binom{2}{1} \binom{1}{1} = 420.$$

In the four tasks above, we first chose the positions for the letter S, then the positions for the letter C, then the position for the letter U, and finally the position for the letter E. If we change the order, then we obtain the same answer. For example, if we choose the positions for the letters in the order C, E, U, S, then we obtain

$$\binom{7}{2} \binom{5}{1} \binom{4}{1} \binom{3}{3},$$

which is indeed equal to 420.

### 3.9.2 Counting Solutions of Linear Equations

Consider the equation

$$x_1 + x_2 + x_3 = 11.$$

We are interested in the number of solutions  $(x_1, x_2, x_3)$ , where  $x_1 \geq 0$ ,  $x_2 \geq 0$ ,  $x_3 \geq 0$  are integers. Examples of solutions are

$$(2, 3, 6), (3, 2, 6), (0, 11, 0), (2, 0, 9).$$

Observe that we consider  $(2, 3, 6)$  and  $(3, 2, 6)$  to be different solutions.

We are going to use the Bijection Rule to determine the number of solutions. For this, we define  $A$  to be the set of all solutions, i.e.,

$$A = \{(x_1, x_2, x_3) : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \text{ are integers, } x_1 + x_2 + x_3 = 11\}.$$

To apply the Bijection Rule, we need a set  $B$  and a bijection  $f : A \rightarrow B$ , such that it is easy to determine the size of  $B$ . This set  $B$  should be chosen such that its elements “encode” the elements of  $A$  in a unique way. Consider the following set  $B$ :

- $B$  is the set of all bitstrings of length 13 that contain exactly 2 many 1s (and, thus, exactly 11 many 0s).

The function  $f : A \rightarrow B$  is defined as follows: If  $(x_1, x_2, x_3)$  is an element of the set  $A$ , then  $f(x_1, x_2, x_3)$  is the bitstring

- that starts with  $x_1$  many 0s,
- is followed by one 1,

- is followed by  $x_2$  many 0s,
- is followed by one 1,
- and ends with  $x_3$  many 0s.

For example, we have

$$f(2, 3, 6) = 0010001000000,$$

$$f(3, 2, 6) = 0001001000000,$$

$$f(0, 11, 0) = 1000000000001,$$

and

$$f(2, 0, 9) = 0011000000000.$$

To show that this function  $f$  maps elements of  $A$  to elements of  $B$ , we have to verify that the string  $f(x_1, x_2, x_3)$  belongs to the set  $B$ . This follows from the following observations:

- The string  $f(x_1, x_2, x_3)$  contains exactly 2 many 1s.
- The number of 0s in the string  $f(x_1, x_2, x_3)$  is equal to  $x_1 + x_2 + x_3$ , which is equal to 11, because  $(x_1, x_2, x_3)$  belongs to the set  $A$ .
- Thus,  $f(x_1, x_2, x_3)$  is a bitstring of length 13 that contains exactly 2 many 1s.

It should be clear that this function  $f$  is one-to-one: If we take two different elements  $(x_1, x_2, x_3)$  in  $A$ , then  $f$  gives us two different bitstrings  $f(x_1, x_2, x_3)$ .

To prove that  $f$  is onto, we have to show that for every bitstring  $b$  in the set  $B$ , there is an element  $(x_1, x_2, x_3)$  in  $A$  such that  $f(x_1, x_2, x_3) = b$ . This element of  $A$  is obtained by taking

- $x_1$  to be the number of 0s to the left of the first 1,
- $x_2$  to be the number of 0s between the two 1s, and
- $x_3$  to be the number of 0s to the right of the second 1.

For example, if  $b = 0000110000000$ , then  $x_1 = 4$ ,  $x_2 = 0$ , and  $x_3 = 7$ . Note that, since  $b$  has length 13 and contains exactly 2 many 1s, we have  $x_1 + x_2 + x_3 = 11$  and, therefore,  $(x_1, x_2, x_3) \in A$ .

Thus, we have shown that  $f : A \rightarrow B$  is indeed a bijection. We know from Theorem 3.6.4 that  $B$  has size  $\binom{13}{2}$ . Therefore, it follows from the Bijection Rule that

$$|A| = |B| = \binom{13}{2} = 78.$$

The following theorem states this result for general linear equations. You are encouraged to come up with the proof.

**Theorem 3.9.1** *Let  $k \geq 1$  and  $n \geq 0$  be integers. The number of solutions to the equation*

$$x_1 + x_2 + \cdots + x_k = n,$$

*where  $x_1 \geq 0, x_2 \geq 0, \dots, x_k \geq 0$  are integers, is equal to*

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

Let us now consider inequalities instead of equations. For example, consider the inequality

$$x_1 + x_2 + x_3 \leq 11.$$

Again, we are interested in the number of solutions  $(x_1, x_2, x_3)$ , where  $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$  are integers. This inequality contains the same solutions as before, but it has additional solutions such as

$$(2, 3, 5), (3, 2, 5), (0, 1, 0), (0, 0, 0).$$

As before, we are going to apply the Bijection Rule. We define

$$A = \{(x_1, x_2, x_3) : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \text{ are integers}, x_1 + x_2 + x_3 \leq 11\}$$

and  $B$  to be the set of all bitstrings of length 14 that contain exactly 3 many 1s (and, thus, exactly 11 many 0s).

The function  $f : A \rightarrow B$  is defined as follows: If  $(x_1, x_2, x_3)$  is an element of  $A$ , then  $f(x_1, x_2, x_3)$  is the bitstring

- that starts with  $x_1$  many 0s,

- is followed by one 1,
- is followed by  $x_2$  many 0s,
- is followed by one 1,
- is followed by  $x_3$  many 0s,
- is followed by one 1,
- and ends with  $14 - (x_1 + x_2 + x_3 + 3)$  many 0s.

For example, we have

$$f(2, 3, 6) = 00100010000001,$$

$$f(2, 3, 5) = 00100010000010,$$

$$f(0, 1, 0) = 10110000000000,$$

and

$$f(0, 0, 0) = 11100000000000.$$

As before, it can be verified that the string  $f(x_1, x_2, x_3)$  belongs to the set  $B$  and the function  $f$  is a bijection. It then follows from the Bijection Rule that

$$|A| = |B| = \binom{14}{3} = 364.$$

The next theorem gives the answer for the general case. As before, you are encouraged to give a proof.

**Theorem 3.9.2** *Let  $k \geq 1$  and  $n \geq 0$  be integers. The number of solutions to the inequality*

$$x_1 + x_2 + \cdots + x_k \leq n,$$

*where  $x_1 \geq 0, x_2 \geq 0, \dots, x_k \geq 0$  are integers, is equal to*

$$\binom{n+k}{k}.$$

## 3.10 The Pigeonhole Principle

In any group of 366 people, there must be two people having the same birthday: Since there are 365 days in a year (ignoring leap years), it is not possible that the birthdays of 366 people are all distinct.

**Pigeonhole Principle:** Let  $k \geq 1$  be an integer. If  $k+1$  or more objects are placed into  $k$  boxes, then there is at least one box containing two or more objects.

Equivalently, if  $A$  and  $B$  are two finite sets with  $|A| > |B|$ , then there is no one-to-one function from  $A$  to  $B$ .

### 3.10.1 India Pale Ale

President of the Carleton Computer Science Society	Favorite Drink
Simon Pratt (2013–2014)	India Pale Ale
Lindsay Bangs (2014–2015)	Wheat Beer
Connor Hillen (2015–2016)	Black IPA
Elisa Kazan (2016–2019)	Cider
William So (2019–2020)	Amber Lager

Simon Pratt loves to drink India Pale Ale (IPA). During each day of the month of April (which has 30 days), Simon drinks at least one bottle of IPA. During this entire month, he drinks exactly 45 bottles of IPA. The claim is that there must be a sequence of consecutive days in April, during which Simon drinks exactly 14 bottles of IPA.

To prove this, let  $b_i$  be the number of bottles that Simon drinks on April  $i$ , for  $i = 1, 2, \dots, 30$ . We are given that each  $b_i$  is a positive integer (i.e.,  $b_i \geq 1$ ) and

$$b_1 + b_2 + \cdots + b_{30} = 45.$$

Define, for  $i = 1, 2, \dots, 30$ ,

$$a_i = b_1 + b_2 + \cdots + b_i,$$

i.e.,  $a_i$  is the total number of bottles of IPA that Simon drinks during the first  $i$  days of April. Consider the sequence of 60 numbers

$$a_1, a_2, \dots, a_{30}, a_1 + 14, a_2 + 14, \dots, a_{30} + 14.$$

Each number in this sequence is an integer that belongs to the set

$$\{1, 2, \dots, 59\}.$$

Therefore, by the Pigeonhole Principle, these 60 numbers cannot all be distinct. Observe that there are no duplicates in the sequence  $a_1, a_2, \dots, a_{30}$ , because all  $b_i$  are at least one. Similarly, there are no duplicates in the sequence  $a_1 + 14, a_2 + 14, \dots, a_{30} + 14$ . It follows that there are two indices  $i$  and  $j$  such that

$$a_i = a_j + 14.$$

Observe that  $j$  must be less than  $i$  and

$$14 = a_i - a_j = b_{j+1} + b_{j+2} + \dots + b_i.$$

Thus, in the period from April  $j + 1$  until April  $i$ , Simon drinks exactly 14 bottles of IPA.

### 3.10.2 Sequences Containing Divisible Numbers

Let  $A = \{1, 2, \dots, 2n\}$  and consider the sequence  $n + 1, n + 2, \dots, 2n$  of elements in  $A$ . This sequence has the property that none of its elements divides any other element in the sequence. Note that the sequence has length  $n$ . The following theorem states that such a sequence of length  $n + 1$  does not exist.

**Theorem 3.10.1** *Let  $n \geq 1$  and consider a sequence  $a_1, a_2, \dots, a_{n+1}$  of  $n+1$  elements from the set  $\{1, 2, \dots, 2n\}$ . Then there are two distinct indices  $i$  and  $j$  such that  $a_i$  divides  $a_j$  or  $a_j$  divides  $a_i$ .*

**Proof.** For each  $i$  with  $1 \leq i \leq n + 1$ , write

$$a_i = 2^{k_i} \cdot q_i,$$

where  $k_i \geq 0$  is an integer and  $q_i$  is an odd integer. For example,

- if  $a_i = 48$ , then  $k_i = 4$  and  $q_i = 3$ , because  $48 = 2^4 \cdot 3$ ,
- if  $a_i = 1$ , then  $k_i = 0$  and  $q_i = 1$ , because  $1 = 2^0 \cdot 1$ ,
- if  $a_i = 7$ , then  $k_i = 0$  and  $q_i = 7$ , because  $7 = 2^0 \cdot 7$ .

Consider the sequence  $q_1, q_2, \dots, q_{n+1}$  of  $n+1$  integers. Each of these numbers is an odd integer that belongs to the set

$$\{1, 3, 5, \dots, 2n - 1\}.$$

Since this set has size  $n$ , the Pigeonhole Principle implies that there must be two numbers in the sequence  $q_1, q_2, \dots, q_{n+1}$  that are equal. In other words, there are two distinct indices  $i$  and  $j$  such that  $q_i = q_j$ . It follows that

$$\frac{a_i}{a_j} = \frac{2^{k_i} \cdot q_i}{2^{k_j} \cdot q_j} = 2^{k_i - k_j}.$$

Thus, if  $k_i \geq k_j$ , then  $a_j$  divides  $a_i$ . Otherwise,  $k_i < k_j$ , and  $a_i$  divides  $a_j$ . ■

### 3.10.3 Long Monotone Subsequences

Let  $n = 3$ , and consider the sequence 20, 10, 9, 7, 11, 2, 21, 1, 20, 31 of  $10 = n^2 + 1$  numbers. This sequence contains an increasing subsequence of length  $4 = n + 1$ , namely 10, 11, 21, 31. The following theorem states this result for arbitrary values of  $n$ .

**Theorem 3.10.2** *Let  $n \geq 1$  be an integer. Every sequence of  $n^2 + 1$  distinct real numbers contains a subsequence of length  $n + 1$  that is either increasing or decreasing.*

**Proof.** Let  $a_1, a_2, \dots, a_{n^2+1}$  be an arbitrary sequence of  $n^2 + 1$  distinct real numbers. For each  $i$  with  $1 \leq i \leq n^2 + 1$ , let  $\text{inc}_i$  denote the length of the longest increasing subsequence that starts at  $a_i$ , and let  $\text{dec}_i$  denote the length of the longest decreasing subsequence that starts at  $a_i$ .

Using this notation, the claim in the theorem can be formulated as follows: There is an index  $i$  such that  $\text{inc}_i \geq n + 1$  or  $\text{dec}_i \geq n + 1$ .

We will prove the claim by contradiction. Thus, we assume that  $\text{inc}_i \leq n$  and  $\text{dec}_i \leq n$  for all  $i$  with  $1 \leq i \leq n^2 + 1$ .

Consider the set

$$B = \{(b, c) : 1 \leq b \leq n, 1 \leq c \leq n\},$$

and think of the elements of  $B$  as being boxes. For each  $i$  with  $1 \leq i \leq n^2 + 1$ , the pair  $(\text{inc}_i, \text{dec}_i)$  is an element of  $B$ . Thus, we have  $n^2 + 1$  elements

$(inc_i, dec_i)$ , which are placed in the  $n^2$  boxes of  $B$ . By the Pigeonhole Principle, there must be a box that contains two (or more) elements. In other words, there exist two integers  $i$  and  $j$  such that  $i < j$  and

$$(inc_i, dec_i) = (inc_j, dec_j).$$

Recall that the elements in the sequence are distinct. Hence,  $a_i \neq a_j$ . We consider two cases.

First assume that  $a_i < a_j$ . Then the length of the longest increasing subsequence starting at  $a_i$  must be at least  $1 + inc_j$ , because we can append  $a_i$  to the longest increasing subsequence starting at  $a_j$ . Therefore,  $inc_i \neq inc_j$ , which is a contradiction.

The second case is when  $a_i > a_j$ . Then the length of the longest decreasing subsequence starting at  $a_i$  must be at least  $1 + dec_j$ , because we can append  $a_i$  to the longest decreasing subsequence starting at  $a_j$ . Therefore,  $dec_i \neq dec_j$ , which is again a contradiction. ■

### 3.10.4 There are Infinitely Many Primes

As a final application of the Pigeonhole Principle, we prove the following result:

**Theorem 3.10.3** *There are infinitely many prime numbers.*

**Proof.** The proof is by contradiction. Thus, we assume that there are, say,  $k$  prime numbers, and denote them by

$$2 = p_1 < p_2 < \cdots < p_k.$$

Note that  $k$  is a fixed integer. Since

$$\lim_{n \rightarrow \infty} \frac{2^n}{(n+1)^k} = \infty,$$

we can choose an integer  $n$  such that

$$2^n > (n+1)^k.$$

Define the function

$$f : \{1, 2, \dots, 2^n\} \rightarrow \mathbb{N}^k$$

as follows: For any integer  $x$  with  $1 \leq x \leq 2^n$ , consider its prime factorization

$$x = p_1^{m_1} \cdot p_2^{m_2} \cdots p_k^{m_k}.$$

We define

$$f(x) = (m_1, m_2, \dots, m_k).$$

Since

$$\begin{aligned} m_i &\leq m_1 + m_2 + \cdots + m_k \\ &\leq m_1 \log p_1 + m_2 \log p_2 + \cdots + m_k \log p_k \\ &= \log(p_1^{m_1} \cdot p_2^{m_2} \cdots p_k^{m_k}) \\ &= \log x \\ &\leq n, \end{aligned}$$

it follows that

$$f(x) \in \{0, 1, 2, \dots, n\}^k.$$

Thus,  $f$  is a function

$$f : \{1, 2, \dots, 2^n\} \rightarrow \{0, 1, 2, \dots, n\}^k.$$

It is easy to see that this function is one-to-one. The set on the left-hand side has size  $2^n$ , whereas the set on the right-hand side has size  $(n+1)^k$ . It then follows from the Pigeonhole Principle that

$$(n+1)^k \geq 2^n,$$

which contradicts our choice for  $n$ . ■

## 3.11 Exercises

**3.1** A licence plate number consists of a sequence of four uppercase letters followed by three digits. How many licence plate numbers are there?

**3.2** A multiple-choice exam consists of 100 questions. Each question has four possible answers  $a$ ,  $b$ ,  $c$ , and  $d$ . How many ways are there to answer the 100 questions (assuming that each question is answered)?

**3.3** For each of the following seven cases, determine how many strings of eight uppercase letters there are.

- Letters can be repeated.
- No letter can be repeated.
- The strings start with PQ (in this order) and letters can be repeated.
- The strings start with PQ (in this order) and no letter can be repeated.
- The strings start and end with PQ (in this order) and letters can be repeated.
- The strings start with XYZ (in this order), end with QP (in this order), and letters can be repeated.
- The strings start with XYZ (in this order) or end with QP (in this order), and letters can be repeated.

**3.4** If  $n$  and  $d$  are positive integers, then  $d$  is a *divisor* of  $n$ , if  $n/d$  is an integer.

Determine the number of divisors of the integer

$$1,170,725,783,076,864 = 2^{17} \cdot 3^{12} \cdot 7^5.$$

**3.5** Let  $k \geq 1$  and  $n \geq 1$  be integers. Consider  $k$  distinct beer bottles and  $n$  distinct students. How many ways are there to hand out the beer bottles to the students, if there is no restriction on how many bottles a student may get?

**3.6** The Carleton Computer Science Society has a Board of Directors consisting of one president, one vice-president, one secretary, one treasurer, and a three-person party committee (whose main responsibility is to buy beer for the other four board members). The entire board consists of seven distinct students. If there are  $n \geq 7$  students in Carleton's Computer Science program, how many ways are there to choose a Board of Directors?

**3.7** The Carleton Computer Science Society has an Academic Events Committee (AEC) consisting of five students and a Beer Committee (BC) consisting of six students (whose responsibility is to buy beer for the AEC).

- Assume there are  $n \geq 6$  students in Carleton's Computer Science program. Also, assume that a student can be both on the AEC and on the BC. What is the total number of ways in which these two committees can be chosen?
- Assume there are  $n \geq 11$  students in Carleton's Computer Science program. Also, assume that a student cannot be both on the AEC and on the BC. What is the total number of ways in which these two committees can be chosen?

**3.8** Let  $f \geq 2$ ,  $m \geq 2$ , and  $k \geq 2$  be integers such that  $k \leq f$  and  $k \leq m$ . The Carleton Computer Science program has  $f$  female students and  $m$  male students. The Carleton Computer Science Society has a Board of Directors consisting of  $k$  students. At least one of the board members is female and at least one of the board members is male. Determine the number of ways in which a Board of Directors can be chosen.

**3.9** Let  $f \geq 4$  and  $m \geq 4$  be integers. The Carleton Computer Science program has  $f$  female students and  $m$  male students that are eligible to be a TA for COMP 2804. Determine the number of ways to choose eight TAs out of these  $f + m$  students, such that the number of female TAs is equal to the number of male TAs

**3.10** Let  $m$  and  $n$  be integers with  $0 \leq m \leq n$ . There are  $n + 1$  students in Carleton's Computer Science program. The Carleton Computer Science Society has a Board of Directors, consisting of one president and  $m$  vice-presidents. The president cannot be vice-president. Prove that

$$(n+1)\binom{n}{m} = (n+1-m)\binom{n+1}{m},$$

by determining, in two different ways, the number of ways to choose a Board of Directors.

**3.11** In Tic-Tac-Toe, we are given a  $3 \times 3$  grid, consisting of unmarked cells. Two players, Xavier and Olivia, take turns marking the cells of this grid. When it is Xavier's turn, he chooses an unmarked cell and marks it with the letter  $X$ . Similarly, when it is Olivia's turn, she chooses an unmarked cell and marks it with the letter  $O$ . The first turn is by Xavier. The players continue making turns until all cells have been marked. Below, you see an example of a completely marked grid.

$O$	$O$	$X$
$X$	$X$	$O$
$X$	$X$	$O$

- What is the number of completely marked grids?
- What is the number of different ways (i.e., ordered sequences) in which the grid can be completely marked, when following the rules given above?

**3.12** In how many ways can you paint 200 chairs, if 33 of them must be painted red, 66 of them must be painted blue, and 101 of them must be painted green?

**3.13** Let  $A$  be the set of all integers  $x > 6543$  such that the decimal representation of  $x$  has distinct digits, none of which is equal to 7, 8, or 9. (The decimal representation does not have leading zeros.) Determine the size of the set  $A$ .

**3.14** Let  $A$  be the set of all integers  $x \in \{1, 2, \dots, 100\}$  such that the decimal representation of  $x$  does not contain the digit 4. (The decimal representation does not have leading zeros.)

- Determine the size of the set  $A$  without using the Complement Rule.
- Use the Complement Rule to determine the size of the set  $A$ .

**3.15** Let  $A$  be a set of size  $m$ , let  $B$  be a set of size  $n$ , and assume that  $n \geq m \geq 1$ . How many functions  $f : A \rightarrow B$  are there that are *not* one-to-one?

**3.16** Consider permutations of the set  $\{a, b, c, d, e, f, g\}$  that do not contain *bge* (in this order) and do not contain *eaf* (in this order). Prove that the number of such permutations is equal to 4806.

**3.17** How many bitstrings of length 8 are there that contain at least 4 consecutive 0s or at least 4 consecutive 1s?

**3.18** How many bitstrings of length 77 are there that start with 010 (i.e., have 010 at positions 1, 2, and 3), or have 101 at positions 2, 3, and 4, or have 010 at positions 3, 4, and 5?

**3.19** Let  $n \geq 12$  be an integer and let  $\{B_1, B_2, \dots, B_n\}$  be a set of  $n$  beer bottles. Consider permutations of these bottles such that there are exactly 10 bottles between  $B_1$  and  $B_n$ . ( $B_1$  can be to the left or right of  $B_n$ .) Prove that the number of such permutations is equal to

$$\binom{n-2}{10} \cdot 10! \cdot 2 \cdot (n-11)!.$$

**3.20** Let  $n \geq 3$  be an integer. The  $Gn$  (or Group of  $n$ ) is an international forum where the  $n$  leaders of the world meet to drink beer together. Two of these leaders are Donald Trump and Justin Trudeau. At the end of their meeting, the  $n$  leaders stand on a line and a group photo is taken.

- Determine the number of ways in which the  $n$  leaders can be arranged on a line, if Donald Trump and Justin Trudeau are standing next to each other.
- Determine the number of ways in which the  $n$  leaders can be arranged on a line, if Donald Trump and Justin Trudeau are not standing next to each other.
- Determine the number of ways in which the  $n$  leaders can be arranged on a line, if Donald Trump is to the left of Justin Trudeau. (Donald does not necessarily stand immediately to the left of Justin.)

**3.21** A string of letters is called a *palindrome*, if reading the string from left to right gives the same result as reading the string from right to left. For example, *madam* and *racecar* are palindromes. Recall that there are five vowels in the English alphabet: *a*, *e*, *i*, *o*, and *u*.

In this exercise, we consider strings consisting of 28 characters, with each character being a lowercase letter. Determine the number of such strings that start and end with the same letter, or are palindromes, or contain vowels only.

**3.22** A *flip* in a bitstring is a pair of adjacent bits that are not equal. For example, the bitstring 010011 has three flips: The first two bits form a flip, the second and third bits form a flip, and the fourth and fifth bits form a flip.

- Determine the number of bitstrings of length 7 that have exactly 3 flips at the following positions: The second and third bits form a flip, the third and fourth bits form a flip, and the fifth and sixth bits form a flip.

- Let  $n \geq 2$  and  $k$  be integers with  $0 \leq k \leq n-1$ . Determine the number of bitstrings of length  $n$  that have exactly  $k$  flips.

**3.23** Let  $m$  and  $n$  be integers with  $m \geq n \geq 1$ . How many ways are there to place  $m$  books on  $n$  shelves, if there must be at least one book on each shelf? As in Section 3.1.3, the order on each shelf matters.

**3.24** You are given  $m$  distinct books  $B_1, B_2, \dots, B_m$  and  $n$  *identical* blocks of wood. How many ways are there to arrange these books and blocks in a straight line?

For example, if  $m = 5$  and  $n = 3$ , then three possible arrangements are ( $W$  stands for a block of wood)

$$WB_3B_1WB_5B_4WB_2,$$

$$WB_1B_3WB_5B_4WB_2,$$

and

$$B_5WB_3B_1WWB_2B_4.$$

**3.25** Let  $n \geq 1$  be an integer and consider  $n$  boys and  $n$  girls. For each of the following three cases, determine how many ways there are to arrange these  $2n$  people on a straight line (the order on the line matters):

- All boys stand next to each other and all girls stand next to each other.
- All girls stand next to each other.
- Boys and girls alternate.

**3.26** Elisa Kazan has a set  $\{C_1, C_2, \dots, C_{50}\}$  consisting of 50 cider bottles. She divides these bottles among 5 friends, so that each friend receives a subset consisting of 10 bottles. Determine the number of ways in which Elisa can divide the bottles.

**3.27** Let  $n \geq 1$  be an integer. Consider a tennis tournament with  $2n$  participants. In the first round of this tournament,  $n$  games will be played and, thus, the  $2n$  people have to be divided into  $n$  pairs. What is the number of ways in which this can be done?

**3.28** The Ottawa Senators and the Toronto Maple Leafs play a best-of-7 series: These two hockey teams play games against each other, and the first team to win 4 games wins the series. Each game has a winner (thus, no game ends in a tie).

A sequence of games can be described by a string consisting of the characters  $S$  (indicating that the Senators win the game) and  $L$  (indicating that the Leafs win the game). Two possible ways for the Senators to win the series are  $(L, S, S, S, S)$  and  $(S, L, S, L, S, S)$ .

Determine the number of ways in which the Senators can win the series.

**3.29** The Beer Committee of the Carleton Computer Science Society has bought large quantities of 10 different types of beer. In order to test which beer students prefer, the committee does the following experiment:

- Out of the  $n \geq 10$  students in Carleton's Computer Science program, 10 students are chosen.
- Each of the 10 students chosen drinks one of the 10 beers; no two students drink the same beer.

What is the number of ways in which this experiment can be done?

**3.30** Let  $m$ ,  $n$ ,  $k$ , and  $\ell$  be integers such that  $m \geq 1$ ,  $n \geq 1$ ,  $1 \leq \ell \leq k \leq \ell+m$  and  $\ell \leq n$ .

After a week of hard work, Elisa Kazan goes to her neighborhood pub. This pub has  $m$  different types of beer and  $n$  different types of cider on tap. Elisa decides to order  $k$  pints: At most one pint of each type, and exactly  $\ell$  pints of cider. Determine the number of ways in which Elisa can order these  $k$  pints. The order in which Elisa orders matters.

**3.31** Let  $m \geq 2$  and  $n \geq 2$  be even integers. You are given  $m$  beer bottles  $B_1, B_2, \dots, B_m$  and  $n$  cider bottles  $C_1, C_2, \dots, C_n$ . Assume you arrange these  $m+n$  bottles on a horizontal line such that

- the leftmost  $m/2$  bottles are all beer bottles, and
- the rightmost  $n/2$  bottles are all cider bottles.

How many such arrangements are there?

**3.32** Consider 10 male students  $M_1, M_2, \dots, M_{10}$  and 7 female students  $F_1, F_2, \dots, F_7$ . Assume these 17 students are arranged on a horizontal line such that no two female students are standing next to each other. We are interested in the number of such arrangements, where the order of the students matters.

- Explain what is wrong with the following argument:

We are going to use the Product Rule:

- Task 1: Arrange the 7 females on a line. There are  $7!$  ways to do this.
- Task 2: Choose 6 males. There are  $\binom{10}{6}$  ways to do this.
- Task 3: Place the 6 males chosen in Task 2 in the 6 “gaps” between the females. There are  $6!$  ways to do this.
- Task 4: At this moment, we have arranged 13 students on a line. We are left with 4 males that have to be placed.
  - \* Task 4.1: Place one male. There are 14 ways to do this.
  - \* Task 4.2: Place one male. There are 15 ways to do this.
  - \* Task 4.3: Place one male. There are 16 ways to do this.
  - \* Task 4.4: Place one male. There are 17 ways to do this.

By the Product Rule, the total number of ways to arrange the 17 students is equal to

$$7! \cdot \binom{10}{6} \cdot 6! \cdot 14 \cdot 15 \cdot 16 \cdot 17 = 43,528,181,760,000.$$

- Determine the number of ways to arrange the 17 students.

*Hint:* Use the Product Rule. What is easier to count: Placing the

female students first and then the male students, or placing the male students first and then the female students?

**3.33** Let  $n \geq 1$  be an integer. A function  $f : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$  is called *awesome*, if there is at least one integer  $i$  in  $\{1, 2, \dots, n\}$  for which  $f(i) = i$ . Determine the number of awesome functions.

**3.34** Let  $n \geq 2$  be an integer. Consider strings consisting of  $n$  digits.

- Determine the number of such strings, in which no two consecutive digits are equal.
- Determine the number of such strings, in which there is at least one pair of consecutive digits that are equal.

**3.35** Consider strings consisting of 12 characters, each character being  $a$ ,  $b$ , or  $c$ . Such a string is called *valid*, if at least one of the characters is missing. For example, *abababababab* is a valid string, whereas *abababacabab* is not a valid string. How many valid strings are there?

**3.36** Consider strings consisting of 40 characters, where each character is an element of  $\{a, b, c\}$ . Such a string is called *cool*, if it contains exactly 8 many  $a$ 's or exactly 7 many  $b$ 's. Determine the number of cool strings.

**3.37** A password consists of 100 characters, each character being a digit or a lowercase letter. A password must contain at least two digits. How many passwords are there?

**3.38** A password is a string of ten characters, where each character is a lowercase letter, a digit, or one of the eight special characters !, @, #, \$, %, &, (, and ).

A password is called *awesome*, if it contains at least one digit or at least one special character. Determine the number of awesome passwords.

**3.39** A password consists of 100 characters, each character being a digit, a lowercase letter, or an uppercase letter. A password must contain at least one digit, at least one lowercase letter, and at least one uppercase letter. How many passwords are there?

*Hint:* Recall De Morgan's Law

$$A \cap B \cap C = \overline{\overline{A} \cup \overline{B} \cup \overline{C}}.$$

**3.40** A password is a string of 100 characters, where each character is a digit or a lowercase letter. A password is called *valid*, if

- it does not start with *abc*, and
- it does not end with *xyz*, and
- it does not start with 3456.

Determine the number of valid passwords.

**3.41** A password is a string of 8 characters, where each character is a lowercase letter or a digit. A password is called *valid*, if it contains at least one digit. In Section 3.3, we have seen that the number of valid passwords is equal to

$$36^8 - 26^8 = 2,612,282,842,880.$$

Explain what is wrong with the following method to count the valid passwords.

We are going to use the Product Rule.

- The procedure is “write a valid password”.
- Since a valid password contains at least one digit, we choose, in the first task, a position for the digit.
- The second task is to write a digit at the chosen position.
- The third task is to write a character (lowercase letter or digit) at each of the remaining 7 positions.

There are 8 ways to do the first task, 10 ways to do the second task, and  $36^7$  ways to do the third task. Therefore, by the Product Rule, the number of valid passwords is equal to

$$8 \cdot 10 \cdot 36^7 = 6,269,133,127,680.$$

**3.42** Consider permutations of the 26 lowercase letters *a, b, c, ..., z*.

- How many such permutations contain the string *wine*?

- How many such permutations do not contain any of the strings *wine*, *vodka*, or *coke*?

**3.43** Determine the number of integers in the set  $\{1, 2, \dots, 1000\}$  that are not divisible by any of 5, 7, and 11.

**3.44** Let  $n \geq 4$  be an integer. Determine the number of permutations of  $\{1, 2, \dots, n\}$ , in which

- 1 and 2 are next to each other, with 1 to the left of 2, or
- 4 and 3 are next to each other, with 4 to the left of 3.

**3.45** Determine the number of functions

$$f : \{1, 2, 3, 4\} \rightarrow \{a, b, c, \dots, z\},$$

such that  $f(1) = f(2)$ , or  $f(3) = f(4)$ , or  $f(1) \neq f(3)$ .

**3.46** Let  $n \geq 3$  be an integer. Determine the number of permutations of  $\{1, 2, \dots, n\}$ , in which

- 1 and 2 are next to each other, with 1 to the left of 2, or
- 2 and 3 are next to each other, with 2 to the left of 3.

Compare your answer with the answer to Exercise 3.44.

**3.47** Let  $n$  and  $k$  be integers with  $2 \leq k \leq n$ , and consider the set  $S = \{1, 2, 3, \dots, 2n\}$ . An ordered sequence of  $k$  elements of  $S$  is called *valid*, if

- this sequence is strictly increasing, or
- this sequence is strictly decreasing, or
- this sequence contains only even numbers (and duplicate elements are allowed).

Determine the number of valid sequences.

**3.48** Let  $n \geq 2$  be an integer.

- Determine the number of strings consisting of  $n$  characters, where each character is an element of the set  $\{a, b, 0\}$ .
- Let  $S$  be a set consisting of  $n$  elements. Determine the number of ordered pairs  $(A, B)$ , where  $A \subseteq S$ ,  $B \subseteq S$ , and  $A \cap B = \emptyset$ .
- Let  $S$  be a set consisting of  $n$  elements. Consider ordered pairs  $(A, B)$ , where  $A \subseteq S$ ,  $B \subseteq S$ , and  $|A \cap B| = 1$ . Prove that the number of such pairs is equal to  $n \cdot 3^{n-1}$ .

**3.49** In a group of 20 people,

- 6 are blond,
- 7 have green eyes,
- 11 are not blond and do not have green eyes.

How many people in this group are blond and have green eyes?

**3.50** Let  $n \geq 1$  be an integer.

- Assume that  $n$  is odd. Determine the number of bitstrings of length  $n$  that contain more 0's than 1's. Justify your answer in plain English.
- Assume that  $n$  is even.
  - Determine the number of bitstrings of length  $n$  in which the number of 0's is equal to the number of 1's.
  - Determine the number of bitstrings of length  $n$  that contain strictly more 0's than 1's.
  - Argue that the binomial coefficient

$$\binom{n}{n/2}$$

is an even integer.

**3.51** Use Pascal's Identity (Theorem 3.7.2) to prove Newton's Binomial Theorem (i.e., Theorem 3.6.5) by induction.

**3.52** Determine the coefficient of  $x^{111}y^{444}$  in the expansion of

$$(-17x + 71y)^{555}.$$

**3.53** Nick is not only your friendly TA<sup>2</sup>, he also has a part-time job in a grocery store. This store sells  $n$  different types of India Pale Ale (IPA) and  $n$  different types of wheat beer, where  $n \geq 2$  is an integer. Prove that

$$\binom{2n}{2} = 2\binom{n}{2} + n^2,$$

by counting, in two different ways, the number of ways to choose two different types of beer.

**3.54** You have won the first prize in the *Louis van Gaal Impersonation Contest*<sup>3</sup>. When you arrive at Louis' home to collect your prize, you see  $n$  beer bottles  $B_1, B_2, \dots, B_n$ ,  $n$  cider bottles  $C_1, C_2, \dots, C_n$ , and  $n$  wine bottles  $W_1, W_2, \dots, W_n$ . Here,  $n$  is an integer with  $n \geq 2$ . Louis tells you that your prize consists of one beer bottle of your choice, one cider bottle of your choice, and one wine bottle of your choice.

Prove that

$$n^3 = (n - 1)^3 + 3(n - 1)^2 + 3(n - 1) + 1,$$

by counting, in two different ways, the number of ways in which you can choose your prize.

**3.55** Let  $n \geq 4$  be an integer and consider the set  $S = \{1, 2, \dots, n\}$ . Let  $k$  be an integer with  $2 \leq k \leq n - 2$ . In this exercise, we consider subsets  $A$  of  $S$  for which  $|A| = k$  and  $\{1, 2\} \not\subseteq A$ . Let  $N$  denote the number of such subsets.

- Use the Sum Rule to determine  $N$ .
- Use the Complement Rule to determine  $N$ .
- Use the above two results to prove that

$$\binom{n}{k} = \binom{n-2}{k} + 2\binom{n-2}{k-1} + \binom{n-2}{k-2}.$$

---

<sup>2</sup>Winter term 2017

<sup>3</sup>Louis van Gaal has been coach of AZ, Ajax, Barcelona, Bayern München, Manchester United, and the Netherlands.

**3.56** Let  $k \geq 1$  be an integer and consider a sequence  $n_1, n_2, \dots, n_k$  of positive integers. Use a combinatorial proof to show that

$$\binom{n_1}{2} + \binom{n_2}{2} + \dots + \binom{n_k}{2} \leq \binom{n_1 + n_2 + \dots + n_k}{2}.$$

*Hint:* For each  $i$  with  $1 \leq i \leq k$ , consider the complete graph on  $n_i$  vertices. How many edges does this graph have?

**3.57** Let  $n \geq 1$  be an integer. Prove that

$$\sum_{k=1}^n \binom{n}{k} \binom{n}{k-1} = \binom{2n}{n+1},$$

by determining, in two different ways, the number of ways to choose  $n+1$  people from a group consisting of  $n$  men and  $n$  women.

**3.58** Let  $n \geq 1$  be an integer. Use Newton's Binomial Theorem (i.e., Theorem 3.6.5) to prove that

$$\sum_{k=1}^n \binom{n}{k} 10^k \cdot 26^{n-k} = 36^n - 26^n. \quad (3.5)$$

In the rest of this exercise, you will give a combinatorial proof of this identity.

Consider passwords consisting of  $n$  characters, each character being a digit or a lowercase letter. A password must contain at least one digit.

- Use the Complement Rule of Section 3.3 to show that the number of passwords is equal to  $36^n - 26^n$ .
- Let  $k$  be an integer with  $1 \leq k \leq n$ . Prove that the number of passwords with exactly  $k$  digits is equal to  $\binom{n}{k} 10^k \cdot 26^{n-k}$ .
- Explain why the above two parts imply the identity in (3.5).

**3.59** Use Newton's Binomial Theorem (i.e., Theorem 3.6.5) to prove that for every integer  $n \geq 1$ ,

$$\sum_{k=0}^n \binom{n}{k} 2^k = 3^n. \quad (3.6)$$

In the rest of this exercise, you will give a combinatorial proof of this identity.

Let  $A = \{1, 2, 3, \dots, n\}$  and  $B = \{a, b, c\}$ . According to Theorem 3.1.2, the number of functions  $f : A \rightarrow B$  is equal to  $3^n$ .

- Consider a fixed integer  $k$  with  $0 \leq k \leq n$  and a fixed subset  $S$  of  $A$  having size  $k$ . Determine the number of functions  $f : A \rightarrow B$  having the property that
  - for all  $x \in S$ ,  $f(x) \in \{a, b\}$ , and
  - for all  $x \in A \setminus S$ ,  $f(x) = c$ .
- Explain why this implies the identity in (3.6).

**3.60** Use Newton's Binomial Theorem (i.e., Theorem 3.6.5) to prove that for every integer  $n \geq 2$ ,

$$\sum_{k=0}^n \binom{n}{k} (n-1)^{n-k} = n^n. \quad (3.7)$$

In the rest of this exercise, you will give a combinatorial proof of this identity.

Consider the set  $A = \{1, 2, \dots, n\}$ . According to Theorem 3.1.2, the number of functions  $f : A \rightarrow A$  is equal to  $n^n$ .

- Consider a fixed integer  $k$  with  $0 \leq k \leq n$  and a fixed subset  $S$  of  $A$  having size  $k$ . Determine the number of functions  $f : A \rightarrow A$  having the property that
  - for all  $x \in S$ ,  $f(x) = x$ , and
  - for all  $x \in A \setminus S$ ,  $f(x) \neq x$ .
- Explain why this implies the identity in (3.7).

**3.61** Let  $n \geq 66$  be an integer and consider the set  $S = \{1, 2, \dots, n\}$ .

- Let  $k$  be an integer with  $66 \leq k \leq n$ . How many 66-element subsets of  $S$  are there whose largest element is equal to  $k$ ?
- Use the result in the first part to prove that

$$\sum_{k=66}^n \binom{k-1}{65} = \binom{n}{66}.$$

**3.62** Let  $a \geq 0$ ,  $b \geq 0$ , and  $n \geq 0$  be integers, and consider the set  $S = \{1, 2, 3, \dots, a+b+n+1\}$ .

- How many subsets of size  $a + b + 1$  does  $S$  have?
- Let  $k$  be an integer with  $0 \leq k \leq n$ . Consider subsets  $T$  of  $S$  such that  $|T| = a + b + 1$  and the  $(a + 1)$ -st smallest element in  $T$  is equal to  $a + k + 1$ . How many such subsets  $T$  are there?
- Use the above results to prove that

$$\sum_{k=0}^n \binom{a+k}{k} \binom{b+n-k}{n-k} = \binom{a+b+n+1}{n}.$$

**3.63** Let  $n \geq 0$  and  $k \geq 0$  be integers.

- How many bitstrings of length  $n + 1$  have exactly  $k + 1$  many 1s?
- Let  $i$  be an integer with  $k \leq i \leq n$ . What is the number of bitstrings of length  $n + 1$  that have exactly  $k + 1$  many 1s and in which the rightmost 1 is at position  $i + 1$ ?
- Use the above two results to prove that

$$\sum_{i=k}^n \binom{i}{k} = \binom{n+1}{k+1}.$$

**3.64** Let  $k$ ,  $m$ , and  $n$  be integers with  $0 \leq k \leq m \leq n$ , and let  $S$  be a set of size  $n$ . Prove that

$$\binom{n}{k} \binom{n-k}{m-k} = \binom{n}{m} \binom{m}{k},$$

by determining, in two different ways, the number of ordered pairs  $(A, B)$  with  $A \subseteq S$ ,  $B \subseteq S$ ,  $A \subseteq B$ ,  $|A| = k$ , and  $|B| = m$ .

**3.65** Let  $m$  and  $n$  be integers with  $0 \leq m \leq n$ , and let  $S$  be a set of size  $n$ . Prove that

$$\sum_{k=m}^n \binom{n}{k} \binom{k}{m} = 2^{n-m} \binom{n}{m},$$

by determining, in two different ways, the number of ordered pairs  $(A, B)$  with  $A \subseteq S$ ,  $|A| = m$ ,  $B \subseteq S$ , and  $A \cap B = \emptyset$ .

*Hint:* The size of  $B$  can be any of the values  $n - m, n - (m + 1), n - (m + 2), \dots, n - n$ . What is the number of pairs  $(A, B)$  having the properties above and for which  $|B| = n - k$ ?

**3.66** Let  $m$  and  $n$  be integers with  $0 \leq m \leq n$ .

- How many bitstrings of length  $n + 1$  have exactly  $m$  many 1s?
- Let  $k$  be an integer with  $0 \leq k \leq m$ . What is the number of bitstrings of length  $n+1$  that have exactly  $m$  many 1s and that start with  $\underbrace{1 \cdots 1}_k 0$ ?
- Use the above two results to prove that

$$\sum_{k=0}^m \binom{n-k}{m-k} = \binom{n+1}{m}.$$

**3.67** Let  $m$  and  $n$  be integers with  $0 \leq m \leq n$ . Use Exercises 3.10, 3.64, and 3.66 to prove that

$$\sum_{k=0}^m \frac{\binom{m}{k}}{\binom{n}{k}} = \frac{n+1}{n+1-m}.$$

**3.68** Let  $n \geq 1$  be an integer. Prove that

$$\sum_{k=1}^n k \binom{n}{k}^2 = n \binom{2n-1}{n-1},$$

by determining, in two different ways, the number of ways a committee can be chosen from a group of  $n$  men and  $n$  women. Such a committee has a woman as the chair and  $n - 1$  other members.

**3.69** Let  $n \geq 2$  be an integer and consider the set  $S = \{1, 2, \dots, n\}$ . An ordered triple  $(A, x, y)$  is called *awesome*, if (i)  $A \subseteq S$ , (ii)  $x \in A$ , and (iii)  $y \in A$ .

- Let  $k$  be an integer with  $1 \leq k \leq n$ . Determine the number of awesome triples  $(A, x, y)$  with  $|A| = k$ .

- Prove that the number of awesome triples  $(A, x, y)$  with  $x = y$  is equal to

$$n \cdot 2^{n-1}.$$

- Determine the number of awesome triples  $(A, x, y)$  with  $x \neq y$ .

- Use the above results to prove that

$$\sum_{k=1}^n k^2 \binom{n}{k} = n(n-1) \cdot 2^{n-2} + n \cdot 2^{n-1}.$$

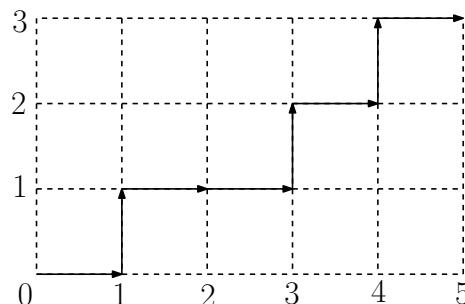
**3.70** Let  $n \geq 1$  be an integer, and let  $X$  and  $Y$  be two disjoint sets, each consisting of  $n$  elements. An ordered triple  $(A, B, C)$  of sets is called *cool*, if

$$A \subseteq X, B \subseteq Y, C \subseteq B, \text{ and } |A| + |B| = n.$$

- Let  $k$  be an integer with  $0 \leq k \leq n$ . Determine the number of cool triples  $(A, B, C)$  for which  $|A| = k$ .
- Let  $k$  be an integer with  $0 \leq k \leq n$ . Determine the number of cool triples  $(A, B, C)$  for which  $|C| = k$ .
- Use the above two results to prove that

$$\sum_{k=0}^n \binom{n}{k}^2 \cdot 2^{n-k} = \sum_{k=0}^n \binom{n}{k} \binom{2n-k}{n}.$$

**3.71** Let  $m \geq 1$  and  $n \geq 1$  be integers. Consider a rectangle whose horizontal side has length  $m$  and whose vertical side has length  $n$ . A path from the bottom-left corner to the top-right corner is called *valid*, if in each step, it either goes one unit to the right or one unit upwards. In the example below, you see a valid path for the case when  $m = 5$  and  $n = 3$ .



How many valid paths are there?

**3.72** Let  $n \geq 1$  be an integer. Prove that

$$\sum_{k=1}^n k \binom{n}{k} = n \cdot 2^{n-1}.$$

*Hint:* Take the derivative of  $(1+x)^n$ .

**3.73** A string consisting of characters is called *cool*, if exactly one character in the string is equal to the letter  $x$  and each other character is a digit. Let  $n \geq 1$  be an integer.

- Determine the number of cool strings of length  $n$ .
- Let  $k$  be an integer with  $1 \leq k \leq n$ . Determine the number of cool strings of length  $n$  that contain exactly  $n-k$  many 0's.
- Use the above two results to prove that

$$\sum_{k=1}^n k \binom{n}{k} 9^{k-1} = n \cdot 10^{n-1}.$$

**3.74** Let  $n \geq 1$  be an integer. We consider binary  $2 \times n$  matrices, i.e., matrices with 2 rows and  $n$  columns, in which each entry is 0 or 1. Any column in such a matrix is of one of four types, based on the bits that occur in this column. We will refer to these types as  $0^0$ -columns,  $0^1$ -columns,  $1^0$ -columns, and  $1^1$ -columns. For example, in the  $2 \times 7$  matrix below, the first, second, and fifth columns are  $0^1$ -columns, the third and seventh columns are  $1^1$ -columns, the fourth column is a  $0^0$ -column, and the sixth column is a  $1^0$ -column.

0	0	1	0	0	1	1
1	1	1	0	1	0	1

For the rest of this exercise, let  $k$  be an integer with  $0 \leq k \leq 2n$ . A binary  $2 \times n$  matrix is called *awesome*, if it contains exactly  $k$  many 0's.

- How many 1's are there in an awesome  $2 \times n$  matrix?
- How many awesome  $2 \times n$  matrices are there?
- Let  $i$  be an integer and consider an arbitrary awesome  $2 \times n$  matrix  $M$  with exactly  $n-i$  many  $1^1$ -columns.

- Prove that  $\lceil k/2 \rceil \leq i \leq k$ .
- Determine the number of  $\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}$ -columns plus the number of  $\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}$ -columns in  $M$ .
- Let  $i$  be an integer. Prove that the number of awesome  $2 \times n$  matrices with exactly  $n - i$  many  $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}$ -columns is equal to

$$2^{2i-k} \binom{n}{n-i} \binom{i}{2i-k}.$$

- Use the above results to prove that

$$\sum_{i=\lceil k/2 \rceil}^k 2^{2i} \binom{n}{i} \binom{i}{k-i} = 2^k \binom{2n}{k}.$$

**3.75** How many different strings can be obtained by reordering the letters of the word **MississippiMills**. (This is a town close to Ottawa. James Naismith, the inventor of basketball, was born there.)

**3.76** In this exercise, we consider strings that can be obtained by reordering the letters of the word **ENGINE**.

- Determine the number of strings that can be obtained.
- Determine the number of strings in which the two letters **E** are next to each other.
- Determine the number of strings in which the two letters **E** are not next to each other and the two letters **N** are not next to each other.

**3.77** Determine the number of elements  $x$  in the set  $\{1, 2, 3, \dots, 99999\}$  for which the sum of the digits in the decimal representation of  $x$  is equal to 8. An example of such an element  $x$  is 3041.

**3.78** In Theorems 3.9.1 and 3.9.2, we have seen how many solutions (in non-negative integers) there are for equations of the type

$$x_1 + x_2 + \cdots + x_k = n$$

and inequalities of the type

$$x_1 + x_2 + \cdots + x_k \leq n.$$

Use this to prove the following identity:

$$\sum_{i=0}^n \binom{i+k-1}{k-1} = \binom{n+k}{k}.$$

**3.79** Let  $n$  and  $k$  be integers with  $n \geq k \geq 1$ . How many solutions are there to the equation

$$x_1 + x_2 + \cdots + x_k = n,$$

where  $x_1 \geq 1, x_2 \geq 1, \dots, x_k \geq 1$  are integers?

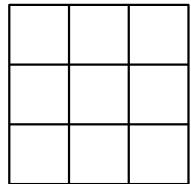
*Hint:* In Theorem 3.9.1, we have seen the answer if  $x_1 \geq 0, x_2 \geq 0, \dots, x_k \geq 0$ .

**3.80** In this exercise, we consider sequences consisting of five digits.

- Determine the number of 5-digit sequences  $d_1d_2d_3d_4d_5$ , whose digits are decreasing, i.e.,  $d_1 > d_2 > d_3 > d_4 > d_5$ .
- Determine the number of 5-digit sequences  $d_1d_2d_3d_4d_5$ , whose digits are non-increasing, i.e.,  $d_1 \geq d_2 \geq d_3 \geq d_4 \geq d_5$ .

*Hint:* Consider the numbers  $x_1 = d_1 - d_2, x_2 = d_2 - d_3, x_3 = d_3 - d_4, x_4 = d_4 - d_5, x_5 = d_5$ . What do you know about  $x_1 + x_2 + x_3 + x_4 + x_5$ ?

**3.81** The square in the left figure below is divided into nine cells. In each cell, we write one of the numbers  $-1, 0$ , and  $1$ .



0	1	0
1	1	-1
-1	0	-1

Use the Pigeonhole Principle to prove that, among the rows, columns, and main diagonals, there exist two that have the same sum. For example, in the right figure above, both main diagonals have sum 0. (Also, the two topmost rows both have sum 1, whereas the bottom row and the right column both have sum  $-2$ .)

**3.82** Let  $S$  be a set consisting of 19 two-digit integers. Thus, each element of  $S$  belongs to the set  $\{10, 11, \dots, 99\}$ .

Use the Pigeonhole Principle to prove that this set  $S$  contains two distinct elements  $x$  and  $y$ , such that the sum of the two digits of  $x$  is equal to the sum of the two digits of  $y$ .

**3.83** Let  $S$  be a set consisting of 9 people. Every person  $x$  in  $S$  has an age  $age(x)$ , which is an integer with  $1 \leq age(x) \leq 60$ .

- Assume that there are two people in  $S$  having the same age. Prove that there exist two distinct subsets  $A$  and  $B$  of  $S$  such that (i) both  $A$  and  $B$  are non-empty, (ii)  $A \cap B = \emptyset$ , and (iii)  $\sum_{x \in A} age(x) = \sum_{x \in B} age(x)$ .
- Assume that all people in  $S$  having different ages. Use the Pigeonhole Principle to prove that there exist two distinct subsets  $A$  and  $B$  of  $S$  such that (i) both  $A$  and  $B$  are non-empty, and (ii)  $\sum_{x \in A} age(x) = \sum_{x \in B} age(x)$ .
- Assume that all people in  $S$  having different ages. Prove that there exist two distinct subsets  $A$  and  $B$  of  $S$  such that (i) both  $A$  and  $B$  are non-empty, (ii)  $A \cap B = \emptyset$ , and (iii)  $\sum_{x \in A} age(x) = \sum_{x \in B} age(x)$ .

**3.84** Let  $n \geq 1$  be an integer. Use the Pigeonhole Principle to prove that in any set of  $n + 1$  integers from  $\{1, 2, \dots, 2n\}$ , there are two elements that are consecutive (i.e., differ by one).

**3.85** Let  $n \geq 1$  be an integer. Use the Pigeonhole Principle to prove that in any set of  $n + 1$  integers from  $\{1, 2, \dots, 2n\}$ , there are two elements whose sum is equal to  $2n + 1$ .

**3.86** Let  $S_1, S_2, \dots, S_{50}$  be a sequence consisting of 50 subsets of the set  $\{1, 2, \dots, 55\}$ . Assume that each of these 50 subsets consists of at least seven elements.

Use the Pigeonhole Principle to prove that there exist two distinct indices  $i$  and  $j$ , such that the largest element in  $S_i$  is equal to the largest element in  $S_j$ .

**3.87** Consider five points in a square with sides of length one. Use the Pigeonhole Principle to prove that there are two of these points having distance at most  $1/\sqrt{2}$ .

**3.88** Let  $S_1, S_2, \dots, S_{26}$  be a sequence consisting of 26 subsets of the set  $\{1, 2, \dots, 9\}$ . Assume that each of these 26 subsets consists of at most three elements. Use the Pigeonhole Principle to prove that there exist two distinct indices  $i$  and  $j$ , such that

$$\sum_{x \in S_i} x = \sum_{x \in S_j} x,$$

i.e., the sum of the elements in  $S_i$  is equal to the sum of the elements in  $S_j$ .

*Hint:* What are the possible values for  $\sum_{x \in S_i} x$ ?

**3.89** Let  $S$  be a set of 90 positive integers, each one having at most 25 digits in decimal notation. Use the Pigeonhole Principle to prove that there are two distinct subsets  $A$  and  $B$  of  $S$  that have the same sum, i.e.,

$$\sum_{x \in A} x = \sum_{x \in B} x.$$

**3.90** Let  $n \geq 2$  be an integer.

- Let  $S$  be a set of  $n + 1$  integers. Prove that  $S$  contains two elements whose difference is divisible by  $n$ .

*Hint:* Use the Pigeonhole Principle.

- Prove that there is an integer that is divisible by  $n$  and whose decimal representation only contains the digits 0 and 5.

*Hint:* Consider the integers 5, 55, 555, 5555, ...

**3.91** In this exercise, we consider the sequence

$$3^0, 3^1, 3^2, \dots, 3^{1000}$$

of integers.

- Prove that this sequence contains two distinct elements whose difference is divisible by 1000. That is, prove that there exist two integers  $m$  and  $n$  with  $0 \leq m < n \leq 1000$ , such that  $3^n - 3^m$  is divisible by 1000.

*Hint:* Consider each element in the sequence modulo 1000 and use the Pigeonhole Principle.

- Use the first part to prove that the sequence

$$3^1, 3^2, \dots, 3^{1000}$$

contains an element whose decimal representation ends with 001. In other words, the last three digits in the decimal representation are 001.

**3.92** Let  $n \geq 2$  be an integer and let  $G = (V, E)$  be a graph whose vertex set  $V$  has size  $n$  and whose edge set  $E$  is non-empty. The degree of any vertex  $u$  is defined to be the number of edges in  $E$  that contain  $u$  as a vertex. Prove that there exist at least two vertices in  $G$  that have the same degree.

*Hint:* Consider the cases when  $G$  is connected and  $G$  is not connected separately. In each case, apply the Pigeonhole Principle. Alternatively, consider a vertex of maximum degree together with its adjacent vertices and, again, apply the Pigeonhole Principle.

**3.93** Let  $d \geq 1$  be an integer. A point  $p$  in  $\mathbb{R}^d$  is represented by its  $d$  real coordinates as  $p = (p_1, p_2, \dots, p_d)$ . The *midpoint* of two points  $p = (p_1, p_2, \dots, p_d)$  and  $q = (q_1, q_2, \dots, q_d)$  is the point

$$\left( \frac{p_1 + q_1}{2}, \frac{p_2 + q_2}{2}, \dots, \frac{p_d + q_d}{2} \right).$$

Let  $P$  be a set of  $2^d + 1$  points in  $\mathbb{R}^d$ , all of which have integer coordinates.

Use the Pigeonhole Principle to prove that this set  $P$  contains two distinct elements whose midpoint has integer coordinates.

*Hint:* The sum of two even integers is even, and the sum of two odd integers is even.

# Chapter 4

## Recursion

In order to understand recursion, you must first understand recursion.

Recursion is the concept where an object (such as a function, a set, or an algorithm) is defined in the following way:

- There are one or more base cases.
- There are one or more rules that define an object in terms of “smaller” objects that have already been defined.

In this chapter, we will see several examples of such recursive definitions and how to use them to solve counting problems.

### 4.1 Recursive Functions

Recall that  $\mathbb{N} = \{0, 1, 2, \dots\}$  denotes the set of natural numbers. Consider the following recursive definition of a function  $f : \mathbb{N} \rightarrow \mathbb{N}$ :

$$\begin{aligned}f(0) &= 3, \\ f(n) &= 2 \cdot f(n-1) + 3, \text{ if } n \geq 1.\end{aligned}$$

These two rules indeed *define* a function, because  $f(0)$  is uniquely defined and for any integer  $n \geq 1$ , if  $f(n-1)$  is uniquely defined, then  $f(n)$  is also uniquely defined, because it is equal to  $2 \cdot f(n-1) + 3$ . Therefore, by induction, for any natural number  $n$ , the function value  $f(n)$  is uniquely defined. We can obtain the values  $f(n)$  in the following way:

- We are given that  $f(0) = 3$ .
- If we apply the recursive rule with  $n = 1$ , then we get

$$f(1) = 2 \cdot f(0) + 3 = 2 \cdot 3 + 3 = 9.$$

- If we apply the recursive rule with  $n = 2$ , then we get

$$f(2) = 2 \cdot f(1) + 3 = 2 \cdot 9 + 3 = 21.$$

- If we apply the recursive rule with  $n = 3$ , then we get

$$f(3) = 2 \cdot f(2) + 3 = 2 \cdot 21 + 3 = 45.$$

- If we apply the recursive rule with  $n = 4$ , then we get

$$f(4) = 2 \cdot f(3) + 3 = 2 \cdot 45 + 3 = 93.$$

Can we “solve” this recurrence relation? That is, can we express  $f(n)$  in terms of  $n$  only? By looking at these values, you may see a pattern, i.e., you may guess that for each  $n \geq 0$ ,

$$f(n) = 3 \cdot 2^{n+1} - 3. \quad (4.1)$$

We prove by induction that this is correct: If  $n = 0$ , then  $f(n) = f(0) = 3$  and  $3 \cdot 2^{n+1} - 3 = 3 \cdot 2^{0+1} - 3 = 3$ . Thus, (4.1) is true for  $n = 0$ . Let  $n \geq 1$  and assume that (4.1) is true for  $n - 1$ , i.e., assume that

$$f(n - 1) = 3 \cdot 2^n - 3.$$

Then

$$\begin{aligned} f(n) &= 2 \cdot f(n - 1) + 3 \\ &= 2(3 \cdot 2^n - 3) + 3 \\ &= 3 \cdot 2^{n+1} - 3. \end{aligned}$$

Thus, we have proved by induction that (4.1) holds for all integers  $n \geq 0$ .

**A recursive definition of factorials:** Consider the following recursive definition of a function  $g : \mathbb{N} \rightarrow \mathbb{N}$ :

$$\begin{aligned} g(0) &= 1, \\ g(n) &= n \cdot g(n-1), \text{ if } n \geq 1. \end{aligned}$$

As in the previous example, a simple induction proof shows that these rules uniquely define the value  $g(n)$  for each  $n \geq 0$ . We leave it to the reader to verify that  $g$  is the factorial function, i.e.,  $g(n) = n!$  for each  $n \geq 0$ .

**A recursive definition of binomial coefficients:** Consider the following recursive definition of a function  $B : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$  with two variables:

$$\begin{aligned} B(n, 0) &= 1, \text{ if } n \geq 0, \\ B(n, n) &= 1, \text{ if } n \geq 0, \\ B(n, k) &= B(n-1, k-1) + B(n-1, k), \text{ if } n \geq 2 \text{ and } 1 \leq k \leq n-1. \end{aligned}$$

The recursive rule has the same form as Pascal's Identity in Theorem 3.7.2. The first base case shows that  $B(n, 0) = 1 = \binom{n}{0}$ , whereas the second base case shows that  $B(n, n) = 1 = \binom{n}{n}$ . From this, it can be shown by induction that  $B(n, k) = \binom{n}{k}$  for all  $n$  and  $k$  with  $0 \leq k \leq n$ .

## 4.2 Fibonacci Numbers

I'll have an order of the Fibonachos.

The *Fibonacci numbers* are defined using the following rules:

$$\begin{aligned} f_0 &= 0, \\ f_1 &= 1, \\ f_n &= f_{n-1} + f_{n-2}, \text{ if } n \geq 2. \end{aligned}$$

In words, there are two base cases (i.e., 0 and 1) and each next element in the sequence is the sum of the previous two elements. This gives the sequence

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \dots$$

The following theorem states that we can “solve” this recurrence relation. That is, we can express the  $n$ -th Fibonacci number  $f_n$  in a non-recursive way, i.e., without using any other Fibonacci numbers.

**Theorem 4.2.1** Let  $\varphi = \frac{1+\sqrt{5}}{2}$  and  $\psi = \frac{1-\sqrt{5}}{2}$  be the two solutions of the quadratic equation  $x^2 = x + 1$ . Then, for all  $n \geq 0$ , we have

$$f_n = \frac{\varphi^n - \psi^n}{\sqrt{5}}.$$

**Proof.** We prove the claim by induction on  $n$ . There are two base cases<sup>1</sup>:

- Both  $f_0$  and  $\frac{\varphi^0 - \psi^0}{\sqrt{5}}$  are equal to 0.
- Both  $f_1$  and  $\frac{\varphi^1 - \psi^1}{\sqrt{5}}$  are equal to 1.

Let  $n \geq 2$  and assume that the claim is true for  $n - 2$  and  $n - 1$ . In other words, assume that

$$f_{n-2} = \frac{\varphi^{n-2} - \psi^{n-2}}{\sqrt{5}}$$

and

$$f_{n-1} = \frac{\varphi^{n-1} - \psi^{n-1}}{\sqrt{5}}.$$

We have to prove that the claim is true for  $n$  as well. Using the definition of  $f_n$ , the two assumptions, and the identities  $\varphi^2 = \varphi + 1$  and  $\psi^2 = \psi + 1$ , we get

$$\begin{aligned} f_n &= f_{n-1} + f_{n-2} \\ &= \frac{\varphi^{n-1} - \psi^{n-1}}{\sqrt{5}} + \frac{\varphi^{n-2} - \psi^{n-2}}{\sqrt{5}} \\ &= \frac{\varphi^{n-2}(\varphi + 1)}{\sqrt{5}} - \frac{\psi^{n-2}(\psi + 1)}{\sqrt{5}} \\ &= \frac{\varphi^{n-2} \cdot \varphi^2}{\sqrt{5}} - \frac{\psi^{n-2} \cdot \psi^2}{\sqrt{5}} \\ &= \frac{\varphi^n - \psi^n}{\sqrt{5}}. \end{aligned}$$

■

---

<sup>1</sup>Do you see why there are two base cases?

### 4.2.1 Counting 00-Free Bitstrings

A bitstring is called *00-free*, if it does not contain two 0's next to each other. Examples of 00-free bitstrings are 10, 010, 01010101, and 1111111. On the other hand, neither of the two bitstrings 101001 and 0100011 is 00-free.

For any integer  $n \geq 1$ , what is the number of 00-free bitstrings having length  $n$ ? Since we do not know the answer yet, we introduce a variable  $B_n$ , one for each  $n \geq 1$ , for the number of such strings. Thus,

- $B_n$  denotes the number of 00-free bitstrings of length  $n$ .

Let us start by determining  $B_n$  for some small values of  $n$ . There are two bitstrings of length 1:

$$0, 1.$$

Since neither of them contains 00, we have  $B_1 = 2$ . There are four bitstrings of length 2:

$$00, 10, 01, 11.$$

Since three of them do not contain 00, we have  $B_2 = 3$ . Similarly, there are eight bitstrings of length 3:

$$000, 001, 010, 100, 011, 101, 110, 111.$$

Since five of them do not contain 00, we have  $B_3 = 5$ .

Let  $n \geq 3$ . We are going to express  $B_n$  in terms of the previous two values  $B_{n-1}$  and  $B_{n-2}$ . This, together with the two base cases  $B_1 = 2$  and  $B_2 = 3$ , will give a recurrence relation for the entire sequence.

Consider a matrix that contains all 00-free bitstrings of length  $n$ , one string per row. Since the number of such strings is equal to  $B_n$ , the matrix has  $B_n$  rows. Also, the matrix has  $n$  columns, because the strings have length  $n$ .

We rearrange the rows of the matrix such that all strings in the *top part* start with 1 and all strings in the *bottom part* start with 0.

- How many rows are there in the top part? Any string in the top part starts with 1 and is followed by a bitstring of length  $n - 1$  that does not contain 00. Thus, if we take the rows in the top part and delete the first bit from each row, then we obtain all 00-free bitstrings of length  $n - 1$ . Since the number of 00-free bitstrings of length  $n - 1$  is equal to  $B_{n-1}$ , it follows that the top part of the matrix consists of  $B_{n-1}$  rows.

- How many rows are there in the bottom part? Any string in the bottom part starts with 0. Since the string does not contain 00, the second bit must be 1. After these first two bits, we have a bitstring of length  $n - 2$  that does not contain 00. Thus, if we take the rows in the bottom part and delete the first two bits from each row, then we obtain all 00-free bitstrings of length  $n - 2$ . Since the number of 00-free bitstrings of length  $n - 2$  is equal to  $B_{n-2}$ , it follows that the bottom part of the matrix consists of  $B_{n-2}$  rows.

Thus, on the one hand, the matrix has  $B_n$  rows. On the other hand, this matrix has  $B_{n-1} + B_{n-2}$  rows. Therefore, we have  $B_n = B_{n-1} + B_{n-2}$ .

To summarize, we have proved that the values  $B_n$ , for  $n \geq 1$ , satisfy the following recurrence relation:

$$\begin{aligned} B_1 &= 2, \\ B_2 &= 3, \\ B_n &= B_{n-1} + B_{n-2}, \text{ if } n \geq 3. \end{aligned}$$

This recurrence relation is the same as the one for the Fibonacci numbers, except that the two base cases are different. The sequence  $B_n$ ,  $n \geq 1$ , consists of the integers

$$2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \dots$$

We obtain this sequence by removing the first three elements (i.e.,  $f_0$ ,  $f_1$ , and  $f_2$ ) from the Fibonacci sequence. We leave it to the reader to verify (using induction) that for all  $n \geq 1$ ,

$$B_n = f_{n+2}.$$

### 4.3 A Recursively Defined Set

Consider the set  $S$  that is defined by the following two rules:

- 5 is an element of the set  $S$ .
- If  $x$  and  $y$  are elements of the set  $S$ , then  $x - y$  is also an element of the set  $S$ .

Thus, if we already know that  $x$  and  $y$  belong to the set  $S$ , then the second rule gives us a new element, i.e.,  $x - y$ , that also belongs to  $S$ .

Can we give a simple description of the set  $S$ ? We are going to use the rules to obtain some elements of  $S$ . From these examples, we then hope to see a pattern from which we guess the simple description of  $S$ . The final step consists of proving that our guess is correct.

- We are given that 5 is an element of  $S$ .
- Applying the rule with  $x = 5$  and  $y = 5$  implies that  $x - y = 0$  is also an element of  $S$ .
- Applying the rule with  $x = 0$  and  $y = 5$  implies that  $x - y = -5$  is also an element of  $S$ .
- Applying the rule with  $x = 5$  and  $y = -5$  implies that  $x - y = 10$  is also an element of  $S$ .
- Applying the rule with  $x = 0$  and  $y = 10$  implies that  $x - y = -10$  is also an element of  $S$ .
- Applying the rule with  $x = 5$  and  $y = -10$  implies that  $x - y = 15$  is also an element of  $S$ .
- Applying the rule with  $x = 0$  and  $y = 15$  implies that  $x - y = -15$  is also an element of  $S$ .

Thus, we have obtained the following elements of  $S$ :

$$-15, -10, -5, 0, 5, 10, 15$$

Since there is clearly a pattern, it is natural to guess that

$$S = \{5n : n \in \mathbb{Z}\}, \quad (4.2)$$

where  $\mathbb{Z}$  is the set of all (positive and negative) integers, including 0. To prove that this is correct, we will first prove that the set on the left-hand side is a subset of the set on the right-hand side. Then we prove that the set on the right-hand side is a subset of the set on the left-hand side.

We start by proving that

$$S \subseteq \{5n : n \in \mathbb{Z}\},$$

which is equivalent to proving that

$$\text{every element of } S \text{ is a multiple of 5.} \quad (4.3)$$

How do we prove this? The set  $S$  is defined using a base case and a recursive rule. The only way to obtain an element of  $S$  is by starting with the base case and then applying the recursive rule a finite number of times. Therefore, the following will prove that (4.3) holds:

- The element in the base case, i.e., 5, is a multiple of 5.
- Let  $x$  and  $y$  be two elements of  $S$  and assume that they are both multiples of 5. Then  $x - y$  (which is the “next” element of  $S$ ) is also a multiple of 5.

Next we prove that

$$\{5n : n \in \mathbb{Z}\} \subseteq S.$$

We will do this by proving that for all  $n \geq 0$ ,

$$5n \in S \text{ and } -5n \in S. \quad (4.4)$$

The proof is by induction on  $n$ . For the base case, i.e., when  $n = 0$ , we observe that, from the definition of  $S$ ,  $x = 5$  and  $y = 5$  are in  $S$  and, therefore,  $x - y = 0$  is also in  $S$ . Therefore, (4.4) is true for  $n = 0$ .

Let  $n \geq 0$  and assume that (4.4) is true for  $n$ , i.e., assume that

$$5n \in S \text{ and } -5n \in S.$$

We have to show that (4.4) is also true for  $n + 1$ , i.e.,

$$5(n+1) \in S \text{ and } -5(n+1) \in S.$$

- It follows from the definition of  $S$  and our assumption that both  $x = 5$  and  $y = -5n$  are in  $S$ . Therefore,  $x - y = 5(n+1)$  is also in  $S$ .
- It follows from the definition of  $S$  and our assumption that both  $x = -5n$  and  $y = 5$  are in  $S$ . Therefore,  $x - y = -5(n+1)$  is also in  $S$ .

Thus, we have shown by induction that (4.4) holds for all  $n \geq 0$ .

Since we have shown that both (4.3) and (4.4) hold, we conclude that (4.2) holds as well. In other words, we have indeed obtained a simple description of the set  $S$ : It is the set of all multiples of 5.

## 4.4 A Gossip Problem

Let  $n \geq 4$  be an integer and consider a group  $P_1, P_2, \dots, P_n$  of  $n$  people. Assume that each person  $P_i$  knows some scandal  $S_i$  that nobody else knows. For any  $i$  and  $j$ , if person  $P_i$  makes a phone call with person  $P_j$ , they exchange the scandals they know at that moment, i.e.,  $P_i$  tells all scandals she knows to  $P_j$ , and  $P_j$  tells all scandals he knows to  $P_i$ . How many phone calls are needed until each of the  $n$  people knows all  $n$  scandals?

An obvious solution is that each pair of people in the group makes one phone call. At the end, each person knows all scandals. The number of phone calls is

$$\binom{n}{2} = \frac{n(n-1)}{2},$$

which is quadratic in the number  $n$  of people. We will see below that only a linear number of phone calls are needed.

Let us first consider the case when  $n = 4$ . At the start, each person  $P_i$  only knows the scandal  $S_i$ , which we visualize in the following table:

$P_1$	$P_2$	$P_3$	$P_4$
$S_1$	$S_2$	$S_3$	$S_4$

Consider the following sequence of phone calls:

1.  $P_1$  calls  $P_2$ . After this phone call, the table looks as follows:

$P_1$	$P_2$	$P_3$	$P_4$
$S_1S_2$	$S_1S_2$	$S_3$	$S_4$

2.  $P_3$  calls  $P_4$ . After this phone call, the table looks as follows:

$P_1$	$P_2$	$P_3$	$P_4$
$S_1S_2$	$S_1S_2$	$S_3S_4$	$S_3S_4$

3.  $P_1$  calls  $P_3$ . After this phone call, the table looks as follows:

$P_1$	$P_2$	$P_3$	$P_4$
$S_1S_2S_3S_4$	$S_1S_2$	$S_1S_2S_3S_4$	$S_3S_4$

4.  $P_2$  calls  $P_4$ . After this phone call, the table looks as follows:

$P_1$	$P_2$	$P_3$	$P_4$
$S_1S_2S_3S_4$	$S_1S_2S_3S_4$	$S_1S_2S_3S_4$	$S_1S_2S_3S_4$

We see that after four phone calls, each person knows all four scandals. Observe that the number of phone calls is  $\binom{4}{2} = 6$  if we would have used the obvious solution mentioned above.

We now have an algorithm that schedules the phone calls for groups of four people. Below, we will extend this “base case” to a recursive algorithm that schedules the phone calls for any group of  $n \geq 4$  people. The approach is as follows:

- We assume that we know how to schedule the phone calls for groups of  $n - 1$  people.
- We use this assumption to schedule the phone calls for groups of  $n$  people.

Let us see how this is done.

- At the start,  $P_1$  knows  $S_1$ ,  $P_2$  knows  $S_2$ ,  $\dots$ ,  $P_n$  knows  $S_n$ .
- $P_{n-1}$  calls  $P_n$ . After this phone call,  $P_1$  knows  $S_1$ ,  $P_2$  knows  $S_2$ ,  $\dots$ ,  $P_{n-2}$  knows  $S_{n-2}$ , and both  $P_{n-1}$  and  $P_n$  know  $S_{n-1}$  and  $S_n$ .
- Consider  $S_{n-1}$  and  $S_n$  to be one scandal  $S'_{n-1}$ .
- Schedule the phone calls for the group  $P_1, P_2, \dots, P_{n-1}$  of  $n - 1$  people, using the scandals  $S_1, S_2, \dots, S_{n-2}, S'_{n-1}$ . (We have assumed that we know how to do this!) At the end, each of  $P_1, P_2, \dots, P_{n-1}$  knows all scandals  $S_1, S_2, \dots, S_n$ .
- At this moment,  $P_n$  only knows  $S_{n-1}$  and  $S_n$ . Therefore,  $P_{n-1}$  again calls  $P_n$  and tells her all scandals  $S_1, S_2, \dots, S_n$ ; the first  $n - 2$  of these are new to  $P_n$ .

Below, you see this recursive algorithm in pseudocode.

**Algorithm** GOSSIP( $n$ ):

```
// n ≥ 4, this algorithm schedules phone calls for P1, P2, ..., Pn
if n = 4
    then P1 calls P2;
        P3 calls P4;
        P1 calls P3;
        P2 calls P4
    else Pn-1 calls Pn;
        GOSSIP(n - 1);
        Pn-1 calls Pn
    endif
```

We are now going to determine the number of phone calls made when running algorithm GOSSIP( $n$ ). Since we do not know the answer yet, we introduce a variable  $C(n)$  to denote this number. It follows from the pseudocode that

$$C(4) = 4.$$

Let  $n \geq 5$ . Algorithm GOSSIP( $n$ ) starts and ends with the same phone call:  $P_{n-1}$  calls  $P_n$ . In between, it runs algorithm GOSSIP( $n - 1$ ), during which, by definition,  $C(n - 1)$  phone calls are made. It follows that

$$C(n) = 2 + C(n - 1) \text{ for } n \geq 5.$$

Thus, we have obtained a recurrence relation for the numbers  $C(n)$ . The first few numbers in the sequence are

$$\begin{aligned} C(4) &= 4, \\ C(5) &= 2 + C(4) = 2 + 4 = 6, \\ C(6) &= 2 + C(5) = 2 + 6 = 8, \\ C(7) &= 2 + C(6) = 2 + 8 = 10. \end{aligned}$$

From this, we guess that

$$C(n) = 2n - 4 \text{ for } n \geq 4.$$

We can easily prove by induction that our guess is correct. Indeed, since both  $C(4)$  and  $2 \cdot 4 - 4$  are equal to 4, the claim is true for  $n = 4$ . If  $n \geq 5$  and  $C(n - 1) = 2(n - 1) - 4$ , then

$$C(n) = 2 + C(n - 1) = 2 + (2(n - 1) - 4) = 2n - 4.$$

This shows that  $C(n) = 2n - 4$  for all  $n \geq 4$ .

It can be shown that algorithm GOSSIP is optimal: Any algorithm that schedules phone calls for  $n \geq 4$  people must make at least  $2n - 4$  phone calls.

You may wonder why the base case for algorithm GOSSIP( $n$ ) is when  $n = 4$ . You will find the reason in Exercise 4.52.

## 4.5 Euclid's Algorithm

We might call Euclid's method the granddaddy of all algorithms, because it is the oldest nontrivial algorithm that has survived to the present day.

— Donald E. Knuth, *The Art of Computer Programming*, Vol. 2, 1997

The *greatest common divisor* of two integers  $a \geq 1$  and  $b \geq 1$  is the largest integer that divides both  $a$  and  $b$ . We denote this largest integer by  $\gcd(a, b)$ . For example, the common divisors of 75 and 45 are 1, 3, 5, and 15. Since 15 is the largest among them,  $\gcd(75, 45) = 15$ . Observe that for any integer  $a \geq 1$ ,  $\gcd(a, a) = a$ .

Assume we are given two large integers  $a$  and  $b$ , say  $a = 371,435,805$  and  $b = 137,916,675$ . How can we compute their greatest common divisor? One approach is to determine the prime factorizations of  $a$  and  $b$ :

$$a = 371,435,805 = 3^2 \cdot 5 \cdot 13^4 \cdot 17^2$$

and

$$b = 137,916,675 = 3^4 \cdot 5^2 \cdot 13^3 \cdot 31.$$

From this, we see that

$$\gcd(a, b) = 3^2 \cdot 5 \cdot 13^3 = 98,865.$$

Unfortunately, it is not known how to obtain, by an efficient algorithm, the prime factorization of a very large integer. As a result, this approach to compute the greatest common divisor of two large integers is not good.

Around 300 BC, Euclid published an algorithm that is both very simple and efficient. This algorithm is based on the *modulo operation*, which we introduce first.

### 4.5.1 The Modulo Operation

Let  $a \geq 1$  and  $b \geq 1$  be integers. If we divide  $a$  by  $b$ , then we obtain a *quotient*  $q$  and a *remainder*  $r$ , which are the unique integers that satisfy

$$a = qb + r, q \geq 0, \text{ and } 0 \leq r \leq b - 1.$$

The modulo operation, denoted by  $a \bmod b$ , is the function that maps the pair  $(a, b)$  to the remainder  $r$ . Thus, we will write

$$a \bmod b = r.$$

For example,

- $17 \bmod 5 = 2$ , because  $17 = 3 \cdot 5 + 2$ ,
- $17 \bmod 17 = 0$ , because  $17 = 1 \cdot 17 + 0$ ,
- $17 \bmod 1 = 0$ , because  $17 = 17 \cdot 1 + 0$ ,
- $17 \bmod 19 = 17$ , because  $17 = 0 \cdot 19 + 17$ .

### 4.5.2 The Algorithm

Euclid's algorithm takes as input two positive integers  $a$  and  $b$ , where  $a \geq b$ , and returns  $\gcd(a, b)$ .

The algorithm starts by computing  $a \bmod b$  and stores the result in a variable  $r$ . If  $r = 0$ , then the algorithm returns the value  $b$ . Otherwise, we have  $r \geq 1$ , in which case the algorithm recursively computes the greatest common divisor of  $b$  and  $r$ . The algorithm is presented in pseudocode below.

**Algorithm** EUCLID( $a, b$ ):

```
// a and b are integers with a ≥ b ≥ 1
r = a mod b;
if r = 0
    then return b
else EUCLID(b, r)
    // observe that b > r ≥ 1
endif
```

Let us look at an example. If we run  $\text{EUCLID}(75, 45)$ , then the algorithm computes  $75 \bmod 45$ , which is 30. Then, it runs  $\text{EUCLID}(45, 30)$ , during which the algorithm computes  $45 \bmod 30$ , which is 15. Next, it runs  $\text{EUCLID}(30, 15)$ , during which the algorithm computes  $30 \bmod 15$ , which is 0. At this moment, the algorithm returns 15, which is indeed the greatest common divisor of the input values 75 and 45.

The following lemma is the basis for a proof that algorithm  $\text{EUCLID}(a, b)$  correctly returns  $\gcd(a, b)$  for any input values  $a \geq b \geq 1$ .

**Lemma 4.5.1** *Let  $a$  and  $b$  be integers with  $a \geq b \geq 1$ , and let  $r = a \bmod b$ .*

1. *If  $r = 0$ , then  $\gcd(a, b) = b$ .*
2. *If  $r \geq 1$ , then  $\gcd(a, b) = \gcd(b, r)$ .*

**Proof.** Let  $q$  and  $r$  be the integers that satisfy  $a = qb + r$ ,  $q \geq 1$ , and  $0 \leq r \leq b - 1$ . (Observe that  $q$  cannot be equal to 0, because  $a \geq b$ .)

If  $r = 0$ , then  $a = qb$ . In this case, it is clear that  $\gcd(a, b) = b$ . Assume that  $r \geq 1$ . We claim that the common divisors of  $a$  and  $b$  are the same as the common divisors of  $b$  and  $r$ :

- Let  $d \geq 1$  be an integer that divides both  $a$  and  $b$ . Since  $r = a - qb$ , it follows that  $d$  divides  $r$ . Thus,  $d$  divides both  $b$  and  $r$ .
- Let  $d \geq 1$  be an integer that divides both  $b$  and  $r$ . Since  $a = qb + r$ , it follows that  $d$  divides  $a$ . Thus,  $d$  divides both  $a$  and  $b$ .

Since the two pairs  $a, b$  and  $b, r$  have the same common divisors, their greatest common divisors are equal as well. ■

**Theorem 4.5.2** *For any two integers  $a$  and  $b$  with  $a \geq b \geq 1$ , algorithm  $\text{EUCLID}(a, b)$  returns  $\gcd(a, b)$ .*

**Proof.** If algorithm  $\text{EUCLID}(a, b)$  generates the recursive call  $\text{EUCLID}(b, r)$ , then  $r < b$ . Thus, in each recursive call to  $\text{EUCLID}$ , the second argument decreases. Since this second argument is a positive integer, the algorithm terminates.

We leave it to the reader to use Lemma 4.5.1 to prove that the output of algorithm  $\text{EUCLID}(a, b)$  is  $\gcd(a, b)$ . ■

### 4.5.3 The Running Time

In the beginning of Section 4.5, we mentioned that Euclid's algorithm is efficient. In this section, we will formalize this.

We are going to bound the total number of modulo operations that are performed when running algorithm  $\text{EUCLID}(a, b)$ . This number will be denoted by  $M(a, b)$ .

For example, when running  $\text{EUCLID}(75, 45)$ , the modulo operation is performed three times: The algorithm computes  $75 \bmod 45$ ,  $45 \bmod 30$ , and  $30 \bmod 15$ . Therefore,  $M(75, 45) = 3$ .

Our goal is to prove an upper bound on  $M(a, b)$  in terms of  $a$  and  $b$ . In fact, as we will see, we will obtain an upper bound in terms of  $b$  only, i.e., the upper bound only depends on the *smaller* of the two input values  $a$  and  $b$ .

As a first upper bound, we have seen in the proof of Theorem 4.5.2 that in each recursive call to algorithm  $\text{EUCLID}$ , the second argument decreases. Since in the initial call  $\text{EUCLID}(a, b)$ , this second argument is equal to  $b$ , the number of modulo operations cannot be larger than  $b$ . It follows that, for all integers  $a$  and  $b$  with  $a \geq b \geq 1$ ,

$$M(a, b) \leq b.$$

This gives an upper bound that is *linear* in  $b$ . Below, we will prove a much better upper bound: The value of  $M(a, b)$  is at most *logarithmic* in  $b$ . We will use the Fibonacci numbers of Section 4.2 to obtain this result. Recall that these numbers are defined by

$$\begin{aligned} f_0 &= 0, \\ f_1 &= 1, \\ f_n &= f_{n-1} + f_{n-2}, \text{ if } n \geq 2. \end{aligned}$$

As mentioned above, we are going to prove an upper bound on  $M(a, b)$  in terms of the logarithm of  $b$ . Usually, when analyzing the running time of an algorithm, we consider a given input and derive an *upper bound* on the running time in terms of the input. For algorithm  $\text{EUCLID}(a, b)$ , we use the opposite approach: We fix a value  $m$  for the running time  $M(a, b)$ , and then prove a *lower bound* on both  $a$  and  $b$  in terms of  $m$ . The following lemma makes this precise.

**Lemma 4.5.3** *Let  $a$  and  $b$  be integers with  $a > b \geq 1$ , and let  $m = M(a, b)$ . Then  $a \geq f_{m+2}$  and  $b \geq f_{m+1}$ .*

**Proof.** The proof is by induction on  $m$ . The base case is when  $m = 1$ . Since  $a \geq b + 1 \geq 2 = f_3$  and  $b \geq 1 = f_2$ , the claim in the lemma holds.

For the induction step, assume that  $m \geq 2$ . Consider the integers  $q$  and  $r$  that satisfy  $a = qb + r$ ,  $q \geq 1$ , and  $0 \leq r \leq b - 1$ . Observe that algorithm EUCLID( $a, b$ ) computes the value  $a \bmod b$ , which is equal to  $r$ . Since  $m \geq 2$ , we have  $r \geq 1$  and the total number of modulo operations performed during the recursive call EUCLID( $b, r$ ) is equal to  $m - 1$ . In other words,  $M(b, r) = m - 1$ . Thus, by induction, we have  $b \geq f_{m+1}$  and  $r \geq f_m$ . We observe that

$$a = qb + r \geq b + r \geq f_{m+1} + f_m = f_{m+2}.$$

This completes the induction step. ■

In Theorem 4.2.1, we have seen that the Fibonacci numbers can be expressed in terms of the numbers  $\varphi = \frac{1+\sqrt{5}}{2}$  and  $\psi = \frac{1-\sqrt{5}}{2}$ . You are encouraged to prove, by induction and using the fact that  $\varphi^2 = \varphi + 1$ , that for any integer  $n \geq 2$ ,

$$f_n \geq \varphi^{n-2}. \quad (4.5)$$

**Theorem 4.5.4** *Let  $a$  and  $b$  be integers with  $a \geq b \geq 1$ . Then*

$$M(a, b) \leq 1 + \log_\varphi b,$$

*i.e., the total number of modulo operations performed by algorithm EUCLID( $a, b$ ) is  $O(\log b)$ .*

**Proof.** If  $a = b$ , then  $M(a, b) = 1$  and the claim obviously holds. Assume that  $a > b$ . Let  $m = M(a, b)$ . By Lemma 4.5.3 and (4.5), we have

$$b \geq f_{m+1} \geq \varphi^{m-1}.$$

By taking logarithms with base  $\varphi$ , we conclude that

$$m - 1 \leq \log_\varphi b,$$

i.e.,

$$M(a, b) = m \leq 1 + \log_\varphi b = O(\log b). \quad \blacksquare$$

## 4.6 The Merge-Sort Algorithm

MERGESORT is a recursive sorting algorithm that works as follows. To sort the sequence  $a_1, a_2, \dots, a_n$  of numbers,

- it recursively sorts the sequence  $a_1, a_2, \dots, a_m$ , where  $m = \lfloor n/2 \rfloor$ , and stores the sorted sequence in a list  $L_1$ ,
- it recursively sorts the sequence  $a_{m+1}, a_{m+2}, \dots, a_n$  and stores the sorted sequence in a list  $L_2$ ,
- it merges the two sorted lists  $L_1$  and  $L_2$  into one sorted list.

Below, you see this recursive algorithm in pseudocode.

```
Algorithm MERGESORT( $L, n$ ):
    //  $L$  is a list of  $n \geq 0$  numbers
    if  $n \geq 2$ 
        then  $m = \lfloor n/2 \rfloor$ ;
             $L_1$  = list consisting of the first  $m$  elements of  $L$ ;
             $L_2$  = list consisting of the last  $n - m$  elements of  $L$ ;
             $L_1$  = MERGESORT( $L_1, m$ );
             $L_2$  = MERGESORT( $L_2, n - m$ );
             $L$  = MERGE( $L_1, L_2$ )
    endif;
    return  $L$ 
```

We still have to specify algorithm MERGE( $L_1, L_2$ ). Of course, this algorithm uses the fact that both  $L_1$  and  $L_2$  are sorted lists. The task is to merge them into one sorted list. This is done in the following way. Initialize an empty list  $L$ . (At the end, this list will contain the final sorted sequence.)

- Let  $x$  be the first element of  $L_1$  and let  $y$  be the first element of  $L_2$ .
- If  $x \leq y$ , then remove  $x$  from  $L_1$  and append it to  $L$  (i.e., add  $x$  at the end of  $L$ ).
- Otherwise (i.e., if  $x > y$ ), remove  $y$  from  $L_2$  and append it to  $L$ .

Repeat these steps until one of  $L_1$  and  $L_2$  is empty. If  $L_1$  is empty, then append  $L_2$  to  $L$ . Otherwise, append  $L_1$  to  $L$ . Here is the algorithm in pseudocode:

```
Algorithm MERGE( $L_1, L_2$ ):
    //  $L_1$  and  $L_2$  are sorted lists
     $L$  = empty list;
    while  $L_1$  is not empty and  $L_2$  is not empty
        do  $x$  = first element of  $L_1$ ;
             $y$  = first element of  $L_2$ ;
            if  $x \leq y$ 
                then remove  $x$  from  $L_1$ ;
                    append  $x$  to  $L$ 
                else remove  $y$  from  $L_2$ ;
                    append  $y$  to  $L$ 
                endif
            endwhile;
            if  $L_1$  is empty
                then append  $L_2$  to  $L$ 
            else append  $L_1$  to  $L$ 
            endif;
        return  $L$ 
```

#### 4.6.1 Correctness of Algorithm MERGESORT

I hope you are convinced that the output  $L$  of algorithm  $\text{MERGE}(L_1, L_2)$  is a sorted list that contains all elements of  $L_1$  and  $L_2$  (and no other elements). How do we prove that algorithm  $\text{MERGESORT}(L, n)$  is correct, i.e., correctly sorts the elements in any list  $L$  of  $n$  numbers? Since the algorithm is recursive, we prove this by induction.

The two base cases are when  $n = 0$  or  $n = 1$ . It follows from the pseudocode for  $\text{MERGESORT}(L, n)$  that it simply returns the input list  $L$ , which is obviously sorted.

Let  $n \geq 2$  and assume that for any integer  $k$  with  $0 \leq k < n$  and for any list  $L'$  of  $k$  numbers, algorithm  $\text{MERGESORT}(L', k)$  returns a list containing the elements of  $L'$  in sorted order. Let  $L$  be a list of  $n$  numbers. By going

through the pseudocode for MERGESORT( $L, n$ ), we observe the following:

- The recursive call MERGESORT( $L_1, m$ ) is on a list with less than  $n$  numbers. Therefore, by the induction hypothesis, its output, which is the list  $L_1$ , is sorted.
- The recursive call MERGESORT( $L_2, n - m$ ) is on a list with less than  $n$  numbers. Again by the induction hypothesis, its output, which is the list  $L_2$ , is sorted.
- Algorithm MERGE( $L_1, L_2$ , ) gets as input the two sorted lists  $L_1$  and  $L_2$ , and returns a list  $L$ . Since algorithm MERGE is correct, it then follows that  $L$  is a sorted list.

It follows that the final list  $L$ , which is returned by algorithm MERGESORT, is sorted.

This proves the correctness of algorithm MERGESORT( $L, n$ ) for any integer  $n \geq 0$  and any list  $L$  of  $n$  numbers.

### 4.6.2 Running Time of Algorithm MERGESORT

We now analyze the running time of algorithm MERGESORT. It follows from the pseudocode that, when running this algorithm together with its recursive calls, several calls are made to algorithm MERGE. We are going to count the total number of *comparisons* that are made. That is, we will determine the total number of times that the line “**if**  $x \leq y$ ” in algorithm MERGE is executed when running algorithm MERGESORT( $L, n$ ).

We first observe that the number of comparisons made by algorithm MERGE( $L_1, L_2$ ) is at most  $|L_1| + |L_2|$ .

Let  $n$  be an integer and assume for simplicity that  $n$  is a power of two, i.e.,  $n = 2^k$  for some integer  $k \geq 0$ . We define  $T(n)$  to be the maximum number of comparisons made when running algorithm MERGESORT( $L, n$ ) on any input list  $L$  of  $n$  numbers. Note that we include in  $T(n)$  all comparisons that are made during all calls to MERGE that are part of all recursive calls that are generated when running MERGESORT( $L, n$ ).

Consider a list  $L$  of  $n$  numbers, where  $n$  is a power of two. For  $n = 1$ , it follows from the pseudocode for MERGESORT( $L, n$ ) that

$$T(1) = 0.$$

Assume that  $n \geq 2$  and consider again the pseudocode for MERGESORT( $L, n$ ). Which parts of the algorithm make comparisons between input elements?

- The call MERGESORT( $L_1, m$ ) is a recursive call on a list of  $m = n/2$  numbers. By definition, the total number of comparisons made in this call (together with all its recursive subcalls) is at most  $T(n/2)$ .
- The call MERGESORT( $L_2, n-m$ ) is a recursive call on a list of  $n-m = n/2$  numbers. By definition, the total number of comparisons made in this call (together with all its recursive subcalls) is at most  $T(n/2)$ .
- Finally, algorithm MERGESORT( $L, n$ ) calls the non-recursive algorithm MERGE( $L_1, L_2$ ). We have seen above that the number of comparisons made in this call is at most  $|L_1| + |L_2| = n$ .

By adding the number of comparisons, we get

$$T(n) \leq T(n/2) + T(n/2) + n = 2 \cdot T(n/2) + n.$$

Thus, we obtain the following recurrence relation:

$$\begin{aligned} T(1) &= 0, \\ T(n) &\leq 2 \cdot T(n/2) + n, \text{ if } n \geq 2 \text{ and } n \text{ is a power of 2.} \end{aligned} \quad (4.6)$$

Our goal was to determine  $T(n)$ , but at this moment, we only have a recurrence relation for this function. We will solve this recurrence relation using a technique called *unfolding*:

Recall that we assume that  $n = 2^k$  for some integer  $k \geq 0$ . We furthermore assume that  $n$  is a large integer. We know from (4.6) that

$$T(n) \leq 2 \cdot T(n/2) + n.$$

If we replace  $n$  by  $n/2$  in (4.6), which is a valid thing to do, we get

$$T(n/2) \leq 2 \cdot T(n/2^2) + n/2.$$

By combining these two inequalities, we get

$$\begin{aligned} T(n) &\leq 2 \cdot T(n/2) + n \\ &\leq 2(2 \cdot T(n/2^2) + n/2) + n \\ &= 2^2 \cdot T(n/2^2) + 2n. \end{aligned}$$

Let us repeat this: Replacing  $n$  by  $n/2^2$  in (4.6) gives

$$T(n/2^2) \leq 2 \cdot T(n/2^3) + n/2^2.$$

By substituting this into the inequality for  $T(n)$ , we get

$$\begin{aligned} T(n) &\leq 2^2 \cdot T(n/2^2) + 2n \\ &\leq 2^2 (2 \cdot T(n/2^3) + n/2^2) + 2n \\ &= 2^3 \cdot T(n/2^3) + 3n. \end{aligned}$$

In the next step, we replace  $n$  by  $n/2^3$  in (4.6), which gives

$$T(n/2^3) \leq 2 \cdot T(n/2^4) + n/2^3.$$

By substituting this into the inequality for  $T(n)$ , we get

$$\begin{aligned} T(n) &\leq 2^3 \cdot T(n/2^3) + 3n \\ &\leq 2^3 (2 \cdot T(n/2^4) + n/2^3) + 3n \\ &= 2^4 \cdot T(n/2^4) + 4n. \end{aligned}$$

At this moment, you will see the pattern and, at the end, we get the inequality

$$T(n) \leq 2^k \cdot T(n/2^k) + kn.$$

Since  $n = 2^k$ , we have  $T(n/2^k) = T(1)$ , which is 0 from the base case of the recurrence relation. Also,  $n = 2^k$  implies that  $k = \log n$ . We conclude that

$$T(n) \leq n \cdot T(1) + n \log n = n \log n.$$

We thus have solved the recurrence relation. In case you have doubts about the validity of the unfolding method, we verify by induction that indeed

$$T(n) \leq n \log n, \text{ for any integer } n \text{ that is a power of 2.}$$

The base case is when  $n = 1$ . In this case, we have  $T(1) = 0$  and  $1 \log 1 = 1 \cdot 0 = 0$ . Let  $n \geq 2$  be a power of 2 and assume that

$$T(n/2) \leq (n/2) \log(n/2).$$

From the recurrence relation, we get

$$T(n) \leq 2 \cdot T(n/2) + n.$$

By substituting the induction hypothesis into this inequality, we get

$$\begin{aligned} T(n) &\leq 2 \cdot (n/2) \log(n/2) + n \\ &= n \log(n/2) + n \\ &= n(\log n - \log 2) + n \\ &= n(\log n - 1) + n \\ &= n \log n. \end{aligned}$$

Thus, by induction,  $T(n) \leq n \log n$  for any integer  $n$  that is a power of 2.

Until now, we have only counted the number of comparisons made by algorithm MERGESORT. It follows from the pseudocode that the total running time, i.e., the total number of “elementary” steps, is within a constant factor of the total number of comparisons. Therefore, if  $n$  is a power of 2, the running time of algorithm MERGESORT( $L, n$ ) is  $O(n \log n)$ .

For general values of  $n$ , the recurrence relation for the number of comparisons becomes the following:

$$\begin{aligned} T(n) &= 0, \text{ if } n = 0 \text{ or } n = 1, \\ T(n) &\leq T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + n, \text{ if } n \geq 2. \end{aligned}$$

It can be shown by induction that this recurrence relation solves to  $T(n) = O(n \log n)$ . We have proved the following result:

**Theorem 4.6.1** *For any list  $L$  of  $n$  numbers, the running time of algorithm MERGESORT( $L, n$ ) is  $O(n \log n)$ .*

## 4.7 Computing the Closest Pair

For a long time researchers felt that there might be a quadratic lower bound on the complexity of the closest-pair problem.

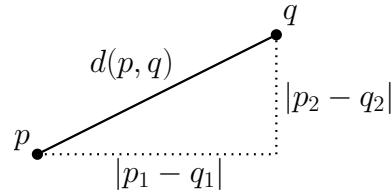
— Jon Louis Bentley,

— *Communications of the ACM*, volume 23, page 226, 1980

If  $p = (p_1, p_2)$  and  $q = (q_1, q_2)$  are two points in  $\mathbb{R}^2$ , then their *distance*  $d(p, q)$  is given by

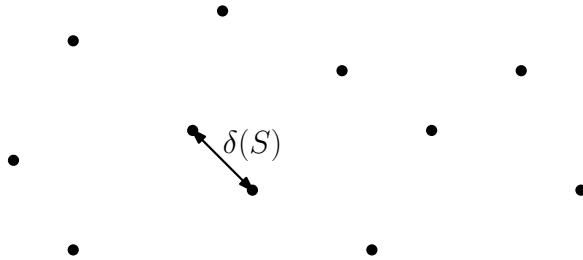
$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

This follows by applying Pythagoras’ Theorem to the right triangle in the following figure.



Let  $S$  be a set of  $n$  points in  $\mathbb{R}^2$ , where  $n \geq 2$  is an integer. The *closest-pair distance* in  $S$ , denoted by  $\delta(S)$ , is the minimum distance between any two distinct points of  $S$ , i.e.,

$$\delta(S) = \min\{d(p, q) : p \in S, q \in S, p \neq q\}.$$



In this section, we consider the problem of designing an efficient algorithm that, when given an arbitrary set  $S$  of  $n$  points in  $\mathbb{R}^2$ , with  $n \geq 2$ , returns the closest-pair distance  $\delta(S)$ .

A trivial algorithm considers all 2-element subsets of  $S$ . For each such subset  $\{p, q\}$ , the algorithm computes the distance  $d(p, q)$ . After all these subsets have been considered, the algorithm returns the smallest distance found. Obviously, the running time of this algorithm is proportional to the number of 2-element subsets of  $S$ , which is

$$\binom{n}{2} = \frac{n(n-1)}{2} = \Theta(n^2).$$

In this section, we will show that the closest pair problem can be solved, by a recursive algorithm, in  $O(n \log n)$  time. In Section 4.7.1, we start by presenting a high-level overview of the basic approach. Then, in Section 4.7.2, we present the details of the recursive algorithm.

### 4.7.1 The Basic Approach

We are given a set  $S$  of  $n$  points in  $\mathbb{R}^2$ , where  $n \geq 2$ . We assume that

- $n$  is a power of two,
- no two points of  $S$  have the same  $x$ -coordinate,
- no two points of  $S$  have the same  $y$ -coordinate.

We remark that neither of these assumptions is necessary. We only make them to simplify the presentation.

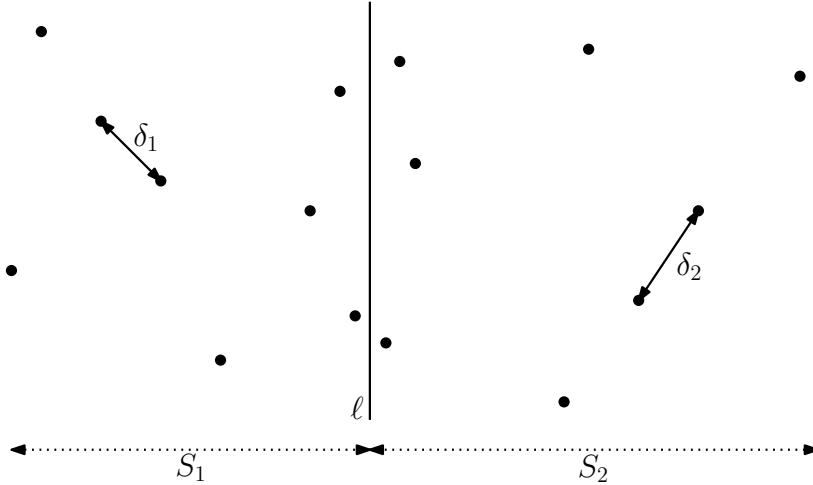
As mentioned above, our algorithm will be recursive. The base case is when  $n = 2$ , i.e., the set  $S$  consists of exactly two points, say  $p$  and  $q$ . In this case, the algorithm simply returns the distance  $d(p, q)$ .

From now on, we assume that  $n \geq 4$ . The algorithm performs the following four steps:

**Step 1:** Let  $\ell$  be a vertical line that splits the set  $S$  into two subsets of equal size. The algorithm computes the set  $S_1$  consisting of all points of  $S$  that are to the left of  $\ell$ , and the set  $S_2$  consisting of all points of  $S$  that are to the right of  $\ell$ . Observe that  $|S_1| = |S_2| = n/2$ .

**Step 2:** The algorithm recursively computes the closest-pair distance  $\delta_1$  in the set  $S_1$ .

**Step 3:** The algorithm recursively computes the closest-pair distance  $\delta_2$  in the set  $S_2$ .



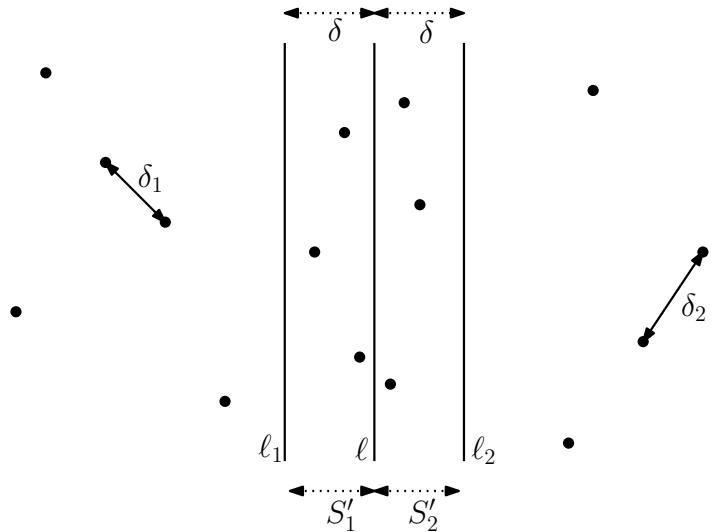
**Step 4:** Let  $\delta = \min(\delta_1, \delta_2)$ . Consider the set

$$A = \{\{p, q\} : p \in S_1, q \in S_2, d(p, q) < \delta\}.$$

- If  $A = \emptyset$ , then the algorithm returns the value of  $\delta$ .
- Assume that  $A \neq \emptyset$ . The algorithm considers all pairs  $\{p, q\} \in A$  and computes the distances  $d(p, q)$ . Let  $\delta_{1,2}$  be the smallest distance found after all these pairs have been considered. Then, the algorithm returns the value of  $\delta_{1,2}$ .

It should be clear that this algorithm correctly returns the closest-pair distance  $\delta(S)$  in the point set  $S$ . What is not clear, however, is how to efficiently perform the last step that involves the set  $A$ . For this, we have to answer two questions: First, how do we efficiently obtain all pairs  $\{p, q\}$  that belong to the set  $A$ ? Second, is there a “small” upper bound on the size of this set  $A$ ?

Let  $\ell_1$  be the vertical line that is at distance  $\delta$  to the left of  $\ell$ , and let  $S'_1$  be the set of all points in  $S_1$  that are between  $\ell_1$  and  $\ell$ . Similarly, let  $\ell_2$  be the vertical line that is at distance  $\delta$  to the right of  $\ell$ , and let  $S'_2$  be the set of all points in  $S_2$  that are between  $\ell$  and  $\ell_2$ . Refer to the figure below for an illustration.



Any point that is on or to the left of  $\ell_1$  has distance at least  $\delta$  to any point that is on or to the right of  $\ell$ . Similarly, any point that is on or to the left of  $\ell$  has distance at least  $\delta$  to any point that is on or to the right of  $\ell_2$ . This implies that the set  $A$  in Step 4 of the algorithm satisfies

$$A = \{\{p, q\} : p \in S'_1, q \in S'_2, d(p, q) < \delta\}. \quad (4.7)$$

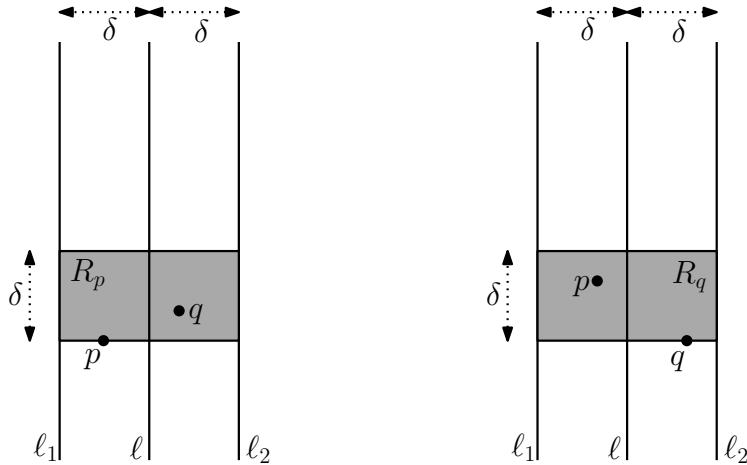
Unfortunately, even using this alternative characterization of the set  $A$ , it is not clear how to obtain all elements of this set in an efficient way. Below, we will define a superset of  $A$ , i.e., a set  $C$  of ordered pairs  $(r, s)$ , with  $r \in S'_1 \cup S'_2$  and  $s \in S'_1 \cup S'_2$ , that contains<sup>2</sup> all elements of  $A$ . As we will see, the size of this new set  $C$  is  $O(n)$  and its elements can be obtained in an efficient way. As a result, the algorithm will use this new set  $C$  in Step 4, instead of  $A$ . If  $C \neq \emptyset$ , let  $\delta'_{1,2}$  be the smallest distance of any pair  $(r, s)$  in  $C$ . The algorithm will return the value

$$\min(\delta, \delta'_{1,2}).$$

Note that, since  $A \subseteq C$ , the algorithm, with the revised Step 4, still correctly returns the closest-pair distance in the set  $S$ .

Before we define the new set  $C$ , we introduce a preliminary set  $B$  that is a superset of  $A$ , i.e.,  $A \subseteq B$ . We will use this set  $B$  to define the set  $C$  that we are looking for. This set  $C$  will satisfy  $B \subseteq C$  and, thus,  $A \subseteq C$ .

We introduce the following notation, which is illustrated in the figure below. Let  $r$  be any point that is between the two lines  $\ell_1$  and  $\ell_2$ . We denote by  $R_r$  the rectangle that has  $r$  on its bottom side, whose left side is on  $\ell_1$ , whose right side is on  $\ell_2$ , and whose height is equal to  $\delta$ . Observe that the width of  $R_r$  is equal to  $2\delta$ .



Consider the set

$$B = B_1 \cup B_2,$$

---

<sup>2</sup>Even though  $A$  consists of unordered pairs and  $C$  consists of ordered pairs, we will cheat a bit and say that  $C$  contains  $A$ .

where

$$B_1 = \{(p, q) : p \in S'_1, q \in S'_2, q \in R_p\}$$

and

$$B_2 = \{(q, p) : p \in S'_1, q \in S'_2, p \in R_q\}.$$

**Lemma 4.7.1** *The set  $B$  is a superset of the set  $A$ , i.e.,  $A \subseteq B$ .*

**Proof.** We have to show that every element of the set  $A$  belongs (as an ordered pair) to the set  $B$ . To prove this, consider an arbitrary element  $\{p, q\}$  of  $A$ . We will show that one of the ordered pairs  $(p, q)$  and  $(q, p)$  is an element of the set  $B$ .

It follows from (4.7) that  $p \in S'_1$  and  $q \in S'_2$ . Thus, to prove that one of  $(p, q)$  and  $(q, p)$  is an element of  $B$ , it remains to be shown that

$$q \in R_p \text{ or } p \in R_q. \quad (4.8)$$

Since  $\{p, q\} \in A$ , we have  $d(p, q) < \delta$ . This implies that the vertical distance between  $p$  and  $q$  is less than  $\delta$ . That is, if  $p = (p_1, p_2)$  and  $q = (q_1, q_2)$ , then  $|p_2 - q_2| < \delta$ .

If  $p_2 < q_2$ , then the point  $q$  is contained in the rectangle  $R_p$  and, therefore, (4.8) holds. Otherwise,  $p_2 > q_2$ , in which case the point  $p$  is contained in the rectangle  $R_q$  and, thus, (4.8) also holds. ■

Is there a non-trivial upper bound on the size of the set  $B$ ? Since each of the two sets  $S'_1$  and  $S'_2$  can have  $n/2$  elements, it is clear that  $|B| \leq n/2 \cdot n/2 = n^2/4$ . In words, the size of  $B$  is *at most* quadratic in  $n$ . The following lemma states that the size of  $B$  is, in fact, at most linear in  $n$ :

**Lemma 4.7.2** *The size of the set  $B$  is at most  $4n$ .*

**Proof.** Let  $p$  be an arbitrary point in  $S'_1$ . We claim that there are at most four points  $q$  such that  $(p, q) \in B_1$ . We will prove this claim by contradiction. Thus, assume that there are at least five such points  $q$ . Observe that for any such point  $q$ , we have  $q \in S'_2$  and  $q \in R_p$ . Therefore, all these points  $q$  are contained in the part of  $R_p$  that is to the right of the line  $\ell$ . This part is a square with sides of length  $\delta$ . By Exercise 3.87, there are two of these points that have distance at most

$$\delta/\sqrt{2} < \delta.$$

Thus, the set  $S'_2$  contains two points having distance less than  $\delta$ . That is, the closest-pair distance in the set  $S_2$  is less than  $\delta$ .

On the other hand, recall that  $\delta = \min(\delta_1, \delta_2)$  and  $\delta_2$  is the closest-pair distance of the set  $S_2$ . It follows that all distances in the set  $S_2$  are at least equal to  $\delta$ . This is a contradiction.

Thus, we have shown that, for this fixed point  $p$  in  $S'_1$ , there are at most four points  $q$  such that  $(p, q) \in B_1$ . Therefore,

$$|B_1| \leq 4|S'_1| \leq 4|S_1| = 4 \cdot n/2 = 2n.$$

By a symmetric argument, for any fixed point  $q$  in  $S'_2$ , there are at most four points  $p$  such that  $(q, p) \in B_2$ . This implies that the set  $B_2$  contains at most  $2n$  elements. We conclude that

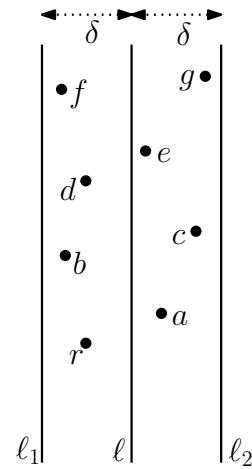
$$|B| = |B_1| + |B_2| \leq 2n + 2n = 4n.$$

■

We are now ready to define the set  $C$  that we are looking for. Let

$$S'_{1,2} = S'_1 \cup S'_2.$$

Imagine that we have the points of this set  $S'_{1,2}$  in increasing order of their  $y$ -coordinates. Consider an arbitrary point  $r$  of  $S'_{1,2}$ . The seven  $y$ -successors of  $r$  are the seven points of  $S'_{1,2}$  that immediately follow  $r$  in this increasing order. In the figure below, these are the points  $a, b, \dots, g$ .



Observe that the number of points that follow  $r$  may be less than seven. In this case, we abuse our terminology a bit and still talk about the seven  $y$ -successors of  $r$ , even though there are fewer of them.

Our final set  $C$  is defined as follows:

$$C = \{(r, s) : r, s \in S'_1 \cup S'_2, s \text{ is one of the seven } y\text{-successors of } r\}. \quad (4.9)$$

**Lemma 4.7.3** *The set  $C$  is a superset of the set  $A$ , i.e.,  $A \subseteq C$ .*

**Proof.** We will prove that  $B \subseteq C$ . It will then follow from Lemma 4.7.1 that  $A \subseteq C$ .

Let  $(p, q)$  be an arbitrary element in the set  $B_1$ . It follows from the definition of  $B_1$  that  $p \in S'_1$ ,  $q \in S'_2$ , and  $q \in R_p$ . To prove that  $(p, q)$  is an element of the set  $C$ , we have to argue that  $q$  is one of the seven  $y$ -successors of  $p$ .

As in the proof of Lemma 4.7.2, (i) the part of  $R_p$  that is to the left of the line  $\ell$  contains at most four points of  $S'_1$  and (ii) the part of  $R_p$  that is to the right of  $\ell$  contains at most four points of  $S'_2$ . Thus, the rectangle  $R_p$  contains at most eight points of  $S'_1 \cup S'_2$ . Since  $p$  is one of them and  $p$  is on the bottom side of  $R_p$ , the point  $q$  must be one of the seven  $y$ -successors of  $p$ .

Thus, we have shown that  $B_1 \subseteq C$ . By a symmetric argument,  $B_2 \subseteq C$ .

Consider the elements  $(r, s)$  of the set  $C$ . There are at most  $n$  choices for the point  $r$ . For each choice of  $r$ , there are at most seven choices for the point  $s$ . This proves the following lemma:

**Lemma 4.7.4** *The size of the set  $C$  is at most  $7n$ .*

## 4.7.2 The Recursive Algorithm

Consider a set of  $n$  points in  $\mathbb{R}^2$ . We make the same assumptions as in Section 4.7.1. Thus,  $n \geq 2$ ,  $n$  is a power of two, no two points have the same  $x$ -coordinate, and no two points have the same  $y$ -coordinate. Our goal is to compute the closest-pair distance in this point set. The base case, i.e., when  $n = 2$ , is easy. Assume that  $n \geq 4$ . In Section 4.7.1, we have seen that the algorithm will make the following steps:

**Step 1:** Determine a vertical line  $\ell$  that splits the point set into two subsets, each having size  $n/2$ . This step is easy to perform if we have the points in sorted order of their  $x$ -coordinates.

**Steps 2 and 3:** Run the algorithm recursively, once for all points to the left of  $\ell$ , and once for all points to the right of  $\ell$ .

**Step 4:** Compute and traverse the set  $C$  that is defined in (4.9). This step is easy to perform if we have the points in sorted order of their  $y$ -coordinates.

We assume that the set of input points is stored in a list  $L$ . The entire algorithm, which we denote by  $\text{CLOSESTPAIR}(L, n)$ , is given in Figure 4.1. In the pseudocode,  $\text{MERGE}(\cdot, \cdot, y)$  refers to the merge algorithm of Section 4.6 that merges two lists, based on the  $y$ -coordinates of the points.

- The input to the call  $\text{CLOSESTPAIR}(L, n)$  is a list  $L$  that stores  $n$  points in  $\mathbb{R}^2$ , where  $n \geq 2$  and  $n$  is a power of two. This list stores the points in increasing order of their  $x$ -coordinates.
- The call  $\text{CLOSESTPAIR}(L, n)$  returns the closest-pair distance between any two distinct points that are stored in  $L$ .
- At termination, the list  $L$  stores the same points, but in sorted order of their  $y$ -coordinates.

The algorithm starts by checking if it is in the base case. Clearly, this base case is easy to handle. Assume that the algorithm is not in the base case, i.e.,  $n \geq 4$ .

- Since  $L$  stores the input points in sorted order of their  $x$ -coordinates, the algorithm obtains the lists  $L_1$  and  $L_2$  by a simple traversal of  $L$ . Observe that, at this moment, the points in both lists  $L_1$  and  $L_2$  are sorted by their  $x$ -coordinates. The value  $z$  that is chosen by the algorithm is the  $x$ -coordinate of the vertical line  $\ell$ .
- In the first recursive call  $\text{CLOSESTPAIR}(L_1, n/2)$ , the algorithm recursively computes the closest-pair distance  $\delta_1$  in  $L_1$ , whereas in the second recursive call  $\text{CLOSESTPAIR}(L_2, n/2)$ , it computes the closest-pair distance  $\delta_2$  in  $L_2$ . After these two recursive calls have terminated, the points in both lists  $L_1$  and  $L_2$  are sorted by their  $y$ -coordinates.

**Algorithm CLOSESTPAIR( $L, n$ ):**

```

if  $n = 2$ 
then  $\delta$  = the distance between the two points in  $L$ ;
    sort the points in  $L$  by their  $y$ -coordinates;
    return  $\delta$ 
else  $L_1$  = list consisting of the first  $n/2$  points in  $L$ ;
     $L_2$  = list consisting of the last  $n/2$  points in  $L$ ;
     $z$  = any value between the  $x$ -coordinates of the last point
        of  $L_1$  and the first point of  $L_2$ ;
    // both  $L_1$  and  $L_2$  are sorted by  $x$ -coordinate
     $\delta_1$  = CLOSESTPAIR( $L_1, n/2$ );
     $\delta_2$  = CLOSESTPAIR( $L_2, n/2$ );
    // both  $L_1$  and  $L_2$  are sorted by  $y$ -coordinate
     $\delta$  =  $\min(\delta_1, \delta_2)$ ;
     $L'_1$  = list consisting of all points  $p$  of  $L_1$  with  $p_1 > z - \delta$ ;
     $L'_2$  = list consisting of all points  $q$  of  $L_2$  with  $q_1 < z + \delta$ ;
    // both  $L'_1$  and  $L'_2$  are sorted by  $y$ -coordinate
     $L'_{1,2}$  = MERGE( $L'_1, L'_2, y$ );
     $L$  = MERGE( $L_1, L_2, y$ );
    // both  $L'_{1,2}$  and  $L$  are sorted by  $y$ -coordinate
    if  $L'_{1,2}$  is empty
    then return  $\delta$ 
    else  $\delta'_{1,2} = \min\{d(r, s) : r, s \in L'_{1,2}, s$  is one of the seven
         $y$ -successors of  $r\}$ ;
        return  $\min(\delta, \delta'_{1,2})$ 
    endif
endif;

```

Figure 4.1: The recursive closest pair algorithm.

- By simple traversals of the lists  $L_1$  and  $L_2$ , the algorithm computes the lists  $L'_1$  and  $L'_2$ . Observe that  $L'_1$  stores all points of  $L_1$  that are to the right of the vertical line  $\ell_1$  that is at distance  $\delta$  to the left of  $\ell$ . Similarly,  $L'_2$  stores all points of  $L_2$  that are to the left of the vertical line  $\ell_2$  that is at distance  $\delta$  to the right of  $\ell$ .

Since both lists  $L'_1$  and  $L'_2$  are in sorted  $y$ -order, the algorithm can use algorithm  $\text{MERGE}(L'_1, L'_2, y)$  to merge these two lists into one list  $L'_{1,2}$  that is also in sorted  $y$ -order. Similarly, the algorithm can run  $\text{MERGE}(L_1, L_2, y)$  to merge the two lists  $L_1$  and  $L_2$  into one list  $L$  that is in sorted  $y$ -order.

In the final step, if the list  $L'_{1,2}$  is non-empty, the algorithm computes the value of  $\delta'_{1,2}$  using a nested for-loop: The outer-loop iterates over all points  $r$  of  $L'_{1,2}$ . For each such point  $r$ , the inner-loop iterates over the seven successors of  $r$  in the list  $L'_{1,2}$ .

Note that the input list  $L$  must contain the points in sorted order of their  $x$ -coordinates. Therefore, before the first call to  $\text{CLOSESTPAIR}$ , we run algorithm  $\text{MERGESORT}(L, n)$  of Section 4.6 to sort the input points by their  $x$ -coordinates. By Theorem 4.6.1, this takes  $O(n \log n)$  time.

We now analyze the running time of algorithm  $\text{CLOSESTPAIR}(L, n)$ . Let  $T(n)$  denote the worst-case running time of this algorithm, when given as input a list of size  $n$  whose points are in sorted  $x$ -order. If  $n = 2$ , then the running time is bounded by some constant, say  $c$ . If  $n \geq 2$ , then the algorithm spends  $2 \cdot T(n/2)$  time for the two recursive calls, whereas the rest of the algorithm takes at most  $c'n$  time, where  $c'$  is some constant. Thus, the function  $T(n)$  satisfies the following recurrence:

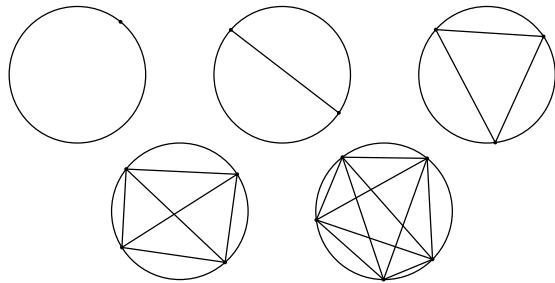
$$\begin{aligned} T(1) &\leq c, \\ T(n) &\leq 2 \cdot T(n/2) + c'n, \text{ if } n \geq 2 \text{ and } n \text{ is a power of 2.} \end{aligned}$$

As in Section 4.6.2, this recurrence solves to  $T(n) = O(n \log n)$ . Thus, we have proved the following result:

**Theorem 4.7.5** *For any list  $L$  of  $n$  points in  $\mathbb{R}^2$ , algorithm  $\text{CLOSESTPAIR}(L, n)$  computes their closest-pair distance in  $O(n \log n)$  time.*

## 4.8 Counting Regions when Cutting a Circle

Take a circle, place  $n$  points on it, and connect each pair of points by a straight-line segment. The points must be placed in such a way that no three segments pass through one point. These segments divide the circle into regions. Define  $R_n$  to be the number of such regions. Can we determine  $R_n$ ?



By looking at the figure above, we see that

$$R_1 = 1, R_2 = 2, R_3 = 4, R_4 = 8, R_5 = 16.$$

There seems to be a clear pattern and it is natural to guess that  $R_n$  is equal to  $2^{n-1}$  for all  $n \geq 1$ . To prove this, we have to argue that the number of regions doubles if we increase  $n$  by 1. If you try to do this, however, then you will fail! The reason is that  $R_n$  is *not* equal to  $2^{n-1}$  for *all*  $n \geq 1$ ; our guess was correct only for  $1 \leq n \leq 5$ .

We will prove below that  $R_n$  grows only *polynomially* in  $n$ . This will imply that  $R_n$  cannot be equal to  $2^{n-1}$  for all  $n$ , because the latter function grows *exponentially*.

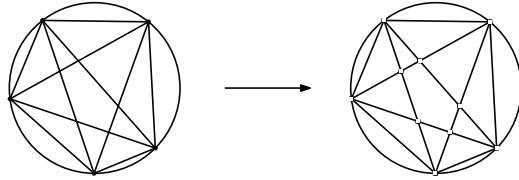
### 4.8.1 A Polynomial Upper Bound on $R_n$

Let  $n$  be a (large) integer, consider a placement of  $n$  points on a circle, and connect each of the  $\binom{n}{2}$  pairs of points by a straight-line segment. Recall that we assume that no three segments pass through one point. We define the following graph:

- Each of the  $n$  points on the circle is a vertex.
- Each intersection point between two segments is a vertex.

- These vertices divide the segments into subsegments and the circle into arcs in a natural way. Each such subsegment and arc is an edge of the graph.

The figure below illustrates this for the case when  $n = 5$ . The graph on the right has  $10 = 5 + 5$  vertices: Each of the 5 points on the circle leads to one vertex and each of the 5 intersection points leads to one vertex. These 10 vertices divide the  $\binom{5}{2} = 10$  segments into 20 straight-line edges and the circle into 5 circular edges. Therefore, the graph has  $20 + 5 = 25$  edges.



Note that, strictly speaking, this process does not define a proper graph, because any two consecutive vertices on the circle are connected by two edges (one straight-line edge and one circular edge), whereas in a proper graph, there can be only one edge between any pair of vertices. For simplicity, however, we will refer to the resulting structure as a graph.

Let  $V_n$  and  $E_n$  be the number of vertices and edges of the graph, respectively. We claim that

$$V_n \leq n + \binom{\binom{n}{2}}{2}. \quad (4.10)$$

This claim follows from the following observations:

- There are exactly  $n$  vertices on the circle.
- The  $n$  points on the circle are connected by  $\binom{n}{2}$  segments, and any two such segments intersect at most once. Therefore, the number of vertices inside the circle is at most the number of pairs of segments. The latter quantity is equal to

$$\binom{\binom{n}{2}}{2}.$$

We next claim that

$$E_n \leq n + \binom{V_n}{2}. \quad (4.11)$$

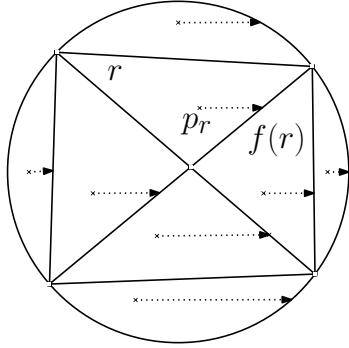
This claim follows from the following observations:

- There are exactly  $n$  edges on the circle.
- Any straight-line edge joins two vertices. Therefore, the number of straight-line edges is at most the number of pairs of vertices, which is  $\binom{V_n}{2}$ .

The final claim is that

$$R_n \leq E_n. \quad (4.12)$$

To prove this claim, we do the following. For each region  $r$ , choose a point  $p_r$  inside  $r$ , such that the  $y$ -coordinate of  $p_r$  is not equal to the  $y$ -coordinate of any vertex. Let  $f(r)$  be the first edge that is reached when walking from  $p_r$  horizontally to the right.



This defines a one-to-one function  $f$  from the set of regions to the set of edges. Therefore, the number of regions, which is  $R_n$ , is at most the number of edges, which is  $E_n$ .

By combining (4.10), (4.11), and (4.12), we get

$$\begin{aligned} R_n &\leq E_n \\ &\leq n + \binom{V_n}{2} \\ &\leq n + \binom{n + \binom{\binom{n}{2}}{2}}{2}. \end{aligned}$$

In order to estimate the last quantity, we are going to use asymptotic notation; see Section 2.3. First observe that

$$\binom{n}{2} = \frac{n(n-1)}{2} = O(n^2).$$

This implies that

$$\binom{\binom{n}{2}}{2} = \binom{O(n^2)}{2} = O(n^4),$$

which implies that

$$n + \binom{\binom{n}{2}}{2} = n + O(n^4) = O(n^4),$$

which implies that

$$\binom{n + \binom{\binom{n}{2}}{2}}{2} = \binom{O(n^4)}{2} = O(n^8),$$

which implies that

$$R_n \leq n + \binom{n + \binom{\binom{n}{2}}{2}}{2} = n + O(n^8) = O(n^8).$$

Thus, we have proved our claim that  $R_n$  grows polynomially in  $n$  and, therefore, for large values of  $n$ ,  $R_n$  is not equal to  $2^{n-1}$ . (Using results on planar graphs that we will see in Section 7.5.1, it can be shown that, in fact,  $R_n = O(n^4)$ .)

We remark that there is a shorter way to prove that  $R_n$  is not equal to  $2^{n-1}$  for all  $n \geq 1$ : You can verify by hand that  $R_6 = 31$ . Still, this single example does not rule out the possibility that  $R_n$  grows exponentially. The analysis that we gave above does rule this out.

We have proved above that  $R_n = O(n^8)$ . We also mentioned that this upper bound can be improved to  $O(n^4)$ . In the following subsections, we will prove that the latter upper bound cannot be improved. That is, we will prove that  $R_n = \Theta(n^4)$ . In fact, we will determine an exact formula, in terms of  $n$ , for the value of  $R_n$ .

### 4.8.2 A Recurrence Relation for $R_n$

Let  $n \geq 2$  be an integer and consider a placement of  $n$  points on a circle. We denote these points by  $p_1, p_2, \dots, p_n$  and assume that they are numbered in counterclockwise order. As before, we connect each of the  $\binom{n}{2}$  pairs of points by a straight-line segment. We assume that no three segments pass through one point. We are going to derive a recurrence relation for the number  $R_n$  of regions in the following way:

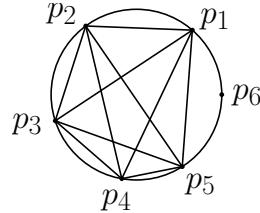
- Remove all segments that have  $p_n$  as an endpoint. At this moment, the number of regions is, by definition, equal to  $R_{n-1}$ .
- Add the  $n - 1$  line segments  $p_1p_n, p_2p_n, \dots, p_{n-1}p_n$  one by one. For each segment  $p_kp_n$  added, determine the *increase*  $I_k$  in the number of regions.
- Take the sum of  $R_{n-1}$  and all increases  $I_k$ , i.e.,

$$R_{n-1} + \sum_{k=1}^{n-1} I_k.$$

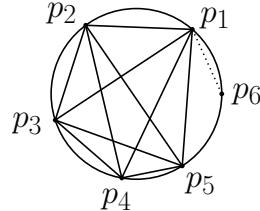
This sum is equal to  $R_n$ , because in the entire process, we have counted each of the regions for  $n$  points exactly once.

- Thus, together with the base case  $R_1 = 1$ , we obtain a recurrence relation for the values  $R_n$ .

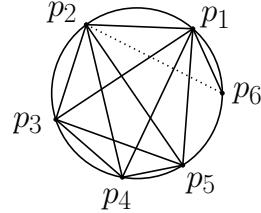
We start by illustrating this process for the case when  $n = 6$ . The figure below shows the situation after we have removed all segments that have  $p_6$  as an endpoint. The number of regions is equal to  $R_5 = 16$ .



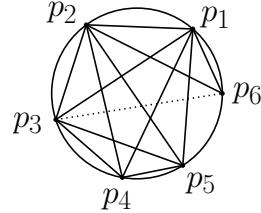
We are going to add, one by one, the five segments that have  $p_6$  as an endpoint. When we add  $p_1p_6$ , one region gets cut into two. Thus, the number of regions increases by one. Using the notation introduced above, we have  $I_1 = 1$ .



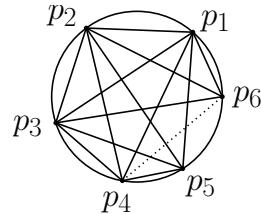
When we add  $p_2p_6$ , four regions get cut into two. Thus, the number of regions increases by four, and we have  $I_2 = 4$ .



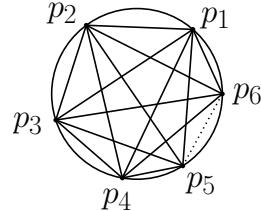
When we add  $p_3p_6$ , five regions get cut into two. Thus, the number of regions increases by five, and we have  $I_3 = 5$ .



When we add  $p_4p_6$ , four regions get cut into two. Thus, the number of regions increases by four, and we have  $I_4 = 4$ .



Finally, when we add  $p_5p_6$ , one region gets cut into two. Thus, the number of regions increases by one, and we have  $I_5 = 1$ .



After having added the five segments with endpoint  $p_6$ , we have accounted for all regions determined by the six points. In other words, the number of regions we have at the end is equal to  $R_6$ . Since the number of regions at the end is also equal to the sum of (i) the number of regions we started with, which is  $R_5$ , and (ii) the total increase, we have

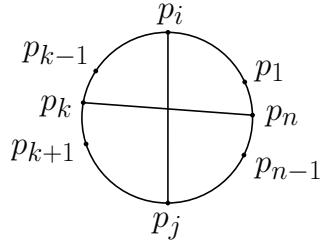
$$R_6 = R_5 + I_1 + I_2 + I_3 + I_4 + I_5 = 31.$$

Let us look at this more carefully. We have seen that  $I_3 = 5$ . That is, when adding the segment  $p_3p_6$ , the number of regions increases by 5. Where does this number 5 come from? The segment  $p_3p_6$  intersects 4 segments, namely  $p_1p_4$ ,  $p_1p_5$ ,  $p_2p_4$ , and  $p_2p_5$ . The increase in the number of regions is one more than the number of intersections. Thus, when adding a segment, if we determine the number  $X$  of intersections between this new segment and existing segments, then the increase in the number of regions is equal to  $1 + X$ .

When we add  $p_3p_6$ , we have  $X = 4$ . Where does this number 4 come from? We make the following observations:

- Any segment that intersects  $p_3p_6$  has one endpoint above  $p_3p_6$  and one endpoint below  $p_3p_6$ .
- Any pair  $(a, b)$  of points on the circle, with  $a$  above  $p_3p_6$  and  $b$  below  $p_3p_6$ , defines a segment  $ab$  that intersects  $p_3p_6$ .
- Thus, the value of  $X$  is equal to the number of pairs  $(a, b)$  of points in  $\{p_1, p_2, p_4, p_5\}$ , where  $a$  is above  $p_3p_6$  and  $b$  is below  $p_3p_6$ . Since there are 2 choices for  $a$  (viz.,  $p_1$  and  $p_2$ ) and 2 choices for  $b$  (viz.,  $p_4$  and  $p_5$ ), it follows from the Product Rule that  $X = 2 \cdot 2 = 4$ .

Now that we have seen the basic approach, we are going to derive the recurrence relation for  $R_n$  for an arbitrary integer  $n \geq 2$ . After having removed all segments that have  $p_n$  as an endpoint, we have  $R_{n-1}$  regions. For each integer  $k$  with  $1 \leq k \leq n-1$ , we add the segment  $p_kp_n$ . What is the number of existing segments that are intersected by this new segment?



We observe that for  $i < j$ ,

- $p_ip_j$  intersects  $p_kp_n$  if and only if  $1 \leq i \leq k - 1$  and  $k + 1 \leq j \leq n - 1$ .

Since there are  $k - 1$  choices for  $i$  and  $n - k - 1$  choices for  $j$ , the Product Rule implies that the number of intersections due to  $p_kp_n$  is equal to  $(k - 1)(n - k - 1)$ . Thus, the segment  $p_kp_n$  goes through  $1 + (k - 1)(n - k - 1)$  regions, and each of them is cut into two. It follows that, when adding  $p_kp_n$ , the increase  $I_k$  in the number of regions is equal to

$$I_k = 1 + (k - 1)(n - k - 1).$$

We conclude that

$$\begin{aligned} R_n &= R_{n-1} + \sum_{k=1}^{n-1} I_k \\ &= R_{n-1} + \sum_{k=1}^{n-1} (1 + (k - 1)(n - k - 1)). \end{aligned}$$

In the summation on the right-hand side

- the term 1 occurs exactly  $n - 1$  times, and
- the term  $(k - 1)(n - k - 1)$  is non-zero only if  $2 \leq k \leq n - 2$ .

It follows that, for  $n \geq 2$ ,

$$R_n = R_{n-1} + (n - 1) + \sum_{k=2}^{n-2} (k - 1)(n - k - 1). \quad (4.13)$$

Thus, together with the base case

$$R_1 = 1, \quad (4.14)$$

we have determined the recurrence relation we were looking for.

### 4.8.3 Simplifying the Recurrence Relation

In this subsection, we will use a combinatorial proof (see Section 3.7) to show that the summation on the right-hand side of (4.13) satisfies

$$\sum_{k=2}^{n-2} (k-1)(n-k-1) = \binom{n-1}{3}, \quad (4.15)$$

for any integer  $n \geq 2$ . (In fact, (4.15) is a special case of the result in Exercise 3.62.)

If  $n \in \{2, 3\}$ , then both sides of (4.15) are equal to zero. Assume that  $n \geq 4$  and consider the set  $S = \{1, 2, \dots, n-1\}$ . We know that the number of 3-element subsets of  $S$  is equal to  $\binom{n-1}{3}$ . As we will see below, the summation on the left-hand side of (4.15) counts exactly the same subsets.

We divide the 3-element subsets of  $S$  into groups based on their middle element. Observe that the middle element can be any of the values  $2, 3, \dots, n-2$ . Thus, for any  $k$  with  $2 \leq k \leq n-2$ , the  $k$ -th group  $G_k$  consists of all 3-element subsets of  $S$  whose middle element is equal to  $k$ . Since the groups are pairwise disjoint, we have

$$\binom{n-1}{3} = \sum_{k=2}^{n-2} |G_k|.$$

What is the size of the  $k$ -th group  $G_k$ ? Any 3-element subset in  $G_k$  consists of

- one element from  $\{1, 2, \dots, k-1\}$ ,
- the element  $k$ , and
- one element from  $\{k+1, k+2, \dots, n-1\}$ .

It then follows from the Product Rule that

$$|G_k| = (k-1) \cdot 1 \cdot (n-k-1) = (k-1)(n-k-1).$$

Thus, we have proved the identity in (4.15), and the recurrence relation in (4.13) and (4.14) becomes

$$\begin{aligned} R_1 &= 1, \\ R_n &= R_{n-1} + (n-1) + \binom{n-1}{3}, \text{ if } n \geq 2. \end{aligned} \quad (4.16)$$

#### 4.8.4 Solving the Recurrence Relation

Now that we have a recurrence relation that looks reasonable, we are going to apply the unfolding technique of Section 4.6 to solve it. Let  $n \geq 2$  be an integer. By repeatedly applying the recurrence relation in (4.16), we get

$$\begin{aligned} R_n &= (n-1) + \binom{n-1}{3} + R_{n-1} \\ &= (n-1) + (n-2) + \binom{n-1}{3} + \binom{n-2}{3} + R_{n-2} \\ &= (n-1) + (n-2) + (n-3) + \binom{n-1}{3} + \binom{n-2}{3} + \binom{n-3}{3} \\ &\quad + R_{n-3}. \end{aligned}$$

By continuing, we get

$$\begin{aligned} R_n &= (n-1) + (n-2) + (n-3) + \cdots + 3 + 2 + 1 \\ &\quad + \binom{n-1}{3} + \binom{n-2}{3} + \binom{n-3}{3} + \cdots + \binom{3}{3} + \binom{2}{3} + \binom{1}{3} \\ &\quad + R_1. \end{aligned}$$

Since  $\binom{2}{3} = \binom{1}{3} = 0$  and  $R_1 = 1$ , we get

$$\begin{aligned} R_n &= (n-1) + (n-2) + (n-3) + \cdots + 3 + 2 + 1 \\ &\quad + \binom{n-1}{3} + \binom{n-2}{3} + \binom{n-3}{3} + \cdots + \binom{3}{3} \\ &\quad + 1. \end{aligned}$$

Since, by Theorem 2.2.10, the first summation is equal to

$$1 + 2 + 3 + \cdots + (n-1) = n(n-1)/2 = \binom{n}{2},$$

we get

$$R_n = 1 + \binom{n}{2} + \sum_{k=3}^{n-1} \binom{k}{3}.$$

The final step is to simplify the summation on the right-hand side. We will use a combinatorial proof to show that

$$\sum_{k=3}^{n-1} \binom{k}{3} = \binom{n}{4}, \tag{4.17}$$

for any integer  $n \geq 2$ . (As was the case for (4.15), the identity in (4.17) is a special case of the result in Exercise 3.62.)

If  $n \in \{2, 3\}$ , then both sides of (4.17) are equal to zero. Assume that  $n \geq 4$  and consider all 4-element subsets of the set  $S = \{1, 2, \dots, n\}$ . We know that there are  $\binom{n}{4}$  many such subsets. We divide these subsets into groups based on their largest element. For any  $k$  with  $3 \leq k \leq n - 1$ , the  $k$ -th group  $G_k$  consists of all 4-element subsets of  $S$  whose largest element is equal to  $k + 1$ . It should be clear that

$$\binom{n}{4} = \sum_{k=3}^{n-1} |G_k|.$$

To determine the size of the group  $G_k$ , we observe that any 4-element subset in  $G_k$  consists of

- three elements from  $\{1, 2, \dots, k\}$  and
- the element  $k + 1$ .

It then follows from the Product Rule that

$$|G_k| = \binom{k}{3} \cdot 1 = \binom{k}{3},$$

completing the proof of (4.17).

After (finally!) having solved and simplified our recurrence relation, we conclude that for any integer  $n \geq 1$ ,

$$R_n = 1 + \binom{n}{2} + \binom{n}{4}.$$

In Exercise 4.76, you will see a shorter way to determine the exact value of  $R_n$ . We went for the long derivation, because it allowed us to illustrate, along the way, several techniques from previous sections.

## 4.9 Exercises

**4.1** The function  $f : \mathbb{N} \rightarrow \mathbb{Z}$  is recursively defined as follows:

$$\begin{aligned} f(0) &= 7, \\ f(n) &= f(n - 1) + 6n - 3 \text{ if } n \geq 1. \end{aligned}$$

Prove that  $f(n) = 3n^2 + 7$  for all integers  $n \geq 0$ .

**4.2** The function  $f : \mathbb{N} \rightarrow \mathbb{Z}$  is recursively defined as follows:

$$\begin{aligned} f(0) &= -18, \\ f(n) &= 9(n-2)(n-3) + f(n-1) \quad \text{if } n \geq 1. \end{aligned}$$

Prove that

$$f(n) = 3(n-1)(n-2)(n-3)$$

for all integers  $n \geq 0$ .

**4.3** The function  $f : \mathbb{N} \rightarrow \mathbb{Z}$  is recursively defined as follows:

$$\begin{aligned} f(0) &= 3, \\ f(n) &= 2 \cdot f(n-1) - (f(n-1))^2 \quad \text{if } n \geq 1. \end{aligned}$$

Prove that  $f(n) = 1 - 2^{2^n}$  for all integers  $n \geq 1$ . (Note that  $2^{2^n}$  denotes 2 to the power of  $2^n$ .)

**4.4** The function  $f : \mathbb{N} \rightarrow \mathbb{N}$  is defined by

$$\begin{aligned} f(0) &= 1, \\ f(n) &= \frac{1}{2} \cdot 4^n \cdot f(n-1) \quad \text{if } n \geq 1. \end{aligned}$$

Prove that for every integer  $n \geq 0$ ,

$$f(n) = 2^{n^2};$$

this reads as 2 to the power  $n^2$ .

**4.5** The function  $f : \mathbb{N} \rightarrow \mathbb{N}$  is defined by

$$\begin{aligned} f(0) &= 0, \\ f(1) &= 0, \\ f(n) &= f(n-2) + 2^{n-1} \quad \text{if } n \geq 2. \end{aligned}$$

- Prove that for every even integer  $n \geq 0$ ,

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

- Prove that for every odd integer  $n \geq 1$ ,

$$f(n) = \frac{2^{n+1} - 4}{3}.$$

**4.6** The function  $f : \{1, 2, 3, \dots\} \rightarrow \mathbb{R}$  is defined by

$$\begin{aligned} f(1) &= 2, \\ f(n) &= \frac{1}{2} \left( f(n-1) + \frac{1}{f(n-1)} \right) \text{ if } n \geq 2. \end{aligned}$$

- Prove that for every integer  $n \geq 1$ ,

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

Note that  $3^{2^{n-1}}$  denotes 3 to the power of  $2^{n-1}$ .

**4.7** You are asked to come up with an exam question about recursive functions. You write down some recurrence, which you then solve. Afterwards, you give the recurrence to the students, together with the solution. The students must then prove that the given solution is indeed correct.

This is a painful process, because you must solve the recurrence yourself. Since you are lazy, you start with the following:

**Exam Question:**

The function  $f : \mathbb{N} \rightarrow \mathbb{N}$  is defined by

$$\begin{aligned} f(0) &= XXX, \\ f(n) &= f(n-1) + YYY \text{ if } n \geq 1. \end{aligned}$$

Prove that for every integer  $n \geq 0$ ,

$$f(n) = 7n^2 - 2n + 9.$$

- Complete the question, i.e., fill in  $XXX$  and  $YYY$ , so that you obtain a complete recurrence that has the given solution.

**4.8** The function  $f : \mathbb{N} \rightarrow \mathbb{Z}$  is defined by

$$f(n) = 2n(n-6)$$

for each integer  $n \geq 0$ . Derive a recursive form of this function.

**4.9** The function  $f : \mathbb{N}^2 \rightarrow \mathbb{N}$  is defined by

$$\begin{aligned} f(0, n) &= 2n && \text{if } n \geq 0, \\ f(m, 0) &= 0 && \text{if } m \geq 1, \\ f(m, 1) &= 2 && \text{if } m \geq 1, \\ f(m, n) &= f(m - 1, f(m, n - 1)) && \text{if } m \geq 1 \text{ and } n \geq 2. \end{aligned}$$

- Determine  $f(2, 2)$ .
- Determine  $f(1, n)$  for  $n \geq 1$ .
- Determine  $f(3, 3)$ .

**4.10** The function  $f : \mathbb{N}^3 \rightarrow \mathbb{N}$  is defined as follows:

$$\begin{aligned} f(k, n, 0) &= k + n && \text{if } k \geq 0 \text{ and } n \geq 0, \\ f(k, 0, 1) &= 0 && \text{if } k \geq 0, \\ f(k, 0, 2) &= 1 && \text{if } k \geq 0, \\ f(k, 0, i) &= k && \text{if } k \geq 0 \text{ and } i \geq 3, \\ f(k, n, i) &= f(k, f(k, n - 1, i), i - 1) && \text{if } k \geq 0, i \geq 1, \text{ and } n \geq 1. \end{aligned}$$

Determine  $f(2, 3, 2)$ .

**4.11** The functions  $f : \mathbb{N} \rightarrow \mathbb{N}$  and  $g : \mathbb{N}^2 \rightarrow \mathbb{N}$  are recursively defined as follows:

$$\begin{aligned} f(0) &= 1, \\ f(n) &= g(n, f(n - 1)) && \text{if } n \geq 1, \\ g(m, 0) &= 0 && \text{if } m \geq 0, \\ g(m, n) &= m + g(m, n - 1) && \text{if } m \geq 0 \text{ and } n \geq 1. \end{aligned}$$

Solve these recurrence relations for  $f$ , i.e., express  $f(n)$  in terms of  $n$ .

**4.12** The functions  $f : \mathbb{N} \rightarrow \mathbb{N}$  and  $g : \mathbb{N}^2 \rightarrow \mathbb{N}$  are recursively defined as follows:

$$\begin{aligned} f(0) &= 1, \\ f(1) &= 2, \\ f(n) &= g(f(n - 2), f(n - 1)) && \text{if } n \geq 2, \\ g(m, 0) &= 2m && \text{if } m \geq 0, \\ g(m, n) &= g(m, n - 1) + 1 && \text{if } m \geq 0 \text{ and } n \geq 1. \end{aligned}$$

Solve these recurrence relations for  $f$ , i.e., express  $f(n)$  in terms of  $n$ .

**4.13** The functions  $f : \mathbb{N} \rightarrow \mathbb{N}$  and  $g : \mathbb{N}^2 \rightarrow \mathbb{N}$  are recursively defined as follows:

$$\begin{aligned} f(0) &= 1, \\ f(n) &= g(f(n-1), 2n) \quad \text{if } n \geq 1, \\ g(0, n) &= 0 \quad \text{if } n \geq 0, \\ g(m, n) &= g(m-1, n) + n \quad \text{if } m \geq 1 \text{ and } n \geq 0. \end{aligned}$$

Solve these recurrence relations for  $f$ , i.e., express  $f(n)$  in terms of  $n$ .

**4.14** The functions  $f : \mathbb{N} \rightarrow \mathbb{N}$ ,  $g : \mathbb{N}^2 \rightarrow \mathbb{N}$ , and  $h : \mathbb{N} \rightarrow \mathbb{N}$  are recursively defined as follows:

$$\begin{aligned} f(n) &= g(n, h(n)) \quad \text{if } n \geq 0, \\ g(m, 0) &= 0 \quad \text{if } m \geq 0, \\ g(m, n) &= g(m, n-1) + m \quad \text{if } m \geq 0 \text{ and } n \geq 1, \\ h(0) &= 1, \\ h(n) &= 2 \cdot h(n-1) \quad \text{if } n \geq 1. \end{aligned}$$

Solve these recurrences for  $f$ , i.e., express  $f(n)$  in terms of  $n$ .

**4.15** The sequence  $a_n$  of numbers, for  $n \geq 0$ , is recursively defined as follows:

$$\begin{aligned} a_0 &= 5, \\ a_1 &= 3, \\ a_n &= 6 \cdot a_{n-1} - 9 \cdot a_{n-2} \quad \text{if } n \geq 2. \end{aligned}$$

- Determine  $a_n$  for  $n = 0, 1, 2, 3, 4, 5$ .
- Prove that for every integer  $n \geq 0$ ,

$$a_n = (5 - 4n) \cdot 3^n.$$

**4.16** Let  $\varphi = \frac{1+\sqrt{5}}{2}$  and  $\psi = \frac{1-\sqrt{5}}{2}$ , and let  $n \geq 0$  be an integer. We have seen in Theorem 4.2.1 that

$$\frac{\varphi^n - \psi^n}{\sqrt{5}} \tag{4.18}$$

is equal to the  $n$ -th Fibonacci number  $f_n$ . Since the Fibonacci numbers are obviously integers, the number in (4.18) is an integer as well.

Prove that the number in (4.18) is a rational number using only Newton's Binomial Theorem (i.e., Theorem 3.6.5).

**4.17** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . In this exercise, you will prove that there exists a Fibonacci number whose 2018 rightmost digits (when written in decimal notation) are all zero.

In the rest of this exercise,  $N$  denotes the number  $10^{4036}$ . For any integer  $n \geq 0$ , define

$$g_n = f_n \bmod 10^{2018}.$$

- Consider the ordered pairs  $(g_n, g_{n+1})$ , for  $n = 0, 1, 2, \dots, N$ . Use the Pigeonhole Principle to prove that these ordered pairs cannot all be distinct. That is, prove that there exist integers  $m \geq 0, p \geq 1$ , such that  $m + p \leq N$  and

$$(g_m, g_{m+1}) = (g_{m+p}, g_{m+p+1}).$$

- Prove that  $(g_{m-1}, g_m) = (g_{m+p-1}, g_{m+p})$ .
- Prove that  $(g_0, g_1) = (g_p, g_{p+1})$ .
- Consider the decimal representation of  $f_p$ . Prove that the 2018 rightmost digits of  $f_p$  are all zero.
- Let  $b \geq 2$  and  $k \geq 1$  be integers. Prove that there exists a Fibonacci number whose  $k$  rightmost digits (when written in base- $b$  notation) are all zero.

**4.18** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . Prove that for each integer  $n \geq 1$ ,

$$\sum_{i=1}^n f_{2i} = f_{2n+1} - 1$$

and

$$f_1^2 + f_2^2 + f_3^2 + \cdots + f_n^2 = f_n f_{n+1}.$$

**4.19** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . Prove that for each integer  $n \geq 0$ ,

- $f_{3n}$  is even,
- $f_{3n+1}$  is odd,

- $f_{3n+2}$  is odd,
- $f_{4n}$  is a multiple of 3.

**4.20** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . In Section 4.2.1, we have seen that for any integer  $m \geq 1$ , the number of 00-free bitstrings of length  $m$  is equal to  $f_{m+2}$ .

Let  $n \geq 2$  be an integer.

- How many 00-free bitstrings of length  $n$  do not contain any 0?
- How many 00-free bitstrings of length  $n$  have the following property: The rightmost 0 is at position 1.
- How many 00-free bitstrings of length  $n$  have the following property: The rightmost 0 is at position 2.
- Let  $k$  be an integer with  $3 \leq k \leq n$ . How many 00-free bitstrings of length  $n$  have the following property: The rightmost 0 is at position  $k$ .
- Use the previous results to prove that

$$f_{n+2} = 1 + \sum_{k=1}^n f_k.$$

**4.21** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . In Section 4.2.1, we have seen that for any integer  $m \geq 1$ , the number of 00-free bitstrings of length  $m$  is equal to  $f_{m+2}$ .

- Let  $n \geq 2$  be an integer. What is the number of 00-free bitstrings of length  $2n - 1$  for which the bit in the middle position is equal to 1?
- Let  $n \geq 3$  be an integer. What is the number of 00-free bitstrings of length  $2n - 1$  for which the bit in the middle position is equal to 0?
- Use the previous results to prove that for any integer  $n \geq 3$ ,

$$f_{2n+1} = f_n^2 + f_{n+1}^2.$$

**4.22** In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . In Section 4.2.1, we have seen that for any integer  $m \geq 1$ , the number of 00-free bitstrings of length  $m$  is equal to  $f_{m+2}$ .

Let  $n \geq 1$  be an integer.

- How many 00-free bitstrings of length  $n + 2$  do not contain any 0?
- How many 00-free bitstrings of length  $n + 2$  contain exactly one 0?
- How many 00-free bitstrings of length  $n + 2$  have the following property: The bitstring contains at least two 0's, and the second rightmost 0 is at position 1.
- How many 00-free bitstrings of length  $n + 2$  have the following property: The bitstring contains at least two 0's, and the second rightmost 0 is at position 2.
- Let  $k$  be an integer with  $3 \leq k \leq n$ . How many 00-free bitstrings of length  $n + 2$  have the following property: The bitstring contains at least two 0's, and the second rightmost 0 is at position  $k$ .
- Let  $k$  be an element of  $\{n + 1, n + 2\}$ . How many 00-free bitstrings of length  $n + 2$  have the following property: The bitstring contains at least two 0's, and the second rightmost 0 is at position  $k$ .
- Use the previous results to prove that

$$\sum_{k=1}^n (n - k + 1) \cdot f_k = f_{n+4} - n - 3,$$

i.e.,

$$n \cdot f_1 + (n - 1) \cdot f_2 + (n - 2) \cdot f_3 + \cdots + 2 \cdot f_{n-1} + 1 \cdot f_n = f_{n+4} - n - 3.$$

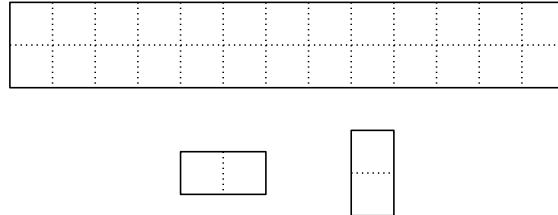
**4.23** Use basic algebra to prove that

$$(x^2 + y^2 + (x + y)^2)^2 = 2(x^4 + y^4 + (x + y)^4).$$

In Section 4.2, we have defined the Fibonacci numbers  $f_0, f_1, f_2, \dots$ . Prove that for each integer  $n \geq 0$ ,

$$(f_n^2 + f_{n+1}^2 + f_{n+2}^2)^2 = 2(f_n^4 + f_{n+1}^4 + f_{n+2}^4).$$

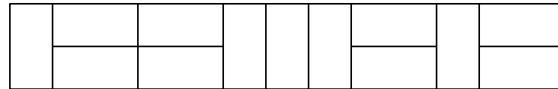
**4.24** Let  $n \geq 1$  be an integer and consider a  $2 \times n$  board  $B_n$  consisting of  $2n$  square cells. The top part of the figure below shows  $B_{13}$ .



A *brick* is a horizontal or vertical board consisting of 2 square cells; see the bottom part of the figure above. A *tiling* of the board  $B_n$  is a placement of bricks on the board such that

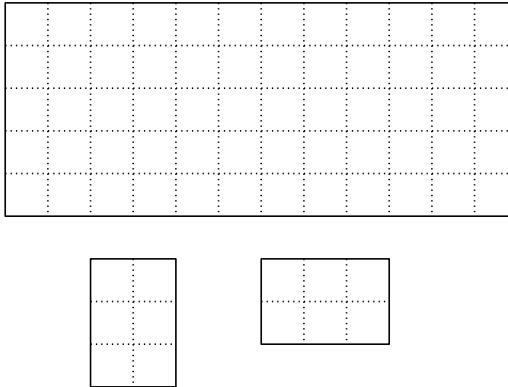
- the bricks exactly cover  $B_n$  and
- no two bricks overlap.

The figure below shows a tiling of  $B_{13}$ .



For  $n \geq 1$ , let  $T_n$  be the number of different tilings of the board  $B_n$ . Determine the value of  $T_n$ , i.e., express  $T_n$  in terms of numbers that we have seen in this chapter.

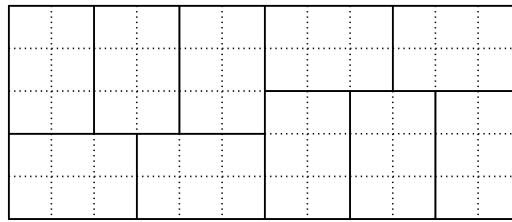
**4.25** Let  $n$  be a positive integer and consider a  $5 \times n$  board  $B_n$  consisting of  $5n$  cells, each one having sides of length one. The top part of the figure below shows  $B_{12}$ .



A *brick* is a horizontal or vertical board consisting of  $2 \times 3 = 6$  cells; see the bottom part of the figure above. A *tiling* of the board  $B_n$  is a placement of bricks on the board such that

- the bricks exactly cover  $B_n$  and
- no two bricks overlap.

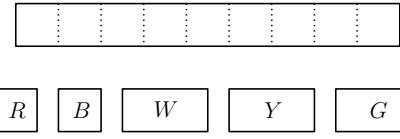
The figure below shows a tiling of  $B_{12}$ .



Let  $T_n$  be the number of different tilings of the board  $B_n$ .

- Let  $n \geq 6$  be a multiple of 6. Determine the value of  $T_n$ .
- Let  $n$  be a positive integer that is not a multiple of 6. Prove that  $T_n = 0$ .

**4.26** Let  $n$  be a positive integer and consider a  $1 \times n$  board  $B_n$  consisting of  $n$  cells, each one having sides of length one. The top part of the figure below shows  $B_9$ .



We have an unlimited supply of *bricks*, which are of the following types (see the bottom part of the figure above):

- There are red ( $R$ ) and blue ( $B$ ) bricks, both of which are  $1 \times 1$  cells.
- There are white ( $W$ ), yellow ( $Y$ ), and green ( $G$ ) bricks, all of which are  $1 \times 2$  cells.

A *tiling* of the board  $B_n$  is a placement of bricks on the board such that

- the bricks exactly cover  $B_n$  and
- no two bricks overlap.

In a tiling, a color can be used more than once and some colors may not be used at all. The figure below shows a tiling of  $B_9$ , in which each color is used and the color red is used twice.

$B$	$W$	$R$	$G$	$R$	$Y$
-----	-----	-----	-----	-----	-----

Let  $T_n$  be the number of different tilings of the board  $B_n$ .

- Determine  $T_1$  and  $T_2$ .
- Let  $n \geq 3$  be an integer. Prove that

$$T_n = 2 \cdot T_{n-1} + 3 \cdot T_{n-2}.$$

- Prove that for any integer  $n$ ,

$$2(-1)^{n-1} + 3(-1)^{n-2} = (-1)^n.$$

- Prove that for any integer  $n \geq 1$ ,

$$T_n = \frac{3^{n+1} + (-1)^n}{4}.$$

**4.27** The sequence of numbers  $a_n$ , for  $n \geq 0$ , is recursively defined as follows:

$$\begin{aligned} a_0 &= 0, \\ a_1 &= 1, \\ a_n &= 2 \cdot a_{n-1} + a_{n-2} \quad \text{if } n \geq 2. \end{aligned}$$

- Determine  $a_n$  for  $n = 0, 1, 2, 3, 4, 5$ .

- Prove that

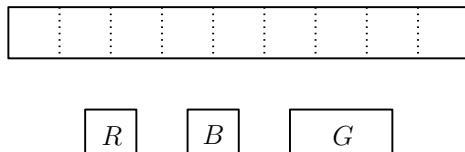
$$a_n = \frac{(1 + \sqrt{2})^n - (1 - \sqrt{2})^n}{2\sqrt{2}} \tag{4.19}$$

for all integers  $n \geq 0$ .

*Hint:* What are the solutions of the equation  $x^2 = 2x + 1$ ?

- Since the numbers  $a_n$ , for  $n \geq 0$ , are obviously integers, the fraction on the right-hand side of (4.19) is an integer as well. Prove that the fraction on the right-hand side of (4.19) is an integer using only Newton's Binomial Theorem (i.e., Theorem 3.6.5).

**4.28** Let  $n$  be a positive integer and consider a  $1 \times n$  board  $B_n$  consisting of  $n$  cells, each one having sides of length one. The top part of the figure below shows  $B_9$ .



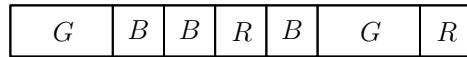
You have an unlimited supply of *bricks*, which are of the following types (see the bottom part of the figure above):

- There are red ( $R$ ) and blue ( $B$ ) bricks, both of which are  $1 \times 1$  cells. We refer to these bricks as *squares*.
- There are green ( $G$ ) bricks, which are  $1 \times 2$  cells. We refer to these as *dominoes*.

A *tiling* of the board  $B_n$  is a placement of bricks on the board such that

- the bricks exactly cover  $B_n$  and
- no two bricks overlap.

In a tiling, a color can be used more than once and some colors may not be used at all. The figure below shows an example of a tiling of  $B_9$ .



Let  $T_n$  be the number of different tilings of the board  $B_n$ .

- Determine  $T_1$ ,  $T_2$ , and  $T_3$ .
- For any integer  $n \geq 1$ , express  $T_n$  in terms of the numbers that appear in Exercise 4.27 .

**4.29** In this exercise, we use the notation of Exercise 4.28. Let  $n \geq 1$  be an integer and consider the  $1 \times n$  board  $B_n$ .

- Consider strings consisting of characters, where each character is  $S$  or  $D$ . Let  $k$  be an integer with  $0 \leq k \leq \lfloor n/2 \rfloor$ . Determine the number of such strings of length  $n - k$ , that contain exactly  $k$  many  $D$ 's.
- Let  $k$  be an integer with  $0 \leq k \leq \lfloor n/2 \rfloor$ . Determine the number of tilings of the board  $B_n$  that use exactly  $k$  dominoes.

*Hint:* How many bricks are used for such a tiling? In the first part, imagine that  $S$  stands for “square” and  $D$  stands for “domino”.

- Use the results of the previous part to prove that

$$T_n = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \cdot 2^{n-2k}.$$

**4.30** In this exercise, we consider strings of characters, where each character is an element of  $\{a, b, c\}$ . For any integer  $n \geq 1$ , let  $E_n$  be the number of such strings of length  $n$  that have an even number of  $c$ 's, and let  $O_n$  be the number of such strings of length  $n$  that have an odd number of  $c$ 's. (Recall that 0 is even.)

- Determine  $E_1$ ,  $O_1$ ,  $E_2$ , and  $O_2$ .
- Explain, in plain English, why

$$E_n + O_n = 3^n.$$

- Prove that for every integer  $n \geq 2$ ,

$$E_n = 2 \cdot E_{n-1} + O_{n-1}.$$

- Prove that for every integer  $n \geq 1$ ,

$$E_n = \frac{1 + 3^n}{2}.$$

**4.31** Consider strings of  $n$  characters, where each character is an element of  $\{a, b, c, d\}$ , that contain an even number of  $a$ s. (Recall that 0 is even.) Let  $E_n$  be the number of such strings. Prove that for any integer  $n \geq 1$ ,

$$E_{n+1} = 2 \cdot E_n + 4^n.$$

**4.32** Let  $A_n$  be the number of bitstrings of length  $n$  that contain 000. Prove that for  $n \geq 4$ ,

$$A_n = A_{n-1} + A_{n-2} + A_{n-3} + 2^{n-3}.$$

**4.33** Let  $n \geq 1$  be an integer and define  $A_n$  to be the number of bitstrings of length  $n$  that do not contain 101.

- Determine  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ .
- Prove that for each integer  $n \geq 4$ ,

$$\begin{aligned} A_n &= 3 + A_1 + A_2 + A_3 + \cdots + A_{n-4} + A_{n-3} + A_{n-1} \\ &= 3 + \sum_{k=1}^{n-3} A_k + A_{n-1}. \end{aligned}$$

*Hint:* Divide the strings into groups depending on the number of leading 1s.

**4.34** Let  $n \geq 1$  be an integer and consider  $n$  people  $P_1, P_2, \dots, P_n$ . Let  $A_n$  be the number of ways these  $n$  people can be divided into groups, such that each group consists of either one or two people.

- Determine  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ .
- Prove that for each integer  $n \geq 3$ ,

$$A_n = A_{n-1} + (n - 1) \cdot A_{n-2}.$$

**4.35** In this exercise, we consider strings of characters, where each character is an element of  $\{a, b, c\}$ . Such a string is called *aa-free*, if it does not contain two consecutive  $a$ 's. For any integer  $n \geq 1$ , let  $F_n$  be the number of *aa-free* strings of length  $n$ .

- Determine  $F_1$ ,  $F_2$ , and  $F_3$ .

- Let  $n \geq 3$  be an integer. Express  $F_n$  in terms of  $F_{n-1}$  and  $F_{n-2}$ .
- Prove that for every integer  $n \geq 1$ ,

$$F_n = \left(\frac{1}{2} + \frac{1}{\sqrt{3}}\right) \left(1 + \sqrt{3}\right)^n + \left(\frac{1}{2} - \frac{1}{\sqrt{3}}\right) \left(1 - \sqrt{3}\right)^n.$$

*Hint:* What are the solutions of the equation  $x^2 = 2x + 2$ ? Using these solutions will simplify the proof.

**4.36** In this exercise, we consider strings of characters, where each character is an element of  $\{a, b, c\}$ . Such a string is called *awesome*, if it does not contain the substring  $ab$  and does not contain the substring  $ba$ . For any integer  $n \geq 1$ , let

1.  $S_n$  denote the number of awesome strings of length  $n$ ,
  2.  $A_n$  denote the number of awesome strings of length  $n$  that start with  $a$ ,
  3.  $B_n$  denote the number of awesome strings of length  $n$  that start with  $b$ ,
  4.  $C_n$  denote the number of awesome strings of length  $n$  that start with  $c$ .
- Determine  $S_1$  and  $S_2$ .
  - Let  $n \geq 1$  be an integer. Express  $S_n$  in terms of  $A_n$ ,  $B_n$ , and  $C_n$ .
  - Let  $n \geq 2$  be an integer. Express  $C_n$  in terms of  $S_{n-1}$ .
  - Let  $n \geq 2$  be an integer. Prove that

$$S_n = (S_{n-1} - B_{n-1}) + (S_{n-1} - A_{n-1}) + S_{n-1}.$$

- Let  $n \geq 3$  be an integer. Prove that

$$S_n = 2 \cdot S_{n-1} + S_{n-2}.$$

- Prove that for every integer  $n \geq 1$ ,

$$S_n = \frac{1}{2} \left(1 + \sqrt{2}\right)^{n+1} + \frac{1}{2} \left(1 - \sqrt{2}\right)^{n+1}.$$

*Hint:* What are the solutions of the equation  $x^2 = 2x + 1$ ? Using these solutions will simplify the proof.

**4.37** A *block* in a bitstring is a maximal consecutive substring of 1's. For example, the bitstring 1100011110100111 has four blocks: 11, 1111, 1, and 111. These blocks have lengths 2, 4, 1, and 3, respectively.

Let  $n \geq 1$  be an integer and let  $B_n$  be the number of bitstrings of length  $n$  that do not contain any block of odd length; in other words, every block in these bitstrings has an even length.

- Determine  $B_1$ ,  $B_2$ ,  $B_3$ , and  $B_4$ .
- Determine the value of  $B_n$ , i.e., express  $B_n$  in terms of numbers that we have seen in this chapter.

**4.38** Let  $n \geq 1$  be an integer and let  $S_n$  be the number of ways in which  $n$  can be written as a sum of 1s and 2s; the order in which the 1s and 2s occur in the sum matters. For example,  $S_3 = 3$ , because

$$3 = 1 + 1 + 1 = 1 + 2 = 2 + 1.$$

- Determine  $S_1$ ,  $S_2$ , and  $S_4$ .
- Determine the value of  $S_n$ , i.e., express  $S_n$  in terms of numbers that we have seen in this chapter.

**4.39** Ever since he was a child, Nick has been dreaming to be like Spiderman. As you all know, Spiderman can climb up the outside of a building; if he is at a particular floor, then, in one step, he can move up several floors. Nick is not that advanced yet. In one step, Nick can move up either one floor or two floors.

Let  $n \geq 1$  be an integer and consider a building with  $n$  floors, numbered 1, 2, . . . ,  $n$ . (The first floor has number 1; this is not the ground floor.) Nick is standing in front of this building, at the ground level. There are different ways in which Nick can climb to the  $n$ -th floor. For example, here are three different ways for the case when  $n = 5$ :

1. move up 2 floors, move up 1 floor, move up 2 floors.
2. move up 1 floor, move up 2 floors, move up 2 floors.
3. move up 1 floor, move up 2 floors, move up 1 floor, move up 1 floor.

Let  $S_n$  be the number of different ways, in which Nick can climb to the  $n$ -th floor.

- Determine,  $S_1$ ,  $S_2$ ,  $S_3$ , and  $S_4$ .
- Determine the value of  $S_n$ , i.e., express  $S_n$  in terms of numbers that we have seen in this chapter.

**4.40** Let  $n \geq 1$  be an integer and consider the set  $S_n = \{1, 2, \dots, n\}$ . A *non-neighbor subset* of  $S_n$  is any subset  $T$  of  $S$  having the following property: If  $k$  is any element of  $T$ , then  $k + 1$  is not an element of  $T$ . (Observe that the empty set is a non-neighbor subset of  $S_n$ .)

For example, if  $n = 3$ , then  $\{1, 3\}$  is a non-neighbor subset, whereas  $\{2, 3\}$  is not a non-neighbor subset.

Let  $N_n$  denote the number of non-neighbor subsets of the set  $S_n$ .

- Determine  $N_1$ ,  $N_2$ , and  $N_3$ .
- Determine the value of  $N_n$ , i.e., express  $N_n$  in terms of numbers that we have seen in this chapter.

**4.41** Let  $n \geq 1$  be an integer and consider the set  $S = \{1, 2, \dots, n\}$ .

- Assume we arrange the elements of  $S$  in sorted order on a horizontal line. Let  $B_n$  be the number of subsets of  $S$  that do not contain any two elements that are neighbors on this line. For example, if  $n = 4$ , then both subsets  $\{1, 3\}$  and  $\{1, 4\}$  are counted in  $B_4$ , but neither of the subsets  $\{2, 3\}$  and  $\{2, 3, 4\}$  is counted.

For each integer  $n \geq 1$ , express  $B_n$  in terms of numbers that we have seen in this chapter.

- Assume we arrange the elements of  $S$  in sorted order along a circle. Let  $C_n$  be the number of subsets of  $S$  that do not contain any two elements that are neighbors on this circle. For example, if  $n = 4$ , then the subset  $\{1, 3\}$  is counted in  $C_4$ , but neither of the subsets  $\{2, 3\}$  and  $\{1, 4\}$  is counted.

For each integer  $n \geq 4$ , express  $C_n$  in terms of numbers that we have seen in this chapter.

**4.42** For any integer  $n \geq 1$ , a permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$  is called *awesome*, if the following condition holds:

- For every  $i$  with  $1 \leq i \leq n$ , the element  $a_i$  in the permutation belongs to the set  $\{i - 1, i, i + 1\}$ .

For example, for  $n = 5$ , the permutation  $2, 1, 3, 5, 4$  is awesome, whereas  $2, 1, 5, 3, 4$  is not an awesome permutation.

Let  $P_n$  denote the number of awesome permutations of the set  $\{1, 2, \dots, n\}$ .

- Determine  $P_1$ ,  $P_2$ , and  $P_3$ .
- Determine the value of  $P_n$ , i.e., express  $P_n$  in terms of numbers that we have seen in this chapter.

*Hint:* Derive a recurrence relation. What are the possible values for the last element  $a_n$  in an awesome permutation?

**4.43** A *block* in a bitstring is a maximal consecutive substring of 1's. For example, the bitstring 1100011110100111 has four blocks: 11, 1111, 1, and 111.

For a given integer  $n \geq 1$ , consider all  $2^n$  bitstrings of length  $n$ . Let  $B_n$  be the total number of blocks in all these bitstrings.

For example, the left part of the table below contains all 8 bitstrings of length 3. Each entry in the rightmost column shows the number of blocks in the corresponding bitstring. Thus,

$$B_3 = 0 + 1 + 1 + 1 + 1 + 2 + 1 + 1 = 8.$$

0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
0	1	1	1
1	0	1	2
1	1	0	1
1	1	1	1

- Determine  $B_1$  and  $B_2$ .
- Let  $n \geq 3$  be an integer.
  - Consider all bitstrings of length  $n$  that start with 0. What is the total number of blocks in these bitstrings?

- Determine the number of blocks in the bitstring

$$\underbrace{1 \cdots 1}_n.$$

- Determine the number of blocks in the bitstring

$$\underbrace{1 \cdots 1}_{n-1} 0.$$

- Let  $k$  be an integer with  $2 \leq k \leq n - 1$ . Consider all bitstrings of length  $n$  that start with

$$\underbrace{1 \cdots 1}_{k-1} 0.$$

Prove that the total number of blocks in these bitstrings is equal to

$$2^{n-k} + B_{n-k}.$$

- Prove that

$$B_n = 2 + B_{n-1} + \sum_{k=2}^{n-1} (2^{n-k} + B_{n-k}).$$

- Use  $1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-2} = 2^{n-1} - 1$ , to prove that

$$B_n = 2^{n-1} + B_1 + B_2 + \cdots + B_{n-1}. \quad (4.20)$$

- Prove that (4.20) also holds for  $n = 2$ .
- Let  $n \geq 3$ . Prove that

$$B_n = 2^{n-2} + 2 \cdot B_{n-1}. \quad (4.21)$$

*Hint:* Write (4.20) on one line. Below this line, write (4.20) with  $n$  replaced by  $n - 1$ .

- Prove that for every  $n \geq 1$ ,

$$B_n = \frac{n+1}{4} \cdot 2^n.$$

- The derivation of the recurrence in (4.21) was quite involved. Prove this recurrence in a direct way.

**4.44** Let  $n \geq 1$  be an integer and consider a set  $S$  consisting of  $n$  elements. A function  $f : S \rightarrow S$  is called *cool*, if for all elements  $x$  of  $S$ ,

$$f(f(f(x))) = x.$$

Let  $A_n$  be the number of cool functions  $f : S \rightarrow S$ .

- Let  $f : S \rightarrow S$  be a cool function, and let  $x$  be an element of  $S$ . Prove that the set

$$\{x, f(x), f(f(x))\}$$

has size 1 or 3.

- Let  $f : S \rightarrow S$  be a cool function, and let  $x$  and  $y$  be two distinct elements of  $S$ . Assume that  $f(y) = y$ . Prove that  $f(x) \neq y$ .
- Prove that for any integer  $n \geq 4$ ,

$$A_n = A_{n-1} + (n-1)(n-2) \cdot A_{n-3}.$$

*Hint:* Let  $y$  be a fixed element in  $S$ . Some cool functions  $f$  have the property that  $f(y) = y$ , whereas some other cool functions  $f$  have the property that  $f(y) \neq y$ .

**4.45** Let  $S$  be the set of ordered pairs of integers that is recursively defined in the following way:

- $(0, 0) \in S$ .
- If  $(a, b) \in S$  then  $(a + 2, b + 3) \in S$ .
- If  $(a, b) \in S$  then  $(a + 3, b + 2) \in S$ .

Prove that for every element  $(a, b)$  in  $S$ ,  $a + b$  is divisible by 5.

**4.46** Let  $S$  be the set of integers that is recursively defined in the following way:

- 4 is an element of  $S$ .

- If  $x$  and  $y$  are elements of  $S$ , then  $x + y^2$  is an element of  $S$ .

Prove that every element of  $S$  is divisible by 4.

**4.47** Let  $S$  be the set of ordered triples of integers that is recursively defined in the following way:

- $(66, 55, 1331) \in S$ .
- If  $(a, b, c) \in S$  then  $(a + 7, b + 5, 14a - 10b + c + 24) \in S$ .

Prove that for every element  $(a, b, c)$  in  $S$ ,

$$a^2 - b^2 = c.$$

**4.48** Let  $S$  be the set of integers that is recursively defined in the following way:

- 1 is an element of  $S$ .
- If  $x$  is an element of  $S$ , then  $x + 2\sqrt{x} + 1$  is also an element of  $S$ .

Give a simple description of the set  $S$  and prove that your answer is correct.

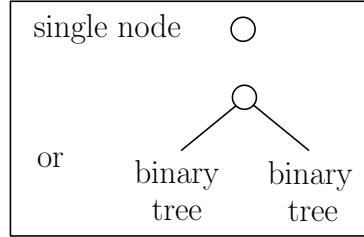
**4.49** The set  $S$  of bitstrings is recursively defined in the following way:

- The string 00 is an element of the set  $S$ .
- The string 01 is an element of the set  $S$ .
- The string 10 is an element of the set  $S$ .
- If the string  $s$  is an element of the set  $S$ , then the string 0s (i.e., the string obtained by adding the bit 0 at the front of  $s$ ) is also an element of the set  $S$ .
- If the string  $s$  is an element of the set  $S$ , then the string 10s (i.e., the string obtained by adding the bits 10 at the front of  $s$ ) is also an element of the set  $S$ .

Let  $s$  be an arbitrary string in the set  $S$ . Prove that  $s$  does not contain the substring 11.

**4.50** A binary tree is

- either one single node
- or a node whose left subtree is a binary tree and whose right subtree is a binary tree.



Prove that any binary tree with  $n$  leaves has exactly  $2n - 1$  nodes.

**4.51** In this exercise, we will denote Boolean variables by lowercase letters, such as  $p$  and  $q$ . A *proposition* is any Boolean formula that can be obtained by applying the following recursive rules:

1. For every Boolean variable  $p$ ,  $p$  is a proposition.
  2. If  $f$  is a proposition, then  $\neg f$  is also a proposition.
  3. If  $f$  and  $g$  are propositions, then  $(f \vee g)$  is also a proposition.
  4. If  $f$  and  $g$  are propositions, then  $(f \wedge g)$  is also a proposition.
- Let  $p$  and  $q$  be Boolean variables. Prove that

$$\neg((p \wedge \neg q) \vee (\neg p \vee q))$$

is a proposition.

- Let  $\uparrow$  denote the *not-and* operator. In other words, if  $f$  and  $g$  are Boolean formulas, then  $(f \uparrow g)$  is the Boolean formula that has the following truth table (0 stands for *false*, and 1 stands for *true*):

$f$	$g$	$(f \uparrow g)$
0	0	1
0	1	1
1	0	1
1	1	0

- Let  $p$  be a Boolean variable. Use a truth table to prove that the Boolean formulas  $(p \uparrow p)$  and  $\neg p$  are equivalent.
  - Let  $p$  and  $q$  be Boolean variables. Use a truth table to prove that the Boolean formulas  $((p \uparrow p) \uparrow (q \uparrow q))$  and  $p \vee q$  are equivalent.
  - Let  $p$  and  $q$  be Boolean variables. Express the Boolean formula  $(p \wedge q)$  as an equivalent Boolean formula that only uses the  $\uparrow$ -operator. Use a truth table to justify your answer.
- Prove that any proposition can be written as an equivalent Boolean formula that only uses the  $\uparrow$ -operator.

**4.52** In Section 4.4, we have seen the recursive algorithm  $\text{GOSSIP}(n)$ , which computes a sequence of phone calls for the persons  $P_1, P_2, \dots, P_n$ . The base case for this algorithm was when  $n = 4$ . Assume we change the base case to  $n = 2$ : In this new base case, there are only two people  $P_1$  and  $P_2$ , and only one phone call is needed. The rest of the algorithm remains unchanged.

Prove that the modified algorithm  $\text{GOSSIP}(n)$  results in a sequence of  $2n - 3$  phone calls for any integer  $n \geq 2$ . (Thus, for  $n \geq 4$ , it makes one more phone call than the algorithm in Section 4.4.)

**4.53** In Section 4.4, we have seen the recursive algorithm  $\text{GOSSIP}(n)$ , which computes a sequence of phone calls for the persons  $P_1, P_2, \dots, P_n$ , for any integer  $n \geq 4$ .

Give an iterative (i.e., non-recursive) version of this algorithm in pseudocode. Your algorithm must produce exactly the same sequence of phone calls as algorithm  $\text{GOSSIP}(n)$ .

**4.54** In Section 4.5, we have seen algorithm  $\text{EUCLID}(a, b)$ , which takes as input two integers  $a$  and  $b$  with  $a \geq b \geq 1$ , and returns their greatest common divisor.

Assume we run algorithm  $\text{EUCLID}(a, b)$  with two input integers  $a$  and  $b$  that satisfy  $b > a \geq 1$ . What is the output of this algorithm?

**4.55** The following recursive algorithm  $\text{FIB}$  takes as input an integer  $n \geq 0$  and returns the  $n$ -th Fibonacci number  $f_n$ :

**Algorithm FIB( $n$ ):**

```

if  $n = 0$  or  $n = 1$ 
then  $f = n$ 
else  $f = \text{FIB}(n - 1) + \text{FIB}(n - 2)$ 
endif;
return  $f$ 
```

Let  $a_n$  be the number of additions made by algorithm FIB( $n$ ), i.e., the total number of times the  $+$ -function in the else-case is called. Prove that for all  $n \geq 0$ ,

$$a_n = f_{n+1} - 1.$$

**4.56** Consider the following recursive algorithm BEER( $n$ ), which takes as input an integer  $n \geq 1$ :

**Algorithm BEER( $n$ ):**

```

if  $n = 1$ 
then eat some peanuts
else choose an arbitrary integer  $m$  with  $1 \leq m \leq n - 1$ ;
    BEER( $m$ );
    drink one pint of beer;
    BEER( $n - m$ )
endif
```

- Explain why, for any integer  $n \geq 1$ , algorithm BEER( $n$ ) terminates.
- Let  $B(n)$  be the number of pints of beer you drink when running algorithm BEER( $n$ ). Determine the value of  $B(n)$ .

**4.57** Consider the following recursive algorithm SILLY, which takes as input an integer  $n \geq 1$  which is a power of 2:

**Algorithm SILLY( $n$ ):**

```
if  $n = 1$ 
then drink one pint of beer
else if  $n = 2$ 
    then fart once
    else fart once;
        SILLY( $n/2$ );
        fart once
    endif
endif
```

For  $n$  a power of 2, let  $F(n)$  be the number of times you fart when running algorithm SILLY( $n$ ). Determine the value of  $F(n)$ .

**4.58** In the fall term of 2015, Nick took the course COMP 2804 at Carleton University. Nick was always sitting in the back of the classroom and spent most of his time eating bananas. Nick uses the following scheme to buy bananas:

- At the start of week 0, there are 2 bananas in Nick's fridge.
- For any integer  $n \geq 0$ , Nick does the following during week  $n$ :
  - At the start of week  $n$ , Nick determines the number of bananas in his fridge and stores this number in a variable  $x$ .
  - Nick goes to Jim's Banana Empire, buys  $x$  bananas, and puts them in his fridge.
  - Nick takes  $n + 1$  bananas out of his fridge and eats them during week  $n$ .

For any integer  $n \geq 0$ , let  $B(n)$  be the number of bananas in Nick's fridge at the start of week  $n$ . Determine the value of  $B(n)$ .

**4.59** Jennifer loves to drink India Pale Ale (IPA). After a week of hard work, Jennifer goes to the pub and runs the following recursive algorithm, which takes as input an integer  $n \geq 1$ , which is a power of 4:

**Algorithm JENNIFERDRINKSIPA( $n$ ):**

```

if  $n = 1$ 
then place one order of chicken wings
else for  $k = 1$  to 4
    do JENNIFERDRINKSIPA( $n/4$ );
        drink  $n$  pints of IPA
    endfor
endif

```

For  $n$  a power of 4, let

- $P(n)$  be the number of pints of IPA that Jennifer drinks when running algorithm JENNIFERDRINKSIPA( $n$ ),
- $C(n)$  be the number of orders of chicken wings that Jennifer places when running algorithm JENNIFERDRINKSIPA( $n$ ).

Determine the values of  $P(n)$  and  $C(n)$ .

**4.60** Elisa Kazan loves to drink cider. During the weekend, Elisa goes to the pub and runs the following recursive algorithm, which takes as input an integer  $n \geq 0$ :

**Algorithm ELISADRINKSCIDER( $n$ ):**

```

if  $n = 0$ 
then order Fibonachos
else if  $n$  is even
    then ELISADRINKSCIDER( $n/2$ );
        drink  $n^2/2$  pints of cider;
        ELISADRINKSCIDER( $n/2$ )
    else for  $i = 1$  to 4
        do ELISADRINKSCIDER( $((n - 1)/2)$ );
            drink  $(n - 1)/2$  pints of cider
        endfor;
        drink 1 pint of cider
    endif
endif

```

For  $n \geq 0$ , let  $C(n)$  be the number of pints of cider that Elisa drinks when running algorithm ELISADRINKSCIDER( $n$ ). Determine the value of  $C(n)$ .

**4.61** Elisa Kazan loves to drink cider. After a week of bossing the Vice-Presidents around, Elisa goes to the pub and runs the following recursive algorithm, which takes as input an integer  $n \geq 0$ :

**Algorithm** ELISAGOESTOTHEPUB( $n$ ):

```
if  $n = 0$ 
then drink one bottle of cider
else for  $k = 0$  to  $n - 1$ 
    do ELISAGOESTOTHEPUB( $k$ );
        drink one bottle of cider
    endfor
endif
```

For  $n \geq 0$ , let  $C(n)$  be the number of bottles of cider that Elisa drinks when running algorithm ELISAGOESTOTHEPUB( $n$ ).

Prove that for every integer  $n \geq 1$ ,

$$C(n) = 3 \cdot 2^{n-1} - 1.$$

*Hint:*  $1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-2} = 2^{n-1} - 1$ .

**4.62** Elisa Kazan loves to drink cider. On Saturday night, Elisa goes to her neighborhood pub and runs the following recursive algorithm, which takes as input an integer  $n \geq 1$ :

**Algorithm** ELISADRINKSCIDER( $n$ ):

```

if  $n = 1$ 
then drink one pint of cider
else if  $n$  is even
    then ELISADRINKSCIDER( $n/2$ );
        drink one pint of cider;
        ELISADRINKSCIDER( $n/2$ )
    else drink one pint of cider;
        ELISADRINKSCIDER( $n - 1$ );
        drink one pint of cider
    endif
endif
```

For any integer  $n \geq 1$ , let  $P(n)$  be the number of pints of cider that Elisa drinks when running algorithm ELISADRINKSCIDER( $n$ ). Determine the value of  $P(n)$ .

**4.63** Let  $n \geq 2$  be an integer and consider a sequence  $s_1, s_2, \dots, s_n$  of  $n$  pairwise distinct numbers. The following algorithm computes the smallest and largest elements in this sequence:

**Algorithm** MINMAX( $s_1, s_2, \dots, s_n$ ):

```

min =  $s_1$ ;
max =  $s_1$ ;
for  $i = 2$  to  $n$ 
do if  $s_i < min$            (1)
    then min =  $s_i$ 
    endif;
    if  $s_i > max$            (2)
    then max =  $s_i$ 
    endif
endwhile;
return ( $min, max$ )
```

This algorithm makes comparisons between input elements in lines (1) and (2). Determine the total number of comparisons as a function of  $n$ .

**4.64** Let  $n \geq 2$  be a power of 2 and consider a sequence  $S$  of  $n$  pairwise distinct numbers. The following algorithm computes the smallest and largest elements in this sequence:

**Algorithm** FASTMINMAX( $S, n$ ):

```

if  $n = 2$ 
then let  $x$  and  $y$  be the two elements in  $S$ ;
    if  $x < y$           (1)
    then  $min = x$ ;
           $max = y$ 
    else  $min = y$ ;
           $max = x$ 
    endif
else divide  $S$  into two subsequences  $S_1$  and  $S_2$ , both of size  $n/2$ ;
     $(min_1, max_1) = \text{FASTMINMAX}(S_1, n/2)$ ;
     $(min_2, max_2) = \text{FASTMINMAX}(S_2, n/2)$ ;
    if  $min_1 < min_2$       (2)
    then  $min = min_1$ 
    else  $min = min_2$ 
    endif;
    if  $max_1 < max_2$       (3)
    then  $max = max_2$ 
    else  $max = max_1$ 
    endif
endif;
return  $(min, max)$ 
```

This algorithm makes comparisons between input elements in lines (1), (2), and (3). Let  $C(n)$  be the total number of comparisons made by algorithm FASTMINMAX on an input sequence of length  $n$ .

- Derive a recurrence relation for  $C(n)$ .
- Use this recurrence relation to prove that  $C(n) = \frac{3}{2}n - 2$  for each  $n \geq 2$  that is a power of 2.

**4.65** Consider the following recursive algorithm, which takes as input a sequence  $(a_1, a_2, \dots, a_n)$  of length  $n$ , where  $n \geq 1$ :

**Algorithm** MYSTERY( $a_1, a_2, \dots, a_n$ ):

```

if  $n = 1$ 
then return the sequence ( $a_1$ )
else  $(b_1, b_2, \dots, b_{n-1}) = \text{MYSTERY}(a_1, a_2, \dots, a_{n-1});$ 
      return the sequence ( $a_n, b_1, b_2, \dots, b_{n-1}$ )
endif
```

- Express the output of algorithm MYSTERY( $a_1, a_2, \dots, a_n$ ) in terms of the input sequence  $(a_1, a_2, \dots, a_n)$ .

**4.66** Consider the following recursive algorithm, which takes as input a sequence  $(a_1, a_2, \dots, a_n)$  of  $n$  numbers, where  $n$  is a power of two, i.e.,  $n = 2^k$  for some integer  $k \geq 0$ :

**Algorithm** MYSTERY( $a_1, a_2, \dots, a_n$ ):

```

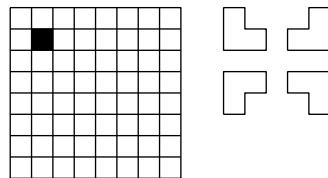
if  $n = 1$ 
then return  $a_1$ 
else for  $i = 1$  to  $n/2$ 
    do  $b_i = \min(a_{2i-1}, a_{2i})$           (*)
    endfor;
    MYSTERY( $b_1, b_2, \dots, b_{n/2}$ )
endif
```

- Express the output of algorithm MYSTERY( $a_1, a_2, \dots, a_n$ ) in terms of the input sequence  $(a_1, a_2, \dots, a_n)$ .
- For any integer  $n \geq 1$  that is a power of two, let  $T(n)$  be the total number of times that line (\*) is executed when running algorithm MYSTERY( $a_1, a_2, \dots, a_n$ ). Derive a recurrence for  $T(n)$  and use it to prove that for any integer  $n \geq 1$  that is a power of two,

$$T(n) = n - 1.$$

**4.67** Let  $k$  be a positive integer and let  $n = 2^k$ . You are given an  $n \times n$  board  $B_n$ , all of whose (square) cells are white, except for one, which is black. (The left part of the figure below gives an example where  $k = 3$  and  $n = 8$ .)

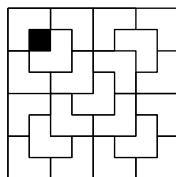
A *tromino* is an L-shaped object consisting of three  $1 \times 1$  cells. Each tromino can appear in four different orientations; see the right part of the figure below.



A *tiling* of the board  $B_n$  is a placement of trominoes on the board such that

- the trominoes cover exactly all white cells (thus, the black cell is not covered by any tromino) and
- no two trominoes overlap.

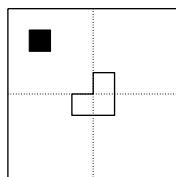
Here is a tiling of the board given above:



Describe a recursive algorithm that

- takes as input a board  $B_n$  having exactly one black cell (which can be anywhere on the board) and
- returns a tiling of this board.

*Hint:* Look at the following figure:

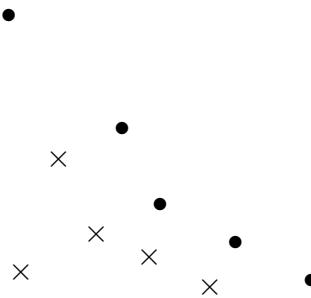


**4.68** Let  $n \geq 1$  be an integer and consider a set  $S$  consisting of  $n$  points in  $\mathbb{R}^2$ . Each point  $p$  of  $S$  is given by its  $x$ - and  $y$ -coordinates  $p_x$  and  $p_y$ , respectively. We assume that no two points of  $S$  have the same  $x$ -coordinate and no two points of  $S$  have the same  $y$ -coordinate.

A point  $p$  of  $S$  is called *maximal* in  $S$  if there is no point in  $S$  that is to the north-east of  $p$ , i.e.,

$$\{q \in S : q_x > p_x \text{ and } q_y > p_y\} = \emptyset.$$

The figure below shows an example, in which the  $\bullet$ -points are maximal and the  $\times$ -points are not maximal. Observe that, in general, there is more than one maximal element in  $S$ .



Describe a recursive algorithm MAXELEM that has the same structure as algorithm MERGESORT in Section 4.6 and has the following specification:

**Algorithm** MAXELEM( $S, n$ ):

**Input:** A set  $S$  of  $n$  points in  $\mathbb{R}^2$ , in sorted order of their  $x$ -coordinates.

**Output:** All maximal elements of  $S$ , in sorted order of their  $x$ -coordinates.

The running time of your algorithm must be  $O(n \log n)$ .

**4.69** The Hadamard matrices  $H_0, H_1, H_2, \dots$  are recursively defined as follows:

$$H_0 = (1)$$

and for  $k \geq 1$ ,

$$H_k = \left( \begin{array}{c|c} H_{k-1} & H_{k-1} \\ \hline H_{k-1} & -H_{k-1} \end{array} \right).$$

Thus,  $H_0$  is a  $1 \times 1$  matrix whose only entry is 1,

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

and

$$H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Observe that  $H_k$  has  $2^k$  rows and  $2^k$  columns.

If  $x$  is a column vector of length  $2^k$ , then  $H_k x$  is the column vector of length  $2^k$  obtained by multiplying the matrix  $H_k$  with the vector  $x$ .

Describe a recursive algorithm MULT that has the following specification:

**Algorithm** MULT( $k, x$ ):

**Input:** An integer  $k \geq 0$  and a column vector  $x$  of length  $n = 2^k$ .

**Output:** The column vector  $H_k x$  (having length  $n$ ).

The running time  $T(n)$  of your algorithm must be  $O(n \log n)$ .

*Hint:* The input only consists of  $k$  and  $x$ . The matrix  $H_k$ , which has  $n^2$  entries, is not given as part of the input. Since you are aiming for an  $O(n \log n)$ -time algorithm, you cannot compute all entries of the matrix  $H_k$ .

**4.70** Let  $m \geq 1$  and  $n \geq 1$  be integers and consider an  $m \times n$  matrix  $A$ . The rows of this matrix are numbered  $1, 2, \dots, m$ , and its columns are numbered  $1, 2, \dots, n$ . Each entry of  $A$  stores one number and, for each row, all numbers in this row are pairwise distinct. For each  $i = 1, 2, \dots, m$ , define

$g(i) =$  the position (i.e., column number) of the smallest number in row  $i$ .

We say that the matrix  $A$  is *awesome*, if

$$g(1) \leq g(2) \leq g(3) \leq \dots \leq g(m).$$

In the matrix below, the smallest number in each row is in boldface. For this example, we have  $m = 4$ ,  $n = 10$ ,  $g(1) = 3$ ,  $g(2) = 3$ ,  $g(3) = 5$ , and  $g(4) = 8$ . Thus, this matrix is awesome.

$$A = \begin{pmatrix} 13 & 12 & \mathbf{5} & 8 & 6 & 9 & 15 & 20 & 19 & 7 \\ 3 & 4 & \mathbf{1} & 17 & 6 & 13 & 7 & 10 & 2 & 5 \\ 19 & 5 & 12 & 7 & \mathbf{2} & 4 & 11 & 13 & 6 & 3 \\ 7 & 4 & 17 & 10 & 5 & 14 & 12 & \mathbf{3} & 20 & 6 \end{pmatrix}.$$

From now on, we assume that the  $m \times n$  matrix  $A$  is awesome.

- Let  $i$  be an integer with  $1 \leq i \leq m$ . Describe an algorithm that computes  $g(i)$  in  $O(n)$  time.
- Describe an algorithm that computes all values  $g(1), g(2), \dots, g(m)$  in  $O(mn)$  total time.

In the rest of this exercise, you will show that all values  $g(1), g(2), \dots, g(m)$  can be computed in  $O(m + n \log m)$  total time.

- Assume that  $m$  is even and assume that you are given the values

$$g(2), g(4), g(6), g(8), \dots, g(m).$$

Describe an algorithm that computes the values

$$g(1), g(3), g(5), g(7), \dots, g(m-1)$$

in  $O(m + n)$  total time.

- Assume that  $m = 2^k$ , i.e.,  $m$  is a power of two. Describe a recursive algorithm **FINDROWMINIMA** that has the following specification:

**Algorithm** FINDROWMINIMA( $A, i$ ):

**Input:** An  $m \times n$  awesome matrix  $A$  and an integer  $i$  with  $0 \leq i \leq k$ .

**Output:** The values  $g(j \cdot m/2^i)$  for  $j = 1, 2, 3, \dots, 2^i$ .

For each  $i$  with  $0 \leq i \leq k$ , let  $T(i)$  denote the running time of algorithm FINDROWMINIMA( $A, i$ ). The running time of your algorithm must satisfy the recurrence

$$\begin{aligned} T(0) &= O(n), \\ T(i) &= T(i-1) + O(2^i + n), \text{ if } 1 \leq i \leq k. \end{aligned}$$

- Assume again that  $m = 2^k$ . Prove that all values  $g(1), g(2), \dots, g(m)$  can be computed in  $O(m + n \log m)$  total time.

*Hint:*  $1 + 2 + 2^2 + 2^3 + \dots + 2^k \leq 2m$ .

**4.71** Prove, for example by induction, that for  $n \geq 1$ ,

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2},$$

and

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

**4.72** Assume you remember that

$$1^2 + 2^2 + 3^2 + \cdots + n^2$$

is equal to a polynomial of degree three, i.e.,

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = An^3 + Bn^2 + Cn + D,$$

but you have forgotten the values of  $A$ ,  $B$ ,  $C$ , and  $D$ . How can you determine these four values?

**4.73** In Section 4.8.3, we have shown that

$$\sum_{k=2}^{n-2} (k-1)(n-k-1) = \binom{n-1}{3}.$$

Use Exercise 4.71 to give an alternative proof.

**4.74** In Section 4.8.4, we have used the fact that

$$\sum_{k=1}^{n-1} k = \binom{n}{2},$$

which follows from Theorem 2.2.10. Give an alternative proof that uses the approach that we used to prove the identity in (4.17).

**4.75** In Section 4.8.4, we have shown that

$$\sum_{k=3}^{n-1} \binom{k}{3} = \binom{n}{4}.$$

Use induction and Pascal's Identity (see Theorem 3.7.2) to give an alternative proof.

**4.76** Consider the numbers  $R_n$  that we defined in Section 4.8. The  $n$  points on the circle define  $\binom{n}{2}$  line segments, one segment for each pair of points. Let  $X$  be the total number of intersections among these  $\binom{n}{2}$  line segments.

- Prove that

$$R_n = 1 + \binom{n}{2} + X.$$

*Hint:* Start with only the circle and the  $n$  points. Then add the  $\binom{n}{2}$  line segments one by one.

- Prove that

$$X = \binom{n}{4}.$$

**4.77** For an integer  $n \geq 1$ , draw  $n$  straight lines, such that no two of them are parallel and no three of them intersect in one single point. These lines divide the plane into regions (some of which are bounded and some of which are unbounded). Denote the number of these regions by  $C_n$ .

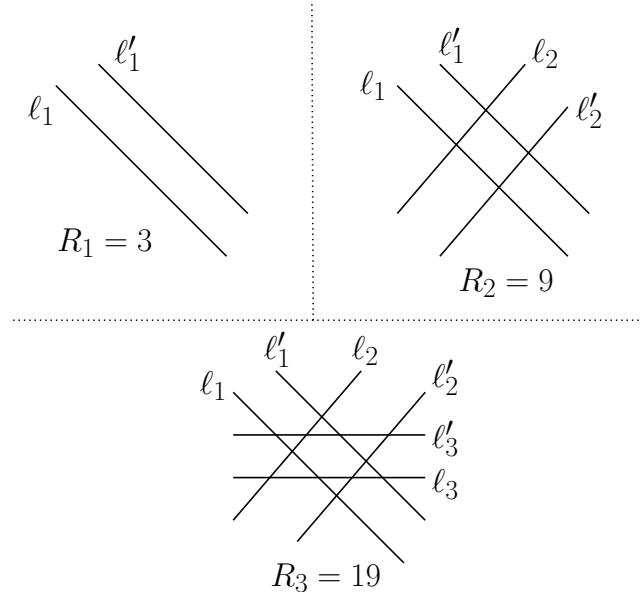
Derive a recurrence relation for the numbers  $C_n$  and use it to prove that for  $n \geq 1$ ,

$$C_n = 1 + \frac{n(n+1)}{2}.$$

**4.78** Let  $n \geq 1$  be an integer. Consider  $2n$  straight lines  $\ell_1, \ell'_1, \dots, \ell_n, \ell'_n$  such that

- for each  $i$  with  $1 \leq i \leq n$ ,  $\ell_i$  and  $\ell'_i$  are parallel,
- no two of the lines  $\ell_1, \dots, \ell_n$  are parallel,
- no two of the lines  $\ell'_1, \dots, \ell'_n$  are parallel,
- no three of the  $2n$  lines intersect in one single point.

These lines divide the plane into regions (some of which are bounded and some of which are unbounded). Denote the number of these regions by  $R_n$ . From the figure below, you can see that  $R_1 = 3$ ,  $R_2 = 9$ , and  $R_3 = 19$ .

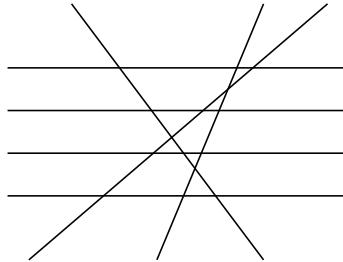


- Derive a recurrence relation for the numbers  $R_n$  and use it to prove that  $R_n = 2n^2 + 1$  for  $n \geq 1$ .

**4.79** Let  $m \geq 1$  and  $n \geq 1$  be integers. Consider  $m$  horizontal lines and  $n$  non-horizontal lines such that

- no two of the non-horizontal lines are parallel,
- no three of the  $m+n$  lines intersect in one single point.

These lines divide the plane into regions (some of which are bounded and some of which are unbounded). Denote the number of these regions by  $R_{m,n}$ . From the figure below, you can see that  $R_{4,3} = 23$ .



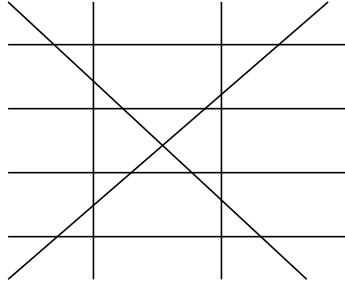
- Derive a recurrence relation for the numbers  $R_{m,n}$  and use it to prove that

$$R_{m,n} = 1 + m(n+1) + \binom{n+1}{2}.$$

**4.80** A line is called *slanted* if it is neither horizontal nor vertical. Let  $k \geq 1$ ,  $m \geq 1$ , and  $n \geq 0$  be integers. Consider  $k$  horizontal lines,  $m$  vertical lines, and  $n$  slanted lines, such that

- no two of the slanted lines are parallel,
- no three of the  $k+m+n$  lines intersect in one single point.

These lines divide the plane into regions (some of which are bounded and some of which are unbounded). Denote the number of these regions by  $R_{k,m,n}$ . From the figure below, you can see that  $R_{4,2,2} = 30$ .



- Prove that

$$R_{k,m,0} = (k+1)(m+1).$$

- Derive a recurrence relation for the numbers  $R_{k,m,n}$  and use it to prove that

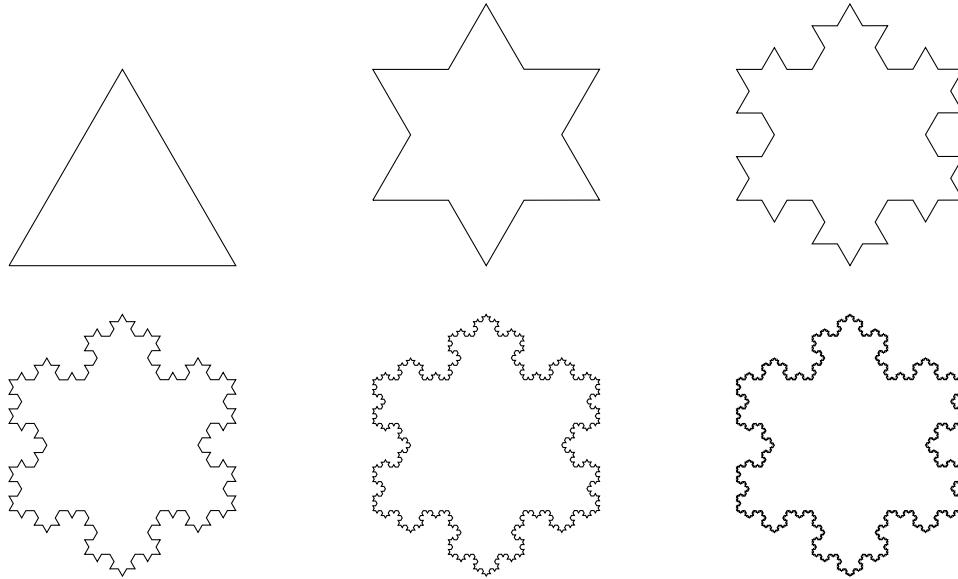
$$R_{k,m,n} = (k+1)(m+1) + (k+m)n + \binom{n+1}{2}.$$

**4.81** The sequence  $SF_0, SF_1, SF_2, \dots$  of *snowflakes* is recursively defined in the following way:

- The snowflake  $SF_0$  is an equilateral triangle with edges of length 1.
- For any integer  $n \geq 1$ , the snowflake  $SF_n$  is obtained by taking the snowflake  $SF_{n-1}$  and doing the following for each of its edges:

- Divide this edge into three edges of equal length.
- Draw an equilateral triangle that has the middle edge from the previous step as its base, and that is outside of  $SF_{n-1}$ .
- Remove the edge that is the base of the equilateral triangle from the previous step.

In the figure below, you see the snowflakes  $SF_0$  up to  $SF_5$ .



- For any integer  $n \geq 0$ , let  $N_n$  be the total number of edges of  $SF_n$ . Determine the value of  $N_n$ , by deriving a recurrence relation and solving it.
- For any integer  $n \geq 0$ , let  $\ell_n$  be the length of one single edge of  $SF_n$ . Determine the value of  $\ell_n$ , by deriving a recurrence relation and solving it.
- For any integer  $n \geq 0$ , let  $L_n$  be the total length of all edges of  $SF_n$ . Prove that
$$L_n = 3 \cdot \left(\frac{4}{3}\right)^n.$$
- Let  $a_0$  be the area of the triangle  $SF_0$ . For any integer  $n \geq 1$ , let  $a_n$  be the area of one single triangle that is added when constructing

$SF_n$  from  $SF_{n-1}$ . Determine the value of  $a_n$ , by deriving a recurrence relation and solving it.

- For any integer  $n \geq 1$ , let  $A_n$  be the total area of all triangles that are added when constructing  $SF_n$  from  $SF_{n-1}$ . Prove that

$$A_n = \frac{3}{4} \cdot \left(\frac{4}{9}\right)^n \cdot a_0.$$

- Let  $n \geq 0$  be an integer. Prove that the total area of  $SF_n$  is equal to

$$\frac{a_0}{5} \cdot \left(8 - 3 \cdot \left(\frac{4}{9}\right)^n\right).$$

*Hint:* For any real number  $x \neq 1$ ,

$$\sum_{k=1}^n x^k = x \cdot \frac{1-x^n}{1-x}.$$

# Chapter 5

## Discrete Probability

We all have some intuitive understanding of the notions of “chance” and “probability”. When buying a lottery ticket, we know that there is a chance of winning the jackpot, but we also know that this chance is very small. Before leaving home in the morning, we check the weather forecast and see that, with probability 80%, we get 15 centimetres of snow in Ottawa. In this chapter, we will give a formal definition of this notion of “probability”. We start by presenting a surprising application of probability and random numbers.

### 5.1 Anonymous Broadcasting

Consider a group of  $n$  people  $P_1, P_2, \dots, P_n$ , for some integer  $n \geq 3$ . One person in this group, say  $P_k$ , would like to broadcast, *anonymously*, a message to all other people in the group. That is,  $P_k$  wants to broadcast a message such that

- everybody in the group receives the message,
- nobody knows that the message was broadcast by  $P_k$ .

In other words, when  $P_i$  (with  $i \neq k$ ) receives the message, she only knows that it was broadcast by one of  $P_1, \dots, P_{i-1}, P_{i+1}, \dots, P_n$ ; she cannot determine who broadcast the message.

At first sight, it seems to be impossible to do this. In 1988, however, David Chaum published, in the Journal of Cryptology, a surprisingly simple

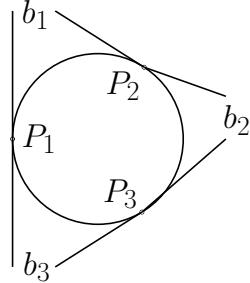
protocol that does achieve this. Chaum referred to the problem as the *Dining Cryptographers Problem*.

We will present and analyze the protocol for the case when  $n = 3$ . Thus, there are three people  $P_1$ ,  $P_2$ , and  $P_3$ . We assume that exactly one of them broadcasts a message and refer to this person as the *broadcaster*. We also assume that the message is a bitstring. The broadcaster will announce the message one bit at a time.

The three people  $P_1$ ,  $P_2$ , and  $P_3$  sit at a table, in clockwise order of their indices. Let  $b$  be the current bit that the broadcaster wants to announce. The protocol for broadcasting this bit is as follows:

**Step 1:** Each person  $P_i$  generates a *random* bit  $b_i$ , for example, by flipping a coin. Thus, with 50% probability,  $b_i = 0$  and with 50% probability,  $b_i = 1$ .

**Step 2:** Each person  $P_i$  shows the bit  $b_i$  to her clockwise neighbor.



At the end of this second step,

- $P_1$  knows  $b_1$  and  $b_3$ , but not  $b_2$ ,
- $P_2$  knows  $b_1$  and  $b_2$ , but not  $b_3$ ,
- $P_3$  knows  $b_2$  and  $b_3$ , but not  $b_1$ .

**Step 3:** Each person  $P_i$  computes the sum  $s_i$  (modulo 2) of the bits that she knows. Thus,

- $P_1$  computes  $s_1 = (b_1 + b_3) \bmod 2$ ,
- $P_2$  computes  $s_2 = (b_1 + b_2) \bmod 2$ ,
- $P_3$  computes  $s_3 = (b_2 + b_3) \bmod 2$ .

**Step 4:** Each person  $P_i$  does the following:

- If  $P_i$  is not the broadcaster, she sets  $t_i = s_i$ .
- If  $P_i$  is the broadcaster, she sets  $t_i = (s_i + b) \bmod 2$ . Recall that  $b$  is the current bit that the broadcaster wants to announce. (Thus, if  $b = 1$ , then  $P_i$  “secretly” flips the bit  $s_i$  and stores the result in  $t_i$ .)

**Step 5:** Each person  $P_i$  shows her bit  $t_i$  to the other two people.

**Step 6:** Each person  $P_i$  computes the sum (modulo 2) of the three bits  $t_1$ ,  $t_2$ , and  $t_3$ , i.e., the value  $(t_1 + t_2 + t_3) \bmod 2$ .

This concludes the description of the protocol for broadcasting one bit  $b$ . Observe that for any bit  $x$ , we have  $(x + x) \bmod 2 = 0$ . Therefore, the bit computed in the last step is equal to

$$\begin{aligned} t_1 + t_2 + t_3 &= s_1 + s_2 + s_3 + b \\ &= (b_1 + b_3) + (b_1 + b_2) + (b_2 + b_3) + b \\ &= (b_1 + b_1) + (b_2 + b_2) + (b_3 + b_3) + b \\ &= b, \end{aligned}$$

where all arithmetic is done modulo 2. In other words, the bit computed in the last step is equal to the bit that the broadcaster wants to announce. This shows that each person in the group receives this bit.

It remains to show that a non-broadcaster cannot determine who broadcast the bit. In the analysis below, we assume that

- $b = 1$ , i.e., the broadcaster announces the bit 1,
- $P_2$  is not the broadcaster.

We have to show that  $P_2$  cannot determine whether  $P_1$  or  $P_3$  is the broadcaster. Note that  $P_2$  knows the values

$$b_1, b_2, t_1, t_2, t_3,$$

but does not know the bit  $b_3$ . We consider the cases when  $b_1 = b_2$  and  $b_1 \neq b_2$  separately.

**Case 1:**  $b_1 = b_2$ . This case has two subcases depending on the value of  $b_3$ .

**Case 1.1:**  $b_3 = b_1$ ; thus, all three  $b$ -bits are equal.

- If  $P_1$  is the broadcaster, then (all arithmetic is done modulo 2)

$$t_1 = s_1 + 1 = b_1 + b_3 + 1 = 1$$

and

$$t_3 = s_3 = b_2 + b_3 = 0.$$

- If  $P_3$  is the broadcaster, then

$$t_1 = s_1 = b_1 + b_3 = 0$$

and

$$t_3 = s_3 + 1 = b_2 + b_3 + 1 = 1.$$

Thus, the broadcaster is the person whose  $t$ -bit is equal to 1.

**Case 1.2:**  $b_3 \neq b_1$  and, thus,  $b_3 \neq b_2$ .

- If  $P_1$  is the broadcaster, then

$$t_1 = s_1 + 1 = b_1 + b_3 + 1 = 0$$

and

$$t_3 = s_3 = b_2 + b_3 = 1.$$

- If  $P_3$  is the broadcaster, then

$$t_1 = s_1 = b_1 + b_3 = 1$$

and

$$t_3 = s_3 + 1 = b_2 + b_3 + 1 = 0.$$

Thus, the broadcaster is the person whose  $t$ -bit is equal to 0.

Since  $P_2$  knows  $b_1$  and  $b_2$ , she knows when Case 1 occurs. Since  $P_2$  does not know  $b_3$ , however, she cannot determine whether Case 1.1 or 1.2 occurs. As a result,  $P_2$  cannot determine whether  $P_1$  or  $P_3$  is the broadcaster.

**Case 2:**  $b_1 \neq b_2$ . This case has two subcases depending on the value of  $b_3$ .

**Case 2.1:**  $b_3 = b_1$  and, thus,  $b_3 \neq b_2$ .

- If  $P_1$  is the broadcaster, then

$$t_1 = s_1 + 1 = b_1 + b_3 + 1 = 1$$

and

$$t_3 = s_3 = b_2 + b_3 = 1.$$

- If  $P_3$  is the broadcaster, then

$$t_1 = s_1 = b_1 + b_3 = 0$$

and

$$t_3 = s_3 + 1 = b_2 + b_3 + 1 = 0.$$

Thus,  $t_1$  is always equal to  $t_3$ , no matter whether  $P_1$  or  $P_3$  is the broadcaster.

**Case 2.2:**  $b_3 \neq b_1$  and, thus,  $b_3 = b_2$ .

- If  $P_1$  is the broadcaster, then

$$t_1 = s_1 + 1 = b_1 + b_3 + 1 = 0$$

and

$$t_3 = s_3 = b_2 + b_3 = 0.$$

- If  $P_3$  is the broadcaster, then

$$t_1 = s_1 = b_1 + b_3 = 1$$

and

$$t_3 = s_3 + 1 = b_2 + b_3 + 1 = 1.$$

Thus,  $t_1$  is always equal to  $t_3$ , no matter whether  $P_1$  or  $P_3$  is the broadcaster.

Since  $P_2$  knows  $b_1$  and  $b_2$ , she knows when Case 2 occurs. Since  $P_2$  does not know  $b_3$ , however, she cannot determine whether Case 2.1 or 2.2 occurs. As in Case 1,  $P_2$  cannot determine whether  $P_1$  or  $P_3$  is the broadcaster.

We conclude from Cases 1 and 2 that the broadcasting of the bit  $b = 1$  is indeed anonymous. Now consider the case when the bit  $b$  to be announced is equal to 0. It follows from the protocol that in this case, there is no “secret bit flipping” done in Step 4 and all three people use the same rules to compute the  $s$ -values and the  $t$ -values. In this case,  $t_1 = s_1$ ,  $t_2 = s_2$ , and

$t_3 = s_3$ , and  $P_2$  can determine the bit  $b_3$ . She cannot, however, determine whether  $P_1$  or  $P_3$  is the broadcaster.

To conclude this section, we remark that for each bit to be announced, the entire protocol must be followed. That is, in each round of the protocol, one bit is broadcast and each person  $P_i$  must flip a coin to determine the bit  $b_i$  that is used in this round. We also remark that the protocol works only if exactly one person is the broadcaster.

## 5.2 Probability Spaces

In this section, we give a formal definition of the notion of “probability” in terms of sets and functions.

**Definition 5.2.1** A *sample space*  $S$  is a non-empty countable set. Each element of  $S$  is called an *outcome* and each subset of  $S$  is called an *event*.

In daily life, we express probabilities in terms of percentages. For example, the weather forecast may tell us that, with 80% probability, we will be getting a snowstorm today. In probability theory, probabilities are expressed in terms of numbers in the interval  $[0, 1]$ . A probability of 80% becomes a probability of 0.8.

**Definition 5.2.2** Let  $S$  be a sample space. A *probability function* on  $S$  is a function  $\Pr : S \rightarrow \mathbb{R}$  such that

- for all  $\omega \in S$ ,  $0 \leq \Pr(\omega) \leq 1$ , and
- $\sum_{\omega \in S} \Pr(\omega) = 1$ .

For any outcome  $\omega$  in the sample space  $S$ , we will refer to  $\Pr(\omega)$  as the probability that the outcome is equal to  $\omega$ .

**Definition 5.2.3** A *probability space* is a pair  $(S, \Pr)$ , where  $S$  is a sample space and  $\Pr : S \rightarrow \mathbb{R}$  is a probability function on  $S$ .

A probability function  $\Pr : S \rightarrow \mathbb{R}$  maps each element of the sample space  $S$  (i.e., each outcome) to a real number in the interval  $[0, 1]$ . It turns

out to be useful to extend this function so that it maps any event to a real number in  $[0, 1]$ . If  $A$  is an event (i.e.,  $A \subseteq S$ ), then we define

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega). \quad (5.1)$$

We will refer to  $\Pr(A)$  as the probability that the event  $A$  occurs.

Note that since  $S \subseteq S$ , the entire sample space  $S$  is an event and

$$\Pr(S) = \sum_{\omega \in S} \Pr(\omega) = 1,$$

where the last equality follows from the second condition in Definition 5.2.2.

### 5.2.1 Examples

**Flipping a coin:** Assume we flip a coin. Since there are two possible outcomes (the coin comes up either *heads* ( $H$ ) or *tails* ( $T$ )), the sample space is the set  $S = \{H, T\}$ . If the coin is *fair*, i.e., the probabilities of  $H$  and  $T$  are equal, then the probability function  $\Pr : S \rightarrow \mathbb{R}$  is given by

$$\begin{aligned}\Pr(H) &= 1/2, \\ \Pr(T) &= 1/2.\end{aligned}$$

Observe that this function  $\Pr$  satisfies the two conditions in Definition 5.2.2. Since this sample space has two elements, there are four events, one event for each subset. These events are

$$\emptyset, \{H\}, \{T\}, \{H, T\},$$

and it follows from (5.1) that

$$\begin{aligned}\Pr(\emptyset) &= 0, \\ \Pr(\{H\}) &= \Pr(H) = 1/2, \\ \Pr(\{T\}) &= \Pr(T) = 1/2, \\ \Pr(\{H, T\}) &= \Pr(H) + \Pr(T) = 1/2 + 1/2 = 1.\end{aligned}$$

**Flipping a coin twice:** If we flip a fair coin twice, then there are four possible outcomes, and the sample space becomes  $S = \{HH, HT, TH, TT\}$ . For example,  $HT$  indicates that the first flip resulted in heads, whereas the

second flip resulted in tails. In this case, the probability function  $\Pr : S \rightarrow \mathbb{R}$  is given by

$$\Pr(HH) = \Pr(HT) = \Pr(TH) = \Pr(TT) = 1/4.$$

Observe again that this function  $\Pr$  satisfies the two conditions in Definition 5.2.2. Since the sample space consists of 4 elements, the number of events is equal to  $2^4 = 16$ . For example,  $A = \{HT, TH\}$  is an event and it follows from (5.1) that

$$\Pr(A) = \Pr(HT) + \Pr(TH) = 1/4 + 1/4 = 1/2.$$

In words, when flipping a fair coin twice, the probability that we see one heads and one tails (without specifying the order) is equal to 1/2.

**Rolling a die twice:** If we roll a fair die, then there are six possible outcomes (1, 2, 3, 4, 5, and 6), each one occurring with probability 1/6. If we roll this die twice, we obtain the sample space

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\},$$

where  $i$  is the result of the first roll and  $j$  is the result of the second roll. Note that  $|S| = 6 \times 6 = 36$ . Since the die is fair, each outcome has the same probability. Therefore, in order to satisfy the two conditions in Definition 5.2.2, we must have

$$\Pr(i, j) = 1/36$$

for each outcome  $(i, j)$  in  $S$ .

If we are interested in the sum of the results of the two rolls, then we define the event

$$A_k = \text{"the sum of the results of the two rolls is equal to } k\text{"},$$

which, using the notation of sets, is the same as

$$A_k = \{(i, j) \in S : i + j = k\}.$$

Consider, for example, the case when  $k = 4$ . There are three possible outcomes of two rolls that result in a sum of 4. These outcomes are (1, 3), (2, 2), and (3, 1). Thus, the event  $A_4$  is equal to

$$A_4 = \{(1, 3), (2, 2), (3, 1)\}. \quad (5.2)$$

In the matrix below, the leftmost column indicates the result of the first roll, the top row indicates the result of the second roll, and each entry is the sum of the results of the two corresponding rolls.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

As can be seen from this matrix, the event  $A_k$  is non-empty only if  $k \in \{2, 3, \dots, 12\}$ . For any other  $k$ , the event  $A_k$  is empty, which means that it can never occur.

It follows from (5.1) that

$$\Pr(A_k) = \sum_{(i,j) \in A_k} \Pr(i,j) = \sum_{(i,j) \in A_k} 1/36 = |A_k|/36.$$

For example, the number 4 occurs three times in the matrix and, therefore, the event  $A_4$  has size three. Observe that we have already seen this in (5.2). It follows that

$$\Pr(A_4) = |A_4|/36 = 3/36 = 1/12.$$

In a similar way, we see that

$$\begin{aligned} \Pr(A_2) &= 1/36, \\ \Pr(A_3) &= 2/36 = 1/18, \\ \Pr(A_4) &= 3/36 = 1/12, \\ \Pr(A_5) &= 4/36 = 1/9, \\ \Pr(A_6) &= 5/36, \\ \Pr(A_7) &= 6/36 = 1/6, \\ \Pr(A_8) &= 5/36, \\ \Pr(A_9) &= 4/36 = 1/9, \\ \Pr(A_{10}) &= 3/36 = 1/12, \\ \Pr(A_{11}) &= 2/36 = 1/18, \\ \Pr(A_{12}) &= 1/36. \end{aligned}$$

A sample space is not necessarily uniquely defined. In the last example, where we were interested in the sum of the results of two rolls of a die, we could also have taken the sample space to be the set

$$S' = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

The probability function  $\Pr'$  corresponding to this sample space  $S'$  is given by

$$\Pr'(k) = \Pr(A_k),$$

because  $\Pr'(k)$  is the probability that we get the outcome  $k$  in the sample space  $S'$ , which is the same as the probability that event  $A_k$  occurs in the sample space  $S$ . You should verify that this function  $\Pr'$  satisfies the two conditions in Definition 5.2.2 and, thus, is a valid probability function on  $S'$ .

### 5.3 Basic Rules of Probability

In this section, we prove some basic properties of probability functions. As we will see, all these properties follow from Definition 5.2.2. Throughout this section,  $(S, \Pr)$  is a probability space.

Recall that an event is a subset of the sample space  $S$ . In particular, the empty set  $\emptyset$  is an event. Intuitively,  $\Pr(\emptyset)$  must be zero, because it is the probability that there is no outcome, which can never occur. The following lemma states that this is indeed the case.

**Lemma 5.3.1**  $\Pr(\emptyset) = 0$ .

**Proof.** By (5.1), we have

$$\Pr(\emptyset) = \sum_{\omega \in \emptyset} \Pr(\omega).$$

Since there are zero terms in this summation, its value is equal to zero. ■

We say that two events  $A$  and  $B$  are *disjoint*, if  $A \cap B = \emptyset$ . A sequence  $A_1, A_2, \dots, A_n$  of events is *pairwise disjoint*, if any pair in this sequence is disjoint. The following lemma is similar to the Sum Rule of Section 3.4.

**Lemma 5.3.2** *If  $A_1, A_2, \dots, A_n$  is a sequence of pairwise disjoint events, then*

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n \Pr(A_i).$$

**Proof.** Let  $A = A_1 \cup A_2 \cup \dots \cup A_n$ . Using (5.1), we have

$$\begin{aligned}\Pr(A) &= \sum_{\omega \in A} \Pr(\omega) \\ &= \sum_{i=1}^n \sum_{\omega \in A_i} \Pr(\omega) \\ &= \sum_{i=1}^n \Pr(A_i).\end{aligned}$$

■

To give an example, assume we roll a fair die twice. What is the probability that the sum of the two results is even? If you look at the matrix in Section 5.2, then you see that there are 18 entries, out of 36, that are even. Therefore, the probability of having an even sum is equal to  $18/36 = 1/2$ . Below we will give a different way to determine this probability.

The sample space is the set

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\},$$

where  $i$  is the result of the first roll and  $j$  is the result of the second roll. Each element of  $S$  has the same probability  $1/36$  of being an outcome of rolling the die twice.

The event we are interested in is

$$A = \{(i, j) \in S : i + j \text{ is even}\}.$$

Observe that  $i + j$  is even if and only if both  $i$  and  $j$  are even or both  $i$  and  $j$  are odd. Therefore, we split the event  $A$  into two disjoint events

$$A_1 = \{(i, j) \in S : \text{both } i \text{ and } j \text{ are even}\}$$

and

$$A_2 = \{(i, j) \in S : \text{both } i \text{ and } j \text{ are odd}\}.$$

By Lemma 5.3.2, we have

$$\Pr(A) = \Pr(A_1) + \Pr(A_2).$$

The set  $A_1$  has  $3 \cdot 3 = 9$  elements, because there are 3 choices for  $i$  and 3 choices for  $j$ . Similarly, the set  $A_2$  has 9 elements. It follows that

$$\Pr(A) = \Pr(A_1) + \Pr(A_2) = 9/36 + 9/36 = 1/2.$$

In the next lemma, we relate the probability that an event occurs to the probability that the event does not occur. If  $A$  is an event, then  $\bar{A}$  denotes its *complement*, i.e.,  $\bar{A} = S \setminus A$ . Intuitively, the sum of  $\Pr(A)$  and  $\Pr(\bar{A})$  must be equal to one, because the event  $A$  either occurs or does not occur. The following lemma states that this is indeed the case. Observe that this is similar to the Complement Rule of Section 3.3.

**Lemma 5.3.3** *For any event  $A$ ,*

$$\Pr(A) = 1 - \Pr(\bar{A}).$$

**Proof.** Since  $A$  and  $\bar{A}$  are disjoint and  $S = A \cup \bar{A}$ , it follows from Lemma 5.3.2 that

$$\Pr(S) = \Pr(A \cup \bar{A}) = \Pr(A) + \Pr(\bar{A}).$$

We have seen in Section 5.2 that  $\Pr(S) = 1$ . ■

Consider again the sample space that we saw after Lemma 5.3.2. We showed that, when rolling a fair die twice, we get an even sum with probability  $1/2$ . It follows from Lemma 5.3.3 that we get an odd sum with probability  $1 - 1/2 = 1/2$ .

The next lemma is similar to the Principle of Inclusion and Exclusion that we have seen in Section 3.5.

**Lemma 5.3.4** *If  $A$  and  $B$  are events, then*

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

**Proof.** Since  $B \setminus A$  and  $A \cap B$  are disjoint and  $B = (B \setminus A) \cup (A \cap B)$ , it follows from Lemma 5.3.2 that

$$\Pr(B) = \Pr(B \setminus A) + \Pr(A \cap B).$$

Next observe that  $A$  and  $B \setminus A$  are disjoint. Since  $A \cup B = A \cup (B \setminus A)$ , we again apply Lemma 5.3.2 and obtain

$$\Pr(A \cup B) = \Pr(A) + \Pr(B \setminus A).$$

By combining these two equations, we obtain the claim in the lemma. ■

To give an example, assume we choose a number  $x$  in the sample space  $S = \{1, 2, \dots, 1000\}$ , such that each element has the same probability  $1/1000$  of being chosen. What is the probability that  $x$  is divisible by 2 or 3? Consider the events

$$A = \{i \in S : i \text{ is divisible by } 2\}$$

and

$$B = \{i \in S : i \text{ is divisible by } 3\}.$$

Then we want to determine  $\Pr(A \cup B)$ , which, by Lemma 5.3.4, is equal to

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Since there are  $\lfloor 1000/2 \rfloor = 500$  even numbers in  $S$ , we have

$$\Pr(A) = 500/1000.$$

Since there are  $\lfloor 1000/3 \rfloor = 333$  elements in  $S$  that are divisible by 3, we have

$$\Pr(B) = 333/1000.$$

Observe that  $i$  belongs to  $A \cap B$  if and only if  $i$  is divisible by 6, i.e.,

$$A \cap B = \{i \in S : i \text{ is divisible by } 6\}.$$

Since there are  $\lfloor 1000/6 \rfloor = 166$  elements in  $S$  that are divisible by 6, we have

$$\Pr(A \cap B) = 166/1000.$$

We conclude that

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= 500/1000 + 333/1000 - 166/1000 \\ &= 667/1000. \end{aligned}$$

**Lemma 5.3.5 (Union Bound)** *For any integer  $n \geq 1$ , if  $A_1, A_2, \dots, A_n$  is a sequence of events, then*

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n \Pr(A_i).$$

**Proof.** The proof is by induction on  $n$ . If  $n = 1$ , we have equality and, thus, the claim obviously holds. Let  $n \geq 2$  and assume the claim is true for  $n - 1$ , i.e., assume that

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_{n-1}) \leq \sum_{i=1}^{n-1} \Pr(A_i).$$

Let  $B = A_1 \cup A_2 \cup \dots \cup A_{n-1}$ . Then it follows from Lemma 5.3.4 that

$$\Pr(B \cup A_n) = \Pr(B) + \Pr(A_n) - \Pr(B \cap A_n) \leq \Pr(B) + \Pr(A_n),$$

because  $\Pr(B \cap A_n) \geq 0$  (this follows from the first condition in Definition 5.2.2). Since we assumed that

$$\Pr(B) \leq \sum_{i=1}^{n-1} \Pr(A_i),$$

it follows that

$$\begin{aligned} \Pr(A_1 \cup A_2 \cup \dots \cup A_n) &= \Pr(B \cup A_n) \\ &\leq \Pr(B) + \Pr(A_n) \\ &\leq \sum_{i=1}^{n-1} \Pr(A_i) + \Pr(A_n) \\ &= \sum_{i=1}^n \Pr(A_i). \end{aligned}$$

■

**Lemma 5.3.6** *If  $A$  and  $B$  are events with  $A \subseteq B$ , then*

$$\Pr(A) \leq \Pr(B).$$

**Proof.** Using (5.1) and the fact that  $\Pr(\omega) \geq 0$  for each  $\omega$  in  $S$ , we have

$$\begin{aligned}\Pr(A) &= \sum_{\omega \in A} \Pr(\omega) \\ &\leq \sum_{\omega \in B} \Pr(\omega) \\ &= \Pr(B).\end{aligned}$$

■

## 5.4 Uniform Probability Spaces

In this section, we consider finite sample spaces  $S$  in which each outcome has the same probability. Since, by Definition 5.2.2, all probabilities add up to one, the probability of each outcome must be equal to  $1/|S|$ .

**Definition 5.4.1** A *uniform probability space* is a pair  $(S, \Pr)$ , where  $S$  is a finite sample space and the probability function  $\Pr : S \rightarrow \mathbb{R}$  satisfies

$$\Pr(\omega) = \frac{1}{|S|},$$

for each outcome  $\omega$  in  $S$ .

The probability spaces that we have seen in Section 5.2.1 are all uniform, except the space  $(S', \Pr')$  that we saw at the end of that section.

To give another example, when playing Lotto 6/49, you choose a 6-element subset of the set  $A = \{1, 2, \dots, 49\}$ . Twice a week, the Ontario Lottery and Gaming Corporation (OLG) draws the six “winning numbers” uniformly at random from  $A$ . If your numbers are equal to those drawn by OLG, then you can withdraw from this course and spend the rest of your life on the beach. Most people find it silly to choose the subset  $\{1, 2, 3, 4, 5, 6\}$ . They think that it is better to choose, for example, the subset  $\{2, 5, 16, 36, 41, 43\}$ . Is this true?

For this example, the sample space is the set  $S$  consisting of all 6-element subsets of  $A$ . Since  $S$  has size  $\binom{49}{6}$  and the subset drawn by OLG is uniform, each outcome (i.e., each 6-element subset of  $S$ ) has a probability of

$$\frac{1}{\binom{49}{6}} = \frac{1}{13,983,816} \approx 0.000000072.$$

In particular, both  $\{1, 2, 3, 4, 5, 6\}$  and  $\{2, 5, 16, 36, 41, 43\}$  have the *same* probability of being the winning numbers. (Still, the latter subset was drawn by OLG on February 8, 2014.)

The lemma below states that in a uniform probability space  $(S, \Pr)$ , the probability of an event  $A$  is the ratio of the size of  $A$  and the size of  $S$ .

**Lemma 5.4.2** *If  $(S, \Pr)$  is a uniform probability space and  $A$  is an event, then*

$$\Pr(A) = \frac{|A|}{|S|}.$$

**Proof.** By using (5.1) and Definition 5.4.1, we get

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega) = \sum_{\omega \in A} \frac{1}{|S|} = \frac{1}{|S|} \sum_{\omega \in A} 1 = \frac{|A|}{|S|}.$$

■

### 5.4.1 The Probability of Getting a Full House

In a standard deck of 52 cards, each card has a *suit* and a *rank*. There are four suits (spades ♠, hearts ♥, clubs ♣, and diamonds ♦), and 13 ranks (Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King).

A hand of five cards is called a *full house*, if three of the cards are of the same rank and the other two cards are also of the same (but necessarily different) rank. For example, the hand

$$7\spadesuit, 7\heartsuit, 7\diamondsuit, Q\spadesuit, Q\clubsuit$$

is a full house, because it consists of three sevens and two Queens.

Assume we get a uniformly random hand of five cards. What is the probability that this hand is a full house? To answer this question, first observe that a hand of five cards is a *subset* of the set of all 52 cards. Thus, the sample space is the set  $S$  consisting of all 5-element subsets of the set of 52 cards and, therefore,

$$|S| = \binom{52}{5} = 2,598,960.$$

Each hand of five cards has a probability of  $1/|S|$  of being chosen.

Since we are interested in the probability of a random hand being a full house, we define the event  $A$  to be the set of all elements in  $S$  that are full houses. By Lemma 5.4.2, we have

$$\Pr(A) = \frac{|A|}{|S|}.$$

Thus, to determine  $\Pr(A)$ , it remains to determine the size of the set  $A$ , i.e., the total number of full houses. For this, we will use the Product Rule of Section 3.1:

- The procedure is “choose a full house”.
- First task: Choose the rank of the three cards in the full house. There are 13 ways to do this.
- Second task: Choose the suits of these three cards. There are  $\binom{4}{3}$  ways to do this.
- Third task: Choose the rank of the other two cards in the full house. There are 12 ways to do this.
- Fourth task: Choose the suits of these two cards. There are  $\binom{4}{2}$  ways to do this.

Thus, the number of full houses is equal to

$$|A| = 13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2} = 3,744.$$

We conclude that the probability of getting a full house is equal to

$$\Pr(A) = \frac{|A|}{|S|} = \frac{3,744}{2,598,960} \approx 0.00144.$$

## 5.5 The Birthday Paradox

Let  $n \geq 2$  be an integer and consider a group of  $n$  people. In this section, we will determine the probability  $p_n$  that at least two of them have the same birthday. We will ignore leap years, so that there are 365 days in one year.

Below, we will show that  $p_2 = 1/365$ . If  $n \geq 366$ , then it follows from the Pigeonhole Principle (see Section 3.10) that there must be at least two people with the same birthday and, therefore,  $p_n = 1$ . Intuitively, if  $n$  increases from 2 to 365, the value of  $p_n$  increases as well. What is the value of  $n$  such that  $p_n$  is larger than  $1/2$  for the first time? That is, what is the value of  $n$  for which  $p_{n-1} \leq 1/2 < p_n$ ? In this section, we will see that this question can be answered using simple counting techniques that we have seen in Chapter 3.

We denote the people by  $P_1, P_2, \dots, P_n$ , we denote the number of days in one year by  $d$ , and we number the days in one year as  $1, 2, \dots, d$ . The sample space is the set

$$S_n = \{(b_1, b_2, \dots, b_n) : b_i \in \{1, 2, \dots, d\} \text{ for each } 1 \leq i \leq n\},$$

where  $b_i$  denotes the birthday of  $P_i$ . Note that

$$|S_n| = d^n.$$

We consider the uniform probability space: For each element  $(b_1, b_2, \dots, b_n)$  in  $S_n$ , we have

$$\Pr(b_1, b_2, \dots, b_n) = \frac{1}{|S_n|} = \frac{1}{d^n}.$$

The event we are interested in is

$$A_n = \text{"at least two of the numbers in } b_1, b_2, \dots, b_n \text{ are equal".}$$

Using the notation of sets, this is the same as

$$A_n = \{(b_1, b_2, \dots, b_n) \in S_n : b_1, b_2, \dots, b_n \text{ contains duplicates}\}.$$

The probability  $p_n$  that we introduced above is equal to

$$p_n = \Pr(A_n).$$

As mentioned above, the Pigeonhole Principle implies that  $p_n = 1$  for  $n > d$ . Therefore, we assume from now on that  $n \leq d$ .

Let us start by determining  $p_2$ . Since we consider the uniform probability space, we have, by Lemma 5.4.2,

$$p_2 = \Pr(A_2) = \frac{|A_2|}{|S_2|}.$$

We know already that  $|S_2| = d^2$ . The event  $A_2$  is equal to

$$A_2 = \{(1, 1), (2, 2), \dots, (d, d)\}.$$

Thus,  $|A_2| = d$  and we obtain

$$p_2 = \frac{|A_2|}{|S_2|} = \frac{d}{d^2} = \frac{1}{d}.$$

To determine  $p_n$  for larger values of  $n$ , it is easier to determine the probability of the complement  $\overline{A}_n$ . The latter probability, together with Lemma 5.3.3, will give us the value of  $p_n$ . Note that

$$\overline{A}_n = \{(b_1, b_2, \dots, b_n) \in S_n : b_1, b_2, \dots, b_n \text{ are pairwise distinct}\}.$$

In words,  $\overline{A}_n$  is the set of all ordered sequences consisting of  $n$  pairwise distinct elements of  $\{1, 2, \dots, d\}$ . In Section 3.6, see (3.1), we have seen that

$$|\overline{A}_n| = \frac{d!}{(d-n)!}.$$

We conclude that, for any  $n$  with  $2 \leq n \leq d$ ,

$$\begin{aligned} p_n &= \Pr(A_n) \\ &= 1 - \Pr(\overline{A}_n) \\ &= 1 - \frac{|\overline{A}_n|}{|S_n|} \\ &= 1 - \frac{d!}{(d-n)!d^n}. \end{aligned}$$

By taking  $d = 365$ , we get  $p_{22} = 0.476$  and  $p_{23} = 0.507$ . Thus, in a random group of 23 people<sup>1</sup>, the probability that at least two of them have the same birthday is more than 50%. Most people are very surprised when they see this for the first time, because our intuition says that a much larger group is needed to have a probability of more than 50%. The values  $p_n$  approach 1 pretty fast. For example,  $p_{40} = 0.891$  and  $p_{100} = 0.9999997$ .

---

<sup>1</sup>two soccer teams plus the referee

### 5.5.1 Throwing Balls into Boxes

When we derived the expression for  $p_n$ , we did not use the fact that the value of  $d$  is equal to 365. In other words, the expression is valid for any value of  $d$ . For general values of  $d$ , we can interpret the birthday problem in the following way: Consider  $d$  boxes  $B_1, B_2, \dots, B_d$ , where  $d$  is a large integer. Assume that we throw  $n$  balls into these boxes so that each ball lands in a uniformly random box. Then  $p_n$  is the probability that there is at least one box that contains more than one ball. Since it is not easy to see how the expression

$$p_n = 1 - \frac{d!}{(d-n)!d^n}$$

depends on  $n$ , we will approximate it by a simpler expression. For this, we will use the inequality

$$1 - x \leq e^{-x}, \quad (5.3)$$

which is valid for any real number  $x$ . If  $x$  is close to zero, then the inequality is very accurate. The easiest way to prove this inequality is by showing that the minimum of the function  $f(x) = x + e^{-x}$  is equal to  $f(0) = 1$ , using techniques from calculus.

If we define  $q_n = 1 - p_n$ , then we have

$$q_n = \frac{d!}{(d-n)!d^n}.$$

Using (5.3), we get

$$\begin{aligned} q_n &= \frac{d}{d} \cdot \frac{d-1}{d} \cdot \frac{d-2}{d} \cdot \frac{d-3}{d} \cdots \frac{d-(n-1)}{d} \\ &= \frac{d-1}{d} \cdot \frac{d-2}{d} \cdot \frac{d-3}{d} \cdots \frac{d-(n-1)}{d} \\ &= (1 - 1/d) \cdot (1 - 2/d) \cdot (1 - 3/d) \cdots (1 - (n-1)/d) \\ &\leq e^{-1/d} \cdot e^{-2/d} \cdot e^{-3/d} \cdots e^{-(n-1)/d} \\ &= e^{-(1+2+3+\cdots+(n-1))/d}. \end{aligned}$$

Using the equality

$$1 + 2 + 3 + \cdots + (n-1) = n(n-1)/2,$$

see Theorem 2.2.10, we thus get

$$q_n \leq e^{-n(n-1)/(2d)},$$

and therefore,

$$p_n = 1 - q_n \geq 1 - e^{-n(n-1)/(2d)}.$$

If  $n$  is large, then  $n(n-1)/(2d)$  is very close to  $n^2/(2d)$  and, thus,

$$p_n \gtrsim 1 - e^{-n^2/(2d)}.$$

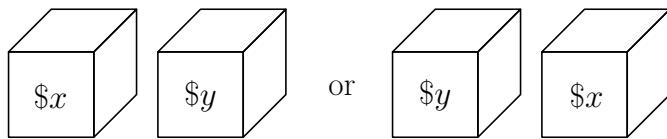
If we take  $n = \sqrt{2d}$ , then we get

$$p_n \gtrsim 1 - e^{-1} \approx 0.632.$$

Thus, for large values of  $d$ , if we throw  $\sqrt{2d}$  balls into  $d$  boxes, then with probability (approximately) at least  $1 - 1/e$ , there is a box that contains more than one ball.

## 5.6 The Big Box Problem

Keith chooses two distinct elements  $x$  and  $y$ , with  $x < y$ , from the set  $A = \{0, 1, 2, \dots, 100\}$ ; he does *not* show these two numbers to us. He takes two identical boxes, and puts  $x$  dollars in one box and  $y$  dollars in the other box. Then Keith closes the two boxes, shuffles them, and puts them on a table. At this moment, we can see the two boxes, they look identical to us, and the only information we have is that they contain different amounts of money, where each amount is an element of the set  $A$ .



We will refer to the box containing  $x$  dollars as the *small box* and to the box containing  $y$  dollars as the *big box*. Our goal is to find the big box. We are allowed to do the following:

1. We can choose one of the two boxes, open it, and determine how much money is inside it.
2. Now we have to make our final decision: Either we keep the box we just opened or we take the other box.

For example, assume that the box we pick in the first step contains \$33. Then we know that the other box contains either less than \$33 or more than \$33. It seems that the only reasonable thing to do is to flip a fair coin when making our final decision. If we do that, then we find the big box with probability 0.5.

In the rest of this section, we will show the surprising result that we can find the big box with probability at least 0.505.

The idea is as follows. *Assume* that we know a number  $z$  such that  $x < z < y$ . (Keep in mind that we do not know  $x$  and we do not know  $y$ . Thus, we assume that we know a number  $z$  that is between the two unknown numbers  $x$  and  $y$ .)

- If the box we choose in the first step contains more than  $z$  dollars, then we know that this is the big box and, therefore, we keep it.
- If the box we choose in the first step contains less than  $z$  dollars, then we know that this is the small box and, therefore, we take the other box.

Thus, if we know this number  $z$  with  $x < z < y$ , then we are guaranteed to find the big box.

Of course, it is not realistic to assume that we know this magic number  $z$ . The trick is to choose a *random*  $z$  and *hope* that it is between  $x$  and  $y$ . If  $z$  is between  $x$  and  $y$ , then we find the big box with probability 1; otherwise, we find the big box with probability 1/2. As we will see later, the overall probability of finding the big box will be at least 0.505.

In order to avoid the case when  $z = x$  or  $z = y$ , we will choose  $z$  from the set

$$B = \{1/2, 3/2, 5/2, \dots, 100 - 1/2\}.$$

Note that  $|B| = 100$ . Our algorithm that attempts to find the big box does the following:

**Algorithm FINDBIGBOX:**

**Step 1:** Choose one of the two boxes uniformly at random, open it, and determine the amount of money inside it; let this amount be  $a$ .

**Step 2:** Choose  $z$  uniformly at random from the set  $B$ .

**Step 3:** Do the following:

- If  $a > z$ , then keep the box chosen in Step 1.
- Otherwise (i.e., if  $a < z$ ), take the other box.

### 5.6.1 The Probability of Finding the Big Box

We are now going to determine the probability that this algorithm finds the big box. First, we have to ask ourselves what the sample space is. There are two places in the algorithm where a random element is chosen:

- In Step 1, we choose the element  $a$ , which is a random element from the set  $\{x, y\}$ . We know that this value  $a$  is equal to one of  $x$  and  $y$ . However, at the end of Step 1, we do not know whether  $a = x$  or  $a = y$ .
- In Step 2, we choose a random element from the set  $B$ .

Based on this, the sample space  $S$  is the Cartesian product

$$S = \{x, y\} \times B = \{(a, z) : a \in \{x, y\}, z \in B\}$$

and Steps 1 and 2 can be replaced by

- choose a uniformly random element  $(a, z)$  in  $S$ .

Note that  $|S| = 200$ .

We say that algorithm FINDBIGBOX is *successful* if it finds the big box. Thus, we want to determine  $\Pr(W)$ , where  $W$  is the event

$$W = \text{“algorithm FINDBIGBOX is successful”}.$$

We are going to write this event as a subset of the sample space  $S$ . For this, we have to determine all elements  $(a, z)$  in  $S$  for which algorithm FINDBIGBOX is successful.

First consider the case when  $a = x$ . In this case, the box we choose in Step 1 is the small box. There are two possibilities for  $z$ :

- If  $x = a > z$ , then the algorithm keeps the small box and, thus, is not successful.
- If  $x = a < z$ , then the algorithm takes the other box (which is the big box) and, thus, is successful.

Thus, the event  $W$  contains the set

$$W_x = \{(x, z) : z \in \{x + 1/2, x + 3/2, \dots, 100 - 1/2\}\}.$$

You can verify that

$$|W_x| = 100 - x.$$

The second case to consider is when  $a = y$ . In this case, the box we choose in Step 1 is the big box. Again, there are two possibilities for  $z$ :

- If  $y = a > z$ , then the algorithm keeps the big box and, thus, is successful.
- If  $y = a < z$ , then the algorithm takes the other box (which is the small box) and, thus, is not successful.

Thus, the event  $W$  contains the set

$$W_y = \{(y, z) : z \in \{1/2, 3/2, \dots, y - 1/2\}\}.$$

You can verify that

$$|W_y| = y.$$

Since  $W = W_x \cup W_y$  and the events  $W_x$  and  $W_y$  are disjoint, we have, by Lemma 5.3.2,

$$\begin{aligned} \Pr(W) &= \Pr(W_x \cup W_y) \\ &= \Pr(W_x) + \Pr(W_y). \end{aligned}$$

Since the element  $(a, z)$  is chosen uniformly at random from the sample space  $S$ , we can use Lemma 5.4.2 to determine the probability that algorithm FINDBIGBOX is successful:

$$\begin{aligned} \Pr(W) &= \Pr(W_x) + \Pr(W_y) \\ &= \frac{|W_x|}{|S|} + \frac{|W_y|}{|S|} \\ &= \frac{100 - x}{200} + \frac{y}{200} \\ &= \frac{1}{2} + \frac{y - x}{200}. \end{aligned}$$

Since  $x$  and  $y$  are distinct integers and  $x < y$ , we have  $y - x \geq 1$ , and we conclude that

$$\Pr(W) \geq \frac{1}{2} + \frac{1}{200} = 0.505.$$

## 5.7 The Monty Hall Problem

The Monty Hall Problem is a well-known puzzle in probability theory. It is named after the host, Monty Hall, of the American television game show *Let's Make a Deal*. The problem became famous in 1990, when (part of) a reader's letter was published in Marilyn vos Savant's column *Ask Marilyn* in the magazine *Parade*:

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Note that the host can always open a door that has a goat behind it. After the host has opened No. 3, we know that the car is either behind No. 1 or No. 2, and it seems that both these doors have the same probability (i.e., 50%) of having the car behind them. We will prove below, however, that this is not true: It is indeed to our advantage to switch our choice.

We assume that the car is equally likely to be behind any of the three doors. Moreover, the host knows what is behind each door.

- We initially choose one of the three doors uniformly at random; this door remains closed.
- The host opens one of the other two doors that has a goat behind it.
- Our final choice is to switch to the other door that is still closed.

Let  $A$  be the event that we win the car and let  $B$  be the event that the initial door has a goat behind it. Then it is not difficult to see that event  $A$  occurs if and only if event  $B$  occurs. Therefore, the probability that we win the car is equal to

$$\Pr(A) = \Pr(B) = 2/3.$$

## 5.8 Conditional Probability

Anil Maheshwari has two children. We are told that one of them is a boy. What is the probability that the other child is also a boy? Most people will say that this probability is  $1/2$ . We will show below that this is not the correct answer.

Since Anil has two children, the sample space is

$$S = \{(b, b), (b, g), (g, b), (g, g)\},$$

where, for example,  $(b, g)$  indicates that the youngest child is a boy and the oldest child is a girl. We assume a uniform probability function, so that each outcome has a probability of  $1/4$ .

We are given the additional information that one of the two children is a boy, or, to be more precise, that at least one of the two children is a boy. This means that the actual sample space is not  $S$ , but

$$\{(b, b), (b, g), (g, b)\}.$$

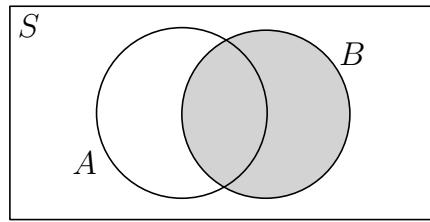
When asking for the probability that the other child is also a boy, we are really asking for the probability that both children are boys. Since there is only one possibility (out of three) for both children to be boys, it follows that this probability is equal to  $1/3$ .

This is an example of a *conditional probability*: We are asking for the probability of an event (both children are boys), given that another event (at least one of the two children is a boy) occurs.

**Definition 5.8.1** Let  $(S, \Pr)$  be a probability space and let  $A$  and  $B$  be two events with  $\Pr(B) > 0$ . The *conditional probability*  $\Pr(A | B)$ , pronounced as “the probability of  $A$  given  $B$ ”, is defined as

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Let us try to understand where this definition comes from. Initially, the sample space is equal to  $S$ . When we are given the additional information that event  $B$  occurs, the sample space “shrinks” to  $B$ , and event  $A$  occurs if and only if event  $A \cap B$  occurs.



You may think that  $\Pr(A | B)$  should therefore be defined to be  $\Pr(A \cap B)$ . However, since the sum of all probabilities must be equal to 1, we have to normalize, i.e., divide by  $\Pr(B)$ . Equivalently, if  $A = B$ , we get  $\Pr(A | A)$ , which is the probability that event  $A$  occurs, given that event  $A$  occurs. This probability should be equal to 1. Indeed, using the definition, we do get

$$\Pr(A | A) = \frac{\Pr(A \cap A)}{\Pr(A)} = \frac{\Pr(A)}{\Pr(A)} = 1.$$

In Exercise 5.24, you are asked to give a formal proof that our definition gives a valid probability function on the sample space  $S$ .

### 5.8.1 Anil's Children

Returning to Anil's two children, we saw that the sample space is

$$S = \{(b, b), (b, g), (g, b), (g, g)\}$$

and we assumed a uniform probability function. The events we considered are

$$A = \text{"both children are boys"}$$

and

$$B = \text{"at least one of the two children is a boy"},$$

and we wanted to know  $\Pr(A | B)$ . Writing  $A$  and  $B$  as subsets of the sample space  $S$ , we get

$$A = \{(b, b)\}$$

and

$$B = \{(b, b), (b, g), (g, b)\}.$$

Using Definition 5.8.1, it follows that

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)} = \frac{|A|/|S|}{|B|/|S|} = \frac{1/4}{3/4} = 1/3,$$

which is the same answer as we got before.

### 5.8.2 Rolling a Die

Assume we roll a fair die, i.e., we choose an element uniformly at random from the sample space

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Consider the events

$$A = \text{"the result is 3"}$$

and

$$B = \text{"the result is an odd integer".}$$

What is the conditional probability  $\Pr(A | B)$ ? To determine this probability, we assume that event  $B$  occurs, i.e., the roll of the die resulted in one of 1, 3, and 5. Given that event  $B$  occurs, event  $A$  occurs in one out of these three possibilities. Thus,  $\Pr(A | B)$  should be equal to  $1/3$ . We are going to verify that this is indeed the answer we get when using Definition 5.8.1: Since

$$A = \{3\}$$

and

$$B = \{1, 3, 5\},$$

we have

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)} = \frac{|A|/|S|}{|B|/|S|} = \frac{1/6}{3/6} = 1/3.$$

Let us now consider the conditional probability  $\Pr(B | A)$ . Thus, we are given that event  $A$  occurs, i.e., the roll of the die resulted in 3. Since 3 is an odd integer, event  $B$  is guaranteed to occur. Therefore,  $\Pr(B | A)$  should be equal to 1. Again, we are going to verify that this is indeed the answer we get when using Definition 5.8.1:

$$\Pr(B | A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A)}{\Pr(A)} = 1.$$

This shows that, in general,  $\Pr(A | B)$  is not equal to  $\Pr(B | A)$ . Observe that this is not surprising. (Do you see why?)

Consider the event

$$C = \text{"the result is a prime number"},$$

which, when written as a subset of the sample space, is

$$C = \{2, 3, 5\}.$$

Then  $\Pr(C | B)$  should be equal to  $2/3$  and  $\Pr(C | A)$  should be equal to 1. Indeed, we have

$$\Pr(C | B) = \frac{\Pr(C \cap B)}{\Pr(B)} = \frac{|C \cap B|/|S|}{|B|/|S|} = \frac{2/6}{3/6} = 2/3$$

and

$$\Pr(C | A) = \frac{\Pr(C \cap A)}{\Pr(A)} = \frac{\Pr(A)}{\Pr(A)} = 1.$$

Recall that  $\bar{B}$  denotes the complement of the event  $B$ . Thus, this is the event

$$\bar{B} = \text{"the result is an even integer"},$$

which, when written as a subset of the sample space, is

$$\bar{B} = \{2, 4, 6\}.$$

Then  $\Pr(C | \bar{B})$  should be equal to  $1/3$ . Indeed, we have

$$\Pr(C | \bar{B}) = \frac{\Pr(C \cap \bar{B})}{\Pr(\bar{B})} = \frac{|C \cap \bar{B}|/|S|}{|\bar{B}|/|S|} = \frac{1/6}{3/6} = 1/3.$$

Observe that

$$\Pr(C | B) + \Pr(C | \bar{B}) = 2/3 + 1/3 = 1.$$

You may think that this is true for any two events  $B$  and  $C$ . This is, however, not the case: Since

$$\bar{A} = \{1, 2, 4, 5, 6\},$$

we have

$$\Pr(C | \bar{A}) = \frac{\Pr(C \cap \bar{A})}{\Pr(\bar{A})} = \frac{|C \cap \bar{A}|/|S|}{|\bar{A}|/|S|} = \frac{2/6}{5/6} = 2/5$$

and, thus,

$$\Pr(C | A) + \Pr(C | \bar{A}) = 1 + 2/5 \neq 1.$$

It should be an easy exercise to verify that

$$\Pr(A | C) + \Pr(\bar{A} | C) = 1.$$

Intuitively, this should be true for any two events  $A$  and  $C$ : When we are given that event  $C$  occurs, then either  $A$  occurs or  $A$  does not occur (in which case  $\bar{A}$  occurs). The following lemma states that this intuition is indeed correct.

**Lemma 5.8.2** *Let  $(S, \Pr)$  be a probability space and let  $A$  and  $B$  be two events with  $\Pr(B) > 0$ . Then*

$$\Pr(A | B) + \Pr(\bar{A} | B) = 1.$$

**Proof.** By definition, we have

$$\begin{aligned}\Pr(A | B) + \Pr(\bar{A} | B) &= \frac{\Pr(A \cap B)}{\Pr(B)} + \frac{\Pr(\bar{A} \cap B)}{\Pr(B)} \\ &= \frac{\Pr(A \cap B) + \Pr(\bar{A} \cap B)}{\Pr(B)}.\end{aligned}$$

Since the events  $A \cap B$  and  $\bar{A} \cap B$  are disjoint, we have, by Lemma 5.3.2,

$$\Pr(A \cap B) + \Pr(\bar{A} \cap B) = \Pr((A \cap B) \cup (\bar{A} \cap B)).$$

By drawing a Venn diagram, you will see that

$$(A \cap B) \cup (\bar{A} \cap B) = B,$$

implying that

$$\Pr(A \cap B) + \Pr(\bar{A} \cap B) = \Pr(B).$$

We conclude that

$$\Pr(A | B) + \Pr(\bar{A} | B) = \frac{\Pr(B)}{\Pr(B)} = 1.$$

■

### 5.8.3 Flip and Flip or Roll

We are given a fair red coin, a fair blue coin, and a fair die. First, we flip the red coin. If the result of this flip is heads, then we flip the blue coin and return the result of this second flip. Otherwise, the red coin came up tails, in which case we roll the die and return the result of this roll.

What is the probability that the value 5 is returned? Our intuition says that this probability is equal to  $1/12$ : The value 5 is returned if and only if the red coin comes up tails (which happens with probability  $1/2$ ) and the result of rolling the die is 5 (which happens with probability  $1/6$ ). We will prove that this is indeed the correct answer.

We start by modifying the above random process so that it better reflects the random choices that are being made:

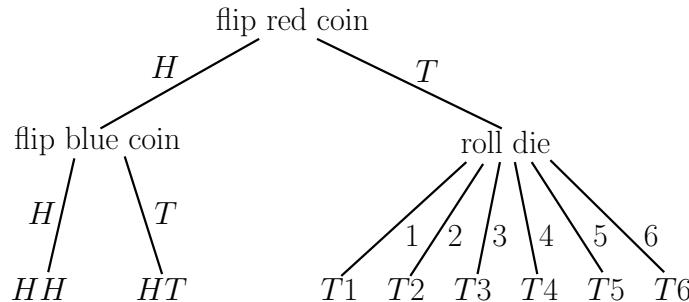
**Algorithm FLIPANDFLIPORROLL:**

```

 $f_r$  = the result of flipping the red coin;
if  $f_r = H$ 
  then  $f_b$  = the result of flipping the blue coin;
    return the ordered pair  $(f_r, f_b)$ 
  else  $d$  = the result of rolling the die;
    return the ordered pair  $(f_r, d)$ 
endif

```

The possible executions of this algorithm are visualized in the following tree diagram:



The sample space is the set  $S$  of all possible values that can be returned by algorithm FLIPANDFLIPORROLL. Thus, we have

$$S = \{(H, H), (H, T), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}.$$

We are interested in the probability that the algorithm returns the value  $(T, 5)$ , i.e., the probability of the event

$$A = \{(T, 5)\}.$$

Since the event  $A$  obviously depends on the result of flipping the red coin, we consider the event

$$R = \text{“the result of flipping the red coin is tails”}.$$

If we write this event as a subset of the sample space, we get

$$R = \{(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}.$$

Observe that  $\Pr(R) = 1/2$  and  $A \cap R = A$ , which implies that

$$\Pr(A) = \Pr(A \cap R).$$

If we rewrite the expression for the conditional probability  $\Pr(A \mid R)$  in Definition 5.8.1, we get

$$\Pr(A) = \Pr(A \cap R) = \Pr(R) \cdot \Pr(A \mid R).$$

We have seen already that  $\Pr(R) = 1/2$ . To determine  $\Pr(A \mid R)$ , we assume that event  $R$  occurs. Under this assumption, event  $A$  occurs if and only if the result of rolling the die is 5, which happens with probability  $1/6$ . Thus,

$$\Pr(A \mid R) = 1/6$$

and we conclude that

$$\Pr(A) = 1/2 \cdot 1/6 = 1/12,$$

which is the answer we were expecting to see.

You may object to this method of determining  $\Pr(A)$ : When we determined  $\Pr(A \mid R)$ , we did not use the definition of conditional probability, i.e.,

$$\Pr(A \mid R) = \frac{\Pr(A \cap R)}{\Pr(R)}.$$

Instead, we used the “informal definition”, by determining the probability that event  $A$  occurs, assuming that event  $R$  occurs. Thus, we do not yet

have a formal justification as to why  $\Pr(A)$  is equal to  $1/12$ . In the rest of this section, we do present a formal justification.

For each integer  $i$  with  $1 \leq i \leq 6$ , we consider the event

$$A_i = \{(T, i)\}$$

and its probability

$$p_i = \Pr(A_i).$$

First observe that

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6,$$

because the die is fair. Let  $p$  denote the common value of the  $p_i$ 's. Next observe that

$$R = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6,$$

where the six events on the right-hand side are pairwise disjoint. We have seen already that  $\Pr(R) = 1/2$ . It follows that

$$\begin{aligned} 1/2 &= \Pr(R) \\ &= \Pr\left(\bigcup_{i=1}^6 A_i\right) \\ &= \sum_{i=1}^6 \Pr(A_i) \\ &= \sum_{i=1}^6 p \\ &= 6p, \end{aligned}$$

implying that  $p = 1/12$ . Since the event  $A$  we are interested in is equal to the event  $A_5$ , we conclude that

$$\Pr(A) = \Pr(A_5) = p_5 = p = 1/12.$$

Thus, we have obtained a formal proof of the fact that the probability of the event  $A$  is equal  $1/12$ .

Using the definition of conditional probability, we can now also formally determine  $\Pr(A | R)$ :

$$\begin{aligned}\Pr(A | R) &= \frac{\Pr(A \cap R)}{\Pr(R)} \\ &= \frac{\Pr(A)}{\Pr(R)} \\ &= \frac{1/12}{1/2} \\ &= 1/6.\end{aligned}$$

## 5.9 The Law of Total Probability

Both Mick and Keith have a random birthday. What is the probability that they have the same birthday? We have seen in Section 5.5 that this probability is equal to  $1/365$ . A common way to determine this probability is as follows: Consider Mick's birthday, which can be any of the 365 days of the year. By symmetry, it does not really matter what Mick's birthday is, so we just assume that it is July 26. Then Mick and Keith have the same birthday if and only if Keith's birthday is also on July 26. Therefore, since Keith has a random birthday, the probability that Mick and Keith have the same birthday is equal to  $1/365$ . The following theorem explains this reasoning.

**Theorem 5.9.1 (Law of Total Probability)** *Let  $(S, \Pr)$  be a probability space and let  $A$  be an event. Assume that  $B_1, B_2, \dots, B_n$  is a sequence of events such that*

1.  $\Pr(B_i) > 0$  for all  $i$  with  $1 \leq i \leq n$ ,
2. the events  $B_1, B_2, \dots, B_n$  are pairwise disjoint, and
3.  $\bigcup_{i=1}^n B_i = S$ .

Then

$$\Pr(A) = \sum_{i=1}^n \Pr(A | B_i) \cdot \Pr(B_i).$$

**Proof.** The assumptions imply that

$$\begin{aligned} A &= A \cap S \\ &= A \cap \left( \bigcup_{i=1}^n B_i \right) \\ &= \bigcup_{i=1}^n (A \cap B_i). \end{aligned}$$

Since the events  $A \cap B_1, A \cap B_2, \dots, A \cap B_n$  are pairwise disjoint, it follows from Lemma 5.3.2 that

$$\begin{aligned} \Pr(A) &= \Pr\left(\bigcup_{i=1}^n (A \cap B_i)\right) \\ &= \sum_{i=1}^n \Pr(A \cap B_i). \end{aligned}$$

The theorem follows by observing that, from Definition 5.8.1,

$$\Pr(A \cap B_i) = \Pr(A | B_i) \cdot \Pr(B_i).$$

■

Let us consider the three conditions in this theorem. The first condition is that  $\Pr(B_i) > 0$ , i.e., there is a positive probability that event  $B_i$  occurs. The second and third conditions, i.e.,

- the events  $B_1, B_2, \dots, B_n$  are pairwise disjoint, and
- $\bigcup_{i=1}^n B_i = S$ ,

are equivalent to

- exactly one of the events  $B_1, B_2, \dots, B_n$  is guaranteed to occur.

In the example in the beginning of this section, we wanted to know  $\Pr(A)$ , where  $A$  is the event

$$A = \text{“Mick and Keith have the same birthday”}.$$

In order to apply Theorem 5.9.1, we *define* a sequence  $B_1, B_2, \dots$  of events that satisfy the conditions in this theorem and for which  $\Pr(A | B_i)$  is easy

to determine. For this example, we define the event  $B_i$ , for each  $i$  with  $1 \leq i \leq 365$ , to be

$$B_i = \text{“Mick’s birthday is on the } i\text{-th day of the year”}.$$

It is clear that (i)  $\Pr(B_i) = 1/365 > 0$  and (ii) exactly one of the events  $B_1, B_2, \dots, B_{365}$  is guaranteed to occur. It follows that

$$\Pr(A) = \sum_{i=1}^{365} \Pr(A | B_i) \cdot \Pr(B_i).$$

To determine  $\Pr(A | B_i)$ , we assume that the event  $B_i$  occurs, i.e., we fix Mick’s birthday to be the  $i$ -th day of the year. Given this event  $B_i$ , event  $A$  occurs if and only if Keith’s birthday is also on the  $i$ -th day. Thus, we have  $\Pr(A | B_i) = 1/365$  and it follows that

$$\begin{aligned} \Pr(A) &= \sum_{i=1}^{365} (1/365) \cdot \Pr(B_i) \\ &= (1/365) \sum_{i=1}^{365} \Pr(B_i) \\ &= (1/365) \cdot 1 \\ &= 1/365, \end{aligned}$$

which is the same answer as we got in the beginning of this section.

### 5.9.1 Flipping a Coin and Rolling Dice

Consider the following experiment:

- We flip a fair coin.
  - If the coin comes up heads, then we roll a fair die. Let  $R$  denote the result of this die.
  - If the coin comes up tails, then we roll two fair dice. Let  $R$  denote the sum of the results of these dice.

What is the probability that the value of  $R$  is equal to 2? That is, if we define the event  $A$  to be

$$A = \text{“the value of } R \text{ is equal to 2”},$$

then we want to know  $\Pr(A)$ . Since the value of  $R$  depends on whether the coin comes up heads or tails, we define the event

$$B = \text{“the coin comes up heads”}.$$

Since (i) both  $B$  and its complement  $\bar{B}$  occur with a positive probability and (ii) exactly one of  $B$  and  $\bar{B}$  is guaranteed to occur, we can apply Theorem 5.9.1 and get

$$\Pr(A) = \Pr(A | B) \cdot \Pr(B) + \Pr(A | \bar{B}) \cdot \Pr(\bar{B}).$$

We determine the four terms on the right-hand side:

- It should be clear that

$$\Pr(B) = \Pr(\bar{B}) = 1/2.$$

- To determine  $\Pr(A | B)$ , we assume that the event  $B$  occurs, i.e., the coin comes up heads. Because of this assumption, we roll one die, and the event  $A$  occurs if and only if the result of this roll is 2. It follows that

$$\Pr(A | B) = 1/6.$$

- To determine  $\Pr(A | \bar{B})$ , we assume that the event  $\bar{B}$  occurs, i.e., the coin comes up tails. Because of this assumption, we roll two dice, and the event  $A$  occurs if and only if both rolls result in 1. Since there are 36 possible outcomes when rolling two dice, it follows that

$$\Pr(A | \bar{B}) = 1/36.$$

We conclude that

$$\begin{aligned}\Pr(A) &= \Pr(A | B) \cdot \Pr(B) + \Pr(A | \bar{B}) \cdot \Pr(\bar{B}) \\ &= 1/6 \cdot 1/2 + 1/36 \cdot 1/2 \\ &= 7/72.\end{aligned}$$

## 5.10 Please Take a Seat

Let  $n \geq 2$  and  $k \geq 0$  be integers. There are  $n + k$  chairs  $C_1, C_2, \dots, C_{n+k}$  inside a room. Outside this room, there are  $n$  people  $P_1, P_2, \dots, P_n$ . These people are told to enter the room one by one, in increasing order of their indices, and each person must sit down in the chair having her index: For  $i = 1, 2, \dots, n$ , person  $P_i$  enters the room and sits down in chair  $C_i$ .

The first person  $P_1$  did not listen to the instructions and, instead of taking chair  $C_1$ , chooses one of the  $n + k$  chairs uniformly at random and sits down in the chosen chair. (Note that chair  $C_1$  may be the chosen chair.) From then on, for  $i = 2, 3, \dots, n$ , person  $P_i$  checks if chair  $C_i$  is available. If this is the case, then  $P_i$  sits down in chair  $C_i$ . Otherwise,  $P_i$  chooses one of the available chairs uniformly at random and sits down in the chosen chair.

We want to determine the probability  $p_{n,k}$  that, at the end, the last person  $P_n$  sits in chair  $C_n$ . Before we analyze this probability, we present this process in pseudocode:

```
Algorithm TAKEASEAT( $n, k$ ):
    //  $n \geq 2$  and  $k \geq 0$ ;
    // the input consists of  $n$  people  $P_1, P_2, \dots, P_n$  and
    //  $n + k$  chairs  $C_1, C_2, \dots, C_{n+k}$ 
     $j =$  uniformly random element in  $\{1, 2, \dots, n + k\}$ ;
    person  $P_1$  sits down in chair  $C_j$ ;
    for  $i = 2$  to  $n$ 
        do if chair  $C_i$  is available
            then person  $P_i$  sits down in chair  $C_i$ 
            else  $j =$  index of a uniformly random available chair;
                person  $P_i$  sits down in chair  $C_j$ 
            endif
        endfor
```

We consider the event

$$A_{n,k} = \text{“after algorithm TAKEASEAT}(n, k)\text{ has terminated, person }P_n \text{ sits in chair }C_n\text{”}.$$

The probability that was mentioned above is given by

$$p_{n,k} = \Pr(A_{n,k}).$$

In the for-loop in algorithm TAKEASEAT( $n, k$ ), the variable  $i$  runs from 2 to  $n$ . We will label the iterations of this loop by the value of the variable  $i$ . Thus, iteration 3 will refer to the iteration in which  $i = 3$ ; observe that this is actually the second time that the algorithm goes through the for-loop.

At this moment, you should convince yourself (for example, by induction on  $i$ ) that the following holds for each  $i = 2, 3, \dots, n$ :

- At the start of iteration  $i$ ,
  - all chairs  $C_2, C_3, \dots, C_{i-1}$  have been taken, and
  - exactly one of the chairs  $C_1, C_i, C_{i+1}, \dots, C_{n+k}$  has been taken.

If we take  $i = n$ , then we see that at the start of iteration  $n$

- all chairs  $C_2, C_3, \dots, C_{n-1}$  have been taken, and
- exactly one of the  $k + 2$  chairs  $C_1, C_n, C_{n+1}, \dots, C_{n+k}$  has been taken.

Event  $A_{n,k}$  occurs if and only if chair  $C_n$  is available (i.e., has not been taken) at the start of iteration  $n$ .

Is it true that the chair among  $C_1, C_n, C_{n+1}, \dots, C_{n+k}$  that has been taken at the start of iteration  $n$  is a uniformly random chair from these  $k + 2$  chairs? If this is the case, then chair  $C_n$  has been taken with probability  $1/(k + 2)$  and, thus,  $C_n$  is available with probability  $1 - 1/(k + 2) = (k + 1)/(k + 2)$ . In other words, if the question above has a positive answer, then

$$p_{n,k} = \Pr(A_{n,k}) = \frac{k+1}{k+2}.$$

In the rest of this section, we will present two ways to prove that this is indeed the correct value of  $p_{n,k}$ . In both proofs, we will use the Law of Total Probability of Section 5.9.

Note that  $p_{n,k}$  does not depend on  $n$ . In particular, if  $k = 0$ , then the probability that person  $P_n$  sits in chair  $C_n$  is equal to  $1/2$ .

### 5.10.1 Determining $p_{n,k}$ Using a Recurrence Relation

Let us start with the case when  $n = 2$ . Thus, there are two people  $P_1$  and  $P_2$ , and  $2 + k$  chairs  $C_1, C_2, \dots, C_{2+k}$ . Event  $A_{2,k}$  occurs if and only if  $P_1$  chooses

one of the  $1 + k$  chairs  $C_1, C_3, C_4, \dots, C_{2+k}$ . Since  $P_1$  chooses a uniformly random chair out of  $2 + k$  chairs, it follows that

$$p_{2,k} = \Pr(A_{2,k}) = \frac{k+1}{k+2}.$$

Assume from now on that  $n \geq 3$ . We are going to derive a recurrence relation that expresses  $p_{n,k}$  in terms of  $p_{2,k}, p_{3,k}, \dots, p_{n-1,k}$ .

Consider the (random) index  $j$  of the chair that  $P_1$  chooses in the first line of algorithm TAKEASEAT( $n, k$ ). We consider three cases, depending on the value of  $j$ .

- Assume that  $j \in \{1, n+1, n+2, \dots, n+k\}$ . Then for each  $i = 2, 3, \dots, n$ , chair  $C_i$  is available at the start of iteration  $i$  and person  $P_i$  sits down in chair  $C_i$ . In particular, during iteration  $n$ ,  $P_n$  sits down in chair  $C_n$  and event  $A_{n,k}$  occurs.
- Assume that  $j = n$ . Then chair  $C_n$  has been taken at the start of iteration  $n$  and event  $A_{n,k}$  does not occur.
- Assume that  $j \in \{2, 3, \dots, n-1\}$ . Then for each  $i = 2, 3, \dots, j-1$ , chair  $C_i$  is available at the start of iteration  $i$  and person  $P_i$  sits down in chair  $C_i$ . At the start of iteration  $j$ , the chairs  $C_1, C_{j+1}, C_{j+2}, \dots, C_{n+k}$  are available and person  $P_j$  chooses one of these chairs uniformly at random. Thus, iterations  $j, j+1, \dots, n$  can be viewed as running algorithm TAKEASEAT( $n-j+1, k$ ), where the  $n-j+1$  people are  $P_j, P_{j+1}, \dots, P_n$  and the  $n-j+1+k$  chairs are  $C_1, C_{j+1}, C_{j+2}, \dots, C_{n+k}$ . In this case, event  $A_{n,k}$  occurs if and only if, after algorithm TAKEASEAT( $n-j+1, k$ ) has terminated, person  $P_n$  sits in chair  $C_n$ , i.e., event  $A_{n-j+1,k}$  occurs.

Thus, we can determine the probability that event  $A_{n,k}$  occurs, if we are given the value of  $j$ ; note that this is a conditional probability. Since  $j$  is a random element in the set  $\{1, 2, \dots, n+k\}$ , we are going to use the Law of Total Probability (Theorem 5.9.1): For each  $j \in \{1, 2, \dots, n+k\}$ , we consider the event

$$B_{n,k,j} = \text{“in the second line of algorithm TAKEASEAT}(n, k)\text{, person } P_1 \text{ sits down in chair } C_j\text{”}.$$

Since exactly one of these events is guaranteed to occur, we can apply Theorem 5.9.1 and obtain

$$\Pr(A_{n,k}) = \sum_{j=1}^{n+k} \Pr(A_{n,k} | B_{n,k,j}) \cdot \Pr(B_{n,k,j}).$$

It follows from the first line in algorithm TAKEASEAT( $n, k$ ) that, for each  $j$  with  $1 \leq j \leq n + k$ ,

$$\Pr(B_{n,k,j}) = \frac{1}{n+k}.$$

- Assume that  $j \in \{1, n+1, n+2, \dots, n+k\}$ . We have seen above that event  $A_{n,k}$  occurs. Thus,

$$\Pr(A_{n,k} | B_{n,k,j}) = 1.$$

- Assume that  $j = n$ . We have seen above that event  $A_{n,k}$  does not occur. Thus,

$$\Pr(A_{n,k} | B_{n,k,n}) = 0.$$

- Assume that  $j \in \{2, 3, \dots, n-1\}$ . We have seen above that event  $A_{n,k}$  occurs if and only if event  $A_{n-j+1,k}$  occurs. Thus,

$$\Pr(A_{n,k} | B_{n,k,j}) = \Pr(A_{n-j+1,k}) = p_{n-j+1,k}.$$

We conclude that

$$\begin{aligned} p_{n,k} &= \Pr(A_{n,k}) \\ &= \sum_{j=1}^{n+k} \Pr(A_{n,k} | B_{n,k,j}) \cdot \Pr(B_{n,k,j}) \\ &= \sum_{j=1}^{n+k} \Pr(A_{n,k} | B_{n,k,j}) \cdot \frac{1}{n+k} \\ &= \frac{1}{n+k} \sum_{j=1}^{n+k} \Pr(A_{n,k} | B_{n,k,j}) \\ &= \frac{1}{n+k} \left( (k+1) + \sum_{j=2}^{n-1} p_{n-j+1,k} \right). \end{aligned}$$

If we write out the terms in this summation, then we get, for each  $n \geq 3$ ,

$$p_{n,k} = \frac{k+1}{n+k} + \frac{1}{n+k} (p_{2,k} + p_{3,k} + \cdots + p_{n-1,k}).$$

As we have seen above, the base case is given by

$$p_{2,k} = \frac{k+1}{k+2}.$$

It remains to solve the recurrence relation. If you use the recurrence to determine  $p_{n,k}$  for some small values of  $n$ , then you will notice that they are all equal to  $(k+1)/(k+2)$ . This suggests that

$$p_{n,k} = \frac{k+1}{k+2}$$

for all integers  $n \geq 2$ . (Recall that we already suspected this.) Using induction on  $n$ , it can easily be proved that this is indeed the case.

### 5.10.2 Determining $p_{n,k}$ by Modifying the Algorithm

Our second solution is obtained by modifying algorithm  $\text{TAKEASEAT}(n, k)$ : Person  $P_1$  again did not listen to the instructions and, instead of taking chair  $C_1$ , chooses one of the  $n+k$  chairs uniformly at random and sits down in the chosen chair. From then on, for  $i = 2, 3, \dots, n-1$ , person  $P_i$  checks if chair  $C_i$  is available. If this is the case, then  $P_i$  sits down in chair  $C_i$ . Otherwise,  $P_1$  is sitting in  $C_i$ , in which case (i)  $P_i$  kicks  $P_1$  out of chair  $C_i$ , (ii)  $P_i$  sits down in chair  $C_i$ , and (iii)  $P_1$  chooses one of the available chairs uniformly at random and sits down in the chosen chair. In the last step, person  $P_n$  checks if chair  $C_n$  is available. If this is the case, then  $P_n$  sits down in chair  $C_n$ . Otherwise,  $P_n$  chooses one of the available chairs uniformly at random and sits down in the chosen chair. In pseudocode, this modified algorithm looks as follows:

**Algorithm** TAKEASEAT'(n, k):

```

// n ≥ 2 and k ≥ 0;
// the input consists of n people  $P_1, P_2, \dots, P_n$  and
//  $n + k$  chairs  $C_1, C_2, \dots, C_{n+k}$ 
j = uniformly random element in  $\{1, 2, \dots, n + k\}$ ;
person  $P_1$  sits down in chair  $C_j$ ;
for  $i = 2$  to  $n - 1$ 
  do //  $P_2$  sits in  $C_2$ ,  $P_3$  sits in  $C_3$ , ...,  $P_{i-1}$  sits in  $C_{i-1}$ 
    if chair  $C_i$  has been taken
      then //  $P_1$  sits in  $C_i$ 
        j = uniformly random element
        in  $\{1, i + 1, i + 2, \dots, n + k\}$ ;
        person  $P_1$  sits down in chair  $C_j$ 
      endif;
    person  $P_i$  sits down in chair  $C_i$ 
  endfor;
//  $P_2$  sits in  $C_2$ ,  $P_3$  sits in  $C_3$ , ...,  $P_{n-1}$  sits in  $C_{n-1}$ 
if chair  $C_n$  is available
  then person  $P_n$  sits down in chair  $C_n$ 
  else  $j$  = uniformly random element
    in  $\{1, n + 1, n + 2, \dots, n + k\}$ ;
    person  $P_n$  sits down in chair  $C_j$ 
  endif
```

As in the previous subsection, we label the iterations of the for-loop by the value of the variable  $i$ . Moreover, we consider the first two lines of the algorithm to be iteration 1. Thus, up to the end of the for-loop, algorithm TAKEASEAT'(n, k) makes iterations that are labeled 1, 2, ...,  $n - 1$ .

It follows from algorithm TAKEASEAT'(n, k) that, after the for-loop has terminated,

- for each  $i = 2, 3, \dots, n - 1$ , person  $P_i$  sits in chair  $C_i$ , and
- person  $P_1$  sits in one of the chairs  $C_1, C_n, C_{n+1}, \dots, C_{n+k}$ .

Recall that  $A_{n,k}$  is the event that person  $P_n$  sits in chair  $C_n$ , after the original algorithm TAKEASEAT(n, k) has terminated. It follows from the modified algorithm TAKEASEAT'(n, k) that event  $A_{n,k}$  occurs if and only if,

after the for-loop of algorithm TAKEASEAT'( $n, k$ ) has terminated, person  $P_1$  sits in one of the chairs  $C_1, C_{n+1}, C_{n+2}, \dots, C_{n+k}$ . In other words,  $p_{n,k} = \Pr(A_{n,k})$  is equal to the probability that, after the for-loop has terminated,  $P_1$  sits in one of  $C_1, C_{n+1}, C_{n+2}, \dots, C_{n+k}$ .

We have seen that, after the for-loop has terminated,  $P_1$  sits in one of the chairs  $C_1, C_n, C_{n+1}, \dots, C_{n+k}$ . Thus, there is a value of  $i$  with  $1 \leq i \leq n-1$ , such that  $P_1$  sits down in one of these chairs during iteration  $i$ . As soon as  $P_1$  sits down in one of these chairs,  $P_1$  stays there until the end of the algorithm. This implies that there is exactly one integer  $i$  having this property. Based on this, and since this integer  $i$  is random, we are again going to use the Law of Total Probability (Theorem 5.9.1): For each  $i \in \{1, 2, \dots, n-1\}$ , we consider the event

$$B_{n,k,i} = \text{"during iteration } i, \text{ person } P_1 \text{ chooses one of the chairs } C_1, C_n, C_{n+1}, \dots, C_{n+k}."$$

Since exactly one of these events is guaranteed to occur, Theorem 5.9.1 implies that

$$\Pr(A_{n,k}) = \sum_{i=1}^{n-1} \Pr(A_{n,k} | B_{n,k,i}) \cdot \Pr(B_{n,k,i}).$$

Consider the event

$$B_{n,k} = \text{"a uniformly random element from the set } \{1, n, n+1, \dots, n+k\} \text{ is not equal to } n."$$

Then for each  $i$  with  $1 \leq i \leq n-1$ , we have

$$\Pr(A_{n,k} | B_{n,k,i}) = \Pr(B_{n,k}) = \frac{k+1}{k+2}.$$

It follows that

$$\begin{aligned} p_{n,k} &= \Pr(A_{n,k}) \\ &= \sum_{i=1}^{n-1} \frac{k+1}{k+2} \cdot \Pr(B_{n,k,i}) \\ &= \frac{k+1}{k+2} \sum_{i=1}^{n-1} \Pr(B_{n,k,i}) \\ &= \frac{k+1}{k+2}, \end{aligned}$$

because the last summation is equal to 1.

## 5.11 Independent Events

Consider two events  $A$  and  $B$  in a sample space  $S$ . In this section, we will define the notion of these two events being “independent”. Intuitively, this should express that (i) the probability that event  $A$  occurs does not depend on whether or not event  $B$  occurs, and (ii) the probability that event  $B$  occurs does not depend on whether or not event  $A$  occurs. Thus, if we assume that  $\Pr(A) > 0$  and  $\Pr(B) > 0$ , then (i)  $\Pr(A)$  should be equal to the conditional probability  $\Pr(A | B)$ , and (ii)  $\Pr(B)$  should be equal to the conditional probability  $\Pr(B | A)$ . As we will show below, the following definition exactly captures this.

**Definition 5.11.1** Let  $(S, \Pr)$  be a probability space and let  $A$  and  $B$  be two events. We say that  $A$  and  $B$  are *independent* if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

In this definition, it is not assumed that  $\Pr(A) > 0$  and  $\Pr(B) > 0$ . If  $\Pr(B) > 0$ , then

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)},$$

and  $A$  and  $B$  are independent if and only if

$$\Pr(A | B) = \Pr(A).$$

Similarly, if  $\Pr(A) > 0$ , then  $A$  and  $B$  are independent if and only if

$$\Pr(B | A) = \Pr(B).$$

### 5.11.1 Rolling Two Dice

Assume we roll a red die and a blue die; thus, the sample space is

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\},$$

where  $i$  is the result of the red die and  $j$  is the result of the blue die. We assume a uniform probability function. Thus, each outcome has a probability of  $1/36$ .

Let  $D_1$  denote the result of the red die and let  $D_2$  denote the result of the blue die. Consider the events

$$A = "D_1 + D_2 = 7"$$

and

$$B = "D_1 = 4".$$

Are these events independent?

- Since

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

we have  $\Pr(A) = 6/36 = 1/6$ .

- Since

$$B = \{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\},$$

we have  $\Pr(B) = 6/36 = 1/6$ .

- Since

$$A \cap B = \{(4, 3)\},$$

we have  $\Pr(A \cap B) = 1/36$ .

- It follows that  $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$  and we conclude that  $A$  and  $B$  are independent.

As an exercise, you should verify that the events

$$A' = "D_1 + D_2 = 11"$$

and

$$B' = "D_1 = 5"$$

are not independent.

Now consider the two events

$$A'' = "D_1 + D_2 = 4"$$

and

$$B'' = "D_1 = 4".$$

Since  $A'' \cap B'' = \emptyset$ , we have

$$\Pr(A'' \cap B'') = \Pr(\emptyset) = 0.$$

On the other hand,  $\Pr(A'') = 1/12$  and  $\Pr(B'') = 1/6$ . Thus,

$$\Pr(A'' \cap B'') \neq \Pr(A'') \cdot \Pr(B'')$$

and the events  $A''$  and  $B''$  are not independent. This is not surprising: If we know that  $B''$  occurs, then  $A''$  does not occur, i.e.,  $\Pr(A'' | B'') = 0$ . Thus, the event  $B''$  has an effect on the probability that the event  $A''$  occurs.

### 5.11.2 A Basic Property of Independent Events

Consider two events  $A$  and  $B$  in a sample space  $S$ . If these events are independent, then the probability that  $A$  occurs does not depend on whether or not  $B$  occurs. Since whether or not  $B$  occurs is the same as whether the complement  $\bar{B}$  does not or does occur, it should not be a surprise that the events  $A$  and  $\bar{B}$  are independent as well. The following lemma states that this is indeed the case.

**Lemma 5.11.2** *Let  $(S, \Pr)$  be a probability space and let  $A$  and  $B$  be two events. If  $A$  and  $B$  are independent, then  $A$  and  $\bar{B}$  are also independent.*

**Proof.** To prove that  $A$  and  $\bar{B}$  are independent, we have to show that

$$\Pr(A \cap \bar{B}) = \Pr(A) \cdot \Pr(\bar{B}).$$

Using Lemma 5.3.3, this is equivalent to showing that

$$\Pr(A \cap \bar{B}) = \Pr(A) \cdot (1 - \Pr(B)). \quad (5.4)$$

Since the events  $A \cap B$  and  $A \cap \bar{B}$  are disjoint and

$$A = (A \cap B) \cup (A \cap \bar{B}),$$

it follows from Lemma 5.3.2 that

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \bar{B}).$$

Since  $A$  and  $B$  are independent, we have

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

It follows that

$$\Pr(A) = \Pr(A) \cdot \Pr(B) + \Pr(A \cap \bar{B}),$$

which is equivalent to (5.4). ■

### 5.11.3 Pairwise and Mutually Independent Events

We have defined the notion of two events being independent. The following definition generalizes this in two ways to sequences of events:

**Definition 5.11.3** Let  $(S, \Pr)$  be a probability space, let  $n \geq 2$ , and let  $A_1, A_2, \dots, A_n$  be a sequence of events.

1. We say that this sequence is *pairwise independent* if for any two distinct indices  $i$  and  $j$ , the events  $A_i$  and  $A_j$  are independent, i.e.,

$$\Pr(A_i \cap A_j) = \Pr(A_i) \cdot \Pr(A_j).$$

2. We say that this sequence is *mutually independent* if for all  $k$  with  $2 \leq k \leq n$  and all indices  $i_1 < i_2 < \dots < i_k$ ,

$$\Pr(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \Pr(A_{i_1}) \cdot \Pr(A_{i_2}) \cdots \Pr(A_{i_k}).$$

Thus, in order to show that the sequence  $A_1, A_2, \dots, A_n$  is pairwise independent, we have to verify  $\binom{n}{2}$  equalities. On the other hand, to show that this sequence is mutually independent, we have to verify  $\sum_{k=2}^n \binom{n}{k} = 2^n - 1 - n$  equalities.

For example, if we want to prove that the sequence  $A, B, C$  of three events is mutually independent, then we have to show that

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B),$$

$$\Pr(A \cap C) = \Pr(A) \cdot \Pr(C),$$

$$\Pr(B \cap C) = \Pr(B) \cdot \Pr(C),$$

and

$$\Pr(A \cap B \cap C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C).$$

To give an example, consider flipping a coin three times and assume that the result is a uniformly random element from the sample space

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\},$$

where, e.g.,  $HHT$  indicates that the first two flips result in heads and the third flip results in tails. For  $i = 1, 2, 3$ , let  $f_i$  denote the result of the  $i$ -th flip, and consider the events

$$A = "f_1 = f_2",$$

$$B = "f_2 = f_3",$$

and

$$C = "f_1 = f_3".$$

If we write these events as subsets of the sample space, then we get

$$A = \{HHH, HHT, TTH, TTT\},$$

$$B = \{HHH, THH, HTT, TTT\},$$

and

$$C = \{HHH, HTH, THT, TTT\}.$$

It follows that

$$\begin{aligned} \Pr(A) &= |A|/|S| = 4/8 = 1/2, \\ \Pr(B) &= |B|/|S| = 4/8 = 1/2, \\ \Pr(C) &= |C|/|S| = 4/8 = 1/2, \\ \Pr(A \cap B) &= |A \cap B|/|S| = 2/8 = 1/4, \\ \Pr(A \cap C) &= |A \cap C|/|S| = 2/8 = 1/4, \\ \Pr(B \cap C) &= |B \cap C|/|S| = 2/8 = 1/4. \end{aligned}$$

Thus, the sequence  $A, B, C$  is pairwise independent. Since

$$A \cap B \cap C = \{HHH, TTT\},$$

we have

$$\Pr(A \cap B \cap C) = |A \cap B \cap C|/|S| = 2/8 = 1/4.$$

Thus,

$$\Pr(A \cap B \cap C) \neq \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$$

and, therefore, the sequence  $A, B, C$  is not mutually independent. Of course, this is not surprising: If both events  $A$  and  $B$  occur, then event  $C$  also occurs.

## 5.12 Describing Events by Logical Propositions

We have defined an event to be a subset of a sample space. In several examples, however, we have described events in plain English or as logical propositions.

- Since the intersection ( $\cap$ ) of sets corresponds to the conjunction ( $\wedge$ ) of propositions, we often write  $A \wedge B$  for the event “both  $A$  and  $B$  occur”.
- Similarly, since the union ( $\cup$ ) of sets corresponds to the disjunction ( $\vee$ ) of propositions, we often write  $A \vee B$  for the event “ $A$  or  $B$  occurs”.

### 5.12.1 Flipping a Coin and Rolling a Die

If we flip a coin and roll a die, the sample space is

$$S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}.$$

The events

$$A = \text{“the coin comes up heads”}$$

and

$$B = \text{“the result of the die is even”}$$

correspond to the subsets

$$A = \{H1, H2, H3, H4, H5, H6\}$$

and

$$B = \{H2, H4, H6, T2, T4, T6\}$$

of the sample space  $S$ , respectively. The event that both  $A$  and  $B$  occur is written as  $A \wedge B$  and corresponds to the subset

$$A \cap B = \{H2, H4, H6\}$$

of  $S$ . The event that  $A$  or  $B$  occurs is written as  $A \vee B$  and corresponds to the subset

$$A \cup B = \{H1, H2, H3, H4, H5, H6, T2, T4, T6\}$$

of  $S$ .

Assume that both the coin and the die are fair, and the results of rolling the die and flipping the coin are independent. The probability that both  $A$  and  $B$  occur, i.e.,  $\Pr(A \wedge B)$ , is equal to  $|A \cap B|/|S| = 3/12 = 1/4$ . We can also use independence to determine this probability:

$$\Pr(A \wedge B) = \Pr(A) \cdot \Pr(B) = 1/2 \cdot 3/6 = 1/4.$$

Observe that when we determine  $\Pr(A)$ , we do not consider the entire sample space  $S$ . Instead, we consider the coin's sample space, which is  $\{H, T\}$ . Similarly, when we determine  $\Pr(B)$ , we consider the die's sample space, which is  $\{1, 2, 3, 4, 5, 6\}$ .

The probability that  $A$  or  $B$  occurs, i.e.,  $\Pr(A \vee B)$ , is equal to

$$\Pr(A \vee B) = |A \cup B|/|S| = 9/12 = 3/4.$$

### 5.12.2 Flipping Coins

Let  $n \geq 2$  be an integer and assume we flip  $n$  fair coins. For each  $i$  with  $1 \leq i \leq n$ , consider the event

$$A_i = \text{"the } i\text{-th coin comes up heads".}$$

We assume that the coin flips are independent of each other, by which we mean that the sequence  $A_1, A_2, \dots, A_n$  of events is mutually independent. Consider the event

$$A = A_1 \wedge A_2 \wedge \cdots \wedge A_n.$$

What is  $\Pr(A)$ , i.e., the probability that all  $n$  coins come up heads? Since there are  $2^n$  many possible outcomes for  $n$  coin flips and only one of them satisfies event  $A$ , this probability is equal to  $1/2^n$ . Alternatively, we can use independence to determine  $\Pr(A)$ :

$$\begin{aligned} \Pr(A) &= \Pr(A_1 \wedge A_2 \wedge \cdots \wedge A_n) \\ &= \Pr(A_1) \cdot \Pr(A_2) \cdots \Pr(A_n). \end{aligned}$$

Since each coin is fair, we have  $\Pr(A_i) = 1/2$  and, thus, we get

$$\Pr(A) = \underbrace{(1/2) \cdot (1/2) \cdots (1/2)}_{n \text{ times}} = (1/2)^n = 1/2^n.$$

### 5.12.3 The Probability of a Circuit Failing

Consider a circuit  $C$  that consists of  $n$  components  $C_1, C_2, \dots, C_n$ . Let  $p$  be a real number with  $0 < p < 1$  and assume that any component fails with probability  $p$ , independently of the other components. For each  $i$  with  $1 \leq i \leq n$ , consider the event

$$A_i = \text{"component } C_i \text{ fails".}$$

Let  $A$  be the event

$$A = \text{“the entire circuit fails”}.$$

- Assume that the entire circuit fails when at least one component fails. What is  $\Pr(A)$ , i.e., the probability that the circuit fails? By our assumption, we have

$$A = A_1 \vee A_2 \vee \cdots \vee A_n$$

and, thus, using De Morgan’s Law,

$$\overline{A} = \overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_n.$$

Using independence and Lemmas 5.3.3 and 5.11.2, we get

$$\begin{aligned}\Pr(A) &= 1 - \Pr(\overline{A}) \\ &= 1 - \Pr(\overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_n) \\ &= 1 - \Pr(\overline{A}_1) \cdot \Pr(\overline{A}_2) \cdots \Pr(\overline{A}_n) \\ &= 1 - \underbrace{(1-p)(1-p) \cdots (1-p)}_{n \text{ times}} \\ &= 1 - (1-p)^n.\end{aligned}$$

Since  $0 < p < 1$ , we have  $\lim_{n \rightarrow \infty} \Pr(A) = 1$ . We conclude that for large values of  $n$ , it is very likely that the circuit fails.

- Now assume that the entire circuit fails when all components fail. Again, we want to know the probability  $\Pr(A)$  that the circuit fails. In this case, we have

$$A = A_1 \wedge A_2 \wedge \cdots \wedge A_n,$$

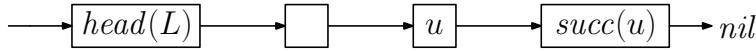
and we get

$$\begin{aligned}\Pr(A) &= \Pr(A_1 \wedge A_2 \wedge \cdots \wedge A_n) \\ &= \Pr(A_1) \cdot \Pr(A_2) \cdots \Pr(A_n) \\ &= \underbrace{p \cdot p \cdots p}_{n \text{ times}} \\ &= p^n.\end{aligned}$$

Since  $0 < p < 1$ , we have  $\lim_{n \rightarrow \infty} \Pr(A) = 0$ . Thus, for large values of  $n$ , it is very likely that the circuit does not fail.

## 5.13 Choosing a Random Element in a Linked List

Consider a linked list  $L$ . Each node  $u$  in  $L$  stores a pointer to its successor node  $\text{succ}(u)$ . If  $u$  is the last node in  $L$ , then  $u$  does not have a successor and  $\text{succ}(u) = \text{nil}$ . We are also given a pointer to the first node  $\text{head}(L)$  of  $L$ .



Our task is to choose, uniformly at random, a node in  $L$ . Thus, if this list has  $n$  nodes, then each node must have a probability of  $1/n$  of being chosen.

We assume that we are given a function `RANDOM`: For any integer  $i \geq 1$ , a call to `RANDOM( $i$ )` returns a uniformly random element from the set  $\{1, 2, \dots, i\}$ ; the value returned is independent of all other calls to this function.

To make the problem interesting, we assume that we do not know the value of  $n$ , i.e., at the start, we do not know the number of nodes in the list  $L$ . Also, we are allowed to only make one pass over this list. We will prove below that the following algorithm solves the problem:

**Algorithm** CHOSERANDOMNODE( $L$ ):

```

 $u = \text{head}(L);$ 
 $i = 1;$ 
while  $u \neq \text{nil}$ 
  do  $r = \text{RANDOM}(i);$ 
    if  $r = 1$ 
      then  $x = u$ 
      endif;
     $u = \text{succ}(u);$ 
     $i = i + 1$ 
  endwhile;
  return  $x$ 
```

In one iteration of the while-loop, the call to `RANDOM( $i$ )` returns a uniformly random element  $r$  from the set  $\{1, 2, \dots, i\}$ . If  $r = 1$ , which happens with probability  $1/i$ , the value of  $x$  is set to the currently visited node. If  $r \neq 1$ , which happens with probability  $1 - 1/i$ , the value of  $x$  does not change during this iteration of the while-loop. Thus,

- in the first iteration,  $x$  is set to the first node of  $L$  with probability 1,
- in the second iteration,  $x$  is set to the second node of  $L$  with probability  $1/2$ , whereas the value of  $x$  does not change with probability  $1/2$ ,
- in the third iteration,  $x$  is set to the third node of  $L$  with probability  $1/3$ , whereas the value of  $x$  does not change with probability  $2/3$ ,
- in the last iteration,  $x$  is set to the last node of  $L$  with probability  $1/|L|$ , whereas the value of  $x$  does not change with probability  $(|L| - 1)/|L|$ .

We now prove that the output  $x$  of algorithm **CHOSERANDOMNODE**( $L$ ) is a uniformly random node of the list  $L$ . Let  $n$  denote the number of nodes in  $L$  and let  $v$  be an arbitrary node in  $L$ . We will prove that, after the algorithm has terminated,  $x = v$  with probability  $1/n$ .

Let  $k$  be the integer such that  $v$  is the  $k$ -th node in  $L$ ; thus,  $1 \leq k \leq n$ . We observe that, after the algorithm has terminated,  $x = v$  if and only if

- during the  $k$ -th iteration, the value of  $x$  is set to  $v$ , and
- for all  $i = k + 1, k + 2, \dots, n$ , during the  $i$ -th iteration, the value of  $x$  does not change.

Consider the event

$$A = \text{“after the algorithm has terminated, } x = v\text{”}.$$

For each  $i$  with  $1 \leq i \leq n$ , consider the event

$$A_i = \text{“the value of } x \text{ changes during the } i\text{-th iteration”}.$$

Then

$$A = A_k \wedge \overline{A}_{k+1} \wedge \overline{A}_{k+2} \wedge \overline{A}_{k+3} \wedge \cdots \wedge \overline{A}_n.$$

Recall that we assume that the output of the function **RANDOM** is independent of all other calls to this function. This implies that the events

$A_1, A_2, \dots, A_n$  are mutually independent. It follows that

$$\begin{aligned}\Pr(A) &= \Pr(A_k \wedge \overline{A}_{k+1} \wedge \overline{A}_{k+2} \wedge \overline{A}_{k+3} \wedge \cdots \wedge \overline{A}_n) \\ &= \Pr(A_k) \cdot \Pr(\overline{A}_{k+1}) \cdot \Pr(\overline{A}_{k+2}) \cdot \Pr(\overline{A}_{k+3}) \cdots \Pr(\overline{A}_n) \\ &= \frac{1}{k} \cdot \left(1 - \frac{1}{k+1}\right) \cdot \left(1 - \frac{1}{k+2}\right) \cdot \left(1 - \frac{1}{k+3}\right) \cdots \left(1 - \frac{1}{n}\right) \\ &= \frac{1}{k} \cdot \frac{k}{k+1} \cdot \frac{k+1}{k+2} \cdot \frac{k+2}{k+3} \cdots \frac{n-1}{n} \\ &= \frac{1}{n}.\end{aligned}$$

## 5.14 Long Runs in Random Bitstrings

Let  $n$  be a large integer and assume we flip a fair coin  $n$  times, where all flips are mutually independent. If we write 0 for heads and 1 for tails, then we obtain a random bitstring

$$R = r_1 r_2 \dots r_n.$$

A *run of length  $k$*  is a substring of  $R$ , all of whose bits are the same. For example, the bitstring

$$00111100101000011000$$

contains, among others, the following substrings in bold,

$$00\mathbf{1111}00101\mathbf{0000}11000,$$

which are runs of lengths 4, 2, and 1, respectively.

Would you be surprised to see a “long” run in the random bitstring  $R$ , say a run of length about  $\log n$ ? Most people will answer this question with “yes”. We will prove below, however, that the correct answer is “no”: The probability that this happens is about  $1 - 1/n^2$ ; thus, it converges to 1 when  $n$  goes to infinity. In other words, you should be surprised if a random bitstring does *not* contain a run of length about  $\log n$ .

We choose a positive integer  $k$  and consider the event

$$A = \text{“}R \text{ contains a run of length at least } k\text{”}.$$

We are going to prove a lower bound on  $\Pr(A)$  in terms of  $n$  and  $k$ . At the end, we will show that by taking  $k$  to be slightly less than  $\log n$ , we have  $\Pr(A) \geq 1 - 1/n^2$ .

For each  $i$  with  $1 \leq i \leq n - k + 1$ , we consider the event

$$A_i = \text{“the substring of length } k \text{ starting at position } i \text{ is a run”}.$$

Since a run of length at least  $k$  can start at any of the positions  $1, 2, \dots, n - k + 1$ , we have

$$A = A_1 \vee A_2 \vee \cdots \vee A_{n-k+1},$$

implying that

$$\Pr(A) = \Pr(A_1 \vee A_2 \vee \cdots \vee A_{n-k+1}).$$

Observe that the events  $A_1, A_2, \dots, A_{n-k+1}$  are not pairwise disjoint. As a result, the probability on the right-hand side is difficult to analyze; it requires the Principle of Inclusion and Exclusion (see Section 3.5). Because of this, we consider the complement of the event  $A$ , i.e., the event

$$\overline{A} = \text{“each run in } R \text{ has length less than } k”.$$

Using De Morgan’s Law, we get

$$\overline{A} = \overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_{n-k+1},$$

where  $\overline{A}_i$  is the complement of  $A_i$ , i.e., the event

$$\overline{A}_i = \text{“the substring of length } k \text{ starting at position } i \text{ is not a run”}.$$

It follows that

$$\Pr(\overline{A}) = \Pr(\overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_{n-k+1}). \quad (5.5)$$

We determine  $\Pr(\overline{A}_i)$ , by first determining  $\Pr(A_i)$ . The event  $A_i$  occurs if and only if

$$r_i = r_{i+1} = \cdots = r_{i+k-1} = 0$$

or

$$r_i = r_{i+1} = \cdots = r_{i+k-1} = 1.$$

Since the coin flips are mutually independent, it follows that

$$\Pr(A_i) = 1/2^k + 1/2^k = 1/2^{k-1}$$

and, therefore,

$$\Pr(\overline{A}_i) = 1 - \Pr(A_i) = 1 - 1/2^{k-1}.$$

Is the probability on the right-hand side of (5.5) equal to the product of the individual probabilities? If the events  $\overline{A}_1, \overline{A}_2, \dots, \overline{A}_{n-k+1}$  are mutually independent, then the answer is “yes”. However, it should be clear that, for example, the events  $\overline{A}_1$  and  $\overline{A}_2$  are not independent: If we are told that event  $A_1$  occurs, then the first  $k$  bits in the bitstring  $R$  are equal; let us say they are all equal to 0. In this case, the probability that event  $A_2$  occurs is equal to the probability that the  $(k+1)$ -st bit in  $R$  is 0, which is equal to  $1/2$  and not  $1/2^{k-1}$  (assuming that  $k \geq 3$ ). It seems that we are stuck. Fortunately, there is a way out:

Let us assume that the integer  $k$  is chosen such that  $n/k$  is an integer. We divide the bitstring  $R = r_1r_2\dots r_n$  into  $n/k$  blocks, each having length  $k$ . Thus,

- the first block is the substring  $r_1r_2\dots r_k$ ,
- the second block is the substring  $r_{k+1}r_{k+2}\dots r_{2k}$ ,
- the third block is the substring  $r_{2k+1}r_{2k+2}\dots r_{3k}$ ,
- the  $(n/k)$ -th block is the substring  $r_{n-k+1}r_{n-k+2}\dots r_n$ .

For each  $i$  with  $1 \leq i \leq n/k$ , we consider the event

$$B_i = \text{“the } i\text{-th block is a run”}.$$

Thus, the complement of  $B_i$  is the event

$$\overline{B}_i = \text{“the } i\text{-th block is not a run”}.$$

Since  $B_i = A_{(i-1)k+1}$  and  $\overline{B}_i = \overline{A}_{(i-1)k+1}$ , we have

$$\Pr(\overline{B}_i) = 1 - 1/2^{k-1}.$$

Observe that

- the events  $\overline{B}_1, \overline{B}_2, \dots, \overline{B}_{n/k}$  are mutually independent, because the blocks do not overlap, and
- if the event  $\overline{A}$  occurs, then the event  $\overline{B}_1 \wedge \overline{B}_2 \wedge \dots \wedge \overline{B}_{n/k}$  also occurs (but, in general, the converse is not true!).

Using Lemma 5.3.6, it follows that

$$\begin{aligned}\Pr(\overline{A}) &\leq \Pr(\overline{B}_1 \wedge \overline{B}_2 \wedge \cdots \wedge \overline{B}_{n/k}) \\ &= \Pr(\overline{B}_1) \cdot \Pr(\overline{B}_2) \cdots \Pr(\overline{B}_{n/k}) \\ &= (1 - 1/2^{k-1}) \cdot (1 - 1/2^{k-1}) \cdots (1 - 1/2^{k-1}) \\ &= (1 - 1/2^{k-1})^{n/k}.\end{aligned}$$

Using the inequality  $1 - x \leq e^{-x}$ , see (5.3), we get

$$1 - 1/2^{k-1} \leq e^{-1/2^{k-1}} = e^{-2/2^k}$$

and, thus,

$$\Pr(\overline{A}) \leq \left(e^{-2/2^k}\right)^{n/k} = e^{-2n/(k2^k)}. \quad (5.6)$$

Note that until now,  $k$  was arbitrary. We choose  $k$  to be

$$k = \log n - 2 \log \log n.$$

Using basic properties of logarithms, see Section 2.4, we will show below that, for this choice of  $k$ , the right-hand side in (5.6) is a “nice” function of  $n$ .

In Section 2.4, we have seen that

$$2^{\log n} = n$$

and

$$2^{2 \log \log n} = \log^2 n.$$

It follows that

$$2^k = 2^{\log n - 2 \log \log n} = \frac{2^{\log n}}{2^{2 \log \log n}} = \frac{n}{\log^2 n}.$$

Thus,

$$\begin{aligned}\frac{2n}{k2^k} &= \frac{2 \log^2 n}{k} \\ &= \frac{2 \log^2 n}{\log n - 2 \log \log n} \\ &\geq \frac{2 \log^2 n}{\log n} \\ &= 2 \log n \\ &= 2 \frac{\ln n}{\ln 2} \\ &\geq 2 \ln n,\end{aligned}$$

implying that

$$\begin{aligned}\Pr(\overline{A}) &\leq e^{-2n/(k2^k)} \\ &\leq e^{-2 \ln n} \\ &= 1/n^2.\end{aligned}$$

We conclude that, for the value of  $k$  chosen above,

$$\Pr(A) = 1 - \Pr(\overline{A}) \geq 1 - 1/n^2.$$

Thus, with probability at least  $1 - 1/n^2$ , a random bitstring of length  $n$  contains a run of length at least  $\log n - 2 \log \log n$ .

We remark that we have been cheating, because we assumed that both  $k$  and  $n/k$  are integers. Assume that  $n$  is of the form  $2^{2^m}$ , for some positive integer  $m$ . Then both  $\log n$  and  $\log \log n$  are integers and, thus,  $k$  is an integer as well. In a correct derivation, we divide the bitstring  $R$  into  $\lfloor n/k \rfloor$  blocks of size  $k$  and, if  $n/k$  is not an integer, one block of length less than  $k$ . We then get

$$\begin{aligned}\Pr(\overline{A}) &\leq (1 - 1/2^{k-1})^{\lfloor n/k \rfloor} \\ &\leq \left(e^{-2/2^k}\right)^{\lfloor n/k \rfloor} \\ &= e^{-2\lfloor n/k \rfloor / 2^k}.\end{aligned}$$

As we have seen before, for  $k = \log n - 2 \log \log n$ , we have  $2^k = n/\log^2 n$ . Since

$$\lfloor n/k \rfloor > n/k - 1,$$

we get

$$\begin{aligned}\frac{2\lfloor n/k \rfloor}{2^k} &> \frac{2(n/k - 1)}{2^k} \\ &= \frac{(2 \log^2 n)(n/k - 1)}{n} \\ &= \frac{2 \log^2 n}{k} - \frac{2 \log^2 n}{n} \\ &\geq 2 \ln n - \frac{2 \log^2 n}{n}\end{aligned}$$

and, thus,

$$\begin{aligned}
 \Pr(\overline{A}) &\leq e^{-2\lfloor n/k \rfloor / 2^k} \\
 &\leq e^{-2\ln n + (2\log^2 n)/n} \\
 &= e^{-2\ln n} \cdot e^{(2\log^2 n)/n} \\
 &= (1/n^2) \cdot (1 + O((\log^2 n)/n)) \\
 &= 1/n^2 + O((\log^2 n)/n^3).
 \end{aligned}$$

This upper bound is larger than the upper bound we had before by only a small additive factor of  $O((\log^2 n)/n^3)$ .

## 5.15 Infinite Probability Spaces

In Section 5.2, we defined a sample space to be any non-empty countable set. All sample spaces that we have seen so far are finite. In some cases, infinite (but countable) sample spaces arise in a natural way. To give an example, assume we flip a fair coin repeatedly and independently until it comes up heads for the first time. The sample space  $S$  is the set of all sequences of coin flips that can occur. If we denote by  $T^n H$  the sequence consisting of  $n$  tails followed by one heads, then

$$\begin{aligned}
 S &= \{H, TH, TTH, TTTH, TTTTH, \dots\} \\
 &= \{T^n H : n \geq 0\},
 \end{aligned}$$

which is indeed an infinite set.

Since the coin is fair and the coin flips are mutually independent, the outcome  $T^n H$  has a probability of  $(1/2)^{n+1}$ , i.e.,

$$\Pr(T^n H) = (1/2)^{n+1}.$$

Recall that according to Definition 5.2.2, in order for this to be a valid probability function, the sum of all probabilities must be equal to 1, i.e., the infinite series

$$\sum_{n=0}^{\infty} \Pr(T^n H) = \sum_{n=0}^{\infty} (1/2)^{n+1}$$

must be equal to 1. Since you may have forgotten about infinite series, we recall the definition in the following subsection.

### 5.15.1 Infinite Series

The divergent series are the invention of the devil, and it is a shame to base on them any demonstration whatsoever.

— Niels Henrik Abel, 1828

**Definition 5.15.1** Let  $a_0, a_1, a_2, \dots$  be an infinite sequence of real numbers. If

$$\lim_{N \rightarrow \infty} \sum_{n=0}^N a_n = \lim_{N \rightarrow \infty} (a_0 + a_1 + a_2 + \dots + a_N)$$

exists, then we say that the infinite series  $\sum_{n=0}^{\infty} a_n$  converges. In this case, the value of this infinite series is equal to

$$\sum_{n=0}^{\infty} a_n = \lim_{N \rightarrow \infty} \sum_{n=0}^N a_n.$$

For example, let  $x$  be a real number with  $x \neq 1$ , and define  $a_n = x^n$  for  $n \geq 0$ . We claim that

$$\sum_{n=0}^N a_n = \sum_{n=0}^N x^n = 1 + x + x^2 + \dots + x^N = \frac{1 - x^{N+1}}{1 - x},$$

which can be proved either by induction on  $N$  or by verifying that

$$(1 - x)(1 + x + x^2 + \dots + x^N) = 1 - x^{N+1}.$$

If  $-1 < x < 1$ , then  $\lim_{N \rightarrow \infty} x^{N+1} = 0$  and it follows that

$$\begin{aligned} \sum_{n=0}^{\infty} x^n &= \lim_{N \rightarrow \infty} \sum_{n=0}^N x^n \\ &= \lim_{N \rightarrow \infty} \frac{1 - x^{N+1}}{1 - x} \\ &= \frac{1}{1 - x}. \end{aligned}$$

We have proved the following result:

**Lemma 5.15.2** *If  $x$  is a real number with  $-1 < x < 1$ , then*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Now we can return to the coin flipping example that we saw in the beginning of Section 5.15. If we take  $x = 1/2$  in Lemma 5.15.2, then we get

$$\begin{aligned} \sum_{n=0}^{\infty} \Pr(T^n H) &= \sum_{n=0}^{\infty} (1/2)^{n+1} \\ &= (1/2) \sum_{n=0}^{\infty} (1/2)^n \\ &= (1/2) \cdot \frac{1}{1 - 1/2} \\ &= 1. \end{aligned}$$

Thus, we indeed have a valid probability function on the infinite sample space  $S = \{T^n H : n \geq 0\}$ .

The limit does not exist.

— Cady Heron (played by Lindsay Lohan),  
— *Mean Girls*, 2004

We have seen in Lemma 5.15.2 that the infinite series  $\sum_{n=0}^{\infty} x^n$  converges if  $-1 < x < 1$ . It is not difficult to see that for all other values of  $x$ , the limit

$$\lim_{N \rightarrow \infty} \sum_{n=0}^N x^n$$

does not exist. As a result, if  $x \geq 1$  or  $x \leq -1$ , the infinite series  $\sum_{n=0}^{\infty} x^n$  does not converge. Another example of an infinite series that does not converge is

$$\sum_{n=1}^{\infty} 1/n = 1 + 1/2 + 1/3 + 1/4 + \dots$$

In Section 6.8.3, we will prove that

$$\sum_{n=1}^N 1/n = 1 + 1/2 + 1/3 + 1/4 + \dots + 1/N$$

is about  $\ln N$ . It follows that

$$\lim_{N \rightarrow \infty} \sum_{n=0}^N 1/n$$

is about

$$\lim_{N \rightarrow \infty} \ln N,$$

which clearly does not exist.

### 5.15.2 Who Flips the First Heads

Consider a game in which two players  $P_1$  and  $P_2$  take turns flipping, independently, a fair coin. Thus, first  $P_1$  flips the coin, then  $P_2$  flips the coin, then  $P_1$  flips the coin, then  $P_2$  flips the coin, etc. The player who flips heads first is the winner of the game.

Who is more likely to win this game? Our intuition says that  $P_1$  has an advantage, because he is the player who starts: If the first flip is heads, then the game is over and  $P_1$  wins. We will prove below that this intuition is correct:  $P_1$  has a probability of  $2/3$  of winning the game and, thus, the winning probability of  $P_2$  is only  $1/3$ .

The sample space  $S$  is the set of all sequences of coin flips that can occur. Since the game is over as soon as a heads is flipped, we have

$$S = \{T^n H : n \geq 0\}.$$

Since  $P_1$  starts, the event

$$A = "P_1 \text{ wins the game}"$$

corresponds to the subset

$$A = \{T^n H : n \geq 0 \text{ and } n \text{ is even}\},$$

which we rewrite as

$$A = \{T^{2m} H : m \geq 0\}.$$

The probability that  $P_1$  wins the game is equal to  $\Pr(A)$ . How do we determine this probability? According to (5.1) in Section 5.2,

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega).$$

Since each outcome  $\omega$  in  $A$  is of the form  $T^{2m}H$ , we have

$$\Pr(A) = \sum_{m=0}^{\infty} \Pr(T^{2m}H).$$

Thus, we have

$$\begin{aligned}\Pr(A) &= \sum_{m=0}^{\infty} \Pr(T^{2m}H) \\ &= \sum_{m=0}^{\infty} (1/2)^{2m+1} \\ &= (1/2) \sum_{m=0}^{\infty} (1/2)^{2m} \\ &= (1/2) \sum_{m=0}^{\infty} (1/4)^m.\end{aligned}$$

By taking  $x = 1/4$  in Lemma 5.15.2, we get

$$\Pr(A) = (1/2) \cdot \frac{1}{1 - 1/4} = 2/3.$$

Let  $B$  be the event

$$B = "P_2 \text{ wins the game}".$$

Since either  $P_1$  or  $P_2$  wins the game, we have

$$\Pr(B) = 1 - \Pr(A) = 1 - 2/3 = 1/3.$$

Let us verify, using an infinite series, that  $\Pr(B)$  is indeed equal to  $1/3$ . The event  $B$  corresponds to the subset

$$B = \{T^nH : n \geq 0 \text{ and } n \text{ is odd}\},$$

which we rewrite as

$$B = \{T^{2m+1}H : m \geq 0\}.$$

The probability that  $P_2$  wins the game is thus equal to

$$\begin{aligned}\Pr(B) &= \sum_{m=0}^{\infty} \Pr(T^{2m+1}H) \\ &= \sum_{m=0}^{\infty} (1/2)^{2m+2} \\ &= (1/4) \sum_{m=0}^{\infty} (1/2)^{2m} \\ &= (1/4) \sum_{m=0}^{\infty} (1/4)^m \\ &= (1/4) \cdot \frac{1}{1 - 1/4} = 1/3.\end{aligned}$$

### 5.15.3 Who Flips the Second Heads

Let us change the game from the previous subsection: Again, the two players  $P_1$  and  $P_2$  take turns flipping, independently, a fair coin, where  $P_1$  starts. The game ends as soon as a second heads comes up. The player who flips the second heads wins the game.

Before you continue reading: Who do you think has a higher probability of winning this game?

In this game, a sequence of coin flips can occur if and only if (i) the sequence contains exactly two heads and (ii) the last element in the sequence is heads. Thus, the sample space  $S$  is given by

$$S = \{T^m HT^n H : m \geq 0, n \geq 0\}.$$

The event

$$A = "P_1 \text{ wins the game}"$$

corresponds to the subset

$$A = \{T^m HT^n H : m \geq 0, n \geq 0, m + n \text{ is odd}\}.$$

Below, we will determine  $\Pr(A)$ , i.e., the probability that  $P_1$  wins the game.

We split the event  $A$  into two events

$$A_1 = "P_1 \text{ flips both the first and the second heads}"$$

and

$$A_2 = \text{“}P_2 \text{ flips the first heads and } P_1 \text{ flips the second heads”}.$$

If we write these two events as subsets of the sample space  $S$ , we get

$$\begin{aligned} A_1 &= \{T^m HT^n H : m \geq 0, n \geq 0, m \text{ is even and } n \text{ is odd}\} \\ &= \{T^{2k} HT^{2\ell+1} H : k \geq 0, \ell \geq 0\} \end{aligned}$$

and

$$\begin{aligned} A_2 &= \{T^m HT^n H : m \geq 0, n \geq 0, m \text{ is odd and } n \text{ is even}\} \\ &= \{T^{2k+1} HT^{2\ell} H : k \geq 0, \ell \geq 0\}. \end{aligned}$$

Observe that  $A_1 \cap A_2 = \emptyset$  and  $A = A_1 \cup A_2$ , implying that

$$\Pr(A) = \Pr(A_1) + \Pr(A_2).$$

We determine the two probabilities on the right-hand side.

We have

$$\begin{aligned} \Pr(A_1) &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \Pr(T^{2k} HT^{2\ell+1} H) \\ &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} (1/2)^{2k+2\ell+3} \\ &= (1/2^3) \sum_{k=0}^{\infty} (1/2)^{2k} \sum_{\ell=0}^{\infty} (1/2)^{2\ell} \\ &= (1/8) \sum_{k=0}^{\infty} (1/4)^k \sum_{\ell=0}^{\infty} (1/4)^{\ell} \\ &= (1/8) \sum_{k=0}^{\infty} (1/4)^k \cdot \frac{1}{1 - 1/4} \\ &= (1/6) \sum_{k=0}^{\infty} (1/4)^k \\ &= (1/6) \cdot \frac{1}{1 - 1/4} \\ &= 2/9 \end{aligned}$$

and

$$\begin{aligned}\Pr(A_2) &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \Pr(T^{2k+1}HT^{2\ell}H) \\ &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} (1/2)^{2k+2\ell+3} \\ &= 2/9.\end{aligned}$$

Thus, the probability that  $P_1$  wins the game is equal to

$$\begin{aligned}\Pr(A) &= \Pr(A_1) + \Pr(A_2) \\ &= 2/9 + 2/9 \\ &= 4/9.\end{aligned}$$

The probability that  $P_2$  wins the game is equal to

$$1 - \Pr(A) = 5/9.$$

Thus,  $P_2$  has a slightly larger probability of winning the game.

You will agree that this was a painful way of determining  $\Pr(A)$ . In Exercise 5.91, you will see an easier way to determine this probability: The game of this subsection can be seen as two rounds of the game in Section 5.15.2. This observation, together with the Law of Total Probability (Theorem 5.9.1) leads to an easier way to prove that  $\Pr(A) = 4/9$ .

## 5.16 Exercises

**5.1** Consider a coin that has 0 on one side and 1 on the other side. We flip this coin once and roll a die twice, and are interested in the product of the three numbers.

- What is the sample space?
- How many possible events are there?
- If both the coin and the die are fair, how would you define the probability function  $\Pr$  for this sample space?

**5.2** Consider the sample space  $S = \{a, b, c, d\}$  and a probability function  $\Pr : S \rightarrow \mathbb{R}$  on  $S$ . Consider the events  $A = \{a\}$ ,  $B = \{a, b\}$ ,  $C = \{a, b, c\}$ , and  $D = \{b, d\}$ . You are given that  $\Pr(A) = 1/10$ ,  $\Pr(B) = 1/2$ , and  $\Pr(C) = 7/10$ . Determine  $\Pr(D)$ .

**5.3** Let  $n$  be a positive integer. We flip a fair coin  $2n$  times and consider the possible outcomes, which are strings of length  $2n$  with each character being  $H$  (= heads) or  $T$  (= tails). Thus, we take the sample space  $S$  to be the set of all such strings. Since our coin is fair, each string of  $S$  should have the same probability. Thus, we define  $\Pr(s) = 1/|S|$  for each string  $s$  in  $S$ . In other words, we have a uniform probability space.

You are asked to determine the probability that in the sequence of  $2n$  flips, the coin comes up heads exactly  $n$  times:

- What is the event  $A$  that describes this?
- Determine  $\Pr(A)$ .

**5.4** A cup contains two pennies (P), one nickel (N), and one dime (D). You choose one coin uniformly at random, and then you choose a second coin from the remaining coins, again uniformly at random.

- Let  $S$  be the sample space consisting of all ordered pairs of letters P, N, and D that represent the possible outcomes. Write out all elements of  $S$ .
- Determine the probability for each element in this sample space.

**5.5** You are given a box that contains the 8 lowercase letters  $a, b, c, d, e, f, g, h$  and the 5 uppercase letters  $V, W, X, Y, Z$ .

In this exercise, we will consider two ways to choose 4 random letters from the box. In the first way, we do uniform sampling without replacement, whereas in the second way, we do uniform sampling with replacement. For each case, you are asked to determine the probability that the 4-th letter chosen is an uppercase letter. Before starting this exercise, spend a few minutes and guess for which case this probability is smaller.

- You choose 4 letters from the box: These letters are chosen in 4 steps, and in each step, you choose a uniformly random letter from the box; this letter is removed from the box.

– What is the sample space?

– Consider the event

$$A = \text{“the 4-th letter chosen is an uppercase letter”}.$$

Determine  $\Pr(A)$ .

- You choose 4 letters from the box: These letters are chosen in 4 steps, and in each step, you choose a uniformly random letter from the box; this letter is *not* removed from the box.

– What is the sample space?

– Consider the event

$$B = \text{“the 4-th letter chosen is an uppercase letter”}.$$

Determine  $\Pr(B)$ .

- 5.6** You flip a fair coin, independently, six times.

- What is the sample space?
- Consider the events

$$A = \text{“the coin comes up heads at least four times”},$$

$$B = \text{“the number of heads is equal to the number of tails”},$$

$$C = \text{“there are at least four consecutive heads”}.$$

Determine  $\Pr(A)$ ,  $\Pr(B)$ ,  $\Pr(C)$ ,  $\Pr(A | B)$ , and  $\Pr(C | A)$ .

- 5.7** Let  $k \geq 2$  be an integer and consider the sample space  $S$  consisting of all sequences of  $k$  characters, where each character is one of the digits  $0, 1, 2, \dots, 9$ .

If we choose a sequence  $s$  uniformly at random from the sample space  $S$ , what is the probability that none of the digits in  $s$  is equal to 5?

- 5.8** You are given a red coin and a blue coin. Both coins have the number 1 on one side and the number 2 on the other side. You flip both coins once (independently of each other) and take the sum of the two results. Consider the events

$$A = \text{“the sum of the results equals 2”},$$

$$B = \text{“the sum of the results equals 3”},$$

$$C = \text{“the sum of the results equals 4”}.$$

- Assume both coins are fair. Determine  $\Pr(A)$ ,  $\Pr(B)$ , and  $\Pr(C)$ .
- Let  $p$  and  $q$  be real numbers with  $0 < p < 1$  and  $0 < q < 1$ . Assume the red coin comes up “1” with probability  $p$  and the blue coin comes up “1” with probability  $q$ . Is it possible to choose  $p$  and  $q$  such that

$$\Pr(A) = \Pr(B) = \Pr(C)?$$

**5.9** Let  $p_1, p_2, \dots, p_6, q_1, q_2, \dots, q_6$  be real numbers such that each  $p_i$  is strictly positive, each  $q_i$  is strictly positive, and  $p_1 + p_2 + \dots + p_6 = q_1 + q_2 + \dots + q_6 = 1$ .

You are given a red die and a blue die. For any  $i$  with  $1 \leq i \leq 6$ , if you roll the red die, then the result is  $i$  with probability  $p_i$ , and if you roll the blue die, then the result is  $i$  with probability  $q_i$ .

You roll each die once (independently of each other) and take the sum of the two results. For any  $s \in \{2, 3, \dots, 12\}$ , consider the event

$$A_s = \text{“the sum of the results equals } s\text{”}.$$

- Let  $x > 0$  and  $y > 0$  be real numbers. Prove that

$$\frac{x}{y} + \frac{y}{x} \geq 2.$$

*Hint:* Rewrite this inequality until you get an equivalent inequality which obviously holds.

- Assume that  $\Pr(A_2) = \Pr(A_{12})$  and denote this common value by  $a$ . Prove that

$$\Pr(A_7) \geq 2a.$$

- Is it possible to choose  $p_1, p_2, \dots, p_6, q_1, q_2, \dots, q_6$  such that for any  $s \in \{2, 3, \dots, 12\}$ ,  $\Pr(A_s) = 1/11$ ?

**5.10** The Fibonacci numbers are defined as follows:  $f_0 = 0$ ,  $f_1 = 1$ , and  $f_n = f_{n-1} + f_{n-2}$  for  $n \geq 2$ .

Let  $n$  be a large integer. A *Fibonacci die* is a die that has  $f_n$  faces. Such a die is fair: If we roll it, each face is on top with the same probability  $1/f_n$ . There are three different types of Fibonacci dice:

- $D_1$ :  $f_{n-2}$  of its faces show the number 1 and the other  $f_{n-1}$  faces show the number 4.

- $D_2$ : Each face shows the number 3.
- $D_3$ :  $f_{n-2}$  of its faces show the number 5 and the other  $f_{n-1}$  faces show the number 2.

Assume we roll each of  $D_1$ ,  $D_2$ , and  $D_3$  once, independently of each other. Let  $R_1$ ,  $R_2$ , and  $R_3$  be the numbers on the top faces of  $D_1$ ,  $D_2$ , and  $D_3$ , respectively. Determine

$$\Pr(R_1 > R_2)$$

and

$$\Pr(R_2 > R_3),$$

and show that

$$\Pr(R_3 > R_1) = \frac{f_{n-2}f_{n+1}}{f_n^2}.$$

**5.11** You are given a fair die. If you roll this die repeatedly, then the results of the rolls are independent of each other.

- You roll the die 6 times. Consider the event

$$A = \text{"there is at least one 6 in this sequence of 6 rolls".}$$

Determine  $\Pr(A)$ .

- You roll the die 12 times. Consider the event

$$B = \text{"there are at least two 6's in this sequence of 12 rolls".}$$

Determine  $\Pr(B)$ .

- You roll the die 18 times. Consider the event

$$C = \text{"there are at least three 6's in this sequence of 18 rolls".}$$

Determine  $\Pr(C)$ .

Before starting this exercise, spend a few minutes and guess which of these three probabilities is the smallest.

**5.12** When Tri is a big boy, he wants to have four children. Assuming that the genders of these children are uniformly random, which of the following three events has the highest probability?

1. All four kids are of the same gender.
2. Exactly three kids are of the same gender.
3. Two kids are boys and two kids are girls.

**5.13** A group of ten people sits down, uniformly at random, around a table. Lindsay and Simon are part of this group. Determine the probability that Lindsay and Simon sit next to each other.

**5.14** Consider five people, each of which has a uniformly random and independent birthday. (We ignore leap years.) Consider the event

$$A = \text{“at least three people have the same birthday”}.$$

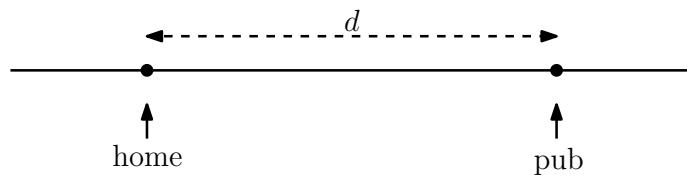
Determine  $\Pr(A)$ .

**5.15** Donald Trump wants to hire two secretaries. There are  $n$  applicants  $a_1, a_2, \dots, a_n$ , where  $n \geq 2$  is an integer. Each of these applicants has a uniformly random birthday, and all birthdays are mutually independent. (We ignore leap years.)

Since Donald is too busy making America great again, he does not have time to interview the applicants. Instead, he uses the following strategy: If there is an index  $i$  such that  $a_i$  and  $a_{i+1}$  have the same birthday, then he chooses the smallest such index  $i$  and hires  $a_i$  and  $a_{i+1}$ . In this case, the hiring process is a *tremendous success*. If such an index  $i$  does not exist, then nobody is hired and the hiring process is a *total disaster*.

Determine the probability that the hiring process is a tremendous success.

**5.16** Let  $d$  and  $n$  be integers such that  $d \geq 1$ ,  $n \geq d$ , and  $n + d$  is even. You live on Somerset Street and want to go to your local pub, which is also located on Somerset Street, at distance  $d$  to the east from your home.



You use the following strategy:

- Initially, you are at your home.
- For each  $i = 1, 2, \dots, n$ , you do the following:
  - You flip a fair and independent coin.
  - If the coin comes up heads, you walk a distance 1 to the east.
  - If the coin comes up tails, you walk a distance 1 to the west.

Consider the event

$$A = \text{“after these } n \text{ steps, you are at your local pub”}.$$

Prove that

$$\Pr(A) = \binom{n}{\frac{n+d}{2}} / 2^n.$$

**5.17** In Section 5.4.1, we have seen the different cards that are part of a standard deck of cards.

- You choose 2 cards uniformly at random from the 13 spades in a deck of 52 cards. Determine the probability that you choose an Ace and a King.
- You choose 2 cards uniformly at random from a deck of 52 cards. Determine the probability that you choose an Ace and a King.
- You choose 2 cards uniformly at random from a deck of 52 cards. Determine the probability that you choose an Ace and a King of the same suit.

**5.18** In Section 5.4.1, we have seen the different cards that are part of a standard deck of cards.

A *hand of cards* is a subset consisting of five cards. A hand of cards is called a *straight*, if the ranks of these five cards are consecutive and the cards are not all of the same suit.

An Ace and a 2 are considered to be consecutive, whereas a King and an Ace are also considered to be consecutive. For example, each of the three hands below is a straight:

$$8\spadesuit, 9\heartsuit, 10\diamondsuit, J\spadesuit, Q\clubsuit$$

$$A\lozenge, 2\heartsuit, 3\spadesuit, 4\spadesuit, 5\clubsuit$$

$$10\lozenge, J\heartsuit, Q\spadesuit, K\spadesuit, A\clubsuit$$

- Assume you get a uniformly random hand of cards. Determine the probability that this hand is a straight.

**5.19** Three people  $P_1$ ,  $P_2$ , and  $P_3$  are in a dark room. Each person has a bag containing one red hat and one blue hat. Each person chooses a uniformly random hat from her bag and puts it on her head. Afterwards, the lights are turned on.

Each person does not know the color of her hat, but can see the colors of the other two hats. Each person  $P_i$  can do one of the following:

- Person  $P_i$  announces “my hat is red”.
- Person  $P_i$  announces “my hat is blue”.
- Person  $P_i$  says “I pass”.

The game is a *success* if at least one person announces the correct color of her hat and no person announces the wrong color of her hat. (If a person passes, then she does not announce any color.)

- Assume person  $P_1$  announces “my hat is red” and both  $P_2$  and  $P_3$  pass. Consider the event

$$A = \text{“the game is a success”}.$$

Determine  $\Pr(A)$ .

- Assume each person  $P_i$  does the following:

- If the two hats that  $P_i$  sees have different colors, then  $P_i$  passes.
- If the two hats that  $P_i$  sees are both red, then  $P_i$  announces “my hat is blue”.
- If the two hats that  $P_i$  sees are both blue, then  $P_i$  announces “my hat is red”.

Consider the event

$$B = \text{“the game is a success”}.$$

Determine  $\Pr(B)$ .

**5.20** Let  $A$  be an event in some probability space  $(S, \Pr)$ . You are given that the events  $A$  and  $\bar{A}$  are independent<sup>2</sup>. Determine  $\Pr(A)$ .

**5.21** You are given three events  $A$ ,  $B$ , and  $C$  in some probability space  $(S, \Pr)$ . Is the following true or false?

$$\Pr(A \cap \bar{B} \cap \bar{C}) = \Pr(A \cup B \cup C) - \Pr(B) - \Pr(C) + \Pr(B \cap C).$$

**5.22** Let  $S$  be a set consisting of 6 positive integers and 8 negative integers. Choose a 4-element subset of  $S$  uniformly at random, and multiply the elements in this subset. Denote the product by  $x$ . Determine the probability that  $x > 0$ .

**5.23** Prove the inequality in (5.3), i.e., prove that

$$1 - x \leq e^{-x}$$

for all real numbers  $x$ .

**5.24** Let  $(S, \Pr)$  be a probability space and let  $B$  be an event with  $\Pr(B) > 0$ . Consider the function  $\Pr' : S \rightarrow \mathbb{R}$  by

$$\Pr'(\omega) = \begin{cases} \frac{\Pr(\omega)}{\Pr(B)} & \text{if } \omega \in B, \\ 0 & \text{if } \omega \notin B. \end{cases}$$

- Prove that  $\Pr'$  is a probability function on  $S$  according to Definition 5.2.2.
- Prove that for any event  $A$ ,

$$\Pr'(A) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

**5.25** Consider two events  $A$  and  $B$  in some probability space  $(S, \Pr)$ .

- Assume that  $\Pr(A) = 1/2$  and  $\Pr(B | \bar{A}) = 3/5$ . Determine  $\Pr(A \cup B)$ .
- Assume that  $\Pr(A \cup B) = 5/6$  and  $\Pr(\bar{A} | \bar{B}) = 1/3$ . Determine  $\Pr(B)$ .

---

<sup>2</sup>This is not a typo.

**5.26** Give an example of a sample space  $S$  and six events  $A, B, C, D, E$ , and  $F$  such that

- $\Pr(A | B) = \Pr(A)$ ,
- $\Pr(C | D) < \Pr(C)$ ,
- $\Pr(E | F) > \Pr(E)$ .

*Hint:* The sequence of six events may contain duplicates. Try to make the sample space  $S$  as small as you can.

**5.27** You roll a fair die twice. Consider the events

$$\begin{aligned} A &= \text{“the sum of the two rolls is 7”}, \\ B &= \text{“the result of the first roll is 4”}. \end{aligned}$$

Determine the conditional probabilities  $\Pr(A | B)$  and  $\Pr(B | A)$ .

**5.28** You flip a fair coin three times. Consider the four events (recall that zero is even)

$$\begin{aligned} A &= \text{“the coin comes up heads an odd number of times”}, \\ B &= \text{“the coin comes up heads an even number of times”}, \\ C &= \text{“the coin comes up tails an odd number of times”}, \\ D &= \text{“the coin comes up tails an even number of times”}. \end{aligned}$$

- Determine  $\Pr(A), \Pr(B), \Pr(C), \Pr(D), \Pr(A | C)$ , and  $\Pr(A | D)$ .
- Are there any two events in the sequence  $A, B, C$ , and  $D$  that are independent?

**5.29** Consider a box that contains four beer bottles  $b_1, b_2, b_3, b_4$  and two cider bottles  $c_1, c_2$ . You choose a uniformly random bottle from the box (and do not put it back), after which you again choose a uniformly random bottle from the box.

Consider the events

$$\begin{aligned} A &= \text{“the first bottle chosen is a beer bottle”}, \\ B &= \text{“the second bottle chosen is a beer bottle”}. \end{aligned}$$

- What is the sample space?
- For each element  $\omega$  in your sample space, determine  $\Pr(\omega)$ .
- Determine  $\Pr(A)$ .
- Determine  $\Pr(B)$ .
- Are the events  $A$  and  $B$  independent?

**5.30** A standard deck of 52 cards contains 13 spades ( $\spadesuit$ ), 13 hearts ( $\heartsuit$ ), 13 clubs ( $\clubsuit$ ), and 13 diamonds ( $\diamondsuit$ ). You choose a uniformly random card from this deck. Consider the events

$$\begin{aligned}A &= \text{"the chosen card is a clubs or a diamonds card"}, \\B &= \text{"the chosen card is a clubs or a hearts card"}, \\C &= \text{"the chosen card is a clubs or a spades card"}.\end{aligned}$$

- Are the events  $A$ ,  $B$ , and  $C$  pairwise independent?
- Are the events  $A$ ,  $B$ , and  $C$  mutually independent?

**5.31** You roll a fair die twice. Consider the events

$$\begin{aligned}A &= \text{"the sum of the results is at least 9"}, \\B &= \text{"at least one of the two rolls results in 2"}, \\C &= \text{"at least one of the two rolls results in 5"}.\end{aligned}$$

- Determine  $\Pr(A)$ ,  $\Pr(B)$ , and  $\Pr(C)$ .
- Determine  $\Pr(B | C)$ .
- Are the events  $A$  and  $B$  independent?
- Are the events  $A$  and  $C$  independent?

**5.32** A hand of 13 cards is chosen uniformly at random from a standard deck of 52 cards. Consider the events

$$\begin{aligned}A &= \text{"the hand has at least one Ace"}, \\B &= \text{"the hand has at least two Aces"}, \\C &= \text{"the hand has the Ace of spades"}.\end{aligned}$$

Determine the conditional probabilities  $\Pr(A | B)$ ,  $\Pr(B | A)$ , and  $\Pr(B | C)$ .

**5.33** We take a uniformly random permutation of a standard deck of 52 cards, so that each permutation has a probability of  $1/52!$ . Consider the events

$$\begin{aligned} A &= \text{"the top card is an Ace"}, \\ B &= \text{"the bottom card is the Ace of spades"}, \\ C &= \text{"the bottom card is the Queen of spades"}. \end{aligned}$$

Determine the conditional probabilities  $\Pr(A | B)$  and  $\Pr(A | C)$ .

**5.34** Consider two dice, each one having one face showing the letter  $a$ , two faces showing the letter  $b$ , and the remaining three faces showing the letter  $c$ . You roll each die once, independently of the other die.

- What is the sample space?

- Consider the events

$$\begin{aligned} A &= \text{"at least one of the two dice shows the letter } b \text{ on its top face"}, \\ B &= \text{"both dice show the same letter on their top faces"}. \end{aligned}$$

Determine  $\Pr(A)$ ,  $\Pr(B)$ , and  $\Pr(A | B)$ .

**5.35** You flip a fair coin, independently, three times. Consider the events

$$\begin{aligned} A &= \text{"the first flip results in heads"}, \\ B &= \text{"the coin comes up heads exactly once"}. \end{aligned}$$

Determine the conditional probabilities  $\Pr(A | B)$  and  $\Pr(B | A)$ .

**5.36** You roll a fair die twice. Consider the events

$$\begin{aligned} A &= \text{"the sum of the results is even"}, \\ B &= \text{"the sum of the results is at least 10"}. \end{aligned}$$

Determine the conditional probability  $\Pr(A | B)$ .

**5.37** You flip a fair coin seven times, independently of each other. Consider the events

$$\begin{aligned} A &= \text{"the number of heads is at least six"}, \\ B &= \text{"the number of heads is at least five"}, \\ C &= \text{"the number of tails is at least two"}, \\ D &= \text{"the number of heads is at least four"}. \end{aligned}$$

Determine the conditional probabilities  $\Pr(A | B)$  and  $\Pr(C | D)$ .

**5.38** Consider the set  $Y = \{1, 2, 3, \dots, 10\}$ . We choose, uniformly at random, a 6-element subset  $X$  of  $Y$ . Consider the events

$$\begin{aligned} A &= \text{“5 is an element of } X\text{”}, \\ B &= \text{“6 is an element of } X\text{”}, \\ C &= \text{“6 is an element of } X \text{ or 7 is an element of } X\text{”}. \end{aligned}$$

- Determine  $\Pr(A)$ ,  $\Pr(B)$ , and  $\Pr(C)$ .
- Determine  $\Pr(A | B)$ ,  $\Pr(A | C)$ , and  $\Pr(B | C)$ .

**5.39** Let  $A$  and  $B$  be two events in some probability space  $(S, \Pr)$  such that  $\Pr(A) = 2/5$  and  $\Pr(\overline{A \cup B}) = 3/10$ .

- Assume that  $A$  and  $B$  are disjoint. Determine  $\Pr(B)$ .
- Assume that  $A$  and  $B$  are independent. Determine  $\Pr(B)$ .

**5.40** In this exercise, we assume that, when a child is born, its gender is uniformly random, its day of birth is uniformly random, the gender and day of birth are independent of each other and independent of other children.

Anil Maheshwari has two children. You are given that at least one of Anil's kids is a boy who was born on a Sunday. Determine the probability that Anil has two boys.

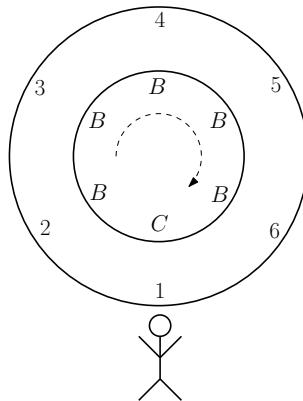
**5.41** Elisa and Nick go to Tan Tran's Darts Bar. When Elisa throws a dart, she hits the dartboard with probability  $p$ . When Nick throws a dart, he hits the dartboard with probability  $q$ . Here,  $p$  and  $q$  are real numbers with  $0 < p < 1$  and  $0 < q < 1$ . Elisa and Nick throw one dart each, independently of each other. Consider the events

$$\begin{aligned} E &= \text{“Elisa's dart hits the dartboard”}, \\ N &= \text{“Nick's dart hits the dartboard”}. \end{aligned}$$

Determine  $\Pr(E | E \cup N)$  and  $\Pr(E \cap N | E \cup N)$ .

**5.42** As everyone knows, Elisa Kazan loves to drink cider. You may not be aware that Elisa is not a big fan of beer.

Consider a round table that has six seats numbered 1, 2, 3, 4, 5, 6. Elisa is sitting in seat 1. On top of the table, there is a rotating tray<sup>3</sup>. On this tray, there are five bottles of beer ( $B$ ) and one bottle of cider ( $C$ ), as in the figure below. After the tray has been spun, there is always a bottle exactly in front of Elisa. (In other words, you can only spin the tray by a multiple of 60 degrees.) Moreover, Elisa can only see the bottle that is in front of her.



Elisa spins the tray uniformly at random in clockwise order. After the tray has come to a rest, there is a bottle of beer in front of her. Since Elisa is obviously not happy, she gets a second chance, i.e., Elisa can choose between one of the following two options:

1. Spin the tray again uniformly at random and independently of the first spin. After the tray has come to a rest, Elisa must drink the bottle that is in front of her.
  2. Rotate the tray one position (i.e., 60 degrees) in clockwise order, after which Elisa must drink the bottle that is in front of her.
- Elisa decides to go for the first option. Determine the probability that she drinks the bottle of cider.
  - Elisa decides to go for the second option. Determine the probability that she drinks the bottle of cider.

**5.43** You are given three dice  $D_1$ ,  $D_2$ , and  $D_3$ :

---

<sup>3</sup>According to Wikipedia, such a tray is called a Lazy Susan or Lazy Suzy. You may have seen them in Chinese restaurants.

- Die  $D_1$  has 0 on two of its faces and 1 on the other four faces.
- Die  $D_2$  has 0 on all six faces.
- Die  $D_3$  has 1 on all six faces.

You throw these three dice in a box so that they end up at uniformly random orientations. You pick a uniformly random die in the box and observe that it has 0 on its top face. Determine the probability that the die that you picked is  $D_1$ .

*Hint:* You want to determine  $\Pr(A | B)$ , where  $A$  is the event that you pick  $D_1$  and  $B$  is the event that you see a 0 on the top face of the die that you picked. There are different ways to define the sample space  $S$ . One way is to take

$$S = \{(D_1, 0), (D_1, 1), (D_2, 0), (D_3, 1)\},$$

where, for example,  $(D_1, 1)$  is the outcome in which you observe 1 on top of die  $D_1$ . Note that this is not a uniform probability space.

**5.44** According to Statistics Canada, a random person in Canada has

- a probability of  $4/5$  to live to at least 70 years old and
- a probability of  $1/2$  to live to at least 80 years old.

John (a random person in Canada) has just celebrated his 70-th birthday. What is the probability that John will celebrate his 80-th birthday?

**5.45** Nick is taking the course SPID 2804 (The Effect of Spiderman on the Banana Industry). The final exam for this course consists of one true/false question. To answer this question, Nick uses the following approach:

1. If Nick knows that the answer to the question is “true”, he answers “true”.
2. If Nick knows that the answer is “false”, he answers “false”.
3. If Nick does not know the answer, he flips a fair coin.
  - (a) If the coin comes up heads, he answers “true”.
  - (b) If the coin comes up tails, he answers “false”.

You are given that Nick knows the answer to the question with probability 0.8. Consider the event

$$A = \text{“Nick gives the correct answer to the question”}.$$

Determine  $\Pr(A)$ .

**5.46** Let  $A$  and  $B$  be events in some probability space  $(S, \Pr)$ , such that  $\Pr(A) \neq 0$  and  $\Pr(B) \neq 0$ . Use the definition of conditional probability to prove *Bayes’ Theorem*:

$$\Pr(A | B) = \frac{\Pr(B | A) \cdot \Pr(A)}{\Pr(B)}.$$

**5.47** Medical doctors have developed a test for detecting disease  $X$ .

- The test is 98% effective on people who have  $X$ : If a person has  $X$ , then with probability 0.98, the test says that the person indeed has  $X$ .
- The test gives a false reading for 3% of the population without the disease: If a person does not have  $X$ , then with probability 0.03, the test says that the person does have  $X$ .
- It is known that 0.1% of the population has  $X$ .

Assume we choose a person uniformly at random from the population and test this person for disease  $X$ .

- Determine the probability that the test says that the person has  $X$ .
- Assume the test says that the person has  $X$ . Use Exercise 5.46 to determine the probability that the person indeed has  $X$ .

**5.48** In this exercise, we consider a standard deck of 52 cards.

- We choose, uniformly at random, one card from the deck. Consider the events

$$\begin{aligned} A &= \text{“the rank of the chosen card is Ace”,} \\ B &= \text{“the suit of the chosen card is diamonds”.} \end{aligned}$$

Are the events  $A$  and  $B$  independent?

- Assume we remove the Queen of hearts from the deck. We choose, uniformly at random, one card from the remaining 51 cards. Consider the events

$$\begin{aligned} C &= \text{"the rank of the chosen card is Ace"}, \\ D &= \text{"the suit of the chosen card is diamonds"}. \end{aligned}$$

Are the events  $C$  and  $D$  independent?

- 5.49** Let  $n \geq 2$  and  $m \geq 1$  be integers and consider two sets  $A$  and  $B$ , where  $A$  has size  $n$  and  $B$  has size  $m$ . We choose a uniformly random function  $f : A \rightarrow B$ . For any two integers  $i$  and  $k$  with  $1 \leq i \leq n$  and  $1 \leq k \leq m$ , consider the event

$$A_{ik} = \{f(i) = k\}.$$

- For two integers  $i$  and  $k$ , determine  $\Pr(A_{ik})$ .
- For two distinct integers  $i$  and  $j$ , and for an integer  $k$ , are the two events  $A_{ik}$  and  $A_{jk}$  independent?

- 5.50** Consider three events  $A$ ,  $B$ , and  $C$  in some probability space  $(S, \Pr)$ , and assume that  $\Pr(B \cap C) \neq 0$  and  $\Pr(C) \neq 0$ . Prove that

$$\Pr(A \cap B \cap C) = \Pr(A | B \cap C) \cdot \Pr(B | C) \cdot \Pr(C).$$

- 5.51** You have a fair die and do the following experiment:

- Roll the die once; let  $x$  be the outcome.
- Roll the die  $x$  times (independently); let  $y$  be the smallest outcome of these  $x$  rolls.
- Roll the die  $y$  times (independently); let  $z$  be the largest outcome of these  $y$  rolls.

Use Exercise 5.50 to determine

$$\Pr(x = 1 \text{ and } y = 2 \text{ and } z = 3).$$

- 5.52** A standard deck of 52 cards has four Aces.

- You get a uniformly random hand of three cards. Consider the event

$$A = \text{“the hand consists of three Aces”}.$$

Determine  $\Pr(A)$ .

- You get three cards, which are chosen one after another. Each of these three cards is chosen uniformly at random from the current deck of cards. (When a card has been chosen, it is removed from the current deck.) Consider the events

$$B = \text{“all three cards are Aces”}$$

and, for  $i = 1, 2, 3$ ,

$$B_i = \text{“the } i\text{-th card is an Ace.”}$$

Express the event  $B$  in terms of  $B_1$ ,  $B_2$ , and  $B_3$ , and use this expression, together with Exercise 5.50, to determine  $\Pr(B)$ .

**5.53** Let  $p$  be a real number with  $0 < p < 1$ . You are given two coins  $C_1$  and  $C_2$ . The coin  $C_1$  is fair, i.e., if you flip this coin, it comes up heads with probability  $1/2$  and tails with probability  $1/2$ . If you flip the coin  $C_2$ , it comes up heads with probability  $p$  and tails with probability  $1 - p$ . You pick one of these two coins uniformly at random, and flip it twice. These two coin flips are independent of each other. Consider the events

$$\begin{aligned} A &= \text{“the first coin flip results in heads”,} \\ B &= \text{“the second coin flip results in heads”.} \end{aligned}$$

- Determine  $\Pr(A)$ .
- Assume that  $p = 1/4$ . Are the events  $A$  and  $B$  independent?
- Determine all values of  $p$  for which the events  $A$  and  $B$  are independent.

**5.54** Let  $n \geq 2$  be an integer. Assume we have  $n$  balls and 10 boxes. We throw the balls independently and uniformly at random in the boxes. Thus, for each  $k$  and  $i$  with  $1 \leq k \leq n$  and  $1 \leq i \leq 10$ ,

$$\Pr(\text{the } k\text{-th ball falls in the } i\text{-th box}) = 1/10.$$

Consider the event

$$A_n = \text{“there is a box that contains at least two balls”}$$

and let  $p_n = \Pr(A_n)$ .

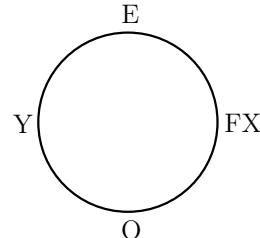
- Determine the smallest value of  $n$  for which  $p_n \geq 1/2$ .
- Determine the smallest value of  $n$  for which  $p_n \geq 2/3$ .

**5.55** Donald Trump wants to hire a secretary and receives  $n$  applications for this job, where  $n \geq 1$  is an integer. Since he is too busy in making important announcements on Twitter, he appoints a three-person hiring committee. After having interviewed the  $n$  applicants, each committee member ranks the applicants from 1 to  $n$ . An applicant is hired for the job if he/she is ranked first by at least two committee members.

Since the committee members do not have the ability to rank the applicants, each member chooses a uniformly random ranking (i.e., permutation) of the applicants, independently of each other.

John is one of the applicants. Determine the probability that John is hired.

**5.56** Edward, Francois-Xavier, Omar, and Yaser are sitting at a round table, as in the figure below.



At 11:59am, they all lower their heads. At noon, each of the boys chooses a uniformly random element from the set  $\{CW, CCW, O\}$ ; these choices are independent of each other. If a boy chooses  $CW$ , then he looks at his clockwise neighbor, if he chooses  $CCW$ , then he looks at his counter-clockwise neighbor, and if he chooses  $O$ , then he looks at the boy at the other side of the table. When two boys make eye contact, they both shout *Vive le Québec libre*.

- Consider the event

$A =$  “both Edward and Francois-Xavier shout *Vive le Québec libre*, whereas neither Omar nor Yaser does”.

Determine  $\Pr(A)$ .

- Consider the event

$B =$  “both Francois-Xavier and Yaser shout *Vive le Québec libre*, whereas neither Edward nor Omar does”.

Determine  $\Pr(B)$ .

- For any integer  $i$  with  $0 \leq i \leq 4$ , consider the event

$C_i =$  “exactly  $i$  boys shout *Vive le Québec libre*”.

Determine

$$\sum_{i=0}^4 \Pr(C_i).$$

Justify your answer in plain English.

- Determine each of the five probabilities  $\Pr(C_0), \Pr(C_1), \dots, \Pr(C_4)$ .

**5.57** You are given a fair die. For any integer  $n \geq 1$ , you roll this die  $n$  times (the rolls are independent). Consider the events

$A_n =$  “the sum of the results of the  $n$  rolls is even”

and

$B_n =$  “the last roll in the sequence of  $n$  rolls results in an even number”,

and their probabilities

$$p_n = \Pr(A_n)$$

and

$$q_n = \Pr(B_n).$$

- Determine  $p_1$ .

- For any integer  $n \geq 1$ , determine  $q_n$ .
- For any integer  $n \geq 2$ , express the event  $A_n$  in terms of the events  $A_{n-1}$  and  $B_n$ .
- Use the previous parts to determine  $p_n$  for any integer  $n \geq 2$ .

**5.58** You are asked to design a random bit generator. You find a coin in your pocket, but, unfortunately, you are not sure if it is a fair coin. After some thought, you come up with the following algorithm  $\text{GENERATEBIT}(n)$ , which takes as input an integer  $n \geq 1$ :

**Algorithm**  $\text{GENERATEBIT}(n)$ :

```
// all coin flips made are mutually independent
flip the coin  $n$  times;
 $k$  = the number of heads in the sequence of  $n$  coin flips;
if  $k$  is odd
then return 0
else return 1
endif
```

In this exercise, you will show that, when  $n \rightarrow \infty$ , the output of algorithm  $\text{GENERATEBIT}(n)$  is a uniformly random bit.

Let  $p$  be the real number with  $0 < p < 1$ , such that, if the coin is flipped once, it comes up heads with probability  $p$  and tails with probability  $1 - p$ . (Note that algorithm  $\text{GENERATEBIT}$  does not need to know the value of  $p$ .) For any integer  $n \geq 1$ , consider the two events

$$A_n = \text{"algorithm } \text{GENERATEBIT}(n) \text{ returns 0"}$$

and

$$B_n = \text{"the } n\text{-th coin flip made by algorithm } \text{GENERATEBIT}(n) \text{ results in heads",}$$

and define

$$P_n = \Pr(A_n)$$

and

$$Q_n = P_n - 1/2.$$

- Determine  $P_1$  and  $Q_1$ .
- For any integer  $n \geq 2$ , prove that

$$P_n = p + (1 - 2p) \cdot P_{n-1}.$$

*Hint:* Express the event  $A_n$  in terms of the events  $A_{n-1}$  and  $B_n$ .

- For any integer  $n \geq 2$ , prove that

$$Q_n = (1 - 2p) \cdot Q_{n-1}.$$

- For any integer  $n \geq 1$ , prove that

$$Q_n = (1 - 2p)^{n-1} \cdot (p - 1/2).$$

- Prove that

$$\lim_{n \rightarrow \infty} Q_n = 0$$

and

$$\lim_{n \rightarrow \infty} P_n = 1/2.$$

**5.59** In this exercise, we will use the product notation. In case you are not familiar with this notation:

- For  $k \leq m$ ,  $\prod_{i=k}^m x_i$  denotes the product

$$x_k \cdot x_{k+1} \cdot x_{k+2} \cdots x_m.$$

- If  $k > m$ , then  $\prod_{i=k}^m x_i$  is an “empty” product, which we define to be equal to 1.

Let  $n \geq 1$  be an integer, and for each  $i = 1, 2, \dots, n$ , let  $p_i$  be a real number such that  $0 < p_i < 1$ . In this exercise, you will prove that

$$\sum_{i=1}^n p_i \prod_{j=i+1}^n (1 - p_j) = 1 - \prod_{i=1}^n (1 - p_i). \quad (5.7)$$

For example,

- for  $n = 1$ , (5.7) becomes

$$p_1 = 1 - (1 - p_1),$$

- for  $n = 2$ , (5.7) becomes

$$p_1(1 - p_2) + p_2 = 1 - (1 - p_1)(1 - p_2),$$

- for  $n = 3$ , (5.7) becomes

$$p_1(1 - p_2)(1 - p_3) + p_2(1 - p_3) + p_3 = 1 - (1 - p_1)(1 - p_2)(1 - p_3).$$

Assume we do an experiment consisting of  $n$  tasks  $T_1, T_2, \dots, T_n$ . Each task is either a success or a failure, independently of the other tasks. For each  $i = 1, 2, \dots, n$ , let  $p_i$  be the probability that  $T_i$  is a success. Consider the event

$$A = \text{“at least one task is a success”}.$$

- Prove (5.7) by determining  $\Pr(A)$  in two different ways.

**5.60** Let  $n \geq 0$  be an integer. In this exercise, you will prove that

$$\sum_{k=0}^n \frac{1}{2^k} \cdot \binom{n+k}{k} = 2^n. \quad (5.8)$$

The Ottawa Senators and the Toronto Maple Leafs play a best-of-( $2n+1$ ) series: These two hockey teams play games against each other, and the first team to win  $n+1$  games wins the series. Assume that

- each game has a winner (thus, no game ends in a tie),
- in any game, the Sens have a probability of  $1/2$  of defeating the Leafs, and
- the results of the games are mutually independent.

Consider the events

$$A = \text{“the Sens win the series”}$$

and

$$B = \text{“the Leafs win the series”}.$$

- Explain in plain English why  $\Pr(A) = \Pr(B) = 1/2$ .
- For each  $k$  with  $0 \leq k \leq n$ , consider the event

$A_k$  = “the Sens win the series after winning the  $(n + k + 1)$ -st game”.

Express the event  $A$  in terms of the events  $A_0, A_1, \dots, A_n$ .

- Consider a fixed value of  $k$  with  $0 \leq k \leq n$ . Prove that

$$\Pr(A_k) = \frac{1}{2^{n+k+1}} \cdot \binom{n+k}{k}.$$

*Hint:* Assume event  $A_k$  occurs. Which team wins the  $(n + k + 1)$ -st game? In the first  $n + k$  games, how many games are won by the Leafs?

- Prove that (5.8) holds by combining the results of the previous parts.

**5.61** Let  $n \geq 0$  be an integer. In this exercise, you will prove that

$$\sum_{k=0}^n \frac{1}{k+1} \binom{n}{k} = \frac{1}{n+1} (2^{n+1} - 1). \quad (5.9)$$

There are  $n + 1$  students in Carleton’s Computer Science program. We denote these students by  $P_1, P_2, \dots, P_{n+1}$ . We play the following game:

1. We choose a uniformly random subset  $X$  of  $\{P_1, P_2, \dots, P_{n+1}\}$ .
2. (a) If  $X \neq \emptyset$ , then we choose a uniformly random student in  $X$ . The chosen student wins a six-pack of cider.  
 (b) If  $X = \emptyset$ , then nobody wins the six-pack.

The random choices made are independent of each other.

- Consider the event

$A_0$  = “nobody wins the six-pack”.

Determine  $\Pr(A_0)$ .

- For each  $i = 1, 2, \dots, n + 1$ , consider the event

$$A_i = \text{“student } P_i \text{ wins the six-pack”}.$$

Explain in plain English why

$$\Pr(A_1) = \Pr(A_2) = \dots = \Pr(A_{n+1}).$$

- Prove that

$$\Pr(A_1) = \frac{1 - 1/2^{n+1}}{n + 1}.$$

- For each  $k$  with  $0 \leq k \leq n$ , consider the event

$$B_k = \text{“}X \text{ has size } k + 1 \text{ and } P_1 \text{ wins the six-pack”}.$$

Prove that

$$\Pr(B_k) = \frac{\binom{n}{k}}{2^{n+1}} \cdot \frac{1}{k + 1}.$$

- Express the event  $A_1$  in terms of the events  $B_0, B_1, \dots, B_n$ .
- Prove that (5.9) holds by combining the results of the previous parts.

**5.62** Let  $n$  and  $k$  be integers with  $1 \leq n \leq k \leq 2n$ . In this exercise, you will prove that

$$\sum_{i=k-n}^n \binom{k}{i} \binom{2n-k}{n-i} = \binom{2n}{n}. \quad (5.10)$$

Jim is working on his assignment for the course COMP 4999 (Computational Aspects of Growing Cannabis). There are  $2n$  questions on this assignment and each of them is worth 1 mark. Two minutes before the deadline, Jim has completed the first  $k$  questions. Jim is very smart and all answers to these  $k$  questions are correct. Jim knows that the instructor, Professor Mary Juana, does not accept late submissions. Because of this, Jim leaves the last  $2n - k$  questions blank and hands in his assignment.

Tri is a teaching assistant for this course. Since Tri is lazy, he does not want to mark all questions. Instead, he chooses a uniformly random subset of  $n$  questions out of the  $2n$  questions, and only marks the  $n$  chosen questions. For each correct answer, Tri gives 2 marks, whereas he gives 0 marks for each wrong (or blank) answer.

For each integer  $i \geq 0$ , consider the event

$$A_i = \text{“Jim receives exactly } 2i \text{ marks for his assignment”}.$$

- Determine the value of the summation  $\sum_i \Pr(A_i)$ . Explain your answer in plain English.
- Determine all values of  $i$  for which the event  $A_i$  is non-empty. For each such value  $i$ , determine  $\Pr(A_i)$ .
- Prove that (5.10) holds by combining the results of the previous parts.

**5.63** Let  $a$  and  $z$  be integers with  $a > z \geq 1$ , and let  $p$  be a real number with  $0 < p < 1$ . Alexa and Zoltan play a game consisting of several rounds. In one round,

1. Alexa receives  $a$  points with probability  $p$  and 0 points with probability  $1 - p$ ,
2. Zoltan receives  $z$  points (with probability 1).

We assume that the results of different rounds are independent.

- Consider the event

$$A = \text{“in one round, Alexa receives more points than Zoltan”}.$$

We say that Alexa is a *better player* than Zoltan, if  $\Pr(A) > 1/2$ .

For which values of  $p$  is Alexa a better player than Zoltan?

- Assume that  $a = 3$ ,  $z = 2$ , and  $p$  is chosen such that  $p > 1/2$  and  $p^2 < 1/2$ . (For example,  $p = (\sqrt{5} - 1)/2$ .)
  - Is Alexa a better player than Zoltan?
  - Alexa and Zoltan play a game consisting of two rounds. We consider the total number of points that each player wins during these two rounds. Consider the event

$$B = \text{“in two rounds, Alexa receives more points than Zoltan”}.$$

Prove that  $\Pr(B) < 1/2$ . (This seems to suggest that Zoltan is a better player than Alexa.)

- Let  $n$  be a large integer, and assume that  $a = n + 1$ ,  $z = n$ , and  $p$  is chosen very close to (but less than) 1. (For example,  $n = 500$  and  $p = 0.99$ .)

- Is Alexa a better player than Zoltan?
- Alexa and Zoltan play a game consisting of  $n$  rounds. We consider the total number of points that each player wins during these  $n$  rounds. Consider the event

$C = \text{“in } n \text{ rounds, Alexa receives more points than Zoltan”}$ .

Prove that  $\Pr(C) = p^n$ . (If  $n = 500$  and  $p = 0.99$ , then  $p^n \approx 0.0066$ . This seems to suggest that Zoltan is a *much* better player than Alexa.)

**5.64** Let  $k \geq 1$  be an integer. Assume we live on a planet on which one year has  $d = 4k^2$  days. Consider  $\sqrt{d} = 2k$  people  $P_1, P_2, \dots, P_{2k}$  living on our planet. Each person has a uniformly random birthday, and the birthdays of these  $2k$  people are mutually independent. Consider the event

$A = \text{“at least two of } P_1, P_2, \dots, P_{2k} \text{ have the same birthday”}$ .

This exercise will lead you through a proof of the claim that

$$0.221 < \Pr(A) < 0.5.$$

Thus, if one year has  $d$  days, then  $\sqrt{d}$  people are enough to have a good chance that not all birthdays are distinct. (This result is similar to the one we obtained in Section 5.5.1.)

- For each  $i$  with  $1 \leq i \leq 2k$ , consider the event

$B_i = \text{“}P_i \text{ has the same birthday as at least one of } P_1, P_2, \dots, P_{i-1}\text{”}$ .

Prove that

$$\Pr(B_i) \leq \frac{i-1}{d}.$$

- Express the event  $A$  in terms of the events  $B_1, B_2, \dots, B_{2k}$ .
- Use the Union Bound (Lemma 5.3.5) to prove that

$$\Pr(A) < 1/2.$$

- Consider the event

$B = \text{“at least two of } P_{k+1}, P_{k+2}, \dots, P_{2k} \text{ have the same birthday”}$

and for each  $i$  with  $1 \leq i \leq k$ , the event

$C_i = \text{“}P_i \text{ has the same birthday as at least one of } P_{k+1}, P_{k+2}, \dots, P_{2k}\text{”}$ .

Prove that

$$\Pr(C_i | \overline{B}) = \frac{1}{4k}.$$

- Prove that if the event  $\overline{A}$  occurs, then the event

$$(\overline{C}_1 \cap \overline{B}) \cap (\overline{C}_2 \cap \overline{B}) \cap \cdots \cap (\overline{C}_k \cap \overline{B})$$

also occurs.

- Prove that

$$\Pr(\overline{A}) \leq \left(1 - \frac{1}{4k}\right)^k.$$

You may use the fact that the events  $\overline{C}_1 \cap \overline{B}$ ,  $\overline{C}_2 \cap \overline{B}$ ,  $\dots$ ,  $\overline{C}_k \cap \overline{B}$  are mutually independent.

- Use the inequality  $1 - x \leq e^{-x}$  to prove that

$$\Pr(A) \geq 1 - e^{-1/4} > 0.221.$$

**5.65** Let  $n$  be a large power of two (thus,  $\log n$  is an integer). Consider a binary string  $s = s_1 s_2 \dots s_n$ , where each bit  $s_i$  is 0 with probability  $1/2$ , and 1 with probability  $1/2$ , independently of the other bits.

A *run of length  $k$*  is a substring of length  $k$ , all of whose bits are equal. In Section 5.14, we have seen that it is very likely that the bitstring  $s$  contains a run of length at least  $\log n - 2 \log \log n$ . In this exercise, you will prove that it is very unlikely that  $s$  contains a run of length more than  $2 \log n$ .

- Let  $k$  be an integer with  $1 \leq k \leq n$ . Consider the event

$A = \text{“the bitstring } s \text{ contains a run of length at least } k\text{”}$ .

For each  $i$  with  $1 \leq i \leq n - k + 1$ , consider the event

$$A_i = \text{“the substring } s_i s_{i+1} \dots s_{i+k-1} \text{ is a run”}.$$

Use the Union Bound (Lemma 5.3.5) to prove that

$$\Pr(A) \leq \frac{n - k + 1}{2^{k-1}}.$$

- Let  $k = 2 \log n$ . Prove that

$$\Pr(A) \leq 2/n.$$

**5.66** A hand of 5 cards is chosen uniformly at random from a standard deck of 52 cards. Consider the event

$$A = \text{“the hand has at least one Ace”}.$$

- Explain what is wrong with the following argument:

We are going to determine  $\Pr(A)$ . Event  $A$  states that the hand has at least one Ace. By symmetry, we may assume that  $A$  is the event that the hand has the Ace of spades. Since there are  $\binom{52}{5}$  hands of five cards and exactly  $\binom{51}{4}$  of them contain the Ace of spades, it follows that

$$\Pr(A) = \frac{\binom{51}{4}}{\binom{52}{5}} = \frac{5}{52}.$$

- Explain what is wrong with the following argument:

We are going to determine  $\Pr(A)$  using the Law of Total Probability (Theorem 5.9.1). For each  $x \in \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$ , consider the event

$$B_x = \text{“the hand has the Ace of suit } x\text{”}.$$

We observe that

$$\Pr(B_x) = \frac{\binom{51}{4}}{\binom{52}{5}} = \frac{5}{52}.$$

We next observe that

$$\Pr(A | B_x) = 1,$$

because if event  $B_x$  occurs, then event  $A$  also occurs. Thus, using the Law of Total Probability, we get

$$\begin{aligned}\Pr(A) &= \sum_x \Pr(A | B_x) \cdot \Pr(B_x) \\ &= \sum_x 1 \cdot \Pr(B_x) \\ &= \sum_x \frac{5}{52} \\ &= 4 \cdot \frac{5}{52} \\ &= \frac{5}{13}.\end{aligned}$$

- Determine the value of  $\Pr(A)$ .

**5.67** You are doing two projects  $P$  and  $Q$ . The probability that project  $P$  is successful is equal to  $2/3$  and the probability that project  $Q$  is successful is equal to  $4/5$ . Whether or not these two projects are successful are independent of each other. What is the probability that both  $P$  and  $Q$  are not successful?

**5.68** Consider two independent events  $A$  and  $B$  in some probability space  $(S, \Pr)$ . Assume that  $A$  and  $B$  are disjoint, i.e.,  $A \cap B = \emptyset$ . What can you say about  $\Pr(A)$  and  $\Pr(B)$ ?

**5.69** You flip three fair coins independently of each other. Let  $A$  be the event “at least two flips in a row are heads” and let  $B$  be the event “the number of heads is even”. (Note that zero is even.) Are  $A$  and  $B$  independent?

**5.70** You flip three fair coins independently of each other. Consider the events

$$A = \text{“there is at most one tails”}$$

and

$$B = \text{“not all flips are identical”}.$$

Are  $A$  and  $B$  independent?

**5.71** Let  $n \geq 2$  be an integer and consider two fixed integers  $a$  and  $b$  with  $1 \leq a < b \leq n$ .

- Use the Product Rule to determine the number of permutations of  $\{1, 2, \dots, n\}$  in which  $a$  is to the left of  $b$ .
- Consider a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ , and define the event

$$A = \text{“in this permutation, } a \text{ is to the left of } b\text{”}.$$

Use your answer to the first part of this exercise to determine  $\Pr(A)$ .

**5.72** Let  $n \geq 4$  be an integer and consider a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ . Consider the event

$$A = \text{“in this permutation, both 3 and 4 are to the left of both 1 and 2”}.$$

Determine  $\Pr(A)$ .

**5.73** Let  $n \geq 3$  be an integer, consider a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ , and define the events

$$A = \text{“in this permutation, 2 is to the left of 3”}$$

and

$$B = \text{“in this permutation, 1 is to the left of 2 and 1 is to the left of 3”}.$$

Are these two events independent?

**5.74** Let  $n \geq 4$  be an integer. Consider a uniformly random permutation of  $\{1, 2, \dots, n\}$  and define the events

$$A = \text{"1 and 2 are next to each other, with 1 to the left of 2, or 4 and 3 are next to each other, with 4 to the left of 3"}$$

and

$$B = \text{"1 and 2 are next to each other, with 1 to the left of 2, or 2 and 3 are next to each other, with 2 to the left of 3".}$$

Determine  $\Pr(A)$  and  $\Pr(B)$ . (Before you determine these probabilities, spend a few minutes and guess which probability is larger.)

**5.75** You flip two fair coins independently of each other. Consider the events

$$\begin{aligned} A &= \text{"the number of heads is odd"}, \\ B &= \text{"the first coin comes up heads"}, \\ C &= \text{"the second coin comes up heads"}. \end{aligned}$$

- Are the events  $A$  and  $B$  independent?
- Are the events  $A$  and  $C$  independent?
- Are the events  $B$  and  $C$  independent?
- Are the events  $A$ ,  $B$ , and  $C$  pairwise independent?
- Are the events  $A$ ,  $B$ , and  $C$  mutually independent?

**5.76** You roll a fair die once. Consider the events

$$\begin{aligned} A &= \text{"the result is an element of } \{1, 3, 4\}\text{"}, \\ B &= \text{"the result is an element of } \{3, 4, 5, 6\}\text{"}. \end{aligned}$$

Are these two events independent?

**5.77** You roll a fair die once. Consider the events

$$\begin{aligned} A &= \text{"the result is even"}, \\ B &= \text{"the result is odd"}, \\ C &= \text{"the result is at most 4"}. \end{aligned}$$

- Are the events  $A$  and  $B$  independent?
- Are the events  $A$  and  $C$  independent?
- Are the events  $B$  and  $C$  independent?

**5.78** You are given a tetrahedron, which is a die with four faces. Each of these faces has one of the bitstrings 110, 101, 011, and 000 written on it. Different faces have different bitstrings.

We roll the tetrahedron so that each face is at the bottom with equal probability  $1/4$ . For  $k = 1, 2, 3$ , consider the event

$$A_k = \text{“the bitstring written on the bottom face has 0 at position } k\text{”}.$$

For example, if the bitstring at the bottom face is 101, then  $A_1$  is false,  $A_2$  is true, and  $A_3$  is false.

- Are the events  $A_1$  and  $A_2$  independent?
- Are the events  $A_1$  and  $A_3$  independent?
- Are the events  $A_2$  and  $A_3$  independent?
- Are the events  $A_1, A_2, A_3$  pairwise independent?
- Are the events  $A_1, A_2, A_3$  mutually independent?

**5.79** In a group of 100 children, 34 are boys and 66 are girls. You are given the following information about the girls:

- Each girl has green eyes or is blond or is left-handed.
- 20 of the girls have green eyes.
- 40 of the girls are blond.
- 50 of the girls are left-handed.
- 10 of the girls have green eyes and are blond.
- 14 of the girls have green eyes and are left-handed.
- 4 of the girls have green eyes, are blond, and are left-handed.

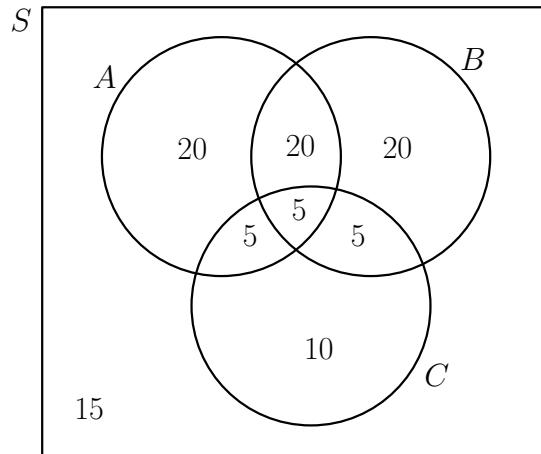
We choose one of these 100 children uniformly at random. Consider the events

$$\begin{aligned} G &= \text{"the child chosen is a girl with green eyes"}, \\ B &= \text{"the child chosen is a blond girl"}, \\ L &= \text{"the child chosen is a left-handed girl"}. \end{aligned}$$

- Are the events  $G$  and  $B$  independent?
- Are the events  $G$  and  $L$  independent?
- Are the events  $B$  and  $L$  independent?
- Verify whether or not the following equation holds:

$$\Pr(G \wedge B \wedge L) = \Pr(G) \cdot \Pr(B) \cdot \Pr(L).$$

**5.80** Let  $S$  be a sample space consisting of 100 elements. Consider three events  $A$ ,  $B$ , and  $C$ , as indicated in the figure below. For example, the event  $A$  consists of 50 elements, 20 of which are only in  $A$ , 20 of which are only in  $A \cap B$ , 5 of which are only in  $A \cap C$ , and 5 of which are in  $A \cap B \cap C$ .



Consider the uniform probability function on this sample space.

- Are the events  $A$  and  $B$  independent?

- Determine whether or not

$$\Pr(A \cap B | C) = \Pr(A | C) \cdot \Pr(B | C).$$

**5.81** Annie, Boris, and Charlie write an exam that consists of only one question: *What is 26 times 26?* Calculators are not allowed during the exam. Both Annie and Boris are pretty clever and each of them gives the correct answer with probability 9/10. Charlie has trouble with two-digit numbers and gives the correct answer with probability 6/10.

- Assume that the three students do not cheat, i.e., each student answers the question independently of the other two students. Determine the probability that at least two of them give the correct answer.
- Assume that Annie and Boris do not cheat, but Charlie copies Annie's answer. Determine the probability that at least two of them give the correct answer.

*Hint:* The answer to the second part is smaller than the answer to the first part.

**5.82** Alexa and Zoltan play the following game:

**AZ-game:**

**Step 1:** Alexa chooses a uniformly random element from the set  $\{1, 2, 3\}$ . Let  $a$  denote the element that Alexa chooses.

**Step 2:** Zoltan chooses a uniformly random element from the set  $\{1, 2, 3\}$ . Let  $z$  denote the element that Zoltan chooses.

**Step 3:** Using one of the three strategies mentioned below, Alexa chooses an element from the set  $\{1, 2, 3\} \setminus \{a\}$ . Let  $a'$  denote the element that Alexa chooses.

**Step 4:** Using one of the three strategies mentioned below, Zoltan chooses an element from the set  $\{1, 2, 3\} \setminus \{z\}$ . Let  $z'$  denote the element that Zoltan chooses.

The AZ-game is a *success* if  $a' \neq z'$ .

- *MinMin Strategy:* In Step 3, Alexa chooses the smallest element in the set  $\{1, 2, 3\} \setminus \{a\}$ , and Zoltan chooses the smallest element in the set  $\{1, 2, 3\} \setminus \{z\}$ .

- Describe the sample space for this strategy.
- For this strategy, determine the probability that the AZ-game is a success.
- *MinMax Strategy:* In Step 3, Alexa chooses the smallest element in the set  $\{1, 2, 3\} \setminus \{a\}$ , and Zoltan chooses the largest element in the set  $\{1, 2, 3\} \setminus \{z\}$ .
  - Describe the sample space for this strategy.
  - For this strategy, determine the probability that the AZ-game is a success.
- *Random Strategy:* In Step 3, Alexa chooses a uniformly random element in the set  $\{1, 2, 3\} \setminus \{a\}$ , and Zoltan chooses a uniformly random element in the set  $\{1, 2, 3\} \setminus \{z\}$ .
  - Describe the sample space for this strategy.
  - For this strategy, determine the probability that the AZ-game is a success.

**5.83** You are given a box that contains one red ball and one blue ball. Consider the following algorithm *RandomRedBlue(n)* that takes as input an integer  $n \geq 3$ :

**Algorithm** *RandomRedBlue(n)*:

```

// n ≥ 3
// initially, the box contains one red ball and one blue ball
// all random choices are mutually independent
for k = 1 to n - 2
  do choose a uniformly random ball in the box;
    if the chosen ball is red
      then put the chosen ball back in the box;
        add one red ball to the box
      else put the chosen ball back in the box;
        add one blue ball to the box
    endif
  endfor
```

For any integers  $n \geq 3$  and  $i$  with  $1 \leq i \leq n - 1$ , consider the event

$$\begin{aligned} A_i^n &= \text{"at the end of algorithm } RandomRedBlue(n), \\ &\quad \text{the number of red balls in the box is equal to } i". \end{aligned}$$

In this exercise, you will prove that for any integers  $n \geq 3$  and  $i$  with  $1 \leq i \leq n - 1$ ,

$$\Pr(A_i^n) = \frac{1}{n-1}. \quad (5.11)$$

- Let  $n \geq 3$  and  $k$  be integers with  $1 \leq k \leq n - 2$ . When running algorithm *RandomRedBlue*( $n$ ),
  - how many balls does the box contain at the start of the  $k$ -th iteration,
  - how many balls does the box contain at the end of the  $k$ -th iteration?
- Let  $n \geq 3$  be an integer. After algorithm *RandomRedBlue*( $n$ ) has terminated, how many balls does the box contain?
- For any integer  $n \geq 3$ , prove that

$$\Pr(A_1^n) = \frac{1}{n-1}.$$

- For any integer  $n \geq 3$ , prove that

$$\Pr(A_{n-1}^n) = \frac{1}{n-1}.$$

- Let  $n = 3$ . Prove that (5.11) holds for all values of  $i$  in the indicated range.
- Let  $n \geq 4$ . Consider the event

$$\begin{aligned} A &= \text{"in the } (n-2)\text{-th iteration of algorithm } RandomRedBlue(n), \\ &\quad \text{a red ball is chosen".} \end{aligned}$$

For any integer  $i$  with  $2 \leq i \leq n - 2$ , express the event  $A_i^n$  in terms of the events  $A_{i-1}^{n-1}$ ,  $A_i^{n-1}$ , and  $A$ .

- Let  $n \geq 4$ . For any integer  $i$  with  $2 \leq i \leq n - 2$ , prove that

$$\Pr(A_i^n) = \Pr(A | A_{i-1}^{n-1}) \cdot \Pr(A_{i-1}^{n-1}) + \Pr(\overline{A} | A_i^{n-1}) \cdot \Pr(A_i^{n-1}).$$

- Let  $n \geq 4$ . Prove that (5.11) holds for all values of  $i$  in the indicated range.

**5.84** Prove that for any real number  $x \neq 1$  and any integer  $N \geq 0$ ,

$$\sum_{n=0}^N x^n = \frac{1 - x^{N+1}}{1 - x}.$$

**5.85** Use the following argumentation to convince yourself that

$$\sum_{n=0}^{\infty} 1/2^n = 2.$$

Take the interval  $I = [0, 2)$  of length 2 on the real line and, for each  $n \geq 0$ , an interval  $I_n$  of length  $1/2^n$ . It is possible to place all intervals  $I_n$  with  $n \geq 0$  in  $I$  such that

- no two intervals  $I_n$  and  $I_m$ , with  $m \neq n$ , overlap and
- all intervals  $I_n$  with  $n \geq 0$  completely cover the interval  $I$ .

**5.86** Alexa, Tri, and Zoltan play the ODDPLAYER game: In one round, each player flips a fair coin.

1. Assume that not all flips are equal. Then the coin flips of exactly two players are equal. The player whose coin flip is different is called the *odd player*. In this case, the odd player wins the game. For example, if Alexa flips tails, Tri flips heads, and Zoltan flips tails, then Tri is the odd player and wins the game.
2. If all three coin flips are equal, then the game is repeated.

Below, this game is presented in pseudocode:

**Algorithm ODDPLAYER:**

```
// all coin flips are mutually independent  
each player flips a fair coin;  
if not all coin flips are equal  
    then the game terminates and the odd player wins  
    else ODDPLAYER  
    endif
```

- What is the sample space?
- Consider the event

$$A = \text{“Alexa wins the game”}.$$

Express this event as a subset of the sample space.

- Use your expression from the previous part to determine  $\Pr(A)$ .
- Use symmetry to determine  $\Pr(A)$ . Explain your answer in plain English and a few sentences.

**Hint:** What is the probability that Tri wins the game? What is the probability that Zoltan wins the game?

**5.87** Two players  $P_1$  and  $P_2$  take turns rolling two fair and independent dice, where  $P_1$  starts the game. The first player who gets a sum of seven wins the game. Determine the probability that player  $P_1$  wins the game.

**5.88** By flipping a fair coin repeatedly and independently, we obtain a sequence of  $H$ 's and  $T$ 's. We stop flipping the coin as soon as the sequence contains either  $HH$  or  $TH$ .

Two players  $P_1$  and  $P_2$  play a game, in which  $P_1$  wins if the last two symbols in the sequence are  $HH$ . Otherwise, the last two symbols in the sequence are  $TH$ , in which case  $P_2$  wins. Determine the probability that player  $P_1$  wins the game.

**5.89** Two players  $P_1$  and  $P_2$  play a game in which they take turns flipping, independently, a fair coin: First  $P_1$  flips the coin, then  $P_2$  flips the coin, then  $P_1$  flips the coin, then  $P_2$  flips the coin, etc. The game ends as soon as the

sequence of coin flips contains either  $HH$  or  $TT$ . The player who flips the coin for the last time is the winner of the game. For example, if the sequence of coin flips is  $HTHTHH$ , then  $P_2$  wins the game.

Determine the probability that player  $P_1$  wins the game.

**5.90** We flip a fair coin repeatedly and independently, and stop as soon as we see one of the two sequences  $HTT$  and  $HHT$ . Let  $A$  be the event that the process stops because  $HTT$  is seen.

- Prove that the event  $A$  is given by the set

$$\{T^m(HT)^nHTT : m \geq 0, n \geq 0\}.$$

In other words, event  $A$  holds if and only if the sequence of coin flips is equal to  $T^m(HT)^nHTT$  for some  $m \geq 0$  and  $n \geq 0$ .

- Prove that  $\Pr(A) = 1/3$ .

**5.91** For  $i \in \{1, 2\}$ , consider the game  $G_i$ , in which two players  $P_1$  and  $P_2$  take turns flipping, independently, a fair coin, where  $P_i$  starts. The game ends as soon as heads comes up. The player who flips heads first is the winner of the game  $G_i$ . For  $j \in \{1, 2\}$ , consider the event

$$B_{ij} = "P_j \text{ wins the game } G_i".$$

In Section 5.15.2, we have seen that

$$\Pr(B_{11}) = \Pr(B_{22}) = 2/3 \quad (5.12)$$

and

$$\Pr(B_{12}) = \Pr(B_{21}) = 1/3. \quad (5.13)$$

Consider the game  $G$ , in which  $P_1$  and  $P_2$  take turns flipping, independently, a fair coin, where  $P_1$  starts. The game ends as soon as a second heads comes up. The player who flips the second heads wins the game. Consider the event

$$A = "P_1 \text{ wins the game } G".$$

In Section 5.15.3, we used an infinite series to show that

$$\Pr(A) = 4/9. \quad (5.14)$$

Use the Law of Total Probability (Theorem 5.9.1) to give an alternative proof of (5.14). You are allowed to use (5.12) and (5.13).

**5.92** Consider two players  $P_1$  and  $P_2$ :

- $P_1$  has one fair coin.
- $P_2$  has two coins. One of them is fair, whereas the other one is 2-headed (Her Majesty is on both sides of this coin).

The two players  $P_1$  and  $P_2$  play a game in which they alternate making turns:  $P_1$  starts, after which it is  $P_2$ 's turn, after which it is  $P_1$ 's turn, after which it is  $P_2$ 's turn, etc.

- When it is  $P_1$ 's turn, she flips her coin once.
- When it is  $P_2$ 's turn, he does the following:
  - $P_2$  chooses one of his two coins uniformly at random. Then he flips the chosen coin once.
  - If the first flip did not result in heads, then  $P_2$  repeats this process one more time:  $P_2$  again chooses one of his two coins uniformly at random and flips the chosen coin once.

The player who flips heads first is the winner of the game.

- Determine the probability that  $P_2$  wins this game, assuming that all random choices and coin flips made are mutually independent.

**5.93** Jennifer loves to drink India Pale Ale (IPA), whereas Connor Hillen prefers Black IPA. Jennifer and Connor decide to go to their favorite pub *Chez Lindsay et Simon*. The beer menu shows that this pub has ten beers on tap:

- Phillips Cabin Fever Imperial Black IPA,
- Big Rig Black IPA,
- Leo's Early Breakfast IPA,
- Goose Island IPA,
- Caboose IPA,
- and five other beers, neither of which is an IPA.

Each of the first five beers is an IPA, whereas each of the first two beers is a Black IPA.

Jennifer and Connor play a game, in which they alternate ordering beer: Connor starts, after which it is Jennifer's turn, after which it is Connor's turn, after which it is Jennifer's turn, etc.

- When it is Connor's turn, he orders two beers; each of these is chosen uniformly at random from the ten beers (thus, these two beers may be equal).
- When it is Jennifer's turn, she orders one of the ten beers, uniformly at random.

The game ends as soon as (i) Connor has ordered at least one Black IPA, in which case he pays the bill, or (ii) Jennifer has ordered at least one IPA, in which case she pays the bill.

- Determine the probability that Connor pays the bill, assuming that all random choices made are mutually independent.

**5.94** You would like to generate a uniformly random bit, i.e., with probability  $1/2$ , this bit is 0, and with probability  $1/2$ , it is 1. You find a coin in your pocket, but you are not sure if it is a fair coin: It comes up heads ( $H$ ) with probability  $p$  and tails ( $T$ ) with probability  $1 - p$ , for some real number  $p$  that is unknown to you. In particular, you do not know if  $p = 1/2$ . In this exercise, you will show that this coin can be used to generate a uniformly random bit.

Consider the following recursive algorithm `GETRANDOMBIT`, which does not take any input:

**Algorithm** `GETRANDOMBIT`:

```
// all coin flips made are mutually independent
flip the coin twice;
if the result is HT
then return 0
else if the result is TH
then return 1
else GETRANDOMBIT
endif
endif
```

- The sample space  $S$  is the set of all sequences of coin flips that can occur when running algorithm GETRANDOMBIT. Determine this sample space  $S$ .
- Prove that algorithm GETRANDOMBIT returns a uniformly random bit.

**5.95** You would like to generate a *biased* random bit: With probability  $2/3$ , this bit is 0, and with probability  $1/3$ , it is 1. You find a *fair* coin in your pocket: This coin comes up heads ( $H$ ) with probability  $1/2$  and tails ( $T$ ) with probability  $1/2$ . In this exercise, you will show that this coin can be used to generate a biased random bit.

Consider the following recursive algorithm GETBIASEDBIT, which does not take any input:

**Algorithm GETBIASEDBIT:**

```
// all coin flips made are mutually independent  
flip the coin;  
if the result is  $H$   
    then return 0  
    else  $b = \text{GETBIASEDBIT};$   
        return  $1 - b$   
endif
```

- The sample space  $S$  is the set of all sequences of coin flips that can occur when running algorithm GETBIASEDBIT. Determine this sample space  $S$ .
- Prove that algorithm GETBIASEDBIT returns 0 with probability  $2/3$ .

**5.96** Both Alexa and Shelly have an infinite bitstring. Alexa's bitstring is denoted by  $a_1a_2a_3\dots$ , whereas Shelly's bitstring is denoted by  $s_1s_2s_3\dots$ . Alexa can see her bitstring, but she cannot see Shelly's bitstring. Similarly, Shelly can see her bitstring, but she cannot see Alexa's bitstring. The bits in both bitstrings are uniformly random and independent.

The ladies play the following game: Alexa chooses a positive integer  $k$  and Shelly chooses a positive integer  $\ell$ . The game is a *success* if  $s_k = 1$  and  $a_\ell = 1$ . In words, the game is a success if Alexa chooses a position in Shelly's

bitstring that contains a 1, and Shelly chooses a position in Alexa's bitstring that contains a 1.

- Assume Alexa chooses  $k = 4$  and Shelly chooses  $\ell = 7$ . Determine the probability that the game is a success.
- Assume Alexa chooses the position, say  $k$ , of the leftmost 1 in her bitstring, and Shelly chooses the position, say  $\ell$ , of the leftmost 1 in her bitstring.
  - If  $k \neq \ell$ , is the game a success?
  - Determine the probability that the game is a success.

**5.97** Alexa and Shelly take turns flipping, independently, a coin, where Alexa starts. The game ends as soon as heads comes up. The lady who flips heads first is the winner of the game.

Alexa proposes that they both use a fair coin. Of course, Shelly does not agree, because she knows from Section 5.15.2 that this gives Alexa a probability of  $2/3$  of winning the game.

The ladies agree on the following: Let  $p$  and  $q$  be real numbers with  $0 < p < 1$  and  $0 \leq q \leq 1$ . Alexa uses a coin that comes up heads with probability  $p$ , and Shelly uses a coin that comes up heads with probability  $q$ .

- Assume that  $p = 1/2$ . Determine the value of  $q$  for which Alexa and Shelly have the same probability of winning the game.
- From now on, assume that  $0 < p < 1$  and  $0 < q < 1$ .
  - Determine the probability that Alexa wins the game.
  - Assume that  $p > 1/2$ . Prove that for any  $q$  with  $0 < q < 1$ , the probability that Alexa wins the game is strictly larger than  $1/2$ .
  - Assume that  $p < 1/2$ . Determine the value of  $q$  for which Alexa and Shelly have the same probability of winning the game.

**5.98** Let  $n \geq 2$  be an integer and consider a uniformly random permutation  $(a_1, a_2, \dots, a_n)$  of the set  $\{1, 2, \dots, n\}$ . For each  $k$  with  $1 \leq k \leq n$ , consider the event

$$A_k = \text{“}a_k \text{ is the largest element among the first } k \text{ elements in the permutation”}.$$

- Let  $k$  and  $\ell$  be two integers with  $1 \leq k < \ell \leq n$ . Prove that the events  $A_k$  and  $A_\ell$  are independent.

*Hint:* Use the Product Rule to determine the number of permutations that define  $A_k$ ,  $A_\ell$ , and  $A_k \cap A_\ell$ , respectively.

- Prove that the sequence  $A_1, A_2, \dots, A_n$  of events is mutually independent.

**5.99** Let  $n \geq 2$  be an integer. We generate a random bitstring  $R = r_1r_2\cdots r_n$ , by setting, for each  $i = 1, 2, \dots, n$ ,  $r_i = 1$  with probability  $1/i$  and, thus,  $r_i = 0$  with probability  $1 - 1/i$ . All random choices made when setting these bits are mutually independent.

For each  $i$  with  $1 \leq i \leq n$ , consider the events

$$B_i = "r_i = 1"$$

and

$$R_i = \text{"the rightmost 1 in the bitstring } R \text{ is at position } i\text{"}.$$

- Determine  $\Pr(R_i)$ .

The following algorithm TRYTOFINDRIGHTMOSTONE( $R, n, m$ ) takes as input the bitstring  $R = r_1r_2\cdots r_n$  of length  $n$  and an integer  $m$  with  $1 \leq m \leq n$ . As the name suggests, this algorithm tries to find the position of the rightmost 1 in the string  $R$ .

```

Algorithm TRYTOFINDRIGHTMOSTONE( $R, n, m$ ):
    for  $i = 1$  to  $m$ 
        do if  $r_i = 1$ 
            then  $k = i$ 
            endif
        endfor;
        //  $k$  is the position of the rightmost 1 in the substring
        //  $r_1r_2 \cdots r_m$ .
        // the next while-loop finds the position of the leftmost 1
        // in the substring  $r_{m+1}r_{m+2} \cdots r_n$ , if this position exists.
         $\ell = m + 1$ ;
        while  $\ell \leq n$  and  $r_\ell = 0$ 
            do  $\ell = \ell + 1$ 
        endwhile;
        // if  $\ell \leq n$ , then  $\ell$  is the position of the leftmost 1 in the
        // substring  $r_{m+1}r_{m+2} \cdots r_n$ .
        if  $\ell \leq n$ 
            then return  $\ell$ 
            else return  $k$ 
        endif

```

Consider the event

$E_m$  = “there is exactly one 1 in the substring  $r_{m+1}r_{m+2} \cdots r_n$ ”.

- Prove that

$$\Pr(E_m) = \frac{m}{n} \left( \frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1} \right).$$

Consider the event

$A$  = “TRYTOFINDRIGHTMOSTONE( $R, n, m$ ) returns the position of the rightmost 1 in the string  $R$ ”.

- Prove that

$$\Pr(A) = \frac{m}{n} \left( 1 + \frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1} \right).$$

**5.100** You realize that it is time to buy a pair of shoes. You look up all  $n$  shoe stores in Ottawa and visit them in *random* order. While shopping, you create a bitstring  $r_1r_2 \cdots r_n$  of length  $n$ : For each  $i$  with  $1 \leq i \leq n$ , you set  $r_i$  to 1 if and only if the  $i$ -th store has the best pair of shoes, among the first  $i$  stores that you have visited.

- Use Exercise 5.98 to prove that this bitstring satisfies the condition in Exercise 5.99.

After you have visited the first  $m$  shoe stores, you are bored of shopping. You keep on visiting shoe stores, but as soon as you visit a store that has a pair of shoes that you like more than the previously best pair you have found, you buy the former pair of shoes.

- Use Exercise 5.99 to determine the probability that you buy the best pair of shoes that is available in Ottawa.



# Chapter 6

## Random Variables and Expectation

A natural question: What is the definition of random variable? Classically, and in many of today's textbooks, you see definitions such as, a random variable is the observed value of a random quantity. What on earth does that mean? How can any sort of theory be built on such vagueness?

— Persi Diaconis and Brian Skyrms, *Ten Great Ideas About Chance*, 2018

### 6.1 Random Variables

We have already seen random variables in Chapter 5, even though we did not use that term there. For example, in Section 5.2.1, we rolled a die twice and were interested in the sum of the results of these two rolls. In other words, we did an “experiment” (rolling a die twice) and asked for a function of the outcome (the sum of the results of the two rolls).

**Definition 6.1.1** Let  $S$  be a sample space. A *random variable* on the sample space  $S$  is a function  $X : S \rightarrow \mathbb{R}$ .

In the example given above, the sample space is

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$$

and the random variable is the function  $X : S \rightarrow \mathbb{R}$  defined by

$$X(i, j) = i + j$$

for all  $(i, j)$  in  $S$ .

Note that the term “random variable” is misleading: A random variable is *not* random, but a function that assigns, to every outcome  $\omega$  in the sample space  $S$ , a real number  $X(\omega)$ . Also, a random variable is *not* a variable, but a function.

A random variable is neither random nor variable.

### 6.1.1 Flipping Three Coins

Assume we flip three coins. The sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

where, e.g.,  $TTH$  indicates that the first two coins come up tails and the third coin comes up heads.

Let  $X : S \rightarrow \mathbb{R}$  be the random variable that maps any outcome (i.e., any element of  $S$ ) to the number of heads in the outcome. Thus,

$$\begin{aligned} X(HHH) &= 3, \\ X(HHT) &= 2, \\ X(HTH) &= 2, \\ X(HTT) &= 1, \\ X(THH) &= 2, \\ X(THT) &= 1, \\ X(TTH) &= 1, \\ X(TTT) &= 0. \end{aligned}$$

If we define the random variable  $Y$  to be the function  $Y : S \rightarrow \mathbb{R}$  that

- maps an outcome to 1 if all three coins come up heads or all three coins come up tails, and
- maps an outcome to 0 in all other cases,

then we have

$$\begin{aligned} Y(HHH) &= 1, \\ Y(HHT) &= 0, \\ Y(HTH) &= 0, \\ Y(HTT) &= 0, \\ Y(THH) &= 0, \\ Y(THT) &= 0, \\ Y(TTH) &= 0, \\ Y(TTT) &= 1. \end{aligned}$$

Since a random variable is a function  $X : S \rightarrow \mathbb{R}$ , it maps any outcome  $\omega$  to a real number  $X(\omega)$ . Usually, we just write  $X$  instead of  $X(\omega)$ . Thus, for any outcome in the sample space  $S$ , we denote the value of the random variable, for this outcome, by  $X$ . In the example above, we flip three coins and write

$X =$  the number of heads

and

$$Y = \begin{cases} 1 & \text{if all three coins come up heads or all three coins come up tails,} \\ 0 & \text{otherwise.} \end{cases}$$

### 6.1.2 Random Variables and Events

Random variables give rise to events in a natural way. In the three-coin example, “ $X = 0$ ” corresponds to the event  $\{TTT\}$ , whereas “ $X = 2$ ” corresponds to the event  $\{HHT, HTH, THH\}$ . The table below gives some values of the random variables  $X$  and  $Y$ , together with the corresponding events.

value	event
$X = 0$	$\{TTT\}$
$X = 1$	$\{HTT, THT, TTH\}$
$X = 2$	$\{HHT, HTH, THH\}$
$X = 3$	$\{HHH\}$
$X = 4$	$\emptyset$
$Y = 0$	$\{HHT, HTH, HTT, THH, THT, TTH\}$
$Y = 1$	$\{HHH, TTT\}$
$Y = 2$	$\emptyset$

Thus, the event “ $X = x$ ” corresponds to the set of all outcomes that are mapped, by the function  $X$ , to the value  $x$ :

**Definition 6.1.2** Let  $S$  be a sample space and let  $X : S \rightarrow \mathbb{R}$  be a random variable. For any real number  $x$ , we define “ $X = x$ ” to be the event

$$\{\omega \in S : X(\omega) = x\}.$$

Let us return to the example in which we flip three coins. Assume that the coins are fair and the three flips are mutually independent. Consider again the corresponding random variables  $X$  and  $Y$ . It should be clear how we determine, for example, the probability that  $X$  is equal to 0, which we will write as  $\Pr(X = 0)$ . Using our interpretation of “ $X = 0$ ” as being the event  $\{TTT\}$ , we get

$$\begin{aligned}\Pr(X = 0) &= \Pr(TTT) \\ &= 1/8.\end{aligned}$$

Similarly, we get

$$\begin{aligned}\Pr(X = 1) &= \Pr(\{HTT, THT, TTH\}) \\ &= 3/8, \\ \Pr(X = 2) &= \Pr(\{HHT, HTH, THH\}) \\ &= 3/8, \\ \Pr(X = 3) &= \Pr(\{HHH\}) \\ &= 1/8, \\ \Pr(X = 4) &= \Pr(\emptyset) \\ &= 0, \\ \Pr(Y = 0) &= \Pr(\{HHT, HTH, HTT, THH, THT, TTH\}) \\ &= 6/8 \\ &= 3/4, \\ \Pr(Y = 1) &= \Pr(\{HHH, TTT\}) \\ &= 2/8 \\ &= 1/4, \\ \Pr(Y = 2) &= \Pr(\emptyset) \\ &= 0.\end{aligned}$$

Consider an arbitrary probability space  $(S, \Pr)$  and let  $X : S \rightarrow \mathbb{R}$  be a random variable. Using (5.1) and Definition 6.1.2, the probability of the event “ $X = x$ ”, i.e., the probability that  $X$  is equal to  $x$ , is equal to

$$\begin{aligned}\Pr(X = x) &= \Pr(\{\omega \in S : X(\omega) = x\}) \\ &= \sum_{\omega : X(\omega) = x} \Pr(\omega).\end{aligned}$$

We have interpreted “ $X = x$ ” as being an event. We extend this to more general statements involving  $X$ . For example, “ $X \geq x$ ” denotes the event

$$\{\omega \in S : X(\omega) \geq x\}.$$

For our three-coin example, the random variable  $X$  can take each of the values 0, 1, 2, and 3 with a positive probability. As a result, “ $X \geq 2$ ” denotes the event “ $X = 2$  or  $X = 3$ ”, and we have

$$\begin{aligned}\Pr(X \geq 2) &= \Pr(X = 2 \vee X = 3) \\ &= \Pr(X = 2) + \Pr(X = 3) \\ &= 3/8 + 1/8 \\ &= 1/2.\end{aligned}$$

## 6.2 Independent Random Variables

In Section 5.11, we have defined the notion of two events being independent. The following definition extends this to random variables.

**Definition 6.2.1** Let  $(S, \Pr)$  be a probability space and let  $X$  and  $Y$  be two random variables on  $S$ . We say that  $X$  and  $Y$  are *independent* if for all real numbers  $x$  and  $y$ , the events “ $X = x$ ” and “ $Y = y$ ” are independent, i.e.,

$$\Pr(X = x \wedge Y = y) = \Pr(X = x) \cdot \Pr(Y = y).$$

Assume we flip three fair coins independently and, as in Section 6.1.1, consider the random variables

$$X = \text{the number of heads}$$

and

$$Y = \begin{cases} 1 & \text{if all three coins come up heads or all three coins come up tails,} \\ 0 & \text{otherwise.} \end{cases}$$

Are these two random variables independent? Observe the following: If  $Y = 1$ , then  $X = 0$  or  $X = 3$ . In other words, if we are given some information about the random variable  $Y$  (in this case,  $Y = 1$ ), then the random variable  $X$  cannot take, for example, the value 2. Based on this, we take  $x = 2$  and  $y = 1$  in Definition 6.2.1. Since the event “ $X = 2 \wedge Y = 1$ ” is equal to  $\emptyset$ , we have

$$\Pr(X = 2 \wedge Y = 1) = \Pr(\emptyset) = 0.$$

On the other hand, we have seen in Section 6.1.2 that  $\Pr(X = 2) = 3/8$  and  $\Pr(Y = 1) = 1/4$ . It follows that

$$\Pr(X = 2 \wedge Y = 1) \neq \Pr(X = 2) \cdot \Pr(Y = 1)$$

and, therefore, the random variables  $X$  and  $Y$  are not independent.

Now consider the random variable

$$Z = \begin{cases} 1 & \text{if the first coin comes up heads,} \\ 0 & \text{if the first coin comes up tails.} \end{cases}$$

We claim that the random variables  $Y$  and  $Z$  are independent. To verify this, we have to show that for all real numbers  $y$  and  $z$ ,

$$\Pr(Y = y \wedge Z = z) = \Pr(Y = y) \cdot \Pr(Z = z). \quad (6.1)$$

Recall from Section 6.1.2 that  $\Pr(Y = 1) = 1/4$  and  $\Pr(Y = 0) = 3/4$ . Since the coin flips are independent, we have  $\Pr(Z = 1) = 1/2$  and  $\Pr(Z = 0) = 1/2$ . Furthermore,

$$\begin{aligned} \Pr(Y = 1 \wedge Z = 1) &= \Pr(HHH) \\ &= 1/8, \\ \Pr(Y = 1 \wedge Z = 0) &= \Pr(TTT) \\ &= 1/8, \\ \Pr(Y = 0 \wedge Z = 1) &= \Pr(HHT, HTH, HTT) \\ &= 3/8, \\ \Pr(Y = 0 \wedge Z = 0) &= \Pr(THH, THT, TTH) \\ &= 3/8. \end{aligned}$$

It follows that

$$\Pr(Y = 1 \wedge Z = 1) = \Pr(Y = 1) \cdot \Pr(Z = 1),$$

$$\Pr(Y = 1 \wedge Z = 0) = \Pr(Y = 1) \cdot \Pr(Z = 0),$$

$$\Pr(Y = 0 \wedge Z = 1) = \Pr(Y = 0) \cdot \Pr(Z = 1),$$

and

$$\Pr(Y = 0 \wedge Z = 0) = \Pr(Y = 0) \cdot \Pr(Z = 0).$$

Thus, (6.1) holds if  $(y, z) \in \{(1, 1), (1, 0), (0, 1), (0, 0)\}$ . For any other pair  $(y, z)$ , such as  $(y, z) = (3, 5)$  or  $(y, z) = (1, 2)$ , at least one of the events “ $Y = y$ ” and “ $Z = z$ ” is the empty set, i.e., cannot occur. Therefore, for such pairs, we have

$$\Pr(Y = y \wedge Z = z) = 0 = \Pr(Y = y) \cdot \Pr(Z = z).$$

Thus, we have indeed verified that (6.1) holds for all real numbers  $y$  and  $z$ . As a result, we have shown that the random variables  $Y$  and  $Z$  are independent.

Are the random variables  $X$  and  $Z$  independent? If  $X = 0$ , then all three coins come up tails and, therefore,  $Z = 0$ . Thus,

$$\Pr(X = 0 \wedge Z = 1) = \Pr(\emptyset) = 0,$$

whereas

$$\Pr(X = 0) \cdot \Pr(Z = 1) = 1/8 \cdot 1/2 \neq 0.$$

As a result, the random variables  $X$  and  $Z$  are not independent.

We have defined the notion of two random variables being independent. As in Definition 5.11.3, there are two ways to generalize this to sequences of random variables:

**Definition 6.2.2** Let  $(S, \Pr)$  be a probability space, let  $n \geq 2$ , and let  $X_1, X_2, \dots, X_n$  be a sequence of random variables on  $S$ .

1. We say that this sequence is *pairwise independent* if for all real numbers  $x_1, x_2, \dots, x_n$ , the sequence “ $X_1 = x_1$ ”, “ $X_2 = x_2$ ”, …, “ $X_n = x_n$ ” of events is pairwise independent.
2. We say that this sequence is *mutually independent* if for all real numbers  $x_1, x_2, \dots, x_n$ , the sequence “ $X_1 = x_1$ ”, “ $X_2 = x_2$ ”, …, “ $X_n = x_n$ ” of events is mutually independent.

## 6.3 Distribution Functions

Consider a random variable  $X$  on a sample space  $S$ . In Section 6.1.2, we have defined  $\Pr(X = x)$ , i.e., the probability of the event “ $X = x$ ”, to be

$$\Pr(X = x) = \Pr(\{\omega \in S : X(\omega) = x\}).$$

This defines a function that maps any real number  $x$  to the real number  $\Pr(X = x)$ . This function is called the distribution function of the random variable  $X$ :

**Definition 6.3.1** Let  $(S, \Pr)$  be a probability space and let  $X : S \rightarrow \mathbb{R}$  be a random variable. The *distribution function* of  $X$  is the function  $D : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$D(x) = \Pr(X = x)$$

for all  $x \in \mathbb{R}$ .

For example, consider a fair red die and a fair blue die, and assume we roll them independently. The sample space is

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\},$$

where  $i$  is the result of the red die and  $j$  is the result of the blue die. Each outcome  $(i, j)$  in  $S$  has the same probability of  $1/36$ .

Let  $X$  be the random variable whose value is equal to the sum of the results of the two dies. The matrix below gives all possible values of  $X$ . The leftmost column gives the result of the red die, the top row gives the result of the blue die, and each other entry is the corresponding value of  $X$ .

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

As can be seen from this matrix, the random variable  $X$  can take any value in  $\{2, 3, 4, \dots, 12\}$ . The distribution function  $D$  of  $X$  is given by

$$\begin{aligned} D(2) &= \Pr(X = 2) = 1/36, \\ D(3) &= \Pr(X = 3) = 2/36, \\ D(4) &= \Pr(X = 4) = 3/36, \\ D(5) &= \Pr(X = 5) = 4/36, \\ D(6) &= \Pr(X = 6) = 5/36, \\ D(7) &= \Pr(X = 7) = 6/36, \\ D(8) &= \Pr(X = 8) = 5/36, \\ D(9) &= \Pr(X = 9) = 4/36, \\ D(10) &= \Pr(X = 10) = 3/36, \\ D(11) &= \Pr(X = 11) = 2/36, \\ D(12) &= \Pr(X = 12) = 1/36, \end{aligned}$$

whereas for all  $x \notin \{2, 3, 4, \dots, 12\}$ ,

$$D(x) = \Pr(X = x) = 0.$$

In Sections 6.6 and 6.7, we will see other examples of distribution functions.

## 6.4 Expected Values

Consider the probability space  $(S, \Pr)$  with sample space  $S = \{1, 2, 3\}$  and probability function  $\Pr$  defined by  $\Pr(1) = 4/5$ ,  $\Pr(2) = 1/10$ , and  $\Pr(3) = 1/10$ . Assume we choose an element in  $S$  according to this probability function. Let  $X$  be the random variable whose value is equal to the element in  $S$  that is chosen. Thus, as a function  $X : S \rightarrow \mathbb{R}$ , we have  $X(1) = 1$ ,  $X(2) = 2$ , and  $X(3) = 3$ .

The “expected value” of  $X$  is the value of  $X$  that we observe “on average”. How should we define this? Since  $X$  has a much higher probability to take the value 1 than the other two values 2 and 3, the value 1 should get a larger “weight” in the expected value of  $X$ . Based on this, it is natural to define the expected value of  $X$  to be

$$1 \cdot \Pr(1) + 2 \cdot \Pr(2) + 3 \cdot \Pr(3) = 1 \cdot \frac{4}{5} + 2 \cdot \frac{1}{10} + 3 \cdot \frac{1}{10} = \frac{13}{10}.$$

**Definition 6.4.1** Let  $(S, \Pr)$  be a probability space and let  $X : S \rightarrow \mathbb{R}$  be a random variable. The *expected value* of  $X$  is defined to be

$$\mathbb{E}(X) = \sum_{\omega \in S} X(\omega) \cdot \Pr(\omega),$$

provided this summation converges absolutely<sup>1</sup>.

### 6.4.1 Some Examples

**Flipping a coin:** Assume we flip a fair coin, in which case the sample space is  $S = \{H, T\}$  and  $\Pr(H) = \Pr(T) = 1/2$ . Define the random variable  $X$  to have value

$$X = \begin{cases} 1 & \text{if the coin comes up heads,} \\ 0 & \text{if the coin comes up tails.} \end{cases}$$

Thus, as a function  $X : S \rightarrow \mathbb{R}$ , we have  $X(H) = 1$  and  $X(T) = 0$ . The expected value  $\mathbb{E}(X)$  of  $X$  is equal to

$$\begin{aligned} \mathbb{E}(X) &= X(H) \cdot \Pr(H) + X(T) \cdot \Pr(T) \\ &= 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} \\ &= \frac{1}{2}. \end{aligned}$$

This example shows that the term “expected value” is a bit misleading:  $\mathbb{E}(X)$  is *not* the value that we expect to observe, because the value of  $X$  is *never* equal to its expected value.

**Rolling a die:** Assume we roll a fair die. Define the random variable  $X$  to be the value of the result. Then,  $X$  takes each of the values in  $\{1, 2, 3, 4, 5, 6\}$  with equal probability  $1/6$ , and we get

$$\begin{aligned} \mathbb{E}(X) &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{7}{2}. \end{aligned}$$

---

<sup>1</sup>The series  $\sum_{n=0}^{\infty} a_n$  converges absolutely if the series  $\sum_{n=0}^{\infty} |a_n|$  converges. If a series converges absolutely, then we can change the order of summation without changing the value of the series.

Now define the random variable  $Y$  to be equal to one divided by the result of the die. In other words,  $Y = 1/X$ . This random variable takes each of the values in  $\{1, 1/2, 1/3, 1/4, 1/5, 1/6\}$  with equal probability  $1/6$ , and we get

$$\begin{aligned}\mathbb{E}(Y) &= 1 \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} \\ &= \frac{49}{120}.\end{aligned}$$

Note that  $\mathbb{E}(Y) \neq 1/\mathbb{E}(X)$ . Thus, this example shows that, in general,  $\mathbb{E}(1/X) \neq 1/\mathbb{E}(X)$ .

**Rolling two dice:** Consider a fair red die and a fair blue die, and assume we roll them independently. The sample space is

$$S = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\},$$

where  $i$  is the result of the red die and  $j$  is the result of the blue die. Each outcome  $(i, j)$  in  $S$  has the same probability of  $1/36$ .

Let  $X$  be the random variable whose value is equal to the sum of the results of the two rolls. As a function  $X : S \rightarrow \mathbb{R}$ , we have  $X(i, j) = i+j$ . The matrix below gives all possible values of  $X$ . The leftmost column indicates the result of the red die, the top row indicates the result of the blue die, and each other entry is the corresponding value of  $X$ .

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The expected value  $\mathbb{E}(X)$  of  $X$  is equal to

$$\begin{aligned}
 \mathbb{E}(X) &= \sum_{(i,j) \in S} X(i,j) \cdot \Pr(i,j) \\
 &= \sum_{(i,j) \in S} (i+j) \cdot \frac{1}{36} \\
 &= \frac{1}{36} \sum_{(i,j) \in S} (i+j) \\
 &= \frac{1}{36} \cdot \text{the sum of all 36 entries in the matrix} \\
 &= \frac{1}{36} \cdot 252 \\
 &= 7.
 \end{aligned}$$

#### 6.4.2 Comparing the Expected Values of Comparable Random Variables

Consider a probability space  $(S, \Pr)$ , and let  $X$  and  $Y$  be two random variables on  $S$ . Recall that  $X$  and  $Y$  are functions that map elements of  $S$  to real numbers. We will write  $X \leq Y$ , if for each element  $\omega \in S$ , we have  $X(\omega) \leq Y(\omega)$ . In other words, the value of  $X$  is at most the value of  $Y$ , no matter which outcome  $\omega$  is chosen. The following lemma should not be surprising:

**Lemma 6.4.2** *Let  $(S, \Pr)$  be a probability space and let  $X$  and  $Y$  be two random variables on  $S$ . If  $X \leq Y$ , then  $\mathbb{E}(X) \leq \mathbb{E}(Y)$ .*

**Proof.** Using Definition 6.4.1 and the assumption that  $X \leq Y$ , we obtain

$$\begin{aligned}
 \mathbb{E}(X) &= \sum_{\omega \in S} X(\omega) \cdot \Pr(\omega) \\
 &\leq \sum_{\omega \in S} Y(\omega) \cdot \Pr(\omega) \\
 &= \mathbb{E}(Y).
 \end{aligned}$$

■

### 6.4.3 An Alternative Expression for the Expected Value

In the last example of Section 6.4.1, we used Definition 6.4.1 to compute the expected value  $\mathbb{E}(X)$  of the random variable  $X$  that was defined to be the sum of the results when rolling two fair and independent dice. This was a painful way to compute  $\mathbb{E}(X)$ , because we added all 36 entries in the matrix. There is a slightly easier way to determine  $\mathbb{E}(X)$ : By looking at the matrix, we see that the value 4 occurs three times. Thus, the event “ $X = 4$ ” has size 3, i.e., if we consider the subset of the sample space  $S$  that corresponds to this event, then this subset has size 3. Similarly, the event “ $X = 7$ ” has size 6, because the value 7 occurs 6 times in the matrix. The table below lists the sizes of all non-empty events, together with their probabilities.

event	size of event	probability
$X = 2$	1	$1/36$
$X = 3$	2	$2/36$
$X = 4$	3	$3/36$
$X = 5$	4	$4/36$
$X = 6$	5	$5/36$
$X = 7$	6	$6/36$
$X = 8$	5	$5/36$
$X = 9$	4	$4/36$
$X = 10$	3	$3/36$
$X = 11$	2	$2/36$
$X = 12$	1	$1/36$

Based on this, we get

$$\begin{aligned} \mathbb{E}(X) &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + \\ &\quad 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} \\ &= 7. \end{aligned}$$

Even though this is still quite painful, less computation is needed. What we have done is the following: In the definition of  $\mathbb{E}(X)$ , i.e.,

$$\mathbb{E}(X) = \sum_{(i,j) \in S} X(i,j) \cdot \Pr(i,j),$$

we rearranged the terms in the summation. That is, instead of taking the sum over all elements  $(i,j)$  in  $S$ ,

- we grouped together all outcomes  $(i, j)$  for which  $X(i, j) = i + j$  has the same value, say,  $k$ ,
- we multiplied this common value  $k$  by the probability that  $X$  is equal to  $k$ ,
- and we took the sum of the resulting products over all possible values of  $k$ .

This resulted in

$$\mathbb{E}(X) = \sum_{k=2}^{12} k \cdot \Pr(X = k).$$

The following lemma states that we can do this for any random variable.

**Lemma 6.4.3** *Let  $(S, \Pr)$  be a probability space and let  $X : S \rightarrow \mathbb{R}$  be a random variable. The expected value of  $X$  is equal to*

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x).$$

**Proof.** Recall that the event “ $X = x$ ” corresponds to the subset

$$A_x = \{\omega \in S : X(\omega) = x\}$$

of the sample space  $S$ . We have

$$\begin{aligned} \mathbb{E}(X) &= \sum_{\omega \in S} X(\omega) \cdot \Pr(\omega) \\ &= \sum_x \sum_{\omega: X(\omega)=x} X(\omega) \cdot \Pr(\omega) \\ &= \sum_x \sum_{\omega: X(\omega)=x} x \cdot \Pr(\omega) \\ &= \sum_x \sum_{\omega \in A_x} x \cdot \Pr(\omega) \\ &= \sum_x x \sum_{\omega \in A_x} \Pr(\omega) \\ &= \sum_x x \cdot \Pr(A_x) \\ &= \sum_x x \cdot \Pr(X = x). \end{aligned}$$

■

When determining the expected value of a random variable  $X$ , it is usually easier to use Lemma 6.4.3 than Definition 6.4.1. To use Lemma 6.4.3, you have to do the following:

- Determine all values  $x$  that  $X$  can take, i.e., determine the *range* of the function  $X$ .
- For each such value  $x$ , determine  $\Pr(X = x)$ .
- Compute the sum of all products  $x \cdot \Pr(X = x)$ .

**Expected value of a random variable  $X : S \rightarrow \mathbb{R}$ :**

- Definition 6.4.1:  $\mathbb{E}(X) = \sum_{\omega \in S} X(\omega) \cdot \Pr(\omega)$ . This is a sum over all elements of the *domain* of  $X$ .
- Lemma 6.4.3:  $\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x)$ . This is a sum over all elements of the *range* of  $X$ .

## 6.5 Linearity of Expectation

In this section, we will present one of the most useful tools for determining expected values. Consider a probability space  $(S, \Pr)$ , and let  $X$  and  $Y$  be two random variables on  $S$ . Recall that  $X$  and  $Y$  are functions that map elements of  $S$  to real numbers. Let  $a$  and  $b$  be two real numbers, and let  $Z : S \rightarrow \mathbb{R}$  be the random variable defined by

$$Z(\omega) = a \cdot X(\omega) + b \cdot Y(\omega)$$

for all elements  $\omega$  in  $S$ . Thus, we combine the random variables  $X$  and  $Y$ , together with the real numbers  $a$  and  $b$ , into a new random variable  $Z$  on the same sample space  $S$ . Usually, we just write this new random variable as  $Z = aX + bY$ .

The *Linearity of Expectation* tells us how to obtain the expected value of  $Z$  from the expected values of  $X$  and  $Y$ :

**Theorem 6.5.1** Let  $(S, \Pr)$  be a probability space. For any two random variables  $X$  and  $Y$  on  $S$ , and for any two real numbers  $a$  and  $b$ ,

$$\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y).$$

**Proof.** We write  $Z = aX + bY$ . Using Definition 6.4.1, we get

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{\omega \in S} Z(\omega) \cdot \Pr(\omega) \\ &= \sum_{\omega \in S} (a \cdot X(\omega) + b \cdot Y(\omega)) \cdot \Pr(\omega) \\ &= a \sum_{\omega \in S} X(\omega) \cdot \Pr(\omega) + b \sum_{\omega \in S} Y(\omega) \cdot \Pr(\omega) \\ &= a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y).\end{aligned}$$

■

Let us return to the example in which we roll two fair and independent dice, one being red and the other being blue. Define the random variable  $X$  to be the sum of the results of the two rolls. We have seen two ways to compute the expected value  $\mathbb{E}(X)$  of  $X$ . We now present a third way, which is the easiest one: We define two random variables

$$Y = \text{the result of the red die}$$

and

$$Z = \text{the result of the blue die.}$$

In Section 6.4.1, we have seen that

$$\mathbb{E}(Y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

By the same computation, we have

$$\mathbb{E}(Z) = \frac{7}{2}.$$

Observe that

$$X = Y + Z.$$

Then, by the Linearity of Expectation (i.e., Theorem 6.5.1), we have

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(Y + Z) \\ &= \mathbb{E}(Y) + \mathbb{E}(Z) \\ &= \frac{7}{2} + \frac{7}{2} \\ &= 7.\end{aligned}$$

We have stated the Linearity of Expectation for two random variables. The proof of Theorem 6.5.1 can easily be generalized to any finite sequence of random variables:

**Theorem 6.5.2** *Let  $(S, \Pr)$  be a probability space, let  $n \geq 2$  be an integer, let  $X_1, X_2, \dots, X_n$  be a sequence of random variables on  $S$ , and let  $a_1, a_2, \dots, a_n$  be a sequence of real numbers. Then,*

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \cdot \mathbb{E}(X_i).$$

The following theorem states that the Linearity of Expectation also holds for infinite sequences of random variables:

**Theorem 6.5.3** *Let  $(S, \Pr)$  be a probability space and let  $X_1, X_2, \dots$  be an infinite sequence of random variables on  $S$  such that the infinite series*

$$\sum_{i=1}^{\infty} \mathbb{E}(|X_i|)$$

*converges. Then,*

$$\mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i).$$

**Proof.** Define the random variable  $X$  to be

$$X = \sum_{i=1}^{\infty} X_i.$$

That is, as a function  $X : S \rightarrow \mathbb{R}$ , we have

$$X(\omega) = \sum_{i=1}^{\infty} X_i(\omega)$$

for all elements  $\omega$  in  $S$ .

The derivation below uses Definition 6.4.1 and the assumption that the infinite series  $\sum_{i=1}^{\infty} \mathbb{E}(|X_i|)$  converges, which allows us to change the order of summation without changing the value of the series:

$$\begin{aligned}\sum_{i=1}^{\infty} \mathbb{E}(X_i) &= \sum_{i=1}^{\infty} \sum_{\omega \in S} X_i(\omega) \cdot \Pr(\omega) \\ &= \sum_{\omega \in S} \sum_{i=1}^{\infty} X_i(\omega) \cdot \Pr(\omega) \\ &= \sum_{\omega \in S} \Pr(\omega) \sum_{i=1}^{\infty} X_i(\omega) \\ &= \sum_{\omega \in S} \Pr(\omega) \cdot X(\omega) \\ &= \mathbb{E}(X) \\ &= \mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right).\end{aligned}$$

■

## 6.6 The Geometric Distribution

Let  $p$  be a real number with  $0 < p < 1$  and consider an experiment that is *successful* with probability  $p$  and *fails* with probability  $1 - p$ . We repeat this experiment independently until it is successful for the first time. What is the expected number of times that we perform the experiment?

We model this problem in the following way: Assume we have a coin that comes up heads with probability  $p$  and, thus, comes up tails with probability  $1 - p$ . We flip this coin repeatedly and independently until it comes up heads for the first time. (We have seen this process in Section 5.15 for the case when  $p = 1/2$ .) Define the random variable  $X$  to be the number of times that we flip the coin; this includes the last coin flip, which resulted in heads. We want to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

The sample space is given by

$$S = \{T^{k-1}H : k \geq 1\},$$

where  $T^{k-1}H$  denotes the sequence consisting of  $k - 1$  tails followed by one heads. Since the coin flips are independent, the outcome  $T^{k-1}H$  has a probability of  $(1 - p)^{k-1}p = p(1 - p)^{k-1}$ , i.e.,

$$\Pr(T^{k-1}H) = p(1 - p)^{k-1}.$$

Let us first verify that all probabilities add up to 1: Using Lemma 5.15.2, we have

$$\begin{aligned}\sum_{k=1}^{\infty} \Pr(T^{k-1}H) &= \sum_{k=1}^{\infty} p(1 - p)^{k-1} \\ &= p \sum_{k=1}^{\infty} (1 - p)^{k-1} \\ &= p \sum_{\ell=0}^{\infty} (1 - p)^{\ell} \\ &= p \cdot \frac{1}{1 - (1 - p)} \\ &= 1.\end{aligned}$$

### 6.6.1 Determining the Expected Value

We are going to use Lemma 6.4.3 to determine the expected value  $\mathbb{E}(X)$ . We first observe that  $X$  can take any value in  $\{1, 2, 3, \dots\}$ . For any integer  $k \geq 1$ ,  $X = k$  if and only if the coin flips give the sequence  $T^{k-1}H$ . It follows that

$$\Pr(X = k) = \Pr(T^{k-1}H) = p(1 - p)^{k-1}. \quad (6.2)$$

By Lemma 6.4.3, we have

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=1}^{\infty} k \cdot \Pr(X = k) \\ &= \sum_{k=1}^{\infty} kp(1 - p)^{k-1} \\ &= p \sum_{k=1}^{\infty} k(1 - p)^{k-1}.\end{aligned}$$

How do we determine the infinite series on the right-hand side?

According to Lemma 5.15.2, we have

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x},$$

for any real number  $x$  with  $-1 < x < 1$ . Both sides of this equation are functions of  $x$  and these two functions are equal to each other. If we differentiate both sides, we get two derivatives that are also equal to each other:

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

Since for  $k = 0$ , the term  $kx^{k-1}$  is equal to 0, we have

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}.$$

If we take  $x = 1 - p$ , we get

$$\begin{aligned}\mathbb{E}(X) &= p \sum_{k=1}^{\infty} k(1-p)^{k-1} \\ &= p \cdot \frac{1}{(1-(1-p))^2} \\ &= \frac{p}{p^2} \\ &= \frac{1}{p}.\end{aligned}$$

In Section 6.3, we have defined the distribution function of a random variable. The distribution function of the coin-flipping random variable  $X$  is given by (6.2). This function is called a geometric distribution:

**Definition 6.6.1** Let  $p$  be a real number with  $0 < p < 1$ . A random variable  $X$  has a *geometric distribution with parameter  $p$* , if its distribution function satisfies

$$\Pr(X = k) = p(1-p)^{k-1}$$

for any integer  $k \geq 1$ .

Our calculation that led to the value of  $\mathbb{E}(X)$  proves the following theorem:

**Theorem 6.6.2** *Let  $p$  be a real number with  $0 < p < 1$  and let  $X$  be a random variable that has a geometric distribution with parameter  $p$ . Then*

$$\mathbb{E}(X) = 1/p.$$

For example, if we flip a fair coin (in which case  $p = 1/2$ ) repeatedly and independently until it comes up heads for the first time, then the expected number of coin flips is equal to 2.

## 6.7 The Binomial Distribution

As in Section 6.6, we choose a real number  $p$  with  $0 < p < 1$ , and consider an experiment that is successful with probability  $p$  and fails with probability  $1-p$ . For an integer  $n \geq 1$ , we repeat the experiment, independently,  $n$  times. What is the expected number of times that the experiment is successful?

We again model this problem using a coin that comes up heads with probability  $p$  and, thus, comes up tails with probability  $1-p$ . We flip the coin, independently,  $n$  times and define the random variable  $X$  to be the number of times the coin comes up heads. We want to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

Since our coin comes up heads with probability  $p$ , it is reasonable to guess that  $\mathbb{E}(X)$  is equal to  $pn$ . For example, if  $p = 1/2$ , then, on average,  $n/2$  of the coin flips should come up heads. We will prove below that  $\mathbb{E}(X)$  is indeed equal to  $pn$ .

### 6.7.1 Determining the Expected Value

Since the random variable  $X$  can take any value in  $\{0, 1, 2, \dots, n\}$ , we have, by Lemma 6.4.3,

$$\mathbb{E}(X) = \sum_{k=0}^n k \cdot \Pr(X = k).$$

Thus, we have to determine  $\Pr(X = k)$ , i.e., the probability that in a sequence of  $n$  independent coin flips, the coin comes up heads exactly  $k$  times.

To give an example, assume that  $n = 4$  and  $k = 2$ . The table below gives all  $\binom{4}{2} = 6$  sequences of 4 coin flips that contain exactly 2 H's, together with their probabilities:

sequence	probability
$HHTT$	$p \cdot p \cdot (1-p) \cdot (1-p) = p^2(1-p)^2$
$HTHT$	$p \cdot (1-p) \cdot p \cdot (1-p) = p^2(1-p)^2$
$HTTH$	$p \cdot (1-p) \cdot (1-p) \cdot p = p^2(1-p)^2$
$THHT$	$(1-p) \cdot p \cdot p \cdot (1-p) = p^2(1-p)^2$
$THTH$	$(1-p) \cdot p \cdot (1-p) \cdot p = p^2(1-p)^2$
$TTHH$	$(1-p) \cdot (1-p) \cdot p \cdot p = p^2(1-p)^2$

As can be seen from this table, each of the  $\binom{4}{2}$  sequences has the same probability  $p^2(1-p)^2$ . It follows that, if  $n = 4$ ,

$$\Pr(X = 2) = \binom{4}{2} p^2(1-p)^2.$$

We now consider the general case. Let  $n \geq 1$  and  $k$  be integers with  $0 \leq k \leq n$ . Then,  $X = k$  if and only if there are exactly  $k$  H's in the sequence of  $n$  coin flips. The number of such sequences is equal to  $\binom{n}{k}$ , and each one of them has probability  $p^k(1-p)^{n-k}$ . Therefore, we have

$$\Pr(X = k) = \binom{n}{k} p^k(1-p)^{n-k}. \quad (6.3)$$

As a sanity check, let us use Newton's Binomial Theorem (i.e., Theorem 3.6.5) to verify that all probabilities add up to 1:

$$\begin{aligned} \sum_{k=0}^n \Pr(X = k) &= \sum_{k=0}^n \binom{n}{k} p^k(1-p)^{n-k} \\ &= ((1-p) + p)^n \\ &= 1. \end{aligned}$$

We are now ready to compute the expected value of the random vari-

able  $X$ :

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^n k \cdot \Pr(X = k) \\ &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k}.\end{aligned}$$

Since

$$\begin{aligned}k \binom{n}{k} &= k \cdot \frac{n!}{k!(n-k)!} \\ &= n \cdot \frac{(n-1)!}{(k-1)!(n-k)!} \\ &= n \binom{n-1}{k-1},\end{aligned}$$

we get

$$\mathbb{E}(X) = \sum_{k=1}^n n \binom{n-1}{k-1} p^k (1-p)^{n-k}.$$

By changing the summation variable from  $k$  to  $\ell + 1$ , we get

$$\begin{aligned}\mathbb{E}(X) &= \sum_{\ell=0}^{n-1} n \binom{n-1}{\ell} p^{\ell+1} (1-p)^{n-1-\ell} \\ &= pn \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^\ell (1-p)^{n-1-\ell}.\end{aligned}$$

By Newton's Binomial Theorem (i.e., Theorem 3.6.5), the summation is equal to

$$((1-p) + p)^{n-1} = 1.$$

Therefore, we get

$$\begin{aligned}\mathbb{E}(X) &= pn \cdot 1 \\ &= pn.\end{aligned}$$

We have done the following: Our intuition told us that  $\mathbb{E}(X) = pn$ . Then, we went through a painful calculation to show that our intuition was correct. There must be an easier way to show that  $\mathbb{E}(X) = pn$ . We will show below that there is indeed a much easier way.

### 6.7.2 Using the Linearity of Expectation

We define a sequence  $X_1, X_2, \dots, X_n$  of random variables as follows: For each  $i$  with  $1 \leq i \leq n$ ,

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th coin flip results in heads,} \\ 0 & \text{if the } i\text{-th coin flip results in tails.} \end{cases}$$

Observe that

$$X = X_1 + X_2 + \cdots + X_n,$$

because

- $X$  counts the number of heads in the sequence of  $n$  coin flips, and
- the summation on the right-hand side is equal to the number of 1's in the sequence  $X_1, X_2, \dots, X_n$ , which, by definition, is equal to the number of heads in the sequence of  $n$  coin flips.

Using the Linearity of Expectation (see Theorem 6.5.2), we get

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n \mathbb{E}(X_i). \end{aligned}$$

Thus, we have to determine the expected value of  $X_i$ . Since  $X_i$  is either 1 or 0, we have, using Lemma 6.4.3,

$$\begin{aligned} \mathbb{E}(X_i) &= 1 \cdot \Pr(X_i = 1) + 0 \cdot \Pr(X_i = 0) \\ &= \Pr(X_i = 1) \\ &= \Pr(\text{the } i\text{-th coin flip results in heads}) \\ &= p. \end{aligned}$$

We conclude that

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \sum_{i=1}^n p \\ &= pn.\end{aligned}$$

I hope you agree that this is much easier than what we did before.

The distribution function of the random variable  $X$  is given by (6.3). This function is called a binomial distribution:

**Definition 6.7.1** Let  $n \geq 1$  be an integer and let  $p$  be a real number with  $0 < p < 1$ . A random variable  $X$  has a *binomial distribution with parameters  $n$  and  $p$* , if its distribution function satisfies

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for any integer  $k$  with  $0 \leq k \leq n$ .

Our calculation that led to the value of  $\mathbb{E}(X)$  proves the following theorem:

**Theorem 6.7.2** Let  $n \geq 1$  be an integer, let  $p$  be a real number with  $0 < p < 1$ , and let  $X$  be a random variable that has a binomial distribution with parameters  $n$  and  $p$ . Then

$$\mathbb{E}(X) = pn.$$

## 6.8 Indicator Random Variables

In Section 6.7, we considered the random variable  $X$  whose value is equal to the number of heads in a sequence of  $n$  independent coin flips. In Section 6.7.2, we defined a sequence  $X_1, X_2, \dots, X_n$  of random variables, where  $X_i = 1$  if the  $i$ -th coin flip results in heads and  $X_i = 0$  otherwise. This random variable  $X_i$  indicates whether or not the  $i$ -th flip in the sequence is heads. Because of this, we call  $X_i$  an indicator random variable.

**Definition 6.8.1** A random variable  $X$  is an *indicator random variable*, if it can only take values in  $\{0, 1\}$ .

As we have already seen in Section 6.7.2, the expected value of an indicator random variable is easy to determine:

**Lemma 6.8.2** *If  $X$  is an indicator random variable, then*

$$\mathbb{E}(X) = \Pr(X = 1).$$

**Proof.** Since  $X$  is either 0 or 1, we have, using Lemma 6.4.3,

$$\begin{aligned}\mathbb{E}(X) &= 0 \cdot \Pr(X = 0) + 1 \cdot \Pr(X = 1) \\ &= \Pr(X = 1).\end{aligned}$$

■

In the following subsections, we will see some examples of how indicator random variables can be used to compute the expected value of non-trivial random variables.

### 6.8.1 Runs in Random Bitstrings

Let  $n$  be a large integer. We generate a random bitstring

$$R = r_1 r_2 \dots r_n$$

by flipping a fair coin, independently,  $n$  times. Let  $k \geq 1$  be an integer. Recall from Section 5.14 that a *run of length  $k$*  is a consecutive subsequence of  $R$ , all of whose bits are equal. Define the random variable  $X$  to be the number of runs of length  $k$ .

For example, if  $R$  is the bitstring

0	0	1	1	1	1	1	0	0	0	1	1	0	0	0	0
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

and  $k = 3$ , then  $X = 6$ , because  $R$  contains 6 runs of length 3, starting at positions 3, 4, 5, 8, 13, and 14.

We want to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

A run of length  $k$  can start at any of the positions  $1, 2, \dots, n - k + 1$ . Our approach will be to define an indicator random variable that tells us whether or not the subsequence of length  $k$  that starts at any such position is a run. Thus, for any  $i$  with  $1 \leq i \leq n - k + 1$ , we define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if the subsequence } r_i r_{i+1} \dots r_{i+k-1} \text{ is a run,} \\ 0 & \text{otherwise.} \end{cases}$$

Using Lemma 6.8.2, we get

$$\mathbb{E}(X_i) = \Pr(X_i = 1).$$

Since  $X_i = 1$  if and only if all bits in the subsequence  $r_i r_{i+1} \dots r_{i+k-1}$  are 0 or all bits in this subsequence are 1, we have

$$\begin{aligned} \mathbb{E}(X_i) &= \Pr(X_i = 1) \\ &= (1/2)^k + (1/2)^k \\ &= 1/2^{k-1}. \end{aligned}$$

Since

$$X = \sum_{i=1}^{n-k+1} X_i,$$

the Linearity of Expectation (see Theorem 6.5.2) implies that

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^{n-k+1} X_i\right) \\ &= \sum_{i=1}^{n-k+1} \mathbb{E}(X_i) \\ &= \sum_{i=1}^{n-k+1} 1/2^{k-1} \\ &= \frac{n - k + 1}{2^{k-1}}. \end{aligned}$$

Observe that the random variables  $X_1, X_2, \dots, X_{n+k-1}$  are not mutually independent. (Do you see why?) Nevertheless, our derivation is correct,

because the Linearity of Expectation is valid for *any* sequence of random variables; independence is *not* needed.

For example, if we take  $k = 1 + \log n$ , then  $2^{k-1} = 2^{\log n} = n$ , so that

$$\mathbb{E}(X) = \frac{n - \log n}{n} = 1 - \frac{\log n}{n}.$$

Thus, for large values of  $n$ , the expected number of runs of length  $1 + \log n$  is very close to 1. This is in line with Section 5.14, because we proved there that it is very likely that the sequence contains a run of length about  $\log n$ .

If we take  $k = 1 + \frac{1}{2} \log n$ , then

$$2^{k-1} = 2^{(\log n)/2} = 2^{\log \sqrt{n}} = \sqrt{n}$$

and

$$\mathbb{E}(X) = \frac{n - \frac{1}{2} \log n}{\sqrt{n}} = \sqrt{n} - \frac{\log n}{2\sqrt{n}}.$$

Thus, for large values of  $n$ , the expected number of runs of length  $1 + \frac{1}{2} \log n$  is very close to  $\sqrt{n}$ .

### 6.8.2 Largest Elements in Prefixes of Random Permutations

Let  $n \geq 1$  be an integer and consider a sequence  $s_1, s_2, \dots, s_n$  of  $n$  numbers. The following algorithm computes the largest element in this sequence:

**Algorithm** FINDMAX( $s_1, s_2, \dots, s_n$ ):

```

max = -∞;
for i = 1 to n
  do if  $s_i > max$ 
    then max =  $s_i$       (*)
    endif
  endfor;
return max

```

We would like to know the number of times that line (\*) is executed, i.e., the number of times that the value of the variable  $max$  changes. For example, if the input sequence is

3, 2, 5, 4, 6, 1,

then the value of  $\max$  changes 3 times, namely when we encounter 3, 5, and 6. On the other hand, for the sequence

$$6, 5, 4, 3, 2, 1,$$

the value of  $\max$  changes only once, whereas for

$$1, 2, 3, 4, 5, 6,$$

it changes 6 times.

Assume that the input sequence  $s_1, s_2, \dots, s_n$  is a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ . Thus, each permutation has probability  $1/n!$  of being the input. We define a random variable  $X$  whose value is equal to the number of times that line (\*) is executed when running algorithm  $\text{FINDMAX}(s_1, s_2, \dots, s_n)$ . We are interested in the expected value  $\mathbb{E}(X)$  of this random variable.

The algorithm makes  $n$  iterations. In each iteration, line (\*) is either executed or not executed. We define, for each iteration, an indicator random variable that tells us whether or not line (\*) is executed during that iteration. That is, for any  $i$  with  $1 \leq i \leq n$ , we define

$$X_i = \begin{cases} 1 & \text{if line (*) is executed in the } i\text{-th iteration,} \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$X = \sum_{i=1}^n X_i,$$

it follows from the Linearity of Expectation (see Theorem 6.5.2) that

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \sum_{i=1}^n \Pr(X_i = 1). \end{aligned}$$

How do we determine  $\Pr(X_i = 1)$ ? Observe that  $X_i = 1$  if and only if the maximum of the subsequence  $s_1, s_2, \dots, s_i$  is at the last position in

this subsequence. Since the entire sequence  $s_1, s_2, \dots, s_n$  is a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ , the elements in the subsequence  $s_1, s_2, \dots, s_i$  are in a uniformly random order as well. The largest element in this subsequence is in any of the  $i$  positions with equal probability  $1/i$ . In particular, the probability that the largest element is at the last position in this subsequence is equal to  $1/i$ . It follows that

$$\Pr(X_i = 1) = 1/i.$$

This can be proved in a more formal way as follows: By the Product Rule, the number of permutations  $s_1, s_2, \dots, s_n$  of  $\{1, 2, \dots, n\}$  for which  $s_i$  is the largest element among  $s_1, s_2, \dots, s_i$  is equal to

$$\binom{n}{i} (i-1)!(n-i)! = n!/i.$$

(Do you see why?) Therefore,

$$\Pr(X_i = 1) = \frac{n!/i}{n!} = 1/i.$$

Thus,

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=1}^n \Pr(X_i = 1) \\ &= \sum_{i=1}^n 1/i \\ &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.\end{aligned}$$

The number on the right-hand side is called the *harmonic number* and denoted by  $H_n$ . In the following subsection, we will show that  $H_n$  is approximately equal to  $\ln n$ . Thus, the expected number of times that line (\*) of algorithm FINDMAX is executed, when given as input a uniformly random permutation of  $\{1, 2, \dots, n\}$ , is about  $\ln n$ .

As a final remark, the indicator random variables  $X_1, X_2, \dots, X_n$  that we have introduced above are mutually independent; see Exercise 5.98. Keep in mind, however, that we do not need this, because the Linearity of Expectation does not require these random variables to be mutually independent.

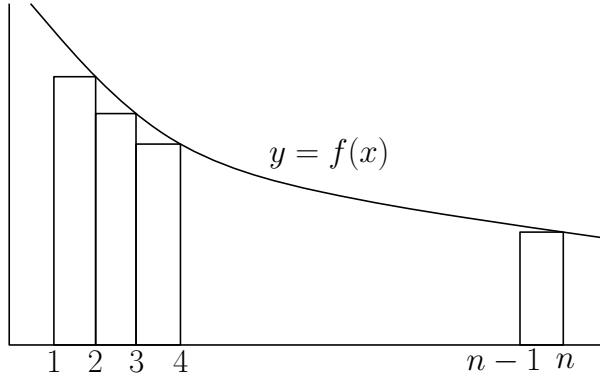
### 6.8.3 Estimating the Harmonic Number

Consider a positive real-valued decreasing function  $f : [1, \infty) \rightarrow \mathbb{R}$ . Thus, if  $1 \leq x < x'$ , then  $f(x) \geq f(x') > 0$ . For any integer  $n \geq 1$ , we would like to estimate the summation

$$\sum_{i=1}^n f(i).$$

For example, if we take  $f(x) = 1/x$ , then the summation is the harmonic number  $H_n$  of the previous subsection.

For each  $i$  with  $2 \leq i \leq n$ , draw the rectangle with bottom-left corner at the point  $(i - 1, 0)$  and top-right corner at the point  $(i, f(i))$ , as in the figure below.



The area of the  $i$ -th rectangle is equal to  $f(i)$  and, thus,

$$\sum_{i=1}^n f(i)$$

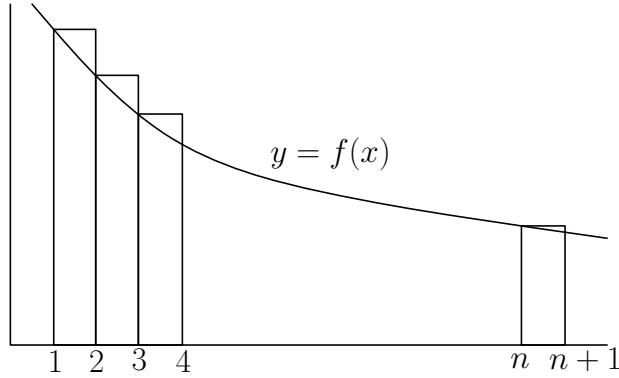
is equal to the sum of

- $f(1)$  and
- the total area of the  $n - 1$  rectangles.

Since  $f$  is decreasing, the rectangles are below the graph  $y = f(x)$ . It follows that the total area of the  $n - 1$  rectangles is less than or equal to the area between  $f$  and the  $x$ -axis, between  $x = 1$  and  $x = n$ . We conclude that

$$\sum_{i=1}^n f(i) \leq f(1) + \int_1^n f(x) dx. \quad (6.4)$$

To obtain a lower bound on the summation, we modify the figure as indicated below: For each  $i$  with  $1 \leq i \leq n$ , draw the rectangle with bottom-left corner at the point  $(i, 0)$  and top-right corner at the point  $(i + 1, f(i))$ ; see the figure below.



In this case, the graph  $y = f(x)$  is below the top sides of the rectangles and, therefore,

$$\sum_{i=1}^n f(i) \geq \int_1^{n+1} f(x) dx. \quad (6.5)$$

If we apply (6.4) and (6.5) to the function  $f(x) = 1/x$ , then we get

$$\begin{aligned} H_n &= \sum_{i=1}^n \frac{1}{i} \\ &\leq 1 + \int_1^n \frac{dx}{x} \\ &= 1 + \ln n \end{aligned}$$

and

$$\begin{aligned} H_n &= \sum_{i=1}^n \frac{1}{i} \\ &\geq \int_1^{n+1} \frac{dx}{x} \\ &= \ln(n+1) \\ &\geq \ln n. \end{aligned}$$

We have proved the following result:

**Lemma 6.8.3** *For any integer  $n \geq 1$ , the harmonic number  $H_n = \sum_{i=1}^n 1/i$  satisfies*

$$\ln n \leq H_n \leq 1 + \ln n.$$

## 6.9 The Insertion-Sort Algorithm

INSERTIONSORT is a simple sorting algorithm that takes as input an array  $A[1 \dots n]$  of numbers. The algorithm uses a for-loop in which a variable  $i$  runs from 2 to  $n$ . At the start of the  $i$ -th iteration,

- the subarray  $A[1 \dots i - 1]$  is sorted, whereas
- the algorithm has not yet seen any of the elements in the subarray  $A[i \dots n]$ .

In the  $i$ -th iteration, the algorithm takes the element  $A[i]$  and repeatedly swaps it with its left neighbor until the subarray  $A[1 \dots i]$  is sorted. The pseudocode of this algorithm is given below.

**Algorithm** INSERTIONSORT( $A[1 \dots n]$ ):

```

for  $i = 2$  to  $n$ 
  do  $j = i$ ;
    while  $j > 1$  and  $A[j] < A[j - 1]$ 
      do swap  $A[j]$  and  $A[j - 1]$ ;
       $j = j - 1$ 
    endwhile
  endfor

```

We are interested in the total number of swaps that are made by this algorithm. The worst-case happens when the input array is sorted in reverse order, in which case the total number of swaps is equal to

$$1 + 2 + 3 + \cdots + (n - 1) = \binom{n}{2}.$$

Thus, in the worst case, each of the  $\binom{n}{2}$  pairs of input elements is swapped.

Assume that the input array  $A[1 \dots n]$  contains a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ . Thus, each permutation has probability

$1/n!$  of being the input. We define the random variable  $X$  to be the total number of swaps made when running algorithm  $\text{INSERTIONSORT}(A[1 \dots n])$ . We will determine the expected value  $\mathbb{E}(X)$  of  $X$ .

Since we want to count the number of pairs of input elements that are swapped, we will use, for each pair of input elements, an indicator random variable that indicates whether or not this pair gets swapped by the algorithm. That is, for each  $a$  and  $b$  with  $1 \leq a < b \leq n$ , we define

$$X_{ab} = \begin{cases} 1 & \text{if } a \text{ and } b \text{ get swapped by the algorithm,} \\ 0 & \text{otherwise.} \end{cases}$$

We observe that, since  $a < b$ , these two elements get swapped if and only if in the input array,  $b$  is to the left of  $a$ . Since the input array contains a uniformly random permutation, the events “ $b$  is to the left of  $a$ ” and “ $a$  is to the left of  $b$ ” are symmetric. Therefore, we have

$$\mathbb{E}(X_{ab}) = \Pr(X_{ab} = 1) = 1/2.$$

A formal proof of this is obtained by showing that there are  $n!/2$  permutations of  $\{1, 2, \dots, n\}$  in which  $b$  appears to the left of  $a$  and, thus,  $n!/2$  permutations in which  $a$  appears to the left of  $b$ . (See also Exercise 5.71.)

Since each pair of input elements is swapped at most once, we have

$$X = \sum_{a=1}^{n-1} \sum_{b=a+1}^n X_{ab}.$$

It follows from the Linearity of Expectation (see Theorem 6.5.2) that

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{a=1}^{n-1} \sum_{b=a+1}^n X_{ab}\right) \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \mathbb{E}(X_{ab}) \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \frac{1}{2} \\ &= \frac{1}{2} \binom{n}{2}. \end{aligned}$$

Thus, the expected number of swaps on a uniformly random input array is one half times the worst-case number of swaps.

## 6.10 The Quick-Sort Algorithm

We have already seen algorithm `QUICKSORT` in Section 1.3. This algorithm takes as input an array  $A[1 \dots n]$  of numbers, which we assume for simplicity to be pairwise distinct. A generic call  $\text{QUICKSORT}(A, i, j)$  takes two indices  $i$  and  $j$  and sorts the subarray  $A[i \dots j]$ . Thus, the call  $\text{QUICKSORT}(A, 1, n)$  sorts the entire array.

**Algorithm** `QUICKSORT`( $A, i, j$ ):

```

if  $i < j$ 
then  $p =$  uniformly random element in  $A[i \dots j]$ ;
    compare  $p$  with all other elements in  $A[i \dots j]$ ;
    rearrange  $A[i \dots j]$  such that it has the following
    form (this rearranging defines the value of  $k$ ):

```

	$< p$	$  p$	$> p$	
$i$		$k$		$j$

```

    QUICKSORT( $A, i, k - 1$ );
    QUICKSORT( $A, k + 1, j$ )
endif

```

The element  $p$  is called the *pivot*. We have seen in Section 1.3 that the worst-case running time of algorithm  $\text{QUICKSORT}(A, 1, n)$  is  $\Theta(n^2)$ . In this section, we will prove that the expected running time is only  $O(n \log n)$ .

We assume for simplicity that the input array is a permutation of the set  $\{1, 2, \dots, n\}$ . We do not make any other assumption about the input. In particular, we do not assume that the input is a random permutation. The only place where randomization is used is when the pivot is chosen: It is chosen uniformly at random in the subarray on which `QUICKSORT` is called.

The quantity that we will analyze is the total number of *comparisons* (between pairs of input elements) that are made during the entire execution of algorithm  $\text{QUICKSORT}(A, 1, n)$ . In such a comparison, the algorithm takes two distinct input elements, say  $a$  and  $b$ , and decides whether  $a < b$  or  $a > b$ . Observe from the pseudocode that the only comparisons being made are between the pivot and all other elements in the subarray that is the input to the current call to `QUICKSORT`. Since the operation “compare  $a$  to  $b$ ” is

the same as the operation “compare  $b$  to  $a$ ” (even though the outcomes are opposite), we will assume below that in such a comparison,  $a < b$ .

We define the random variable  $X$  to be the total number of comparisons that are made by algorithm  $\text{QUICKSORT}(A, 1, n)$ . We will prove that the expected value of  $X$  satisfies  $\mathbb{E}(X) = O(n \log n)$ .

For each  $a$  and  $b$  with  $1 \leq a < b \leq n$ , we consider the indicator random variable

$$X_{ab} = \begin{cases} 1 & \text{if } a \text{ and } b \text{ are compared to each other when} \\ & \text{running } \text{QUICKSORT}(A, 1, n), \\ 0 & \text{otherwise.} \end{cases}$$

Since each pair of input elements is compared at most once, we have

$$X = \sum_{a=1}^{n-1} \sum_{b=a+1}^n X_{ab}.$$

It follows from the Linearity of Expectation (see Theorem 6.5.2) that

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{a=1}^{n-1} \sum_{b=a+1}^n X_{ab}\right) \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \mathbb{E}(X_{ab}) \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \Pr(X_{ab} = 1). \end{aligned}$$

We consider two input elements  $a$  and  $b$  with  $1 \leq a < b \leq n$ . We are going to determine  $\Pr(X_{ab} = 1)$ , i.e., the probability that the elements  $a$  and  $b$  are compared to each other when running algorithm  $\text{QUICKSORT}(A, 1, n)$ . Consider the set

$$S_{ab} = \{a, a + 1, \dots, b\}.$$

At the start of algorithm  $\text{QUICKSORT}(A, 1, n)$ , all elements of the set  $S_{ab}$  are part of the input. Consider the first pivot  $p$  that is chosen. We observe the following:

- Assume that  $p \notin S_{ab}$ .

- If  $p < a$ , then after the algorithm has rearranged the input array, all elements of the set  $S_{ab}$  are to the right of  $p$  and, thus, all these elements are part of the input for the recursive call  $\text{QUICKSORT}(A, k+1, n)$ . During the rearranging,  $a$  and  $b$  have not been compared to each other. However, they *may* be compared to each other during later recursive calls.
- If  $p > b$ , then after the algorithm has rearranged the input array, all elements of the set  $S_{ab}$  are to the left of  $p$  and, thus, all these elements are part of the input for the recursive call  $\text{QUICKSORT}(A, 1, k-1)$ . During the rearranging,  $a$  and  $b$  have not been compared to each other. However, they *may* be compared to each other during later recursive calls.
- Assume that  $p \in S_{ab}$ .
  - If  $p \neq a$  and  $p \neq b$ , then after the algorithm has rearranged the input array,  $a$  is to the left of  $p$  and  $b$  is to the right of  $p$ . During the rearranging,  $a$  and  $b$  have not been compared to each other. Also, since  $a$  and  $b$  have been “separated”, they will not be compared to each other during later recursive calls. Thus, we have  $X_{ab} = 0$ .
  - If  $p = a$  or  $p = b$ , then during the rearranging,  $a$  and  $b$  have been compared to each other. Thus, we have  $X_{ab} = 1$ . (Note that in later recursive calls,  $a$  and  $b$  will not be compared to each other again.)

We conclude that whether or not  $a$  and  $b$  are compared to each other is completely determined by the element of the set  $S_{ab}$  that is the first element in this set to be chosen as a pivot. If this element is equal to  $a$  or  $b$ , then  $X_{ab} = 1$ . On the other hand, if this element belongs to  $S_{ab} \setminus \{a, b\}$ , then  $X_{ab} = 0$ . Since

- in any recursive call, the pivot is chosen uniformly at random from the subarray that is the input for this call, and
- at the start of the first recursive call in which the pivot belongs to the set  $S_{ab}$ , all elements of this set are part of the input for this call,

each of the  $b - a + 1$  elements of  $S_{ab}$  has the same probability of being the

first element of  $S_{ab}$  that is chosen as a pivot. It follows that

$$\Pr(X_{ab} = 1) = \frac{2}{b - a + 1}.$$

We conclude that

$$\begin{aligned}\mathbb{E}(X) &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \Pr(X_{ab} = 1) \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \frac{2}{b - a + 1} \\ &= 2 \sum_{a=1}^{n-1} \left( \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-a+1} \right) \\ &\leq 2 \sum_{a=1}^{n-1} \left( \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \right) \\ &= 2 \sum_{a=1}^{n-1} (H_n - 1) \\ &= 2(n-1)(H_n - 1) \\ &\leq 2n(H_n - 1),\end{aligned}$$

where  $H_n$  is the harmonic number that we have seen in Sections 6.8.2 and 6.8.3. Using Lemma 6.8.3, it follows that

$$\mathbb{E}(X) \leq 2n \ln n.$$

## 6.11 Skip Lists

Consider a set  $S$  of  $n$  numbers. We would like to store these numbers in a data structure that supports the following operations:

- **SEARCH( $x$ ):** This operation returns the largest element in the set  $S$  that is less than or equal to  $x$ .
- **INSERT( $x$ ):** This operation inserts the number  $x$  into the set  $S$ .
- **DELETE( $x$ ):** This operation deletes the number  $x$  from the set  $S$ .

A standard data structure for this problem is a balanced binary search tree (such as a red-black tree or an AVL-tree), which allows each of these three operations to be performed in  $O(\log n)$  time. Searching in a binary search tree is straightforward, but keeping the tree balanced after an insertion or deletion is cumbersome.

In this section, we introduce *skip lists* as an alternative data structure. A skip list is constructed using the outcomes of coin flips, which result in a structure that is balanced in the expected sense. Because of this, the insertion and deletion algorithms become straightforward: We, as a programmer, do not have to take care of rebalancing operations, because the coin flips take care of this.

To define a skip list for the set  $S$  of  $n$  numbers, we first construct a sequence  $S_0, S_1, S_2, \dots$  of subsets of  $S$ :

- Let  $S_0 = S$ .
- For  $i = 0, 1, 2, \dots$ , assume that the set  $S_i$  has already been constructed. If  $S_i$  is non-empty, we do the following:
  - Initialize an empty set  $S_{i+1}$ .
  - For each element  $y$  in the set  $S_i$ , flip a fair and independent coin. If the coin comes up heads, element  $y$  is added to the set  $S_{i+1}$ .

The process terminates as soon as the next set  $S_{i+1}$  is empty.

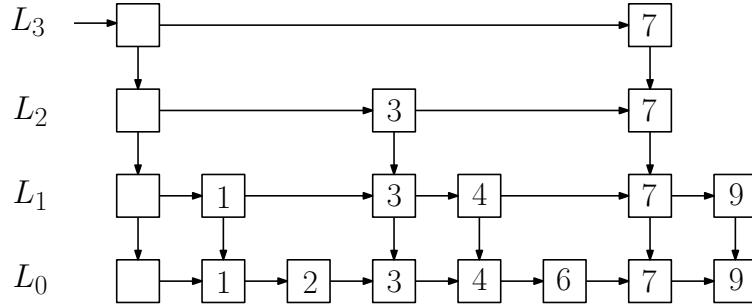
Let  $h$  be the number of non-empty sets that are constructed by this process, and consider the sequence  $S_0, S_1, \dots, S_h$  of sets. Observe that  $h$  is a random variable and each of the sets  $S_1, S_2, \dots, S_h$  is a random subset of  $S$ .

The skip list for  $S$  consists of the following:

- For each  $i$  with  $0 \leq i \leq h$ , we store the sorted sequence of elements of the set  $S_i$  in a linked list  $L_i$ .
  - Each node  $u$  of  $L_i$  stores one element of  $S_i$ , which is denoted by  $\text{key}(u)$ .
  - Each node  $u$  of  $L_i$  stores a pointer to its successor node in  $L_i$ , which is denoted by  $\text{right}(u)$ . If  $u$  is the rightmost node in  $L_i$ , then  $\text{right}(u) = \text{nil}$ .
  - We add a *dummy node* at the beginning of  $L_i$ . The key of this node is  $\text{nil}$  and its successor is the node of  $L_i$  whose key is the smallest element in  $S_i$ .

- For each  $i$  with  $1 \leq i \leq h$  and each node  $u$  of  $L_i$ ,  $u$  stores a pointer to the node  $u'$  in  $L_{i-1}$  for which  $\text{key}(u') = \text{key}(u)$ . The node  $u'$  is denoted by  $\text{down}(u)$ .
- There is a pointer to the dummy node in the list  $L_h$ . We will refer to this node as the *root* of the skip list.

The value of  $h$  is called the *height* of the skip list. An example of a skip list of height  $h = 3$  for the set  $S = \{1, 2, 3, 4, 6, 7, 9\}$  is shown in the figure below.



### 6.11.1 Algorithm SEARCH

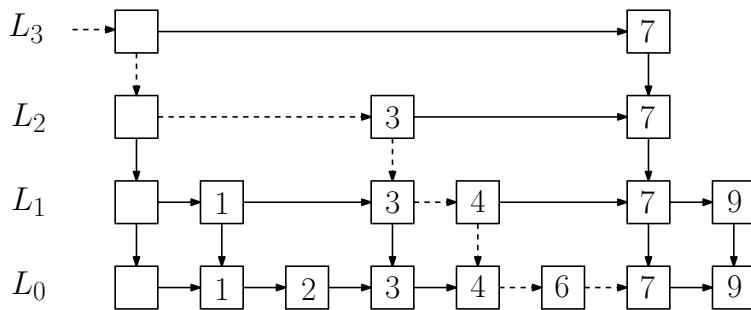
The algorithm that searches for a number  $x$  keeps track of the current node  $u$  and the index  $i$  of the list  $L_i$  that contains  $u$ . Initially,  $u$  is the root of the skip list and  $i = h$ . At any moment, if  $i \geq 1$ , the algorithm tests if the key of  $\text{right}(u)$  is less than  $x$ . If this is the case, then  $u$  moves one node to the right in the list  $L_i$ ; otherwise,  $u$  moves to the node  $\text{down}(u)$  in the list  $L_{i-1}$ . Once  $i = 0$ , node  $u$  moves to the right in the list  $L_0$  and stops at the last node whose key is at most equal to  $x$ . The pseudocode of this algorithm  $\text{SEARCH}(x)$  is given below.

**Algorithm** SEARCH( $x$ ):

```

// returns the rightmost node  $u$  in  $L_0$  such that  $\text{key}(u) \leq x$ 
 $u =$  root of the skip list;
 $i = h;$ 
while  $i \geq 1$ 
  do if  $\text{right}(u) \neq \text{nil}$  and  $\text{key}(\text{right}(u)) < x$ 
    then  $u = \text{right}(u)$ 
    else  $u = \text{down}(u);$ 
       $i = i - 1$ 
    endif
  endwhile;
  while  $\text{right}(u) \neq \text{nil}$  and  $\text{key}(\text{right}(u)) \leq x$ 
    do  $u = \text{right}(u)$ 
  endwhile;
  return  $u$ 
```

The dashed arrows in the figure below show the path that is followed when running algorithm SEARCH(7). Note that if we replace “ $\text{key}(\text{right}(u)) < x$ ” in the first while-loop by “ $\text{key}(\text{right}(u)) \leq x$ ”, we obtain a different path that ends in the same node: This path moves from the root to the node in  $L_3$  whose key is 7, and then it moves down to the list  $L_0$ . As we will see later, using the condition “ $\text{key}(\text{right}(u)) < x$ ” simplifies the algorithm for deleting an element from the skip list.

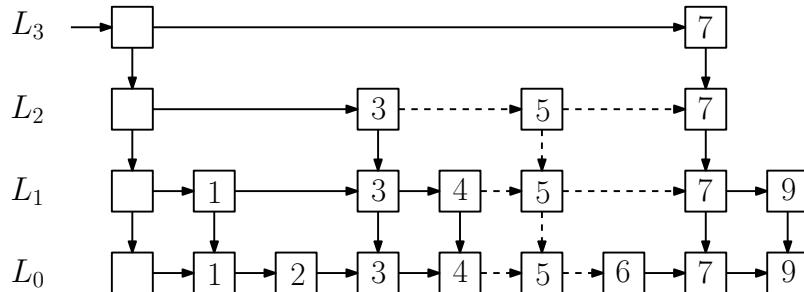


### 6.11.2 Algorithms INSERT and DELETE

Algorithm INSERT( $x$ ) takes as input a number  $x$  and inserts it into the skip list. This algorithm works as follows:

- Run algorithm  $\text{SEARCH}(x)$  and consider the node  $u$  that is returned. We assume that  $\text{key}(u) \neq x$  and, thus,  $x$  is not in the skip list yet. Observe that the new number  $x$  belongs between the nodes  $u$  and  $\text{right}(u)$ .
- Flip a fair and independent coin repeatedly until it comes up tails for the first time. Let  $k$  be the number of flips.
- Add the new number  $x$  to the lists  $L_0, L_1, \dots, L_{k-1}$ . Note that if  $k \geq h+2$ , we have to add new lists  $L_{h+1}, \dots, L_{k-1}$  to the skip list (each one containing a dummy node and a node storing  $x$ ), set  $h = k - 1$ , and update the pointer to the root of the new skip list.
- When adding  $x$  to a list  $L_i$ , we have to know its predecessor in this list.
  - To find these predecessors, we modify algorithm  $\text{SEARCH}(x)$  as follows: Each time the current node  $u$  moves down, we push  $u$  onto an initially empty stack. In this way, the predecessors that we need are stored, in the correct order, on the stack.
  - An easier way that avoids using a stack is to flip the coin and, thus, determine  $k$ , before running algorithm  $\text{SEARCH}(x)$ . We then modify algorithm  $\text{SEARCH}(x)$ : If  $i < k$  and the current node  $u$  moves down, we add the new number  $x$  to  $L_i$  between the nodes  $u$  and  $\text{right}(u)$ .

The figure below shows the skip list that results when inserting the number 5 into our example skip list. In this case,  $k = 3$  and the new number is added to the lists  $L_0$ ,  $L_1$ , and  $L_2$ . The dashed arrows indicate the pointers that are changed during this insertion.



Algorithm  $\text{DELETE}(x)$  takes as input a number  $x$  and deletes it from the skip list. This algorithm does the following:

- Run a modified version of algorithm  $\text{SEARCH}(x)$ : Each time the current node  $u$  moves down, test if  $\text{key}(\text{right}(u)) = x$ . If this is the case, delete the node  $\text{right}(u)$  by setting  $\text{right}(u) = \text{right}(\text{right}(u))$ . Finally, delete the node in  $L_0$  whose key is equal to  $x$ .
- At this moment, it may happen that some of the lists  $L_h, L_{h-1}, \dots$  only consist of dummy nodes. If this is the case, delete these lists, and update the height  $h$  and the root of the new skip list.

Implementation details of skip lists and algorithms  $\text{SEARCH}$ ,  $\text{INSERT}$ , and  $\text{DELETE}$  can be found in Pat Morin's free textbook *Open Data Structures*, which is available at <http://opendatastructures.org/>

### 6.11.3 Analysis of Skip Lists

In this subsection, we will prove that the expected size of a skip list is  $O(n)$  and the expected running time of algorithm  $\text{SEARCH}$  is  $O(\log n)$ . This will imply that the expected running times of algorithms  $\text{INSERT}$  and  $\text{DELETE}$  are  $O(\log n)$  as well. Throughout this subsection, we assume for simplicity that  $n$  is a power of 2, so that  $\log n$  is an integer.

Consider again the lists  $L_0, L_1, \dots, L_h$  in the skip list. For the purpose of analysis, we define, for each integer  $i > h$ ,  $L_i$  to be an empty list.

For each number  $x$  that is stored in the list  $L_0$ , we define the random variable  $h(x)$  to be the largest value of  $i$  such that  $x$  is contained in the list  $L_i$ . Thus,  $x$  occurs in the lists  $L_0, L_1, \dots, L_{h(x)}$ , but not in the list  $L_{h(x)+1}$ .

**Lemma 6.11.1** *For any number  $x$  that is stored in the list  $L_0$ ,*

$$\mathbb{E}(h(x)) = 1.$$

**Proof.** The value of  $h(x)$  is determined by the following process: flip a fair coin repeatedly and independently until it comes up tails for the first time. The value of  $h(x)$  is then equal to the number of flips minus one. For example, if we flip the coin three times (i.e., obtain the sequence *HHT*), then  $x$  is contained in the lists  $L_0, L_1$ , and  $L_2$ , but not in  $L_3$ ; thus,  $h(x) = 2$ . By Theorem 6.6.2, the expected number of coin flips is equal to two. As a result, the expected value of  $h(x)$  is equal to one. ■

**Lemma 6.11.2** *For any number  $x$  that is stored in the list  $L_0$  and for any  $i \geq 0$ ,*

$$\Pr(x \in L_i) = 1/2^i.$$

**Proof.** The claim follows from the fact that  $x$  is contained in the list  $L_i$  if and only if the first  $i$  coin flips for  $x$  all result in heads. ■

**Lemma 6.11.3** *Let  $i \geq 0$  and let  $|L_i|$  denote the number of nodes in the list  $L_i$ , ignoring the dummy node. Then,*

$$\mathbb{E}(|L_i|) = n/2^i.$$

**Proof.** We know from Lemma 6.11.2 that each number  $x$  in  $L_0$  is contained in  $L_i$  with probability  $1/2^i$ , independently of the other numbers in  $L_i$ . Therefore,  $|L_i|$  is a random variable that has a binomial distribution with parameters  $n$  and  $p = 1/2^i$ . The claim then follows from Theorem 6.7.2. ■

**Lemma 6.11.4** *Let  $X$  be the random variable whose value is equal to the total number of nodes in all lists  $L_0, L_1, L_2, \dots$ , ignoring the dummy nodes. Then,*

$$\mathbb{E}(X) = 2n.$$

**Proof.** We will give two proofs. In the first proof, we observe that

$$X = \sum_{i=0}^h |L_i|$$

and, thus,

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=0}^h |L_i|\right).$$

Observe that the number of terms in the summation on the right-hand side is equal to  $h + 1$ , which is a random variable. In general, the Linearity of Expectation does *not* apply to summations consisting of a *random* number of terms; see Exercise 6.64 for an example. Therefore, we proceed as follows.

Recall that, for the purpose of analysis, we have defined, for each integer  $i > h$ ,  $L_i$  to be an empty list. It follows that

$$X = \sum_{i=0}^{\infty} |L_i|.$$

Using the Linearity of Expectation (i.e., Theorem 6.5.3) and Lemmas 6.11.3 and 5.15.2, we get

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=0}^{\infty} |L_i|\right) \\ &= \sum_{i=0}^{\infty} \mathbb{E}(|L_i|) \\ &= \sum_{i=0}^{\infty} n/2^i \\ &= n \sum_{i=0}^{\infty} (1/2)^i \\ &= 2n.\end{aligned}$$

In the second proof, we use the fact that each number  $x$  in  $L_0$  occurs in exactly  $1 + h(x)$  lists, namely  $L_0, L_1, \dots, L_{h(x)}$ . Thus, we have

$$X = \sum_x (1 + h(x)).$$

Using the Linearity of Expectation (i.e., Theorem 6.5.2) and Lemma 6.11.1, we get

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}\left(\sum_x (1 + h(x))\right) \\ &= \sum_x \mathbb{E}(1 + h(x)) \\ &= \sum_x (1 + \mathbb{E}(h(x))) \\ &= \sum_x 2 \\ &= 2n.\end{aligned}$$

■

**Lemma 6.11.5** *Recall that  $h$  is the random variable whose value is equal to the height of the skip list. We have*

$$\mathbb{E}(h) \leq \log n + 1.$$

**Proof.** Since

$$h = \max_x h(x),$$

we have

$$\mathbb{E}(h) = \mathbb{E} \left( \max_x h(x) \right).$$

It is tempting, but wrong, to think that this is equal to

$$\max_x \mathbb{E}(h(x)),$$

which is equal to 1 by Lemma 6.11.1. (In Exercise 6.63, you will find a simple example showing that, in general, the expected value of a maximum is *not* equal to the maximum of the expected values.)

To prove a correct upper bound on  $\mathbb{E}(h)$ , we introduce, for each integer  $i \geq 1$ , an indicator random variable

$$X_i = \begin{cases} 1 & \text{if the list } L_i \text{ stores at least one number,} \\ 0 & \text{otherwise.} \end{cases}$$

We observe that

$$h = \sum_{i=1}^{\infty} X_i.$$

Since  $X_i$  is either 0 or 1, it is obvious that

$$\mathbb{E}(X_i) \leq 1. \tag{6.6}$$

We next claim that

$$X_i \leq |L_i|. \tag{6.7}$$

To justify this, if the list  $L_i$  does not store any number, then (6.7) becomes  $0 \leq 0$ , which is a true statement. On the other hand, if the list  $L_i$  stores

at least one number, then (6.7) becomes  $1 \leq |L_i|$ , which is again a true statement. Combining (6.7) with Lemmas 6.4.2 and 6.11.3, we obtain

$$\mathbb{E}(X_i) \leq \mathbb{E}(|L_i|) = n/2^i. \quad (6.8)$$

Using the Linearity of Expectation (i.e., Theorem 6.5.3), we get

$$\begin{aligned} \mathbb{E}(h) &= \mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) \\ &= \sum_{i=1}^{\infty} \mathbb{E}(X_i) \\ &= \sum_{i=1}^{\log n} \mathbb{E}(X_i) + \sum_{i=\log n+1}^{\infty} \mathbb{E}(X_i). \end{aligned}$$

If we apply (6.6) to the first summation and (6.8) to the second summation, we get

$$\begin{aligned} \mathbb{E}(h) &\leq \sum_{i=1}^{\log n} 1 + \sum_{i=\log n+1}^{\infty} \frac{n}{2^i} \\ &= \log n + \sum_{j=0}^{\infty} \frac{n}{2^{\log n+1+j}} \\ &= \log n + \sum_{j=0}^{\infty} \frac{n}{n \cdot 2^{1+j}} \\ &= \log n + \sum_{j=0}^{\infty} \frac{1}{2^{1+j}} \\ &= \log n + \frac{1}{2} \sum_{j=0}^{\infty} \frac{1}{2^j} \\ &= \log n + \frac{1}{2} \cdot 2 \\ &= \log n + 1. \end{aligned}$$

■

**Lemma 6.11.6** Let  $Y$  be the random variable whose value is equal to the total number of nodes in all lists  $L_0, L_1, L_2, \dots$ , including the dummy nodes. Then

$$\mathbb{E}(Y) \leq 2n + \log n + 2.$$

**Proof.** The total number of dummy nodes is equal to  $h + 1$ . Using the notation of Lemma 6.11.4, we have

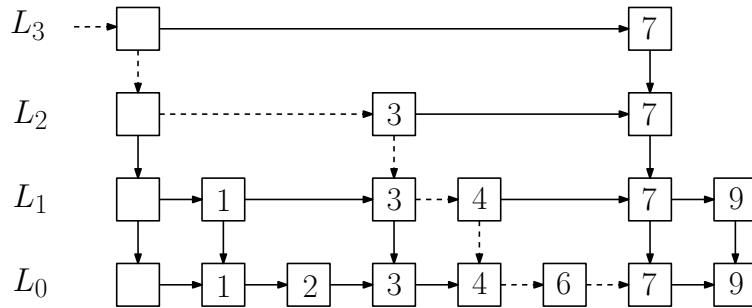
$$Y = X + h + 1.$$

Thus, using the Linearity of Expectation (i.e., Theorem 6.5.2) and Lemmas 6.11.4 and 6.11.5, we get

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(X + h + 1) \\ &= \mathbb{E}(X) + \mathbb{E}(h) + 1 \\ &\leq 2n + (\log n + 1) + 1 \\ &= 2n + \log n + 2. \end{aligned}$$

■

Consider any number  $x$ . As we have seen in Section 6.11.1, algorithm  $\text{SEARCH}(x)$  starts at the root of the skip list and follows a path to the right-most node  $u$  in the bottom list  $L_0$  for which  $\text{key}(u) \leq x$ . We will refer to this path as the *search path* of the algorithm. In the figure below, you see the same skip list as we have seen before. The dashed arrows indicate the search path of algorithm  $\text{SEARCH}(7)$ .



**Lemma 6.11.7** For any number  $x$ , let  $N$  be the random variable whose value is equal to the number of nodes on the search path of algorithm  $\text{SEARCH}(x)$ . Then,

$$\mathbb{E}(N) \leq 2 \log n + 5.$$

**Proof.** Consider the node  $u$  that is returned by algorithm  $\text{SEARCH}(x)$ , let  $v$  be the second last node on the search path, and let  $P$  be the part of this search path from the root to  $v$ . In the example above,  $u$  is the node in  $L_0$  whose key is 7,  $v$  is the node in  $L_0$  whose key is 6, and  $P$  is the part of the dashed path from the root to  $v$ .

Let  $M$  be the random variable whose value is equal to the number of nodes on  $P$ . Then,  $N = M + 1$  and

$$\mathbb{E}(N) = \mathbb{E}(M + 1) = \mathbb{E}(M) + 1.$$

Thus, it suffices to prove that

$$\mathbb{E}(M) \leq 2 \log n + 4.$$

Consider the following path  $P'$  in the skip list:

- $P'$  starts at node  $v$ .
- At any node on  $P'$ , the path  $P'$  moves up one level if this is possible, and moves one node to the left otherwise.

You should convince yourself that this path  $P'$  is the reverse of  $P$  and, therefore,  $M$  is equal to the number of nodes on  $P'$ . You should also convince yourself that this may not be true, if we take for  $P$  the path from the root to  $u$ .

For each  $i \geq 0$ , let  $M_i$  be the random variable whose value is equal to the number of nodes in the list  $L_i$  at which the path  $P'$  moves one node to the left. Then,  $M$  is the sum of

- $h$ : these are the nodes on  $P'$  at which  $P'$  moves up one level,
- 1: this accounts for the last node on  $P'$ , which is the root, and
- $\sum_{i=0}^h M_i$ .

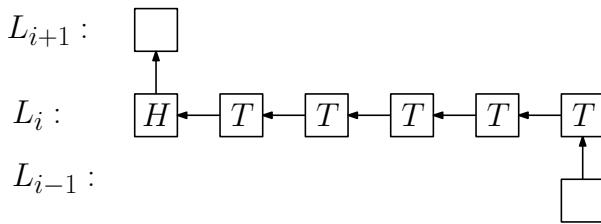
Thus,

$$\begin{aligned}\mathbb{E}(M) &= \mathbb{E} \left( h + 1 + \sum_{i=0}^h M_i \right) \\ &= \mathbb{E}(h) + 1 + \mathbb{E} \left( \sum_{i=0}^h M_i \right).\end{aligned}$$

As in the proof of Lemma 6.11.4, the number of terms in the latter summation is equal to  $h + 1$ , which is a random variable. Therefore, we cannot apply the Linearity of Expectation to this sum. As in the proof of Lemma 6.11.4, we proceed as follows. We first observe that

$$M = h + 1 + \sum_{i=0}^{\infty} M_i.$$

As the figure below indicates, the random variable  $M_i$  can be interpreted as being the number of tails obtained when flipping a fair coin until it comes up heads for the first time. Since (i) the list  $L_i$  may be empty (in which case  $M_i = 0$ ) or (ii) the portion of the path  $P'$  in  $L_i$  may terminate because it reaches the dummy node,  $M_i$  is in fact less than or equal to the number of tails.



Therefore, by Lemma 6.4.2 and Theorem 6.6.2,

$$\mathbb{E}(M_i) \leq 1. \quad (6.9)$$

Also, since  $M_i$  is less than or equal to the size  $|L_i|$  of the list  $L_i$  (ignoring the dummy node), we have, using Lemmas 6.4.2 and 6.11.3,

$$\mathbb{E}(M_i) \leq \mathbb{E}(|L_i|) = n/2^i. \quad (6.10)$$

Using the Linearity of Expectation (i.e., Theorem 6.5.3), we get

$$\begin{aligned}\mathbb{E}(M) &= \mathbb{E} \left( h + 1 + \sum_{i=0}^{\infty} M_i \right) \\ &= \mathbb{E}(h) + 1 + \sum_{i=0}^{\infty} \mathbb{E}(M_i) \\ &= \mathbb{E}(h) + 1 + \sum_{i=0}^{\log n} \mathbb{E}(M_i) + \sum_{i=\log n+1}^{\infty} \mathbb{E}(M_i).\end{aligned}$$

We know from Lemma 6.11.5 that  $\mathbb{E}(h) \leq \log n + 1$ . If we apply (6.9) to the first summation and (6.10) to the second summation, we get

$$\begin{aligned}\mathbb{E}(M) &\leq (\log n + 1) + 1 + \sum_{i=0}^{\log n} 1 + \sum_{i=\log n+1}^{\infty} n/2^i \\ &= 2\log n + 3 + \sum_{i=\log n+1}^{\infty} n/2^i.\end{aligned}$$

We have seen the infinite series in the proof of Lemma 6.11.5 and showed that it is equal to 1. Thus, we conclude that

$$\mathbb{E}(M) \leq 2\log n + 4.$$

■

## 6.12 Exercises

**6.1** Consider a fair coin that has 0 on one side and 1 on the other side. We flip this coin once and roll a fair die twice. Consider the following random variables:

$$\begin{aligned}X &= \text{the result of the coin,} \\ Y &= \text{the sum of the two dice,} \\ Z &= X \cdot Y.\end{aligned}$$

- Determine the distribution functions of  $X$ ,  $Y$ , and  $Z$ .
- Are  $X$  and  $Y$  independent random variables?
- Are  $X$  and  $Z$  independent random variables?
- Are  $Y$  and  $Z$  independent random variables?
- Are  $X$ ,  $Y$  and  $Z$  mutually independent random variables?

**6.2** Consider the set  $S = \{2, 3, 5, 30\}$ . We choose a uniformly random element  $x$  from this set. Consider the random variables

$$\begin{aligned} X &= \begin{cases} 1 & \text{if } x \text{ is divisible by 2,} \\ 0 & \text{otherwise,} \end{cases} \\ Y &= \begin{cases} 1 & \text{if } x \text{ is divisible by 3,} \\ 0 & \text{otherwise,} \end{cases} \\ Z &= \begin{cases} 1 & \text{if } x \text{ is divisible by 5,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- Is the sequence  $X, Y, Z$  of random variables pairwise independent?
- Is the sequence  $X, Y, Z$  of random variables mutually independent?

**6.3** Let  $a$  and  $b$  be real numbers. You flip a fair and independent coin three times. For  $i = 1, 2, 3$ , let

$$f_i = \begin{cases} a & \text{if the } i\text{-th coin flip results in heads,} \\ b & \text{if the } i\text{-th coin flip results in tails.} \end{cases}$$

Consider the random variables

$$\begin{aligned} X &= f_1 \cdot f_2, \\ Y &= f_2 \cdot f_3. \end{aligned}$$

- Assume that  $a = b$ . Are the random variables  $X$  and  $Y$  independent?
- Assume that  $a = 0$  and  $b \neq a$ . Are the random variables  $X$  and  $Y$  independent?
- Assume that  $a \neq 0$  and  $b = -a$ . Are the random variables  $X$  and  $Y$  independent?
- Assume that  $a \neq 0$ ,  $b \neq 0$ ,  $a \neq b$ , and  $b \neq -a$ . Are the random variables  $X$  and  $Y$  independent?

**6.4** Lindsay and Simon want to play a game in which the expected amount of money that each of them wins is equal to zero. After having chosen a number  $x$ , the game is played as follows: Lindsay rolls a fair die, independently, three times.

- If none of the three rolls results in 6, then Lindsay pays one dollar to Simon.
- If exactly one of the rolls results in 6, then Simon pays one dollar to Lindsay.
- If exactly two rolls result in 6, then Simon pays two dollars to Lindsay.
- If all three rolls result in 6, then Simon pays  $x$  dollars to Lindsay.

Determine the value of  $x$ .

**6.5** You are given a fair coin.

- You flip this coin twice; the two flips are independent. For each heads, you win 3 dollars, whereas for each tails, you lose 2 dollars. Consider the random variable

$$X = \text{the amount of money that you win.}$$

- Use the definition of expected value to determine  $\mathbb{E}(X)$ .
- Use the linearity of expectation to determine  $\mathbb{E}(X)$ .
- You flip this coin 99 times; these flips are mutually independent. For each heads, you win 3 dollars, whereas for each tails, you lose 2 dollars. Consider the random variable

$$Y = \text{the amount of money that you win.}$$

Determine the expected value  $\mathbb{E}(Y)$  of  $Y$ .

**6.6** Let  $r$  and  $b$  be positive integers and define  $\alpha = \frac{r}{r+b}$ . A bowl contains  $r$  red balls and  $b$  blue balls; thus,  $\alpha$  is the fraction of the balls that are red. Consider the following experiment:

- Choose one ball uniformly at random.
  - If the chosen ball is red, then put it back, together with an additional red ball.
  - If the chosen ball is blue, then put it back, together with an additional blue ball.

Define the random variable  $X$  to be the fraction of the balls that are red, after this experiment. Prove that  $\mathbb{E}(X) = \alpha$ .

**6.7** The Ontario Lottery and Gaming Corporation (OLG) offers the following lottery game:

- OLG chooses a winning number  $w$  in the set  $S = \{0, 1, 2, \dots, 999\}$ .
- If John wants to play, he pays \$1 and chooses a number  $x$  in  $S$ .
  - If  $x = w$ , then John receives \$700 from OLG. In this case, John wins \$699.
  - Otherwise,  $x \neq w$  and John does not receive anything. In this case, John loses \$1.

Assume that

- John plays this game once per day for one year (i.e., for 365 days),
- each day, OLG chooses a new winning number,
- each day, John chooses  $x$  uniformly at random from the set  $S$ , independently from previous choices.

Define the random variable  $X$  to be the total amount of dollars that John wins during one year. Determine the expected value  $\mathbb{E}(X)$ .

*Hint:* Use the Linearity of Expectation.

**6.8** Assume we flip a fair coin twice, independently of each other. Consider the following random variables:

$$\begin{aligned} X &= \text{the number of heads,} \\ Y &= \text{the number of tails,} \\ Z &= \text{the number of heads times the number of tails.} \end{aligned}$$

- Determine the expected values of these three random variables.
- Are  $X$  and  $Y$  independent random variables?
- Are  $X$  and  $Z$  independent random variables?
- Are  $Y$  and  $Z$  independent random variables?

**6.9** As of this writing<sup>2</sup>, Ma Long is the number 1 ranked ping pong player in the world. Simon Bose<sup>3</sup> also plays ping pong, but he is not at Ma's level yet. If you play a game of ping pong against Ma, then you win with probability  $p$ . If you play a game against Simon, you win with probability  $q$ . Here,  $p$  and  $q$  are real numbers such that  $0 < p < q < 1$ . (Of course,  $p$  is much smaller than  $q$ .) If you play several games against Ma and Simon, then the results are mutually independent.

You have the choice between the following two series of games:

1. *MSM*: First, play against Ma, then against Simon, then against Ma.
2. *SMS*: First, play against Simon, then against Ma, then against Simon.

For each  $s \in \{\text{MSM}, \text{SMS}\}$ , consider the event

$$A_s = \text{"you play series } s \text{ and beat Ma at least once and beat Simon at least once"}$$

and the random variable

$X_s$  = the number of games you win when playing series  $s$ .

- Determine  $\Pr(A_{\text{MSM}})$  and  $\Pr(A_{\text{SMS}})$ . Which of these two probabilities is larger?
- Determine  $\mathbb{E}(X_{\text{MSM}})$  and  $\mathbb{E}(X_{\text{SMS}})$ . Which of these two expected values is larger?

**6.10** In order to attract more customers, the Hyacintho Cactus Bar and Grill in downtown Ottawa organizes a game night, hosted by their star employee Tan Tran.

After paying \$26, a player gets two questions  $P$  and  $Q$ . If the player gives the correct answer to question  $P$ , this player wins \$30; if the player gives the correct answer to question  $Q$ , this player wins \$60. A player can choose between the following two options:

1. Start with question  $P$ . In this case, the player is allowed to answer question  $Q$  only if the answer to question  $P$  is correct.

---

<sup>2</sup>November 2016

<sup>3</sup>Jit's son

2. Start with question  $Q$ . In this case, the player is allowed to answer question  $P$  only if the answer to question  $Q$  is correct.

Elisa decides to play this game. The probability that Elisa correctly answers question  $P$  is equal to  $1/2$ , whereas she correctly answers question  $Q$  with probability  $1/3$ . The events of correctly answering are independent.

- Assume Elisa chooses the first option. Define the random variable  $X$  to be the amount of money that Elisa wins (this includes the \$26 that she has to pay in order to play the game). Determine the expected value  $\mathbb{E}(X)$ .
- Assume Elisa chooses the second option. Define the random variable  $Y$  to be the amount of money that Elisa wins (this includes the \$26 that she has to pay in order to play the game). Determine the expected value  $\mathbb{E}(Y)$ .

**6.11** Assume we roll two fair and independent dice, where one die is red and the other die is blue. Let  $(i, j)$  be the outcome, where  $i$  is the result of the red die and  $j$  is the result of the blue die. Consider the random variables

$$X = i + j$$

and

$$Y = i - j.$$

Are  $X$  and  $Y$  independent random variables?

**6.12** Assume we roll two fair and independent dice, where one die is red and the other die is blue. Let  $(i, j)$  be the outcome, where  $i$  is the result of the red die and  $j$  is the result of the blue die. Consider the random variables

$$X = |i - j|$$

and

$$Y = \max(i, j).$$

Are  $X$  and  $Y$  independent random variables?

**6.13** Consider the sample space  $S = \{1, 2, 3, \dots, 10\}$ . We choose a uniformly random element  $x$  in  $S$ . Consider the following random variables:

$$X = \begin{cases} 0 & \text{if } x \in \{1, 2\}, \\ 1 & \text{if } x \in \{3, 4, 5, 6\}, \\ 2 & \text{if } x \in \{7, 8, 9, 10\} \end{cases}$$

and

$$Y = \begin{cases} 0 & \text{if } x \text{ is even,} \\ 1 & \text{if } x \text{ is odd.} \end{cases}$$

Are  $X$  and  $Y$  independent random variables?

**6.14** Consider the 8-element set  $A = \{a, b, c, d, e, f, g, h\}$ . We choose a uniformly random 5-element subset  $B$  of  $A$ . Consider the following random variables:

$$\begin{aligned} X &= |B \cap \{a, b, c, d\}|, \\ Y &= |B \cap \{e, f, g, h\}|. \end{aligned}$$

- Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .
- Are  $X$  and  $Y$  independent random variables?

**6.15** You roll a fair die repeatedly and independently until the result is an even number. Consider the random variables

$$X = \text{the number of times you roll the die}$$

and

$$Y = \text{the result of the last roll.}$$

For example, if the results of the rolls are 5, 1, 3, 3, 5, 2, then  $X = 6$  and  $Y = 2$ .

Prove that the random variables  $X$  and  $Y$  are independent.

**6.16** Consider two random variables  $X$  and  $Y$ . If  $X$  and  $Y$  are independent, then it can be shown that

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

In this exercise, you will show that the converse of this statement is, in general, not true.

Let  $X$  be the random variable that takes each of the values  $-1, 0$ , and  $1$  with probability  $1/3$ . Let  $Y$  be the random variable with value  $Y = X^2$ .

- Prove that  $X$  and  $Y$  are not independent.
- Prove that  $\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$ .

**6.17** You are given two independent random variables  $X$  and  $Y$ , where

$$\Pr(X = 1) = \Pr(X = -1) = \Pr(Y = 1) = \Pr(Y = -1) = 1/2.$$

Consider the random variable

$$Z = X \cdot Y.$$

Are  $X$  and  $Z$  independent random variables?

**6.18** Jennifer loves to drink India Pale Ale (IPA), whereas Lindsay Bangs prefers wheat beer. Jennifer and Lindsay decide to go to their favorite pub *Chez Connor et Simon*. The beer menu shows that this pub has ten beers on tap:

- Five of these beers are of the IPA style.
- Three of these beers are of the wheat beer style.
- Two of these beers are of the pilsner style.

Jennifer and Lindsay order a uniformly random subset of seven beers (thus, there are no duplicates). Consider the following random variables:

$$\begin{aligned} J &= \text{the number of IPAs in this order,} \\ L &= \text{the number of wheat beers in this order.} \end{aligned}$$

- Determine the expected value  $\mathbb{E}(L)$  of the random variable  $L$ .
- Are  $J$  and  $L$  independent random variables?

**6.19** You roll a fair die five times, where all rolls are independent of each other. Consider the random variable

$$X = \text{the largest value in these five rolls.}$$

Prove that the expected value  $\mathbb{E}(X)$  of the random variable  $X$  is equal to

$$\mathbb{E}(X) = \frac{14077}{2592}.$$

*Hint:* What are the possible values for  $X$ ? What is  $\Pr(X = k)$ ?

**6.20** Consider the following algorithm, which takes as input a large integer  $n$  and returns a random subset  $A$  of the set  $\{1, 2, \dots, n\}$ :

**Algorithm** RANDOMSUBSET( $n$ ):

```
// all coin flips are mutually independent
A = ∅;
for i = 1 to n
do flip a fair coin;
    if the result of the coin flip is heads
    then A = A ∪ {i}
    endif
endfor;
return A
```

Define

$$\max(A) = \begin{cases} \text{the largest element in } A & \text{if } A \neq \emptyset, \\ 0 & \text{if } A = \emptyset, \end{cases}$$

$$\min(A) = \begin{cases} \text{the smallest element in } A & \text{if } A \neq \emptyset, \\ 0 & \text{if } A = \emptyset, \end{cases}$$

and the random variable

$$X = \max(A) - \min(A).$$

- Prove that the expected value  $\mathbb{E}(X)$  of the random variable  $X$  satisfies

$$\mathbb{E}(X) = n - 3 + f(n),$$

where  $f(n)$  is some function that converges to 0 when  $n \rightarrow \infty$ .

*Hint:* Introduce random variables  $Y = \min(A)$  and  $Z = \max(A)$  and compute their expected values. You may use

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

- Give an intuitive explanation why  $\mathbb{E}(X)$  is approximately equal to  $n-3$ .

**6.21** Let  $n \geq 1$  be an integer and let  $A[1 \dots n]$  be an array that stores a permutation of the set  $\{1, 2, \dots, n\}$ . If the array  $A$  is sorted, then  $A[k] = k$  for  $k = 1, 2, \dots, n$  and, thus,

$$\sum_{k=1}^n |A[k] - k| = 0. \quad (6.11)$$

If the array  $A$  is not sorted and  $A[k] = i$ , where  $i \neq k$ , then  $|A[k] - k|$  is equal to the “distance” between the position of the value  $i$  in  $A$  and the position of  $i$  in case the array were sorted. Thus, the summation in (6.11) is a measure for the “sortedness” of the array  $A$ : If the summation is small, then  $A$  is “close” to being sorted. On the other hand, if the summation is large, then  $A$  is “far away” from being sorted. In this exercise, you will determine the expected value of the summation in (6.11).

Assume that the array stores a uniformly random permutation of the set  $\{1, 2, \dots, n\}$ . For each  $k = 1, 2, \dots, n$ , consider the random variable

$$X_k = |A[k] - k|,$$

and let

$$X = \sum_{k=1}^n X_k.$$

- Assume that  $n = 1$ . Determine the expected value  $\mathbb{E}(X)$ .
- Assume that  $n \geq 2$ . Is the sequence  $X_1, X_2, \dots, X_n$  of random variables pairwise independent?
- Assume that  $n \geq 1$ . Let  $k$  be an integer with  $1 \leq k \leq n$ . Prove that

$$\mathbb{E}(X_k) = \frac{n+1}{2} + \frac{k^2 - k - kn}{n}.$$

**Hint:** Assume  $A[k] = i$ . If  $1 \leq i \leq k$ , then  $|A[k] - k| = k - i$ . If  $k+1 \leq i \leq n$ , then  $|A[k] - k| = i - k$ . For any integer  $m \geq 1$ ,

$$1 + 2 + 3 + \dots + m = \frac{m(m+1)}{2}.$$

- Assume that  $n \geq 1$ . Prove that

$$\mathbb{E}(X) = \frac{n^2 - 1}{3}.$$

**Hint:**

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

**6.22** Let  $n \geq 2$  be an integer. You are given  $n$  cider bottles  $C_1, C_2, \dots, C_n$  and two beer bottles  $B_1$  and  $B_2$ . Consider a uniformly random permutation of these  $n+2$  bottles. The positions in this permutation are numbered  $1, 2, \dots, n+2$ . Consider the following two random variables:

- |     |   |  |
|-----|---|--|
| $X$ | = | the position of the first cider bottle,          |
| $Y$ | = | the position of the first bottle having index 1. |

For example, if  $n = 5$  and the permutation is

$$B_2, C_5, C_2, C_4, B_1, C_3, C_1,$$

then  $X = 2$  and  $Y = 5$ .

- Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

- Determine the expected value  $\mathbb{E}(Y)$  of the random variable  $Y$ .

*Hint:*  $\sum_{k=1}^{n+1} k = (n+1)(n+2)/2$  and  $\sum_{k=1}^{n+1} k^2 = (n+1)(n+2)(2n+3)/6$ .

- Are  $X$  and  $Y$  independent random variables?

**6.23** Let  $m \geq 1$  and  $n \geq 1$  be integers. You are given  $m$  cider bottles  $C_1, C_2, \dots, C_m$  and  $n$  beer bottles  $B_1, B_2, \dots, B_n$ . Consider a uniformly random permutation of these  $m+n$  bottles. The positions in this permutation are numbered  $1, 2, \dots, m+n$ . Consider the random variable

$X$  = the position of the leftmost cider bottle.

- Determine the possible values for  $X$ .

- For any value  $k$  that  $X$  can take, prove that

$$\Pr(X = k) = \frac{m}{k} \cdot \frac{\binom{n}{k-1}}{\binom{m+n}{k}}.$$

*Hint:* Use the Product Rule to determine the number of permutations for which  $X = k$ . Rewrite your answer using basic properties of binomial coefficients.

- For each  $i = 1, 2, \dots, n$ , consider the indicator random variable

$$X_i = \begin{cases} 1 & \text{if } B_i \text{ is to the left of all cider bottles,} \\ 0 & \text{otherwise.} \end{cases}$$

Prove that

$$\mathbb{E}(X_i) = \frac{1}{m+1}.$$

- Express  $X$  in terms of  $X_1, X_2, \dots, X_n$ .
- Use the expression from the previous part to determine  $\mathbb{E}(X)$ .
- Prove that

$$\sum_{k=1}^{n+1} \frac{\binom{n}{k-1}}{\binom{m+n}{k}} = \frac{m+n+1}{m(m+1)}.$$

**6.24** Let  $b \geq 1$ ,  $c \geq 1$ , and  $w \geq 1$  be integers, and let  $n = b + c + w$ . You are given  $b$  beer bottles  $B_1, B_2, \dots, B_b$ ,  $c$  cider bottles  $C_1, C_2, \dots, C_c$ , and  $w$  wine bottles  $W_1, W_2, \dots, W_w$ . Let  $m \geq 1$  be an integer with  $m \leq b$  and  $m \leq n - b$ .

All  $n$  bottles are in a box. From this box, you choose a uniformly random subset consisting of  $m$  bottles. Consider the random variables

$$\begin{aligned} X &= \text{the number of beer bottles in the chosen subset,} \\ Y &= \text{the number of cider bottles in the chosen subset,} \\ Z &= \text{the number of wine bottles in the chosen subset.} \end{aligned}$$

- Determine the expected value  $\mathbb{E}(X + Y + Z)$ .
- Let  $k$  be an integer with  $0 \leq k \leq m$ . Prove that

$$\Pr(X = k) = \frac{\binom{b}{k} \binom{n-b}{m-k}}{\binom{n}{m}}.$$

- For each  $i = 1, 2, \dots, b$  and  $j = 1, 2, \dots, c$ , consider the indicator random variables

$$X_i = \begin{cases} 1 & \text{if } B_i \text{ is in the chosen subset,} \\ 0 & \text{otherwise.} \end{cases}$$

and

$$Y_j = \begin{cases} 1 & \text{if } C_j \text{ is in the chosen subset,} \\ 0 & \text{otherwise.} \end{cases}$$

Prove that

$$\mathbb{E}(X_i) = \mathbb{E}(Y_j) = \frac{m}{n}.$$

- Prove that

$$\sum_{k=0}^m k \binom{b}{k} \binom{n-b}{m-k} = \frac{bm}{n} \binom{n}{m}.$$

- Let  $i$  and  $j$  be integers with  $1 \leq i \leq b$  and  $1 \leq j \leq c$ . Are the random variables  $X_i$  and  $Y_j$  independent?
- Let  $i$  and  $j$  be integers with  $1 \leq i \leq b$  and  $1 \leq j \leq c$ . Determine  $\mathbb{E}(X_i \cdot Y_j)$ .
- Let  $i$  and  $j$  be integers with  $1 \leq i \leq b$  and  $1 \leq j \leq c$ . Is the following true or false?

$$\mathbb{E}(X_i \cdot Y_j) = \mathbb{E}(X_i) \cdot \mathbb{E}(Y_j).$$

**6.25** Let  $m \geq 1$ ,  $n \geq 1$ , and  $k \geq 1$  be integers with  $k \leq m + n$ . Consider a set  $P$  consisting of  $m$  men and  $n$  women. We choose a uniformly random  $k$ -element subset  $Q$  of  $P$ . Consider the random variables

$$\begin{aligned} X &= \text{the number of men in the chosen subset } Q, \\ Y &= \text{the number of women in the chosen subset } Q, \\ Z &= X - Y. \end{aligned}$$

- Prove that

$$\mathbb{E}(Z) = 2 \cdot \mathbb{E}(X) - k.$$

- Determine the expected value  $\mathbb{E}(X)$ .

**Hint:** Denote the men as  $M_1, M_2, \dots, M_m$ . Use indicator random variables.

- Prove that

$$\mathbb{E}(Z) = k \cdot \frac{m-n}{m+n}.$$

**6.26** You are given four fair and independent dice, each one having six faces:

1. One die is red and has the numbers 7, 7, 7, 7, 1, 1 on its faces.
2. One die is blue and has the numbers 5, 5, 5, 5, 5, 5 on its faces.
3. One die is green and has the numbers 9, 9, 3, 3, 3, 3 on its faces.
4. One die is yellow and has the numbers 8, 8, 8, 2, 2, 2 on its faces.

Let  $c$  be a color in the set {red, blue, green, yellow}. You roll the die of color  $c$ . Define the random variable  $X_c$  to be the result of this roll.

- For each  $c \in \{\text{red, blue, green, yellow}\}$ , determine the expected value  $\mathbb{E}(X_c)$  of the random variable  $X_c$ .
- Let  $c$  and  $c'$  be two distinct colors in the set {red, blue, green, yellow}. Determine

$$\Pr(X_c < X_{c'}) + \Pr(X_c > X_{c'}).$$

- Let  $c$  and  $c'$  be two distinct colors in the set {red, blue, green, yellow}. We say that the die of color  $c$  is *better* than the die of color  $c'$ , if

$$\Pr(X_c > X_{c'}) > 1/2.$$

- Is the red die better than the blue die?
- Is the blue die better than the green die?
- Is the green die better than the yellow die?
- Is the yellow die better than the red die?
- Explain why these dice are called *non-transitive dice*.

**6.27** In this exercise, you are given a fair and independent coin. Let  $n \geq 1$  be an integer. Farah flips the coin  $n$  times, whereas May flips the coin  $n+1$  times. Consider the following two random variables:

- |     |   |  |
|-----|---|--|
| $X$ | = | the number of heads in Farah's sequence of coin flips, |
| $Y$ | = | the number of heads in May's sequence of coin flips.   |

Let  $A$  be the event

$$A = "X < Y".$$

- Prove that

$$\Pr(A) = \frac{1}{2^{2n+1}} \sum_{k=0}^n \sum_{\ell=k+1}^{n+1} \binom{n}{k} \cdot \binom{n+1}{\ell}.$$

- Consider the following two random variables:

$$\begin{aligned} X' &= \text{the number of tails in Farah's sequence of coin flips,} \\ Y' &= \text{the number of tails in May's sequence of coin flips.} \end{aligned}$$

- What is  $X + X'$ ?
- What is  $Y + Y'$ ?
- Let  $B$  be the event

$$B = "X' < Y'".$$

Explain in plain English why

$$\Pr(A) = \Pr(B).$$

- Express the event  $B$  in terms of the event  $A$ .
- Use the results of the previous parts to determine  $\Pr(A)$ .

- Prove that

$$\sum_{k=0}^n \sum_{\ell=k+1}^{n+1} \binom{n}{k} \cdot \binom{n+1}{\ell} = 2^{2n}.$$

**6.28** Elisa Kazan's neighborhood pub serves three types of drinks: cider, wine, and beer. Elisa likes cider and wine, but does not like beer.

After a week of hard work, Elisa goes to this pub and repeatedly orders a random drink (the results of the orders are mutually independent). If she gets a glass of cider or a glass of wine, then she drinks it and places another order. As soon as she gets a pint of beer, she drinks it and takes a taxi home.

When Elisa orders one drink, she gets a glass of cider with probability  $2/5$ , a glass of wine with probability  $2/5$ , and a pint of beer with probability  $1/5$ .

Consider the random variables

$$\begin{aligned} X &= \text{the number of drinks that Elisa orders,} \\ Y &= \text{the number of different types that Elisa drinks.} \end{aligned}$$

If we denote cider by  $C$ , wine by  $W$ , and beer by  $B$ , then a possible sequence of drinks is  $CCWCB$ ; for this case  $X = 5$  and  $Y = 3$ . For the sequence  $WWWB$ , we have  $X = 4$  and  $Y = 2$ .

- Determine the expected value  $\mathbb{E}(X)$ .
- Describe the sample space in terms of strings consisting of characters  $C$ ,  $W$ , and  $B$ .
- Describe the event “ $Y = 1$ ” in terms of a subset of the sample space.
- Use the result of the previous part to determine  $\Pr(Y = 1)$ .
- Describe the event “ $Y = 2$ ” in terms of a subset of the sample space.
- Use the result of the previous part to determine  $\Pr(Y = 2)$ .
- Determine  $\Pr(Y = 3)$ .
- Use the results of the previous five parts to determine the expected value  $\mathbb{E}(Y)$ .
- Consider the random variable

$$Y_c = \begin{cases} 1 & \text{if Elisa drinks at least one glass of cider,} \\ 0 & \text{otherwise.} \end{cases}$$

Determine the expected value  $\mathbb{E}(Y_c)$ .

- Consider the random variable

$$Y_w = \begin{cases} 1 & \text{if Elisa drinks at least one glass of wine,} \\ 0 & \text{otherwise.} \end{cases}$$

Determine the expected value  $\mathbb{E}(Y_w)$ .

- Express  $Y$  in terms of  $Y_c$  and  $Y_w$ .
- Use the results of the previous three parts to determine the expected value  $\mathbb{E}(Y)$ .

**6.29** You repeatedly flip a fair coin and stop as soon as you get tails followed by heads. (All coin flips are mutually independent.) Consider the random variable

$$X = \text{the total number of coin flips.}$$

For example, if the sequence of coin flips is *HHHTTTTH*, then  $X = 8$ .

- Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use the linearity of expectation.

**6.30** In Section 6.6, we have shown that for  $-1 < x < 1$ ,

$$\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}.$$

In this exercise, you will prove this identity in a different way.

Consider the following infinite matrix:

$$\begin{pmatrix} x & 0 & 0 & 0 & 0 & 0 & \dots \\ x^2 & x^2 & 0 & 0 & 0 & 0 & \dots \\ x^3 & x^3 & x^3 & 0 & 0 & 0 & \dots \\ x^4 & x^4 & x^4 & x^4 & 0 & 0 & \dots \\ x^5 & x^5 & x^5 & x^5 & x^5 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

We are going to add all elements in this matrix in two different ways. A *row-sum* is the sum of all elements in one row, whereas a *column-sum* is the sum of all elements in one column.

Note that the sum of all row-sums is equal to

$$x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + \dots = \sum_{k=1}^{\infty} kx^k.$$

- Using only the identity  $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$  and algebraic manipulation, prove that the sum of all column-sums is equal to

$$\frac{x}{(1-x)^2}.$$

**6.31** Let  $X$  be a random variable that takes values in  $\{0, 1, 2, 3, \dots\}$ . By Lemma 6.4.3, we have

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \cdot \Pr(X = k).$$

As in Exercise 6.30, define an infinite matrix and use it to prove that

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \Pr(X \geq k).$$

**6.32** Let  $0 < p < 1$  and consider a coin that comes up heads with probability  $p$  and tails with probability  $1 - p$ . We flip the coin independently until it comes up heads for the first time. Define the random variable  $X$  to be the number of times that we flip the coin. In Section 6.6, we have shown that  $\mathbb{E}(X) = 1/p$ . Below, you will prove this in a different way.

- Let  $k \geq 1$  be an integer. Determine  $\Pr(X \geq k)$ .
- Using only the identity  $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ , the expression for  $\mathbb{E}(X)$  from Exercise 6.31, and your answer for  $\Pr(X \geq k)$ , prove that  $\mathbb{E}(X) = 1/p$ .

**6.33** By flipping a fair coin repeatedly and independently, we obtain a sequence of  $H$ 's and  $T$ 's. We stop flipping the coin as soon as the sequence contains either  $HH$  or  $TT$ . Define the random variable  $X$  to be the number of times that we flip the coin. For example, if the sequence of coin flips is  $HTHTT$ , then  $X = 5$ .

- Let  $k \geq 2$  be an integer. Determine  $\Pr(X = k)$ .
- Determine the expected value  $\mathbb{E}(X)$  of  $X$  using the expression

$$\mathbb{E}(X) = \sum_k k \cdot \Pr(X = k).$$

*Hint:* Recall that, for  $-1 < x < 1$ ,  $\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}$ .

- Determine  $\Pr(X \geq 1)$ .
- Let  $k \geq 2$  be an integer. Determine  $\Pr(X \geq k)$ .

- According to Exercise 6.31, we have

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \Pr(X \geq k).$$

Use this expression to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

**6.34** Consider an experiment that is successful with probability 0.8. We repeat this experiment (independently) until it is successful for the first time. The first 5 times we do the experiment, we have to pay \$10 per experiment. After this, we have to pay \$5 per experiment. Define the random variable  $X$  to be the total amount of money that we have to pay during all experiments. Determine the expected value  $\mathbb{E}(X)$ .

*Hint:* Recall that  $\sum_{k=1}^{\infty} kx^{k-1} = 1/(1-x)^2$ .

**6.35** When Lindsay and Simon have a child, this child is a boy with probability  $1/2$  and a girl with probability  $1/2$ , independently of the gender of previous children. Lindsay and Simon stop having children as soon as they have a girl. Consider the random variables

$B$  = the number of boys that Lindsay and Simon have

and

$G$  = the number of girls that Lindsay and Simon have.

Determine the expected values  $\mathbb{E}(B)$  and  $\mathbb{E}(G)$ .

**6.36** Let  $p$  be a real number with  $0 < p < 1$ . When Lindsay and Simon have a child, this child is a boy with probability  $p$  and a girl with probability  $1 - p$ , independently of the gender of previous children. Lindsay and Simon stop having children as soon as they have a child that has the same gender as their first child. Define the random variable  $X$  to be the number of children that Lindsay and Simon have. Determine the expected value  $\mathbb{E}(X)$ .

*Hint:* Recall that  $\sum_{k=1}^{\infty} kx^{k-1} = 1/(1-x)^2$ .

**6.37** Let  $X_1, X_2, \dots, X_n$  be a sequence of mutually independent random variables. For each  $i$  with  $1 \leq i \leq n$ , assume that

- the variable  $X_i$  is either equal to 0 or equal to  $n + 1$ , and

- $\mathbb{E}(X_i) = 1$ .

Determine

$$\Pr(X_1 + X_2 + \cdots + X_n \leq n).$$

**6.38** The Ottawa Senators and the Toronto Maple Leafs play a best-of-seven series: These two hockey teams play games against each other, and the first team to win four games wins the series. Assume that

- each game has a winner (thus, no game ends in a tie),
- in any game, the Sens have a probability of  $3/4$  of defeating the Leafs,
- the results of the games are mutually independent.

Determine the probability that seven games are played in this series.

**6.39** Let  $n \geq 1$  be an integer, let  $p$  be a real number with  $0 < p < 1$ , and let  $X$  be a random variable that has a binomial distribution with parameters  $n$  and  $p$ . In Section 6.7.1, we have seen that the expected value  $\mathbb{E}(X)$  of  $X$  satisfies

$$\mathbb{E}(X) = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k}. \quad (6.12)$$

Recall Newton's Binomial Theorem (i.e., Theorem 3.6.5):

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

- Use (6.12) to prove that  $\mathbb{E}(X) = pn$ , by taking the derivative, with respect to  $y$ , in Newton's Binomial Theorem.

**6.40** A *block* in a bitstring is a maximal consecutive substring of 1's. For example, the bitstring 100011110100111 has four blocks: 1, 11111, 1, and 111.

Let  $n \geq 1$  be an integer and consider a random bitstring of length  $n$  that is obtained by flipping a fair coin, independently,  $n$  times. Define the random variable  $X$  to be the number of blocks in this bitstring.

- Use Exercise 4.43 to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

- Use indicator random variables to determine the expected value  $\mathbb{E}(X)$  of  $X$ .

**6.41** Let  $n \geq 1$  be an integer and consider a uniformly random permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$ . Define the random variable  $X$  to be the number of indices  $i$  for which  $1 \leq i < n$  and  $a_i < a_{i+1}$ .

Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use indicator random variables.

**6.42** Let  $n \geq 2$  be an integer and let  $a_1, a_2, \dots, a_n$  be a permutation of the set  $\{1, 2, \dots, n\}$ . Define  $a_0 = 0$  and  $a_{n+1} = 0$ , and consider the sequence

$$a_0, a_1, a_2, a_3, \dots, a_n, a_{n+1}.$$

A position  $i$  with  $1 \leq i \leq n$  is called *awesome*, if  $a_i > a_{i-1}$  and  $a_i > a_{i+1}$ . In words,  $i$  is awesome if the value at position  $i$  is larger than both its neighboring values.

For example, if  $n = 6$  and the permutation is  $2, 5, 4, 3, 1, 6$ , we get the sequence

value	0	2	5	4	3	1	6	0
position	0	1	2	3	4	5	6	7

In this case, the positions 2 and 6 are awesome, whereas the positions 1, 3, 4, and 5 are not awesome.

Consider a uniformly random permutation of the set  $\{1, 2, \dots, n\}$  and define the random variable  $X$  to be the number of awesome positions. Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

*Hint:* Use indicator random variables.

**6.43** Let  $n \geq 1$  be an integer and consider a permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$ . We partition this permutation into *increasing subsequences*. For example, for  $n = 10$ , the permutation

$$3, 5, 8, 1, 2, 4, 10, 7, 6, 9$$

is partitioned into four increasing subsequences: (i) 3, 5, 8, (ii) 1, 2, 4, 10, (iii) 7, and (iv) 6, 9.

Let  $a_1, a_2, \dots, a_n$  be a uniformly random permutation of  $\{1, 2, \dots, n\}$ . Define the random variable  $X$  to be the number of increasing subsequences in the partition of this permutation. For the example above, we have  $X = 4$ . In this exercise, you will determine the expected value  $\mathbb{E}(X)$  of  $X$  in two different ways.

- For each  $i$  with  $1 \leq i \leq n$ , let

$$X_i = \begin{cases} 1 & \text{if an increasing subsequence starts at position } i, \\ 0 & \text{otherwise.} \end{cases}$$

For the example above, we have  $X_1 = 1$ ,  $X_2 = 0$ ,  $X_3 = 0$ , and  $X_8 = 1$ .

- Determine  $\mathbb{E}(X_1)$ .
- Let  $i$  be an integer with  $2 \leq i \leq n$ . Use the Product Rule to determine the number of permutations of  $\{1, 2, \dots, n\}$  for which  $X_i = 1$ .
- Use these indicator random variables to determine  $\mathbb{E}(X)$ .
- For each  $i$  with  $1 \leq i \leq n$ , let

$$Y_i = \begin{cases} 1 & \text{if the value } i \text{ is the leftmost element of an increasing} \\ & \text{subsequence,} \\ 0 & \text{otherwise.} \end{cases}$$

For the example above, we have  $Y_1 = 1$ ,  $Y_3 = 1$ ,  $Y_5 = 0$ , and  $Y_7 = 1$ .

- Determine  $\mathbb{E}(Y_1)$ .
- Let  $i$  be an integer with  $2 \leq i \leq n$ . Use the Product Rule to determine the number of permutations of  $\{1, 2, \dots, n\}$  for which  $Y_i = 1$ .
- Use these indicator random variables to determine  $\mathbb{E}(X)$ .

**6.44** Lindsay Bangs and Simon Pratt visit their favorite pub that has 10 different beers on tap. Both Lindsay and Simon order, independently of each other, a uniformly random subset of 5 beers.

- One of the beers available is *Leo's Early Breakfast IPA*. Determine the probability that this is one of the beers that Lindsay orders.
- Let  $X$  be the random variable whose value is the number of beers that are ordered by both Lindsay and Simon. Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use indicator random variables.

**6.45** Lindsay and Simon have discovered a new pub that has  $n$  different beers  $B_1, B_2, \dots, B_n$  on tap, where  $n \geq 1$  is an integer. They want to try all different beers in this pub and agree on the following approach: During a period of  $n$  days, they visit the pub every day. On each day, they drink one of the beers. Lindsay drinks the beers in order, i.e., on the  $i$ -th day, she drinks beer  $B_i$ . Simon takes a uniformly random permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$  and drinks beer  $B_{a_i}$  on the  $i$ -th day.

Let  $X$  be the random variable whose value is the number of days during which Lindsay and Simon drink the same beer. Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use indicator random variables.

**6.46** Consider the following recursive algorithm TwoTAILS, which takes as input a positive integer  $n$ :

**Algorithm** TwoTAILS( $n$ ):

```
// all coin flips are mutually independent
flip a fair coin twice;
if the coin came up tails exactly twice
then return 2n
else TwoTAILS( $n + 1$ )
endif
```

- You run algorithm TwoTAILS(1), i.e., with  $n = 1$ . Define the random variable  $X$  to be the value of the output of this algorithm. Let  $k \geq 1$  be an integer. Determine  $\Pr(X = 2^k)$ .
- Is the expected value  $\mathbb{E}(X)$  of the random variable  $X$  finite or infinite?

**6.47** Let  $A[1 \dots n]$  be an array of  $n$  numbers. Consider the following two algorithms, which take as input the array  $A$  and a number  $x$ . If  $x$  is not present in  $A$ , then these algorithms return the message “not present”. Otherwise, they return an index  $i$  such that  $A[i] = x$ . The first algorithm runs linear search from left to right, whereas the second algorithm runs linear search from right to left.

**Algorithm** LINEARSEARCHLEFTTORIGHT( $A, x$ ):

```
i := 1;
while i ≤ n and A[i] ≠ x do i := i + 1 endwhile;
if i = n + 1 then return "not present" else return i endif
```

**Algorithm** LINEARSEARCHRIGHTTOLEFT( $A, x$ ):

```
i := n;
while i ≥ 1 and A[i] ≠ x do i := i - 1 endwhile;
if i = 0 then return "not present" else return i endif
```

Consider the following algorithm, which again take as input the array  $A$  and a number  $x$ . If  $x$  is not present in  $A$ , then it returns the message “not present”. Otherwise, it returns an index  $i$  such that  $A[i] = x$ .

**Algorithm** RANDOMLINEARSEARCH( $A, x$ ):

```
flip a fair coin;
if the coin comes up heads
then LINEARSEARCHLEFTTORIGHT(A, x)
else LINEARSEARCHRIGHTTOLEFT(A, x)
endif
```

Assume that the number  $x$  occurs exactly once in the array  $A$  and let  $k$  be the index such that  $A[k] = x$ . Let  $X$  be the random variable whose value is the number of times the test “ $A[i] \neq x$ ” is made in algorithm RANDOMLINEARSEARCH( $A, x$ ). (In words,  $X$  is the number of comparisons made by algorithm RANDOMLINEARSEARCH( $A, x$ ).) Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

**6.48** Let  $n \geq 3$  be an integer and let  $p$  be a real number with  $0 < p < 1$ . Consider the set  $V = \{1, 2, \dots, n\}$ . We construct a graph  $G = (V, E)$  with vertex set  $V$ , whose edge set  $E$  is determined by the following random process: Each unordered pair  $\{i, j\}$  of vertices, where  $i \neq j$ , occurs as an edge in  $E$  with probability  $p$ , independently of the other unordered pairs.

A *triangle* in  $G$  is an unordered triple  $\{i, j, k\}$  of distinct vertices, such that  $\{i, j\}$ ,  $\{j, k\}$ , and  $\{k, i\}$  are edges in  $G$ .

Define the random variable  $X$  to be the total number of triangles in the graph  $G$ . Determine the expected value  $\mathbb{E}(X)$ .

*Hint:* Use indicator random variables.

- 6.49** In Section 6.9, we have seen the following algorithm `INSERTIONSORT`, which sorts any input array  $A[1 \dots n]$ :

```
Algorithm INSERTIONSORT( $A[1 \dots n]$ ):
  for  $i = 2$  to  $n$ 
    do  $j = i$ ;
      while  $j > 1$  and  $A[j] < A[j - 1]$ 
        do swap  $A[j]$  and  $A[j - 1]$ ;
         $j = j - 1$ 
      endwhile
    endfor
```

Consider an input array  $A[1 \dots n]$ , where each element  $A[i]$  is chosen independently and uniformly at random from the set  $\{1, 2, \dots, m\}$ .

- Let  $i$  and  $j$  be two indices with  $1 \leq i < j \leq n$ , and consider the values  $A[i]$  and  $A[j]$  (just before the algorithm starts). Prove that

$$\Pr(A[i] > A[j]) = \frac{1}{2} - \frac{1}{2m}.$$

- Let  $X$  be the random variable that is equal to the number of times the swap-operation is performed when running `INSERTIONSORT( $A[1 \dots n]$ )`. Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

- 6.50** Let  $n \geq 2$  be an integer. Consider the following random process that divides the integers  $1, 2, \dots, n$  into two sorted lists  $L_1$  and  $L_2$ :

1. Initialize both  $L_1$  and  $L_2$  to be empty.
2. For each  $i = 1, 2, \dots, n$ , flip a fair coin. If the coin comes up heads, then add  $i$  at the end of list  $L_1$ . Otherwise, add  $i$  at the end of the list  $L_2$ . (All coin flips during this process are mutually independent.)

We now run algorithm  $\text{MERGE}(L_1, L_2)$  of Section 4.6. Define the random variable  $X$  to be the total number of comparisons made when running this algorithm: As in Section 4.6,  $X$  counts the number of times the line “if  $x \leq y$ ” in algorithm  $\text{MERGE}(L_1, L_2)$  is executed. In this exercise, you will determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

- Prove that  $\mathbb{E}(X) = 1/2$  for the case when  $n = 2$ .
- Prove that  $\mathbb{E}(X) = 5/4$  for the case when  $n = 3$ .
- Assume that  $n \geq 2$ . For each  $i$  and  $j$  with  $1 \leq i < j \leq n$ , consider the indicator random variable

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are compared,} \\ 0 & \text{otherwise.} \end{cases}$$

Prove that  $\mathbb{E}(X_{ij}) = (1/2)^{j-i}$ .

*Hint:* Assume that  $i$  and  $j$  are compared. Can  $i$  and  $j$  be in the same list? What about the elements  $i, i+1, \dots, j-1$  and the element  $j$ ?

- Determine  $\mathbb{E}(X)$ .

$$\text{Hint: } 1 + x + x^2 + x^3 + \cdots + x^k = \frac{1-x^{k+1}}{1-x}.$$

**6.51** Assume we have  $n$  balls and  $m$  boxes. We throw the balls independently and uniformly at random in the boxes. Thus, for each  $k$  and  $i$  with  $1 \leq k \leq n$  and  $1 \leq i \leq m$ ,

$$\Pr(\text{the } k\text{-th ball falls in the } i\text{-th box}) = 1/m.$$

Consider the following three random variables:

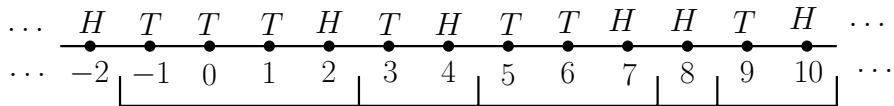
$$\begin{aligned} X &= \text{the number of boxes that do not contain any ball,} \\ Y &= \text{the number of boxes that contain at least one ball,} \\ Z &= \text{the number of boxes that contain exactly one ball.} \end{aligned}$$

- Determine the expected values  $\mathbb{E}(X)$ ,  $\mathbb{E}(Y)$ , and  $\mathbb{E}(Z)$ .
- Assuming that  $m = n$ , determine the limits
  1.  $\lim_{n \rightarrow \infty} \mathbb{E}(X)/n$ ,

2.  $\lim_{n \rightarrow \infty} \mathbb{E}(Y)/n,$
3.  $\lim_{n \rightarrow \infty} \mathbb{E}(Z)/n.$

*Hint:*  $\lim_{n \rightarrow \infty} (1 - 1/n)^n = 1/e.$

- 6.52** Let  $0 < p < 1$  and consider a coin that comes up heads with probability  $p$  and tails with probability  $1 - p$ . For each integer  $n$ , let  $b_n$  be the outcome when flipping this coin; thus,  $b_n \in \{H, T\}$ . The values  $b_n$  partition the set of integers into intervals, where each interval is a maximal consecutive sequence of zero or more  $T$ 's followed by one  $H$ :



- Consider the interval that contains the integer 0, and let  $X$  be its length. (In the example above,  $X = 4$ .) Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use the Linearity of Expectation. The answer is not  $1/p$ , which is the expected number of coin flips until the first  $H$ .

- 6.53** Your friend Mick takes a permutation of  $1, 2, \dots, n$ , stores it in boxes  $B_1, B_2, \dots, B_n$  (so that each box stores exactly one number), and then closes all boxes. You have no idea what the permutation is.

Mick opens the boxes  $B_1, B_2, \dots, B_n$ , one after another. For each  $i$  with  $1 \leq i \leq n$ , just before opening box  $B_i$ , you have to guess which number is stored in it.

- Assume that, when you guess the number in box  $B_i$ , you do not remember the numbers stored in  $B_1, B_2, \dots, B_{i-1}$ . Then, the only reasonable thing you can do is to take a random element in  $\{1, 2, \dots, n\}$  and guess that this random element is stored in  $B_i$ .

Assume that you do this for each  $i$  with  $1 \leq i \leq n$ . Let  $X$  be the random variable whose value is equal to the number of times that your guess is correct. Compute the expected value  $\mathbb{E}(X)$  of  $X$ .

- Now assume that your memory is perfect, so that, when you guess the number in box  $B_i$ , you know the numbers stored in  $B_1, B_2, \dots, B_{i-1}$ .

How would you make the  $n$  guesses such that the following is true: If  $Y$  is the random variable whose value is equal to the number of times that your guess is correct, then the expected value  $\mathbb{E}(Y)$  of  $Y$  satisfies  $\mathbb{E}(Y) = \Omega(\log n)$ .

**6.54** Let  $n \geq 1$  be an integer.

- Consider a fixed integer  $i$  with  $1 \leq i \leq n$ . How many permutations  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$  have the property that  $a_i = i$ ?
- We choose a permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$  uniformly at random. Consider the random variable

$$X = |\{i : 1 \leq i \leq n \text{ and } a_i = i\}|.$$

Determine the expected value  $\mathbb{E}(X)$ .

*Hint:* Use indicator random variables.

**6.55** Let  $n \geq 2$  be an integer. Consider a uniformly random permutation  $a_1, a_2, \dots, a_n$  of the set  $\{1, 2, \dots, n\}$ . Define the random variable  $X$  to be the number of ordered pairs  $(i, j)$  with  $1 \leq i < j \leq n$  for which  $a_i = j$  and  $a_j = i$ . Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use indicator random variables.

**6.56** Let  $n \geq 2$  be an integer and consider  $n$  people  $P_1, P_2, \dots, P_n$ . Each of these people has a uniformly random birthday, and all birthdays are mutually independent. (We ignore leap years.) Consider the random variable

$X =$  the number of indices  $i$  such that  $P_i$  and  $P_{i+1}$  have the same birthday.

Determine the expected value  $\mathbb{E}(X)$ .

**Hint:** Use indicator random variables.

**6.57** Let  $d \geq 1$  be the number of days in one year, let  $n \geq 2$  be an integer, and consider a group  $P_1, P_2, \dots, P_n$  of  $n$  people. Assume that each person has a uniformly random and independent birthday. Define the random variable  $X$  to be the number of pairs  $\{P_i, P_j\}$  of people that have the same birthday. Prove that

$$\mathbb{E}(X) = \frac{1}{d} \binom{n}{2}.$$

*Hint:* Use indicator random variables.

**6.58** Nick wants to know how many students cheat on the assignments. One approach is to ask every student “Did you cheat?”. This obviously does not work, because every student will answer “I did not cheat”. Instead, Nick uses the following ingenious scheme, which gives a reasonable estimate of the number of cheaters, without identifying them.

We denote the students by  $S_1, S_2, \dots, S_n$ . Let  $k$  denote the number of cheaters. Nick knows the value of  $n$ , but he does not know the value of  $k$ .

For each  $i$  with  $1 \leq i \leq n$ , Nick does the following:

1. Nick meets student  $S_i$  and asks “Did you cheat?”.
2. Student  $S_i$  flips a fair coin twice, independently of each other;  $S_i$  does not show the results of the coin flips to Nick.
  - (a) If the coin flips are  $HH$  or  $HT$ , then  $S_i$  is honest in answering the question: If  $S_i$  is a cheater, then he answers “I cheated”; otherwise, he answers “I did not cheat”.
  - (b) If the coin flips are  $TH$ , then  $S_i$  answers “I cheated”.
  - (c) If the coin flips are  $TT$ , then  $S_i$  answers “I did not cheat”.
- Define the random variable  $X$  to be the number of students who answer “I cheated”. Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* For each  $i$ , use an indicator random variable  $X_i$  which indicates whether or not  $S_i$  answers “I cheated”. If  $S_i$  is a cheater, what is  $\mathbb{E}(X_i)$ ? If  $S_i$  is not a cheater, what is  $\mathbb{E}(X_i)$ ?

- Consider the random variable

$$Y = 2X - n/2.$$

Prove that  $\mathbb{E}(Y) = k$ . In words, the expected value of  $Y$  is equal to the number of cheaters.

**6.59** You roll a fair die repeatedly, and independently, until you have seen all of the numbers 1, 2, 3, 4, 5, 6 at least once. Consider the random variable

$$X = \text{the number of times you roll the die.}$$

For example, if you roll the sequence

$$5, 5, 3, 5, 1, 3, 4, 2, 5, 2, 1, 3, 6,$$

then  $X = 13$ .

Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

*Hint:* Use the Linearity of Expectation. If you have seen exactly  $i$  different elements from the set  $\{1, 2, 3, 4, 5, 6\}$ , how many times do you expect to roll the die until you see a new element from this set?

**6.60** Michiel's Craft Beer Company (MCBC) sells  $n$  different brands of India Pale Ale (IPA). When you place an order, MCBC sends you one bottle of IPA, chosen uniformly at random from the  $n$  different brands, independently of previous orders.

Simon Pratt wants to try all different brands of IPA. He repeatedly places orders at MCBC (one bottle per order) until he has received at least one bottle of each brand.

Define the random variable  $X$  to be the total number of orders that Simon places. Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

*Hint:* Use the Linearity of Expectation. If Simon has received exactly  $i$  different brands of IPA, how many orders does he expect to place until he receives a new brand?

**6.61** MCBC still sells  $n$  different brands of IPA. As in Exercise 6.60, when you place an order, MCBC sends you one bottle of IPA, chosen uniformly at random from the  $n$  different brands, independently of previous orders.

Simon Pratt places  $m$  orders at MCBC. Define the random variable  $X$  to be the total number of distinct brands that Simon receives. Determine the expected value  $\mathbb{E}(X)$  of  $X$ .

*Hint:* Use indicator random variables.

**6.62** You are given an array  $A[0 \dots n-1]$  of  $n$  numbers. Let  $D$  be the number of *distinct* numbers that occur in this array. For each  $i$  with  $0 \leq i \leq n-1$ , let  $N_i$  be the number of elements in the array that are equal to  $A[i]$ .

- Show that  $D = \sum_{i=0}^{n-1} 1/N_i$ .

Consider the following algorithm:

**Algorithm ESTIMATED( $A[1 \dots n]$ ):**

**Step 1:** Choose an integer  $k$  in  $\{0, 1, 2, \dots, n - 1\}$  uniformly at random, and let  $a = A[k]$ .

**Step 2:** Traverse the array and compute the number  $N_k$  of times that  $a$  occurs.

**Step 3:** Return the value  $X = n/N_k$ .

- Determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

*Hint:* Use the definition of expected value, i.e., Definition 6.4.1.

**6.63** One of Jennifer and Thomas is chosen uniformly at random. The person who is chosen wins \$100. Consider the random variables

$$\begin{aligned} J &= \text{the amount that Jennifer wins,} \\ T &= \text{the amount that Thomas wins.} \end{aligned}$$

Prove that

$$\mathbb{E}(\max(J, T)) \neq \max(\mathbb{E}(J), \mathbb{E}(T)).$$

**6.64** Consider the sample space

$$S = \{(123), (132), (213), (231), (312), (321), (111), (222), (333)\}.$$

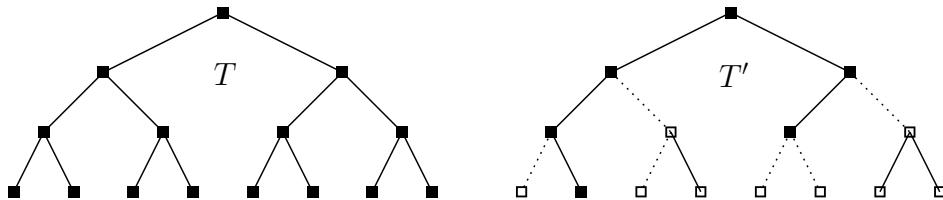
We choose an element  $u$  from  $S$  uniformly at random. For each  $i = 1, 2, 3$ , let  $X_i$  be the random variable whose value is the  $i$ -th number in  $u$ . (For example, if  $u = (312)$ , then  $X_1 = 3$ ,  $X_2 = 1$ , and  $X_3 = 2$ .) Let  $N$  be the random variable whose value is equal to that of  $X_2$ .

- Verify that  $\Pr(X_i = k) = 1/3$  for any  $i$  and  $k$  with  $1 \leq i \leq 3$  and  $1 \leq k \leq 3$ .
- Verify that  $X_1, X_2$  and  $X_3$  are pairwise independent.
- Verify that  $X_1, X_2$  and  $X_3$  are not mutually independent.
- Verify that  $\sum_{i=1}^{\mathbb{E}(N)} \mathbb{E}(X_i) = 4$ .
- Verify that  $\mathbb{E}\left(\sum_{i=1}^N X_i\right) \neq \sum_{i=1}^{\mathbb{E}(N)} \mathbb{E}(X_i)$ .

**6.65** Let  $k \geq 0$  be an integer and let  $T$  be a full binary tree, whose levels are numbered  $0, 1, 2, \dots, k$ . (The root is at level 0, whereas the leaves are at level  $k$ .) Assume that each edge of  $T$  is removed with probability  $1/2$ , independently of other edges. Denote the resulting graph by  $T'$ .

Define the random variable  $X$  to be the number of nodes that are connected to the root by a path in  $T'$ ; the root itself is included in  $X$ .

In the left figure below, the tree  $T$  is shown for the case when  $k = 3$ . The right figure shows the tree  $T'$ : The dotted edges are those that have been removed from  $T$ , the black nodes are connected to the root by a path in  $T'$ , whereas the white nodes are not connected to the root by a path in  $T'$ . For this case,  $X = 6$ .

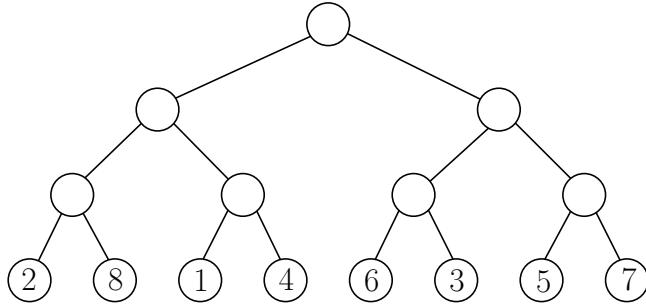


- Let  $n$  be the number of nodes in the tree  $T$ . Express  $n$  in terms of  $k$ .
- Prove that the expected value  $\mathbb{E}(X)$  of the random variable  $X$  is equal to

$$\mathbb{E}(X) = \log(n + 1).$$

*Hint:* For any  $\ell$  with  $0 \leq \ell \leq k$ , how many nodes of  $T$  are at level  $\ell$ ? Use indicator random variables to determine the expected number of level- $\ell$  nodes of  $T$  that are connected to the root by a path in  $T'$ .

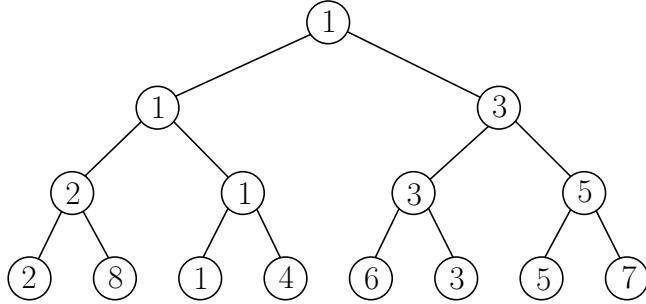
**6.66** Let  $n \geq 2$  be a power of two and consider a full binary tree with  $n$  leaves. Let  $a_1, a_2, \dots, a_n$  be a random permutation of the numbers  $1, 2, \dots, n$ . Store this permutation at the leaves of the tree, in the order  $a_1, a_2, \dots, a_n$ , from left to right. For example, if  $n = 8$  and the permutation is  $2, 8, 1, 4, 6, 3, 5, 7$ , then we obtain the following tree:



Perform the following process on the tree:

- Visit the levels of the tree from bottom to top.
- At each level, take all pairs of consecutive nodes that have the same parent. For each such pair, compare the numbers stored at the two nodes, and store the smaller of these two numbers at the common parent.

For our example tree, we obtain the following tree:



It is clear that at the end of this process, the root stores the number 1. Define the random variable  $X$  to be the number that is not equal to 1 and that is stored at a child of the root; think of  $X$  being the “loser of the final game”. For our example tree,  $X = 3$ .

In this exercise, you will determine the expected value  $\mathbb{E}(X)$  of the random variable  $X$ .

- Prove that  $2 \leq X \leq 1 + n/2$ .
- Prove that the following is true for each  $k$  with  $1 \leq k \leq n/2$ :  $X \geq k + 1$  if and only if

- all numbers  $1, 2, \dots, k$  are stored in the left subtree of the root
  - or all numbers  $1, 2, \dots, k$  are stored in the right subtree of the root.
- Prove that for each  $k$  with  $1 \leq k \leq n/2$ ,

$$\Pr(X \geq k+1) = 2 \cdot \frac{\binom{n/2}{k} k!(n-k)!}{n!} = 2 \cdot \frac{\binom{n/2}{k}}{\binom{n}{k}}.$$

- According to Exercise 6.31, we have

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \Pr(X \geq k).$$

Prove that

$$\mathbb{E}(X) = \Pr(X \geq 1) + \sum_{k=1}^{n/2} \Pr(X \geq k+1).$$

- Use Exercise 3.67 to prove that

$$\mathbb{E}(X) = 3 - \frac{4}{n+2}.$$

**6.67** If  $X$  is a random variable that can take any value in  $\{1, 2, 3, \dots\}$ , and  $A$  is an event, then the *conditional expected value*  $\mathbb{E}(X | A)$  is given by

$$\mathbb{E}(X | A) = \sum_{k=1}^{\infty} k \cdot \Pr(X = k | A).$$

In words,  $\mathbb{E}(X | A)$  is the expected value of  $X$ , when you are given that the event  $A$  occurs.

You roll a fair die repeatedly, and independently, until you see the number 6. Define the random variable  $X$  to be the number of times you roll the die (this includes the last roll, in which you see the number 6). It follows from Theorem 6.6.2 that  $\mathbb{E}(X) = 6$ . Let  $A$  be the event

$$A = \text{“the results of all rolls are even numbers”}.$$

Determine the conditional expected value  $\mathbb{E}(X | A)$ .

*Hint:*  $\mathbb{E}(X | A) \neq 3$ . Recall that  $\sum_{k=1}^{\infty} k \cdot x^{k-1} = 1/(1-x)^2$ .

**6.68** For any integer  $n \geq 0$  and any real number  $x$  with  $0 < x < 1$ , define the function

$$F_n(x) = \sum_{k=n}^{\infty} \binom{k}{n} x^k.$$

(Using the *ratio test* from calculus, it can be shown that this infinite series converges for any fixed integer  $n$ .)

- Determine a closed form expression for  $F_0(x)$ .
- Let  $n \geq 1$  be an integer and let  $x$  be a real number with  $0 < x < 1$ . Prove that

$$F_n(x) = \frac{x}{n} \cdot F_{n-1}(x) + \frac{x^2}{n} \cdot F'_{n-1}(x),$$

where  $F'_{n-1}$  denotes the derivative of  $F_{n-1}$ .

*Hint:* If  $k \geq n \geq 1$ , then  $\binom{k}{n} = \frac{k}{n} \binom{k-1}{n-1}$ .

- Prove that for any integer  $n \geq 0$  and any real number  $x$  with  $0 < x < 1$ ,

$$F_n(x) = \frac{x^n}{(1-x)^{n+1}},$$

and

$$F'_n(x) = \frac{x^n + n \cdot x^{n-1}}{(1-x)^{n+2}}.$$

- Let  $n \geq 0$  and  $m$  be integers with  $m \geq n+1$ . Prove that

$$\sum_{\ell=0}^{\min(n+1,m-n)} (-1)^\ell \binom{n+1}{\ell} \binom{m-\ell}{n} = 0.$$

*Hint:* You have shown above that

$$(1-x)^{n+1} \sum_{k=n}^{\infty} \binom{k}{n} x^k = (1-x)^{n+1} \cdot F_n(x) = x^n. \quad (6.13)$$

Use Newton's Binomial Theorem to expand  $(1-x)^{n+1}$ . Then consider the expansion of the left-hand side in (6.13). What is the coefficient of  $x^m$  in this expansion?

**6.69** Consider a fair red coin and a fair blue coin. We repeatedly flip both coins, and keep track of the number of times that the red coin comes up heads. As soon as the blue coin comes up tails, the process terminates.

A formal description of this process is given in the pseudocode below. The value of the variable  $i$  is equal to the number of iterations performed so far, the value of the variable  $h$  is equal to the number of times that the red coin came up heads so far, whereas the Boolean variable  $stop$  is used to decide when the while-loop terminates.

**Algorithm** RANDOMCOINFLIPS:

```
// both the red coin and the blue coin are fair
// all coin flips are mutually independent
i = 0;
h = 0;
stop = false;
while stop = false
  do i = i + 1;
    flip the red coin;
    if the result of the red coin is heads
      then h = h + 1
      endif;
    flip the blue coin;
    if the result of the blue coin is tails
      then stop = true
      endif
  endwhile;
  return i and h
```

Consider the random variables

- $X$  = the value of  $i$  that is returned by algorithm RANDOMCOINFLIPS,
- $Y$  = the value of  $h$  that is returned by algorithm RANDOMCOINFLIPS.

Assume that the value of the random variable  $Y$  is equal to some integer  $n \geq 0$ . In this exercise, you will determine the expected value of the random variable  $X$ .

Thus, we are interested in the *conditional expected value*  $\mathbb{E}(X | Y = n)$ , which is the expected value of  $X$  (i.e., the number of iterations of the while-

loop), when you are given that the event “ $Y = n$ ” (i.e., during the while-loop, the red coin comes up heads  $n$  times) occurs. Formally, we have

$$\mathbb{E}(X \mid Y = n) = \sum_k k \cdot \Pr(X = k \mid Y = n),$$

where the summation ranges over all values of  $k$  that  $X$  can take.

The functions  $F_n$  and  $F'_n$  that are used below are the same as those in Exercise 6.68.

- Let  $n \geq 1$  be an integer. Prove that

$$\Pr(Y = n) = \sum_{k=n}^{\infty} \Pr(Y = n \mid X = k) \cdot \Pr(X = k).$$

- Prove that

$$\Pr(Y = 0) = \sum_{k=1}^{\infty} \Pr(Y = 0 \mid X = k) \cdot \Pr(X = k).$$

- Let  $n \geq 1$  be an integer. Prove that

$$\Pr(Y = n) = F_n(1/4).$$

- Prove that

$$\Pr(Y = 0) = \frac{1}{3}.$$

- Let  $n \geq 1$  be an integer. Prove that

$$\mathbb{E}(X \mid Y = n) = \frac{F'_n(1/4)}{4 \cdot F_n(1/4)}.$$

- Let  $n \geq 1$  be an integer. Prove that

$$\mathbb{E}(X \mid Y = n) = \frac{4n + 1}{3}.$$

- Prove that

$$\mathbb{E}(X \mid Y = 0) = \frac{4}{3}.$$

**6.70** Let  $(S, \Pr)$  be a probability space, and let  $X$  and  $Y$  be two identical non-negative random variables on  $S$ . Thus, for all  $\omega$  in  $S$ ,  $X(\omega) = Y(\omega) \geq 0$ .

Consider the new probability space  $(S^2, \Pr)$ , where  $S^2$  is the Cartesian product  $S \times S$  and

$$\Pr(\omega_1, \omega_2) = \Pr(\omega_1) \cdot \Pr(\omega_2)$$

for all elements  $(\omega_1, \omega_2)$  in  $S^2$ . (In words, we choose two elements  $\omega_1$  and  $\omega_2$  in  $S$ , independently of each other.)

Consider the random variable  $Z$  on  $S^2$  defined by

$$Z(\omega_1, \omega_2) = \min((X(\omega_1))^2, (Y(\omega_2))^2)$$

for all  $(\omega_1, \omega_2)$  in  $S^2$ . Observe that the expected value of  $Z$  is equal to

$$\mathbb{E}(Z) = \sum_{\omega_1 \in S} \sum_{\omega_2 \in S} Z(\omega_1, \omega_2) \cdot \Pr(\omega_1, \omega_2).$$

- Let  $a$  and  $b$  be two non-negative real numbers. Prove that

$$\min(a^2, b^2) \leq ab.$$

- Prove that

$$\mathbb{E}(Z) \leq (\mathbb{E}(X))^2.$$

**6.71** Carleton University has implemented a new policy for students who cheat on assignments:

1. When a student is caught cheating, the student meets with the Dean.
2. The Dean has a box that contains  $n$  coins. One of these coins has the number  $n$  written on it, whereas each of the other  $n - 1$  coins has the number 1 written on it. Here,  $n$  is a very large integer.
3. The student chooses a uniformly random coin from the box.
4. If  $x$  is the number written on the chosen coin, then the student gives  $x^2$  bottles of cider to Elisa Kazan.

Consider the random variables

- $X$  = the number written on the chosen coin,
- $Z$  = the number of bottles of cider that Elisa gets.

(Note that  $Z = X^2$ .)

- Prove that

$$\mathbb{E}(X) = 2 - 1/n \leq 2.$$

- Prove that

$$\mathbb{E}(Z) = n + 1 - 1/n \geq n.$$

- Prove that

$$\mathbb{E}(X^2) \neq O((\mathbb{E}(X))^2).$$

- By the arguments above, Elisa gets, on average, a very large amount of cider. Since she cannot drink all these bottles, Carleton University changes their policy:

1. The student chooses a uniformly random coin from the box (and puts it back in the box).
2. Again, the student chooses a uniformly random coin from the box (and puts it back in the box).
3. If  $x$  is the number written on the first chosen coin, and  $y$  is the number written on the second chosen coin, then the student gives  $\min(x^2, y^2)$  bottles of cider to Elisa.

Consider the random variables

- $U$  = the number written on the first chosen coin,
- $V$  = the number written on the second chosen coin,
- $W$  = the number of bottles of cider that Elisa gets.

Use Exercise 6.70 to prove that

$$\mathbb{E}(W) \leq 4.$$



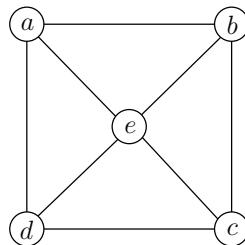
# Chapter 7

## The Probabilistic Method

The Probabilistic Method is a very powerful and surprising tool that uses probability theory to prove results in discrete mathematics. In this chapter, we will illustrate this method using several examples.

### 7.1 Large Bipartite Subgraphs

Recall that a *graph* is a pair  $G = (V, E)$ , where  $V$  is a finite set whose elements are called *vertices* and  $E$  is a set whose elements are unordered pairs of distinct vertices. The elements of  $E$  are called *edges*. Assume we partition the vertex set  $V$  of  $G$  into two subsets  $A$  and  $B$  (thus,  $A \cap B = \emptyset$  and  $A \cup B = V$ ). We say that an edge of  $E$  is *between*  $A$  and  $B$ , if one vertex of this edge is in  $A$  and the other vertex is in  $B$ .



For example, in the graph above, let  $A = \{a, d\}$  and  $B = \{b, c, e\}$ . Then four of the eight edges are between  $A$  and  $B$ , namely  $\{a, b\}$ ,  $\{a, e\}$ ,  $\{d, c\}$ , and  $\{d, e\}$ . Thus, the vertex set of this graph can be partitioned into two subsets  $A$  and  $B$ , such that at least half of  $G$ 's edges are between  $A$  and  $B$ . The following theorem states that this is true for any graph.

**Theorem 7.1.1** Let  $G = (V, E)$  be a graph with  $m$  edges. The vertex set  $V$  of  $G$  can be partitioned into two subsets  $A$  and  $B$  such that the number of edges between  $A$  and  $B$  is at least  $m/2$ .

**Proof.** Consider the following random process: Initialize  $A = \emptyset$  and  $B = \emptyset$ . For each vertex  $u$  of  $G$ , flip a fair and independent coin. If the coin comes up heads, add  $u$  to  $A$ ; otherwise, add  $u$  to  $B$ .

Define the random variable  $X$  to be the number of edges of  $G$  that are between  $A$  and  $B$ . We will determine the expected value  $\mathbb{E}(X)$  of  $X$ .

Number the edges of  $G$  arbitrarily as  $e_1, e_2, \dots, e_m$ . For each  $i$  with  $1 \leq i \leq m$ , consider the indicator random variable

$$X_i = \begin{cases} 1 & \text{if } e_i \text{ is an edge between } A \text{ and } B, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$X = \sum_{i=1}^m X_i$$

and

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^m X_i\right) \\ &= \sum_{i=1}^m \mathbb{E}(X_i) \\ &= \sum_{i=1}^m \Pr(X_i = 1). \end{aligned}$$

To determine  $\Pr(X_i = 1)$ , let  $e_i$  have vertices  $a$  and  $b$ . The following table shows the four possibilities for  $a$  and  $b$ ; each one of them occurs with probability  $1/4$ .

$a \in A, b \in A$	$X_i = 0$
$a \in A, b \in B$	$X_i = 1$
$a \in B, b \in A$	$X_i = 1$
$a \in B, b \in B$	$X_i = 0$

Since  $X_i = 1$  in two out of the four cases, we have

$$\Pr(X_i = 1) = 2/4 = 1/2,$$

and it follows that

$$\mathbb{E}(X) = \sum_{i=1}^m 1/2 = m/2.$$

Assume the claim in the theorem does not hold. Then, no matter how we partition the vertex set  $V$  into  $A$  and  $B$ , the number of edges between  $A$  and  $B$  will be less than  $m/2$ . In particular, the random variable  $X$  will always be less than  $m/2$ . But then,  $\mathbb{E}(X) < m/2$  as well, contradicting that  $\mathbb{E}(X) = m/2$ .  $\blacksquare$

## 7.2 Ramsey Theory

We return to a problem that we have seen in Section 1.1. Consider a complete graph with  $n$  vertices, in which each vertex represents a person. Any pair of distinct vertices is connected by an edge. Such an edge is *solid* if the two persons representing the vertices of this edge are friends. If these persons are strangers, the edge is *dashed*. Consider a subset  $S$  of  $k$  vertices. We say that  $S$  is a *solid  $k$ -clique*, if any two distinct vertices in  $S$  are connected by a solid edge. Thus, a solid  $k$ -clique represents a group of  $k$  mutual friends. If any two distinct vertices of  $S$  are connected by a dashed edge, then we say that  $S$  is a *dashed  $k$ -clique*; this represents a group of  $k$  mutual strangers.

In Section 1.1, we stated, without proof, Theorem 1.1.3. We repeat the statement of this theorem and use the Probabilistic Method to prove it.

**Theorem 7.2.1** *Let  $k \geq 3$  and  $n \geq 3$  be integers with  $n \leq \lfloor 2^{k/2} \rfloor$ . There exists a complete graph with  $n$  vertices, in which each edge is either solid or dashed, such that this graph does not contain a solid  $k$ -clique and does not contain a dashed  $k$ -clique.*

**Proof.** We denote the complete graph with  $n$  vertices by  $K_n$ . Consider the following random process: For each edge  $e$  of  $K_n$ , flip a fair and independent coin. If the coin comes up heads, make  $e$  a solid edge; otherwise, make  $e$  a dashed edge.

Consider the event

$$A = \text{“there is a solid } k\text{-clique or there is a dashed } k\text{-clique”}.$$

We will prove below that  $\Pr(A) < 1$ . This will imply that  $\Pr(\overline{A}) > 0$ , i.e., the event

$$\overline{A} = \text{“there is no solid } k\text{-clique and there is no dashed } k\text{-clique”}$$

has a positive probability. This, in turn, will imply that the statement in the theorem holds: If the statement would not hold, then  $\Pr(\overline{A})$  would be zero.

Thus, it remains to prove that  $\Pr(A) < 1$ . The vertex set of  $K_n$  has exactly  $\binom{n}{k}$  many subsets of size  $k$ . We denote these subsets by  $V_i$ ,  $i = 1, 2, \dots, \binom{n}{k}$ . For each  $i$  with  $1 \leq i \leq \binom{n}{k}$ , consider the event

$$A_i = \text{“}V_i \text{ is a solid } k\text{-clique or a dashed } k\text{-clique”}.$$

Since the event  $A_i$  occurs if and only if the edges joining the  $\binom{k}{2}$  pairs of vertices of  $V_i$  are either all solid or all dashed, we have

$$\Pr(A_i) = \frac{2}{2^{\binom{k}{2}}};$$

note that the denominator is equal to 2 to the power  $\binom{k}{2}$ .

Since  $A$  occurs if and only if  $A_1 \vee A_2 \vee \dots \vee A_{\binom{n}{k}}$  occurs, the Union Bound (i.e., Lemma 5.3.5) implies that

$$\begin{aligned} \Pr(A) &= \Pr\left(A_1 \vee A_2 \vee \dots \vee A_{\binom{n}{k}}\right) \\ &\leq \sum_{i=1}^{\binom{n}{k}} \Pr(A_i) \\ &= \sum_{i=1}^{\binom{n}{k}} \frac{2}{2^{\binom{k}{2}}} \\ &= \frac{2^{\binom{n}{k}}}{2^{\binom{k}{2}}}. \end{aligned}$$

If we can show that the quantity in the last line is less than one, then the proof is complete. We have

$$\begin{aligned} \frac{2\binom{n}{k}}{2\binom{k}{2}} &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \cdot \frac{2}{2^{(k^2-k)/2}} \\ &\leq \frac{n^k}{k!} \cdot \frac{2^{1+k/2}}{2^{k^2/2}}. \end{aligned}$$

Since  $n \leq \lfloor 2^{k/2} \rfloor \leq 2^{k/2}$ , we get

$$\begin{aligned} \frac{2\binom{n}{k}}{2\binom{k}{2}} &\leq \frac{(2^{k/2})^k}{k!} \cdot \frac{2^{1+k/2}}{2^{k^2/2}} \\ &= \frac{2^{1+k/2}}{k!}. \end{aligned}$$

By Exercise 2.8, we have  $k! > 2^{1+k/2}$  for  $k \geq 3$ . Thus, we conclude that

$$\frac{2\binom{n}{k}}{2\binom{k}{2}} < 1.$$

■

Take, for example,  $k = 20$  and  $n = 1024$ . Theorem 7.2.1 states that there exists a group of 1024 people that does not contain a subgroup of 20 mutual friends and does not contain a subgroup of 20 mutual strangers. In fact, the proof shows more: Consider a group of 1024 people such that any two are friends with probability 1/2, and strangers with probability 1/2. The above proof shows that  $\Pr(A)$ , i.e., the probability that there is a subgroup of 20 mutual friends or there is a subgroup of 20 mutual strangers, satisfies

$$\Pr(A) \leq \frac{2^{1+k/2}}{k!} = \frac{2^{11}}{20!}.$$

Therefore, with probability at least

$$1 - \frac{2^{11}}{20!} = 0.99999999999999158,$$

(there are 15 nines) this group does not contain a subgroup of 20 mutual friends and does not contain a subgroup of 20 mutual strangers.

### 7.3 Sperner's Theorem

In Section 1.2, we considered the following problem. Let  $S$  be a set of size  $n$  and consider a sequence  $S_1, S_2, \dots, S_m$  of  $m$  subsets of  $S$ , such that for all  $i$  and  $j$  with  $i \neq j$ ,

$$S_i \not\subseteq S_j \text{ and } S_j \not\subseteq S_i. \quad (7.1)$$

What is the largest possible value of  $m$  for which such a sequence exists?

The sequence consisting of all subsets of  $S$  having size  $\lfloor n/2 \rfloor$  satisfies (7.1). This sequence has length  $m = \binom{n}{\lfloor n/2 \rfloor}$ . In Section 1.2, we stated, without proof, that this is the largest possible value of  $m$ ; see Theorem 1.2.1. After stating this theorem again, we will prove it using the Probabilistic Method.

**Theorem 7.3.1 (Sperner)** *Let  $n \geq 1$  be an integer and let  $S$  be a set with  $n$  elements. Let  $S_1, S_2, \dots, S_m$  be a sequence of  $m$  subsets of  $S$ , such that for all  $i$  and  $j$  with  $i \neq j$ ,*

$$S_i \not\subseteq S_j \text{ and } S_j \not\subseteq S_i.$$

*Then*

$$m \leq \binom{n}{\lfloor n/2 \rfloor}.$$

**Proof.** We assume that none of the subsets in the sequence  $S_1, S_2, \dots, S_m$  is empty, because otherwise,  $m$  must be equal to 1, in which case the theorem clearly holds.

We assume that  $S = \{1, 2, \dots, n\}$ . We choose a uniformly random permutation  $a_1, a_2, \dots, a_n$  of the elements of  $S$ ; thus, each permutation has probability  $1/n!$  of being chosen. Consider the following sequence of subsets  $A_1, A_2, \dots, A_n$  of  $S$ : For  $j = 1, 2, \dots, n$ ,

$$A_j = \{a_1, a_2, \dots, a_j\}.$$

For example, if  $n = 4$  and the permutation is 3, 1, 4, 2, then

$$\begin{aligned} A_1 &= \{3\}, \\ A_2 &= \{1, 3\}, \\ A_3 &= \{1, 3, 4\}, \\ A_4 &= \{1, 2, 3, 4\}. \end{aligned}$$

Observe that the subsets  $A_1, A_2, \dots, A_n$  are random subsets of  $S$ , because the permutation was randomly chosen.

Consider a subset  $S_i$  in the statement of the theorem. We say that  $S_i$  occurs in the sequence  $A_1, A_2, \dots, A_n$  if there is an index  $j$  such that  $S_i = A_j$ .

Define the random variable  $X$  to be the number of subsets in the sequence  $S_1, S_2, \dots, S_m$  that occur in  $A_1, A_2, \dots, A_n$ . Since the subsets  $A_1, A_2, \dots, A_n$  are properly nested, i.e.,

$$A_1 \subset A_2 \subset \cdots \subset A_n,$$

the assumption in the theorem implies that  $X$  is either 0 or 1. It follows that the expected value of  $X$  satisfies

$$\mathbb{E}(X) \leq 1.$$

We now derive an exact expression for the value of  $\mathbb{E}(X)$ . For each  $i$  with  $1 \leq i \leq m$ , consider the indicator random variable

$$X_i = \begin{cases} 1 & \text{if } S_i \text{ occurs in the sequence } A_1, A_2, \dots, A_n, \\ 0 & \text{otherwise.} \end{cases}$$

Let  $k$  denote the size of the subset  $S_i$ , i.e.,  $k = |S_i|$ . Then

$$X_i = 1 \text{ if and only if } S_i = A_k.$$

Since  $A_k = \{a_1, a_2, \dots, a_k\}$ ,  $X_i = 1$  if and only if the first  $k$  values in the permutation form a permutation of the subset  $S_i$ :

$a_1, a_2, \dots, a_k$	$a_{k+1}, a_{k+2}, \dots, a_n$

permutation of  $S_i$

The Product Rule of Section 3.1 shows that there are  $k!(n - k)!$  many permutations of  $S$  that have this property. Therefore, since we chose a random permutation of  $S$ , we have

$$\begin{aligned} \mathbb{E}(X_i) &= \Pr(X_i = 1) \\ &= \frac{k!(n - k)!}{n!} \\ &= \frac{1}{\binom{n}{k}} \\ &= \frac{1}{\binom{n}{|S_i|}}. \end{aligned}$$

Thus, since

$$X = \sum_{i=1}^m X_i,$$

we get

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^m X_i\right) \\ &= \sum_{i=1}^m \mathbb{E}(X_i) \\ &= \sum_{i=1}^m \frac{1}{\binom{n}{|S_i|}}.\end{aligned}$$

If we combine this with our upper bound  $\mathbb{E}(X) \leq 1$ , we get

$$\sum_{i=1}^m \frac{1}{\binom{n}{|S_i|}} \leq 1.$$

For a fixed value of  $n$ , the binomial coefficient  $\binom{n}{k}$  is maximized when  $k = \lfloor n/2 \rfloor$ ; i.e., the largest value in the  $n$ -th row of Pascal's Triangle (see Section 3.8) is in the middle. Thus,

$$\binom{n}{|S_i|} \leq \binom{n}{\lfloor n/2 \rfloor},$$

implying that

$$\begin{aligned}1 &\geq \sum_{i=1}^m \frac{1}{\binom{n}{|S_i|}} \\ &\geq \sum_{i=1}^m \frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \\ &= \frac{m}{\binom{n}{\lfloor n/2 \rfloor}}.\end{aligned}$$

We conclude that

$$m \leq \binom{n}{\lfloor n/2 \rfloor}.$$

■

## 7.4 The Jaccard Distance between Finite Sets

Let  $X$  and  $Y$  be two finite and non-empty sets. We want to define a measure that indicates how “close together” these two sets are. This measure should be equal to 0 if the two sets are the same (i.e.,  $X = Y$ ), it should be equal to 1 if the two sets are disjoint (i.e.,  $X \cap Y = \emptyset$ ), and it should be in the open interval  $(0, 1)$  in all other cases.

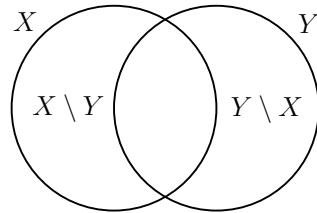
The *symmetric difference*  $X \ominus Y$  is defined to be the “union minus the intersection”, i.e.,

$$X \ominus Y = (X \cup Y) \setminus (X \cap Y).$$

From the Venn diagram below, it should be clear that

$$X \ominus Y = (X \setminus Y) \cup (Y \setminus X),$$

i.e., the set consisting of all elements in  $X$  that are not in  $Y$  and all elements in  $Y$  that are not in  $X$ .



If the symmetric difference  $X \ominus Y$  is “small” compared to the union  $X \cup Y$ , then the two sets  $X$  and  $Y$  are “pretty much the same”. On the other hand, if  $X \ominus Y$  is “large” compared to  $X \cup Y$ , then the sets  $X$  and  $Y$  are “very different”.

Based on this, the *Jaccard distance*  $d_J(X, Y)$  between the two finite and non-empty sets  $X$  and  $Y$  is defined as

$$d_J(X, Y) = \frac{|X \ominus Y|}{|X \cup Y|}. \quad (7.2)$$

Since

$$|X \ominus Y| = |X \cup Y| - |X \cap Y|,$$

we have

$$d_J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}. \quad (7.3)$$

The following claims are easy to verify:

- $0 \leq d_J(X, Y) \leq 1$ .
- $d_J(X, Y) = d_J(Y, X)$ .
- $d_J(X, X) = 0$ .
- If  $X \cap Y = \emptyset$ , then  $d_J(X, Y) = 1$ .
- If  $X \neq Y$  and  $X \cap Y \neq \emptyset$ , then  $0 < d_J(X, Y) < 1$ .

In the rest of this section, we will prove that the Jaccard distance satisfies the *triangle inequality*:

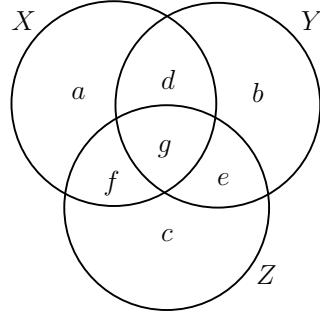
**Theorem 7.4.1** *Let  $X$ ,  $Y$ , and  $Z$  be finite and non-empty sets. Then*

$$d_J(X, Z) \leq d_J(X, Y) + d_J(Y, Z).$$

We will present two proofs of this result. The first proof uses “brute force”: We consider the Venn diagram for the sets  $X$ ,  $Y$ , and  $Z$ . Based on this diagram, we transform the inequality in Theorem 7.4.1 into an equivalent algebraic inequality. We then argue that the algebraic inequality is valid. In the second proof, we show that the inequality in Theorem 7.4.1 can be rephrased as an inequality involving probabilities. The result then follows by straightforward applications of Lemma 5.3.6 and the Union Bound (Lemma 5.3.5).

### 7.4.1 The First Proof

In the figure below, you see the Venn diagram for the three sets  $X$ ,  $Y$ , and  $Z$ . The variables  $a, b, \dots, g$  denote the number of elements in the different parts of this diagram. For example,  $d$  denotes the number of elements that are in  $X$  and in  $Y$ , but not in  $Z$ , whereas  $g$  denotes the number of elements that are in all three sets. Note that some of these variables may be equal to 0. However, since none of the three sets  $X$ ,  $Y$ , and  $Z$  is empty, we have  $a + d + f + g > 0$ ,  $b + d + e + g > 0$ , and  $c + e + f + g > 0$ .



Using the definition of Jaccard distance in (7.2), the inequality in Theorem 7.4.1 is equivalent to

$$\frac{a+c+d+e}{a+c+d+e+f+g} \leq \frac{a+b+e+f}{a+b+d+e+f+g} + \frac{b+c+d+f}{b+c+d+e+f+g},$$

which we rewrite as

$$\frac{a+b+e+f}{a+b+d+e+f+g} + \frac{b+c+d+f}{b+c+d+e+f+g} - \frac{a+c+d+e}{a+c+d+e+f+g} \geq 0.$$

After combining the three fractions into one fraction, and expanding the three products in the numerator of the resulting fraction, we get<sup>1</sup>

$$\frac{N}{D} \geq 0,$$

where

$$\begin{aligned} N = & a^2b + ab^2 + a^2c + 2abc + b^2c + ac^2 + bc^2 + a^2d + 2abd + b^2d + \\ & 2acd + 2bcd + ad^2 + bd^2 + 2abe + b^2e + 2ace + 2bce + c^2e + \\ & ade + 2bde + cde + be^2 + ce^2 + a^2f + 4abf + 2b^2f + 4acf + \\ & 4bcf + c^2f + 4adf + 5bdf + 3cdf + 2d^2f + 3ae f + 5bef + 4cef + \\ & 4def + 2e^2f + 3af^2 + 4bf^2 + 3cf^2 + 4df^2 + 4ef^2 + 2f^3 + 2abg + \\ & 2b^2g + 2acg + 2bcg + adg + 3bdg + 3beg + ceg + 3afg + 6bfg + \\ & 3cfg + 4dfg + 4efg + 4f^2g + 2bg^2 + 2fg^2 \end{aligned}$$

and

$$D = (a+b+d+e+f+g)(b+c+d+e+f+g)(a+c+d+e+f+g).$$

Observe that  $D > 0$ . Moreover, all terms in the equation for  $N$  are non-negative and they are connected by plus signs. It follows that  $N \geq 0$  and, therefore,  $N/D \geq 0$ . Thus, we have proved Theorem 7.4.1.

---

<sup>1</sup>with some help from Wolfram Alpha

### 7.4.2 The Second Proof

Consider the set  $X \cup Y \cup Z$ . Note that this is a set, so that there are no duplicates. Let  $n = |X \cup Y \cup Z|$  and consider a uniformly random permutation

$$x_1, x_2, x_3, \dots, x_n$$

of the elements of  $X \cup Y \cup Z$ . Consider the random variables

$$\begin{aligned} i &= \min\{\ell : x_\ell \in X\}, \\ j &= \min\{\ell : x_\ell \in Y\}, \\ k &= \min\{\ell : x_\ell \in Z\}. \end{aligned}$$

In words,  $i$  is determined by walking along the sequence  $x_1, x_2, x_3, \dots, x_n$ , from left to right. The value of  $i$  is the index of the first element that belongs to the set  $X$ .

Consider the event

$$A_{XY} = "i \neq j".$$

We are going to determine the probability  $\Pr(A_{XY})$  that this event occurs. Observe that

$$\Pr(A_{XY}) = 1 - \Pr(\overline{A}_{XY}),$$

where  $\overline{A}_{XY}$  is the event

$$\overline{A}_{XY} = "i = j".$$

To determine  $\Pr(\overline{A}_{XY})$ , we do the following. Remove from the sequence  $x_1, x_2, \dots, x_n$  all elements that do not belong to  $X$  and do not belong to  $Y$ . Then we are left with a uniformly random permutation of the set  $X \cup Y$ . The event  $\overline{A}_{XY}$  occurs if and only if the first element of this new sequence belongs to both  $X$  and  $Y$ . Since each element of  $X \cup Y$  has the same probability of being the first element in this new sequence, it follows that

$$\Pr(\overline{A}_{XY}) = \frac{|X \cap Y|}{|X \cup Y|}$$

and, thus, using (7.3),

$$\Pr(A_{XY}) = 1 - \frac{|X \cap Y|}{|X \cup Y|} = d_J(X, Y).$$

If we consider the events

$$A_{XZ} = \text{“}i \neq k\text{”}$$

and

$$A_{YZ} = \text{“}j \neq k\text{”},$$

then we have, by the same arguments,

$$\Pr(A_{XZ}) = d_J(X, Z)$$

and

$$\Pr(A_{YZ}) = d_J(Y, Z).$$

Thus, the inequality in Theorem 7.4.1 is equivalent to

$$\Pr(A_{XZ}) \leq \Pr(A_{XY}) + \Pr(A_{YZ}).$$

Since

$$i \neq k \Rightarrow i \neq j \vee j \neq k,$$

Lemma 5.3.6 implies that

$$\Pr(A_{XZ}) \leq \Pr(A_{XY} \vee A_{YZ}).$$

By applying the Union Bound (Lemma 5.3.5), we conclude that

$$\Pr(A_{XZ}) \leq \Pr(A_{XY}) + \Pr(A_{YZ}).$$

Thus, we have completed our second proof of Theorem 7.4.1.

## 7.5 Planar Graphs and the Crossing Lemma

Consider a graph  $G = (V, E)$ . Any one-to-one function  $f : V \rightarrow \mathbb{R}^2$  gives an *embedding* of  $G$ :

1. Each vertex  $a$  of  $V$  is drawn as the point  $f(a)$  in the plane.
2. Each edge  $\{a, b\}$  of  $E$  is drawn as the straight-line segment  $f(a)f(b)$  between the points  $f(a)$  and  $f(b)$ .

Besides the function  $f$  being one-to-one, we assume that it satisfies the following three properties:

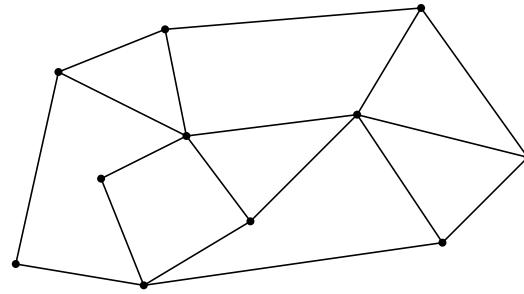
1. For any two edges  $\{a, b\}$  and  $\{a', b'\}$  of  $E$ , the intersection of the line segments  $f(a)f(b)$  and  $f(a')f(b')$  is empty or consists of exactly one point.
2. For any edge  $\{a, b\}$  in  $E$  and any vertex  $c$  in  $V$ , the point  $f(c)$  is not in the interior of the line segment  $f(a)f(b)$ .
3. For any three edges  $\{a, b\}$ ,  $\{a', b'\}$ , and  $\{a'', b''\}$  of  $E$ , the line segments  $f(a)f(b)$ ,  $f(a')f(b')$ , and  $f(a'')f(b'')$  do not have a point in common that is in the interior of any of these line segments.

For simplicity, we do not distinguish any more between a graph and its embedding. That is, a vertex  $a$  refers to both an element of  $V$  and the point in the plane that represents  $a$ . Similarly, an edge refers to both an element of  $E$  and the line segment that represents it.

### 7.5.1 Planar Graphs

**Definition 7.5.1** An embedding of a graph  $G = (V, E)$  is called *plane*, if no two edges of  $E$  intersect, except possibly at their endpoints. A graph  $G$  is called *planar* if there is a plane embedding of  $G$ .

Consider a plane embedding of a planar graph  $G$ . Again for simplicity, we denote this embedding by  $G$ . This embedding consists of vertices, edges, and faces (one of them being the unbounded face). For example, in the following plane embedding, there are 11 vertices, 18 edges, and 9 faces.



In the rest of this section, we will use the following notation:

- $G$  denotes a plane embedding of a planar graph.
- $v$  denotes the number of vertices of  $G$ .

- $e$  denotes the number of edges of  $G$ .
- $f$  denotes the number of faces in the embedding of  $G$ .

How many edges can  $G$  have? Since  $G$  has  $v$  vertices, we obviously have  $e \leq \binom{v}{2} = \Theta(v^2)$ , an upper bound which holds for any graph with  $v$  vertices. Since our graph  $G$  is planar, we expect a much smaller upper bound on  $e$ : If  $G$  has  $\Theta(v^2)$  edges, then it seems to be impossible to draw  $G$  without edge crossings. Below, we will prove that  $e$  is, in fact, at most linear in  $v$ . The proof will use Euler's Theorem for planar graphs:

**Theorem 7.5.2 (Euler)** *Consider any plane embedding of a planar graph  $G$ . Let  $v$ ,  $e$ , and  $f$  be the number of vertices, edges, and faces of this embedding, respectively. Moreover, let  $c$  be the number of connected components of  $G$ . Then*

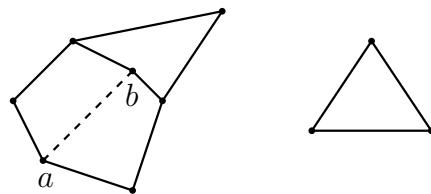
$$v - e + f = c + 1. \quad (7.4)$$

**Proof.** The idea of the proof is as follows. We start by removing all edges from  $G$  (but keep all vertices), and show that (7.4) holds. Then we add back the edges of  $G$ , one by one, and show that (7.4) remains valid throughout this process.

After having removed all edges, we have  $e = 0$  and the embedding consists of a collection of  $v$  points. Since  $f = 1$  and  $c = v$ , the relation  $v - e + f = c + 1$  holds.

Assume the relation  $v - e + f = c + 1$  holds and consider what happens when we add an edge  $ab$ . There are two possible cases.

**Case 1:** Before adding the edge  $ab$ , the vertices  $a$  and  $b$  belong to the same connected component.



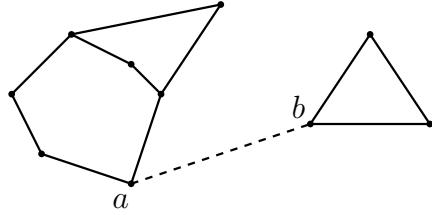
When adding the edge  $ab$ ,

- the number  $v$  of vertices does not change,

- the number  $e$  of edges increases by one,
- the number  $f$  of faces increases by one (because the edge  $ab$  splits one face into two),
- the number  $c$  of connected components does not change.

It follows that the relation  $v - e + f = c + 1$  still holds after  $ab$  has been added.

**Case 2:** Before adding the edge  $ab$ , the vertices  $a$  and  $b$  belong to different connected components.



When adding the edge  $ab$ ,

- the number  $v$  of vertices does not change,
- the number  $e$  of edges increases by one,
- the number  $f$  of faces does not change,
- the number  $c$  of connected components decreases by one.

It again follows that the relation  $v - e + f = c + 1$  still holds after  $ab$  has been added. ■

Usually, Euler's Theorem is stated for connected planar graphs, i.e., planar graphs for which  $c = 1$ :

**Theorem 7.5.3 (Euler)** *Consider any plane embedding of a connected planar graph  $G$ . If  $v$ ,  $e$ , and  $f$  denote the number of vertices, edges, and faces of this embedding, respectively, then*

$$v - e + f = 2.$$

We now show how to use Euler's Theorem to prove an upper bound on the number of edges and faces of any connected planar graph:

**Theorem 7.5.4** *Let  $G$  be any plane embedding of a connected planar graph with  $v \geq 3$  vertices. Then*

1.  *$G$  has at most  $3v - 6$  edges and*
2. *this embedding has at most  $2v - 4$  faces.*

**Proof.** As before, let  $e$  and  $f$  denote the number of edges and faces of  $G$ , respectively. If  $v = 3$ , then  $e \leq 3$  and  $f \leq 2$ . Hence, in this case, we have  $e \leq 3v - 6$  and  $f \leq 2v - 4$ .

Assume that  $v \geq 4$ . We number the faces of  $G$  arbitrarily from 1 to  $f$ . For each  $i$  with  $1 \leq i \leq f$ , let  $m_i$  denote the number of edges on the  $i$ -th face of  $G$ . Since each edge lies on the boundary of at most two faces, the summation  $\sum_{i=1}^f m_i$  counts each edge at most twice. Thus,

$$\sum_{i=1}^f m_i \leq 2e.$$

On the other hand, since  $G$  is connected and  $v \geq 4$ , each face has at least three edges on its boundary, i.e.,  $m_i \geq 3$ . It follows that

$$\sum_{i=1}^f m_i \geq 3f.$$

Combining these two inequalities implies that  $3f \leq 2e$ , which we rewrite as

$$f \leq 2e/3.$$

Using Euler's formula (with  $c = 1$ , because  $G$  is connected), we obtain

$$e = v + f - 2 \leq v + 2e/3 - 2,$$

which is equivalent to

$$e \leq 3v - 6.$$

We also obtain

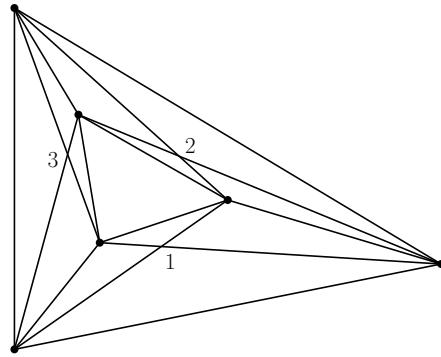
$$f \leq 2e/3 \leq 2(3v - 6)/3 = 2v - 4.$$

This completes the proof. ■

### 7.5.2 The Crossing Number of a Graph

Consider an embedding of a graph  $G = (V, E)$ . We say that two distinct edges of  $E$  cross, if their interiors have a point in common. In this case, we call this common point a *crossing*.

The example below shows an embedding of the complete graph  $K_6$  on six vertices, which are denoted by black dots. In this embedding, there are three crossings, which are numbered 1, 2, and 3.



**Definition 7.5.5** The *crossing number*  $cr(G)$  of a graph  $G$  is defined to be the minimum number of crossings in any embedding of  $G$ .

Thus, a graph  $G$  is planar if and only if  $cr(G) = 0$ . The example above shows that  $cr(K_6) \leq 3$ .

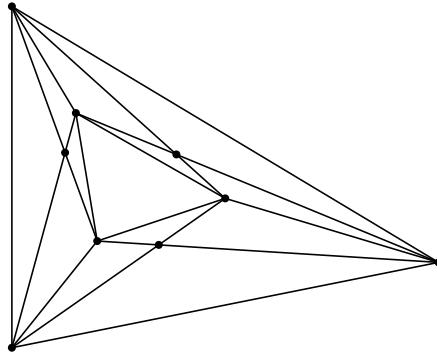
In the rest of this section, we consider the following problem: Given a graph  $G$  with  $v$  vertices and  $e$  edges, can we prove good bounds, in terms of  $v$  and  $e$ , on the crossing number  $cr(G)$  of  $G$ ?

#### A simple lower bound on the crossing number

Let  $G$  be any graph with  $v \geq 3$  vertices and  $e$  edges. Consider an embedding of  $G$  having  $cr(G)$  crossings; hence, this embedding is the “best” one.

We “make”  $G$  planar, by defining all crossings to be vertices. That is, let  $H$  be the graph whose vertex set is the union of the vertex set of  $G$  and the set of all crossings in the embedding. Edges of  $G$  are cut by the crossings into smaller edges, which are edges in the graph  $H$ .

The figure below shows the planar version of the embedding of  $K_6$  that we saw before. This new graph has 9 vertices and 21 edges.



We make the following observations:

- The graph  $H$  is planar, because it is embedded without any crossings.
- The graph  $H$  has  $v + cr(G)$  vertices.
- How many edges does  $H$  have? Any crossing in  $G$  is the intersection of exactly two edges of  $G$ ; these two edges contribute four edges to  $H$ . Hence, for any crossing in  $G$ , the number of edges in  $H$  increases by two. It follows that  $H$  has  $e + 2 \cdot cr(G)$  edges.

Since  $H$  is planar, we know from Theorem 7.5.4 that the number of its edges is bounded from above by three times the number of its vertices minus six, i.e.,

$$e + 2 \cdot cr(G) \leq 3(v + cr(G)) - 6.$$

By rewriting this inequality, we obtain the following result:

**Theorem 7.5.6** *For any graph  $G$  with  $v \geq 3$  vertices and  $e$  edges, we have*

$$cr(G) \geq e - 3v + 6.$$

For example, consider the complete graph  $K_n$  on  $n$  vertices, where  $n \geq 3$ . Since this graph has  $\binom{n}{2}$  edges, we obtain

$$cr(K_n) \geq \binom{n}{2} - 3n + 6 = \frac{1}{2}n^2 - \frac{7}{2}n + 6. \quad (7.5)$$

For  $n = 6$ , we get  $cr(K_6) \geq 3$ . Since we have seen before that  $cr(K_6) \leq 3$ , it follows that  $cr(K_6) = 3$ .

Since  $K_n$  has  $\binom{n}{2}$  edges and any two of them cross at most once, we have the following obvious upper bound on the crossing number of  $K_n$ :

$$cr(K_n) \leq \binom{\binom{n}{2}}{2} = O(n^4). \quad (7.6)$$

(Of course (7.6) holds for any graph with  $n$  vertices.)

To conclude this subsection, (7.5) gives an  $n^2$ -lower bound, whereas (7.6) gives an  $n^4$ -upper bound on the crossing number of  $K_n$ . In the next section, we will determine the true asymptotic behavior of  $cr(K_n)$ .

### A better lower bound on the crossing number

As before, let  $G$  be any graph with  $v \geq 3$  vertices and  $e$  edges. Again we consider an embedding of  $G$  having  $cr(G)$  crossings. In the rest of this subsection, we will use the Probabilistic Method to prove a lower bound on the crossing number of  $G$ .

We choose a real number  $p$  such that  $0 < p \leq 1$ . Consider a coin that comes up heads with probability  $p$  and comes up tails with probability  $1 - p$ . Let  $G_p$  be the random subgraph of  $G$ , that is obtained as follows.

- For each vertex  $a$  of  $G$ , flip the coin (independently of the other coin flips) and add  $a$  as a vertex to  $G_p$  if and only if the coin comes up heads.
- Each edge  $ab$  of  $G$  appears as an edge in  $G_p$  if and only if both  $a$  and  $b$  are vertices of  $G_p$ .

Recall that we fixed the embedding of  $G$ . As a result, this random process gives us an embedding of  $G_p$  (which may not be the best one in terms of the number of crossings.)

We denote the number of vertices, edges, and crossings in the embedding of  $G_p$  by  $v_p$ ,  $e_p$ , and  $x_p$ , respectively. Observe that these three quantities are random variables.

It follows from Theorem 7.5.6 that

$$cr(G_p) - e_p + 3v_p \geq 6,$$

provided that  $v_p \geq 3$ . This implies that

$$cr(G_p) - e_p + 3v_p \geq 0,$$

for any value of  $v_p$  that results from our random choices.

Since  $cr(G_p) \leq x_p$ , it follows that

$$x_p - e_p + 3v_p \geq 0.$$

The left-hand side is a random variable, which is always non-negative, no matter what graph  $G_p$  results from our random choices. Therefore, its expected value is also non-negative, i.e.,

$$\mathbb{E}(x_p - e_p + 3v_p) \geq 0.$$

Using the Linearity of Expectation (i.e., Theorem 6.5.2), we get

$$\mathbb{E}(x_p) - \mathbb{E}(e_p) + 3 \cdot \mathbb{E}(v_p) \geq 0. \quad (7.7)$$

We are now going to compute these three expected values separately.

The random variable  $v_p$  is equal to the number of successes in  $v$  independent trials, each one having success probability  $p$ . In other words,  $v_p$  has a binomial distribution with parameters  $v$  and  $p$ , and, therefore, by Theorem 6.7.2,

$$\mathbb{E}(v_p) = pv.$$

To compute  $\mathbb{E}(e_p)$ , we number the edges of  $G$  arbitrarily from 1 to  $e$ . For each  $i$  with  $1 \leq i \leq e$ , define  $X_i$  to be the indicator random variable with value

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th edge is an edge in } G_p, \\ 0 & \text{otherwise.} \end{cases}$$

Since an edge of  $G$  is in  $G_p$  if and only if both its vertices are in  $G_p$ , it follows that

$$\mathbb{E}(X_i) = \Pr(X_i = 1) = p^2.$$

Then, since  $e_p = \sum_{i=1}^e X_i$ , we get

$$\mathbb{E}(e_p) = \mathbb{E}\left(\sum_{i=1}^e X_i\right) = \sum_{i=1}^e \mathbb{E}(X_i) = \sum_{i=1}^e p^2 = p^2 e.$$

Finally, we compute the expected value of the random variable  $x_p$ . Number the crossings in the embedding of  $G$  arbitrarily from 1 to  $cr(G)$ . For each  $i$  with  $1 \leq i \leq cr(G)$ , define  $Y_i$  to be the indicator random variable with value

$$Y_i = \begin{cases} 1 & \text{if the } i\text{-th crossing is a crossing in } G_p, \\ 0 & \text{otherwise.} \end{cases}$$

Let  $ab$  and  $cd$  be the edges of  $G$  that cross in the  $i$ -th crossing<sup>2</sup>. This crossing appears as a crossing in  $G_p$  if and only if both  $ab$  and  $cd$  are edges in  $G_p$ . Since the points  $a, b, c$ , and  $d$  are pairwise distinct, it follows that the  $i$ -th crossing of  $G$  appears as a crossing in  $G_p$  with probability  $p^4$ . Thus,

$$\mathbb{E}(Y_i) = \Pr(Y_i = 1) = p^4.$$

Since  $x_p = \sum_{i=1}^{cr(G)} Y_i$ , it follows that

$$\mathbb{E}(x_p) = \mathbb{E}\left(\sum_{i=1}^{cr(G)} Y_i\right) = \sum_{i=1}^{cr(G)} \mathbb{E}(Y_i) = \sum_{i=1}^{cr(G)} p^4 = p^4 \cdot cr(G).$$

Substituting the three expected values into (7.7), we get

$$p^4 \cdot cr(G) - p^2 e + 3 \cdot pv \geq 0,$$

which we rewrite as

$$cr(G) \geq \frac{p^2 e - 3pv}{p^4}. \quad (7.8)$$

Observe that this inequality holds for *any* real number  $p$  with  $0 < p \leq 1$ .

If we assume that  $e \geq 4v$ , and take  $p = 4v/e$  (so that  $0 < p \leq 1$ ), then we obtain a new lower bound on the crossing number:

**Theorem 7.5.7 (Crossing Lemma)** *Let  $G$  be any graph with  $v$  vertices and  $e$  edges. If  $e \geq 4v$ , then*

$$cr(G) \geq \frac{1}{64} \frac{e^3}{v^2}.$$

Applying this lower bound to the complete graph  $K_n$  gives  $cr(K_n) = \Omega(n^4)$ . This lower bound is much better than the quadratic lower bound in (7.5) and it matches the upper bound in (7.6). Hence, we have shown that  $cr(K_n) = \Theta(n^4)$ .

**Remark 7.5.8** Let  $n$  be a very large integer and consider the complete graph  $K_n$  with  $v = n$  vertices and  $e = \binom{n}{2}$  edges. Let us see what happens if we repeat the proof for this graph. We choose a random subgraph  $G_p$  of

---

<sup>2</sup>By our definition of embedding, see Section 7.5.1, there are exactly two edges that determine the  $i$ -th crossing.

$K_n$ , where  $p = 4v/e = 8/(n - 1)$ . The expected number of vertices in  $G_p$  is equal to  $pn$ , which is approximately equal to 8. Thus, the random graph  $G_p$  is, expected, extremely small. Then we apply the *weak* lower bound of Theorem 7.5.6 to this, again expected, *extremely small* graph. The result is a proof that in any embedding of the *huge* graph  $K_n$ , there are  $\Omega(n^4)$  crossings!

## 7.6 Exercises

**7.1** Prove that, for any graph  $G$  with  $m$  edges, the sequence  $X_1, X_2, \dots, X_m$  of random variables in the proof of Theorem 7.1.1 is pairwise independent.

Give an example of a graph for which this sequence is not mutually independent.

**7.2** Prove that Theorem 7.5.4 also holds if  $G$  is not connected.

**7.3** Let  $K_5$  be the complete graph on 5 vertices. In this graph, each pair of vertices is connected by an edge. Prove that  $K_5$  is not planar.

**7.4** Let  $G$  be any embedding of a connected planar graph with  $v \geq 4$  vertices. Assume that this embedding has no triangles, i.e., there are no three vertices  $a, b$ , and  $c$ , such that  $ab, bc$ , and  $ac$  are edges of  $G$ .

- Prove that  $G$  has at most  $2v - 4$  edges.
- Let  $K_{3,3}$  be the complete bipartite graph on 6 vertices. The vertex set of this graph consists of two sets  $A$  and  $B$ , both of size three, and each vertex of  $A$  is connected by an edge to each vertex of  $B$ . Prove that  $K_{3,3}$  is not planar.

**7.5** Consider the numbers  $R_n$  that were defined in Section 4.8. In Section 4.8.1, we proved that  $R_n = O(n^8)$ . Prove that  $R_n = O(n^4)$ .

**7.6** Let  $n$  be a sufficiently large positive integer and consider the complete graph  $K_n$ . This graph has vertex set  $V = \{1, 2, \dots, n\}$ , and each pair of distinct vertices is connected by an undirected edge. (Thus,  $K_n$  has  $\binom{n}{2}$  edges.)

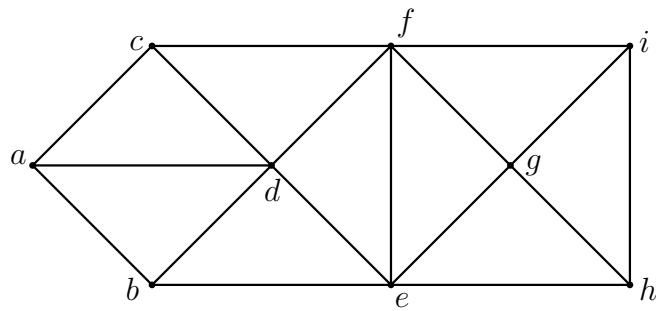
Let  $\vec{K}_n$  be the directed graph obtained by making each edge  $\{i, j\}$  of  $K_n$  a directed edge; thus, in  $\vec{K}_n$ , this edge either occurs as the directed edge  $(i, j)$  from  $i$  to  $j$  or as the directed edge  $(j, i)$  from  $j$  to  $i$ .

We say that three pairwise distinct vertices  $i, j$ , and  $k$  define a *directed triangle* in  $\vec{K}_n$ , if

- $(i, j)$ ,  $(j, k)$ , and  $(k, i)$  are edges in  $\vec{K}_n$  or
  - $(i, k)$ ,  $(k, j)$ , and  $(j, i)$  are edges in  $\vec{K}_n$ .

Prove that there exists a way to direct the edges of  $K_n$ , such that the number of directed triangles in  $\vec{K}_n$  is at least  $\frac{1}{4} \binom{n}{3}$ .

**7.7** Let  $G = (V, E)$  be a graph with vertex set  $V$  and edge set  $E$ . A subset  $I$  of  $V$  is called an *independent set* if for any two distinct vertices  $u$  and  $v$  in  $I$ ,  $(u, v)$  is not an edge in  $E$ . For example, in the following graph,  $I = \{a, e, i\}$  is an independent set.



Let  $n$  and  $m$  denote the number of vertices and edges in  $G$ , respectively, and assume that  $m \geq n/2$ . This exercise will lead you through a proof of the fact that  $G$  contains an independent set of size at least  $n^2/(4m)$ .

Consider the following algorithm, in which all random choices made are mutually independent:

---

**Algorithm INDEPSET( $G$ ):**

**Step 1:** Set  $H = G$ .

**Step 2:** Let  $d = 2m/n$ . For each vertex  $v$  of  $H$ , with probability  $1 - 1/d$ , delete the vertex  $v$ , together with its incident edges, from  $H$ .

**Step 3:** As long as the graph  $H$  contains edges, do the following: Pick an arbitrary edge  $(u, v)$  in  $H$ , and remove the vertex  $u$ , together with its incident edges, from  $H$ .

**Step 4:** Let  $I$  be the vertex set of the graph  $H$ . Return  $I$ .

- Argue that the set  $I$  that is returned by this algorithm is an independent set in  $G$ .
- Let  $X$  and  $Y$  be the random variables whose values are the number of vertices and edges in the graph  $H$  after Step 2, respectively. Prove that

$$\mathbb{E}(X) = n^2/(2m)$$

and

$$\mathbb{E}(Y) = n^2/(4m).$$

- Let  $Z$  be the random variable whose value is the size of the independent set  $I$  that is returned by the algorithm. Argue that

$$Z \geq X - Y.$$

- Prove that

$$\mathbb{E}(Z) \geq n^2/(4m).$$

- Argue that this implies that the graph  $G$  contains an independent set of size at least  $n^2/(4m)$ .

**7.8** Elisa Kazan is having a party at her home. Elisa has a round table that has 52 seats numbered  $0, 1, 2, \dots, 51$  in clockwise order. Elisa invites 51 friends, so that the total number of people at the party is 52. Of these 52 people, 15 drink cider, whereas the other 37 drink beer.

In this exercise, you will prove the following claim: No matter how the 52 people sit at the table, there is always a consecutive group of 7 people such that at least 3 of them drink cider.

From now on, we consider an arbitrary (which is not random) arrangement of the 52 people sitting at the table.

- Let  $k$  be a uniformly random element of the set  $\{0, 1, 2, \dots, 51\}$ . Consider the consecutive group of 7 people that sit in seats  $k, k+1, k+2, \dots, k+6$ ; these seat numbers are to be read modulo 52. Define the random variable  $X$  to be the number of people in this group that drink cider. Prove that  $\mathbb{E}(X) > 2$ .

*Hint:* Number the 15 cider drinkers arbitrarily as  $P_1, P_2, \dots, P_{15}$ . For each  $i$  with  $1 \leq i \leq 15$ , consider the indicator random variable

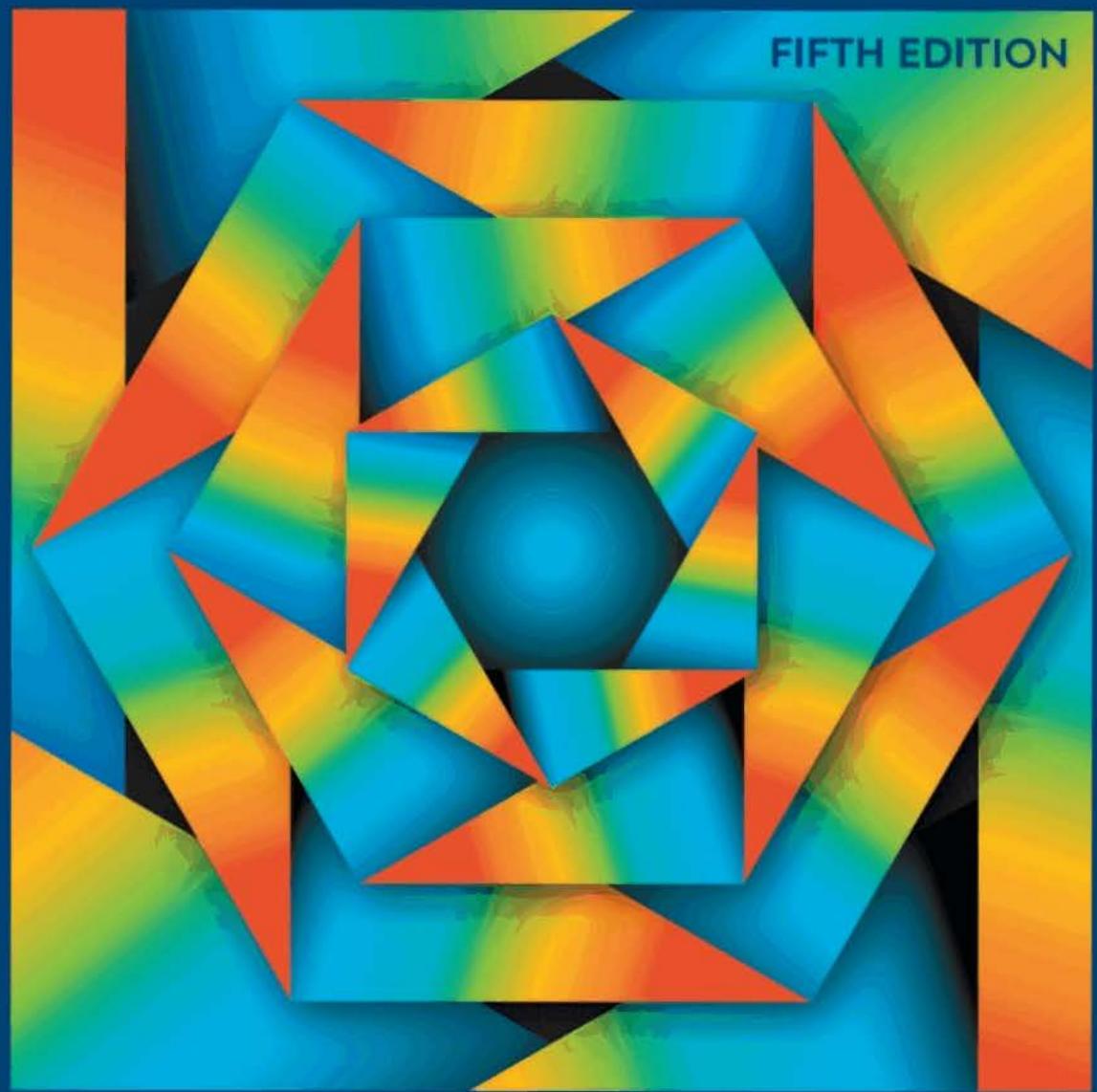
$$X_i = \begin{cases} 1 & \text{if } P_i \text{ sits in one of the seats } k, k+1, k+2, \dots, k+6, \\ 0 & \text{otherwise.} \end{cases}$$

- For the given arrangement of the 52 people sitting at the table, prove that there is a consecutive group of 7 people such that at least 3 of them drink cider.

*Hint:* Assume the claim is false. What is an upper bound on  $\mathbb{E}(X)$ ?

# LINEAR ALGEBRA

FIFTH EDITION



Stephen H.  
**FRIEDBERG**

Arnold J.  
**INSEL**

Lawrence E.  
**SPENCE**

---

Fifth Edition

# Linear Algebra

Stephen H. Friedberg  
Arnold J. Insel  
Lawrence E. Spence

*Illinois State University*

---



Director, Portfolio Management: Deirdre Lynch  
Executive Editor: Jeff Weidenaar  
Editorial Assistant: Jonathan Krebs  
Content Producer: Tara Corpuz  
Managing Producer: Scott Disanno  
Product Marketing Manager: Yvonne Vannatta  
Field Marketing Manager: Evan St. Cyr  
Field Marketing Assistant: Jon Bryant  
Senior Author Support/Technology Specialist: Joe Vetere  
Manager, Rights and Permissions: Gina Cheselka  
Cover Design: Pearson CSC  
Manufacturing Buyer: Carol Melville, LSC Communications  
Cover Image: Mark Grenier/Shutterstock

Copyright ©2019, 2003, 1997 by Pearson Education, Inc. All Rights Reserved. Printed in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit [www.pearsoned.com/permissions/](http://www.pearsoned.com/permissions/).

PEARSON, ALWAYS LEARNING, and MYLAB are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

**Library of Congress Cataloging-in-Publication Data**

Names: Friedberg, Stephen H., author. | Insel, Arnold J., author. | Spence, Lawrence E., author.

Title: Linear algebra / Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence (Illinois State University).

Description: Fifth edition. | Upper Saddle River, New Jersey : Pearson Education, Inc., 2018. | Includes indexes.

Identifiers: LCCN 2018016171 | ISBN 9780134860244 (alk. paper) | ISBN 0134860241 (alk. paper)

Subjects: LCSH: Algebras, Linear--Textbooks.

Classification: LCC QA184 .F75 2018 | DDC 512/.5--dc23  
LC record available at <https://lccn.loc.gov/2018016171>

To our families:  
Ruth Ann, Rachel, Jessica, and Jeremy  
Barbara, Thomas, and Sara  
Linda, Stephen, and Alison

*This page intentionally left blank*

# Contents

---

<b>Preface</b>	<b>viii</b>
<b>To the Student</b>	<b>xii</b>
<b>1 Vector Spaces</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Vector Spaces . . . . .	6
1.3 Subspaces . . . . .	17
1.4 Linear Combinations and Systems of Linear Equations . . . . .	24
1.5 Linear Dependence and Linear Independence . . . . .	36
1.6 Bases and Dimension . . . . .	43
1.7* Maximal Linearly Independent Subsets . . . . .	59
Index of Definitions . . . . .	63
<b>2 Linear Transformations and Matrices</b>	<b>64</b>
2.1 Linear Transformations, Null Spaces, and Ranges . . . . .	64
2.2 The Matrix Representation of a Linear Transformation . . . . .	79
2.3 Composition of Linear Transformations and Matrix Multiplication . . . . .	87
2.4 Invertibility and Isomorphisms . . . . .	100
2.5 The Change of Coordinate Matrix . . . . .	110
2.6* Dual Spaces . . . . .	119
2.7* Homogeneous Linear Differential Equations with Constant Coefficients . . . . .	128
Index of Definitions . . . . .	145

---

\*Sections denoted by an asterisk are optional.

**3 Elementary Matrix Operations and Systems of Linear Equations 147**

3.1	Elementary Matrix Operations and Elementary Matrices . . . . .	148
3.2	The Rank of a Matrix and Matrix Inverses . . . . .	152
3.3	Systems of Linear Equations—Theoretical Aspects . . . . .	168
3.4	Systems of Linear Equations—Computational Aspects . . . . .	181
	Index of Definitions . . . . .	197

**4 Determinants 199**

4.1	Determinants of Order 2 . . . . .	199
4.2	Determinants of Order $n$ . . . . .	209
4.3	Properties of Determinants . . . . .	222
4.4	Summary—Important Facts about Determinants . . . . .	232
4.5*	A Characterization of the Determinant . . . . .	238
	Index of Definitions . . . . .	244

**5 Diagonalization 245**

5.1	Eigenvalues and Eigenvectors . . . . .	246
5.2	Diagonalizability . . . . .	261
5.3*	Matrix Limits and Markov Chains . . . . .	282
5.4	Invariant Subspaces and the Cayley–Hamilton Theorem . . . . .	311
	Index of Definitions . . . . .	325

**6 Inner Product Spaces 327**

6.1	Inner Products and Norms . . . . .	327
6.2	The Gram–Schmidt Orthogonalization Process and Orthogonal Complements . . . . .	339
6.3	The Adjoint of a Linear Operator . . . . .	354
6.4	Normal and Self-Adjoint Operators . . . . .	366
6.5	Unitary and Orthogonal Operators and Their Matrices . . . . .	376
6.6	Orthogonal Projections and the Spectral Theorem . . . . .	395
6.7*	The Singular Value Decomposition and the Pseudoinverse . . . . .	402
6.8*	Bilinear and Quadratic Forms . . . . .	419
6.9*	Einstein’s Special Theory of Relativity . . . . .	448
6.10*	Conditioning and the Rayleigh Quotient . . . . .	459
6.11*	The Geometry of Orthogonal Operators . . . . .	466

<b>Table of Contents</b>	<b>vii</b>
Index of Definitions . . . . .	473
<b>7 Canonical Forms</b>	<b>475</b>
7.1 The Jordan Canonical Form I . . . . .	475
7.2 The Jordan Canonical Form II . . . . .	490
7.3 The Minimal Polynomial . . . . .	509
7.4* The Rational Canonical Form . . . . .	517
Index of Definitions . . . . .	541
<b>Appendices</b>	<b>542</b>
A Sets . . . . .	542
B Functions . . . . .	544
C Fields . . . . .	546
D Complex Numbers . . . . .	549
E Polynomials . . . . .	555
<b>Answers to Selected Exercises</b>	<b>565</b>
<b>Index</b>	<b>582</b>

# Preface

---

The language and concepts of matrix theory and, more generally, of linear algebra have come into widespread usage in the social and natural sciences, computer science, and statistics. In addition, linear algebra continues to be of great importance in modern treatments of geometry and analysis.

The primary purpose of this fifth edition of *Linear Algebra* is to present a careful treatment of the principal topics of linear algebra and to illustrate the power of the subject through a variety of applications. Throughout, we emphasize the symbiotic relationship between linear transformations and matrices. However, where appropriate, theorems are stated in the more general infinite-dimensional case. This enables us to apply the basic theory of vector spaces and linear transformations to find solutions to a homogeneous linear differential equation as well as the best approximation by a trigonometric polynomial to a continuous function.

Although the only formal prerequisite for this book is a one-year course in calculus, it requires the mathematical sophistication of typical junior and senior mathematics majors. This book is especially suited to a second course in linear algebra that emphasizes abstract vector spaces, although it can be used in a first course with a strong theoretical emphasis.

## SUGGESTED COURSE OUTLINES

The book is organized for use in a number of different courses (ranging from three to eight semester hours in length). The core material (vector spaces, linear transformations and matrices, systems of linear equations, determinants, diagonalization, and inner product spaces) is found in Chapters 1 through 5 and Sections 6.1 through 6.5. Chapters 6 and 7, on inner product spaces and canonical forms, are completely independent and may be studied in either order. In addition, throughout the book there are applications to such areas as differential equations, economics, geometry, and physics. These applications are not central to the mathematical development, however, and may be excluded at the discretion of the instructor.

We have attempted to make it possible for many of the important topics of linear algebra to be covered in a one-semester course. This goal has led us to develop the major topics with fewer preliminaries than in a traditional approach. (Our treatment of the Jordan canonical form, for instance, does not require any theory of polynomials.) The resulting economy permits us to cover the core material of the book (omitting many of the optional sections

and a detailed discussion of determinants) in a one-semester four-hour course for students who have had some prior exposure to linear algebra.

## OVERVIEW OF CONTENTS

Chapter 1 of the book presents the basic theory of vector spaces: subspaces, linear combinations, linear dependence and independence, bases, and dimension. The chapter concludes with an optional section in which we prove that every infinite-dimensional vector space has a basis.

Linear transformations and their relationships to matrices are the subjects of Chapter 2. We discuss the null space, range, and matrix representations of a linear transformation, isomorphisms, and change of coordinates. The chapter ends with optional sections on dual spaces and homogeneous linear differential equations.

The application of vector space theory and linear transformations to systems of linear equations is found in Chapter 3. We have chosen to defer this important subject so that it can be presented as a consequence of the preceding material. This approach allows the familiar topic of linear systems to illuminate the abstract theory and permits us to avoid messy matrix computations in the presentation of Chapters 1 and 2. There are occasional examples in these chapters, however, where we solve systems of linear equations. (Of course, these examples are not a part of the theoretical development.) The necessary background is contained in Section 1.4.

Determinants, the subject of Chapter 4, are of much less importance than they once were. In a short course (less than one year), we prefer to treat determinants lightly so that more time may be devoted to the material in Chapters 5 through 7. Consequently, we have presented two alternatives in Chapter 4—a complete development of the theory (Sections 4.1 through 4.3) and a summary of important facts that are needed for the remaining chapters (Section 4.4). Optional Section 4.5 presents an axiomatic development of the determinant.

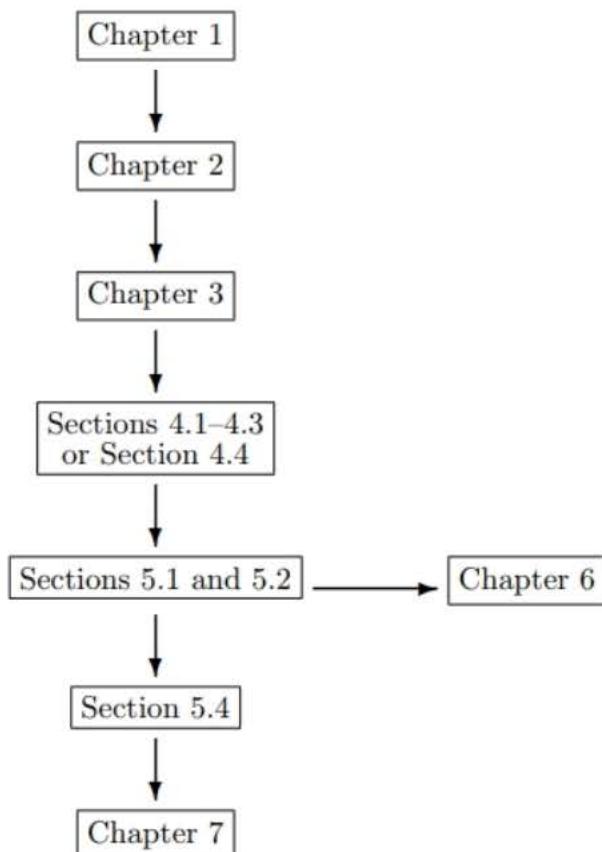
Chapter 5 discusses eigenvalues, eigenvectors, and diagonalization. One of the most important applications of this material occurs in computing matrix limits. We have therefore included an optional section on matrix limits and Markov chains in this chapter even though the most general statement of some of the results requires a knowledge of the Jordan canonical form. Section 5.4 contains material on invariant subspaces and the Cayley–Hamilton theorem.

Inner product spaces are the subject of Chapter 6. The basic mathematical theory (inner products; the Gram–Schmidt process; orthogonal complements; the adjoint of an operator; normal, self-adjoint, orthogonal and unitary operators; orthogonal projections; and the spectral theorem) is contained in Sections 6.1 through 6.6. Sections 6.7 through 6.11 contain diverse applications of the rich inner product space structure.

Canonical forms are treated in Chapter 7. Sections 7.1 and 7.2 develop the Jordan canonical form, Section 7.3 presents the minimal polynomial, and Section 7.4 discusses the rational canonical form.

There are five appendices. The first four, which discuss sets, functions, fields, and complex numbers, respectively, are intended to review basic ideas used throughout the book. Appendix E on polynomials is used primarily in Chapters 5 and 7, especially in Section 7.4. We prefer to cite particular results from the appendices as needed rather than to discuss the appendices independently.

The following diagram illustrates the dependencies among the various chapters.



One final word is required about our notation. Sections and subsections labeled with an asterisk (\*) are optional and may be omitted as the instructor sees fit. An exercise accompanied by the dagger symbol (†) is not optional, however—we use this symbol to identify an exercise that is cited in some later section that is not optional.

## DIFFERENCES BETWEEN THE FOURTH AND FIFTH EDITIONS

The organization of the fifth edition is essentially the same as its predecessor. Nevertheless, this edition contains many significant local changes that improve the book. Several of these streamline the presentation, and others clarify expositions that had led to student misunderstandings. Further improvements include revised proofs of some theorems, additional examples, new exercises, and literally hundreds of minor editorial changes. Many of these changes were prompted by George Bergman (University of California, Berkeley), who provided detailed comments about his experiences teaching from earlier editions, as well as numerous other professors and students who have written us with questions and comments about the fourth edition.

As an additional aid to students, we have made available online a solution to one theoretical exercise in each section of the book. These exercises each have their exercise number printed within a gray colored box, and the last sentence of each of these exercises gives a short URL for its online solution. See, for example, Exercise 5 on page 6.

We have also made available online four applications of the subject matter in Sections 2.3, 5.3, 6.5, and 6.6. These applications are new to the fifth edition.

To find the latest information about this book, consult our website. We encourage comments, which can be sent to us by email. Our website and email addresses are listed below.

website: [goo.gl/y1jz4Y](http://goo.gl/y1jz4Y)

email: [linearalgebra@ilstu.edu](mailto:linearalgebra@ilstu.edu)

*Stephen H. Friedberg  
Arnold J. Insel  
Lawrence E. Spence*

# To the Student

---

In a general sense, we can think of linear algebra as the mathematics of linear processes. It is concerned with sets of objects, called *vector spaces*, in which a concept of linearity exists, and with special functions defined on vector spaces, called *linear transformations*, that preserve linearity. Although linear algebra is a branch of pure mathematics, it plays an essential role in many areas of applied mathematics, statistics, engineering, and physics. With the recent development of computers, the importance of computations, simulations, and modeling has greatly expanded the use of linear algebra to other disciplines also. In this book, we present applications of linear algebra to the study of differential equations, statistics, genetics, and physics, among others.

This book is intended to be used as a textbook for an advanced linear algebra course, but it assumes no previous coursework in linear algebra. The emphasis in this book is on understanding the conceptual ideas that are presented here and using them to justify other results, that is, to be able to write mathematical proofs.

One of the first obstacles to overcome in order to communicate in any advanced field of study is mastering the vocabulary of that field. In mathematics, a definition is a very precise statement about the properties that must be possessed by an object being defined. These properties tell you exactly what you must verify in order to show that an object satisfies its definition. For example, on pages 6–7, the definition of a *vector space* is given. To prove that a certain set  $V$  with specific operations of addition and scalar multiplication is a vector space, you must verify that conditions (VS1) through (VS8) hold for  $V$  with these operations. Exercises 10–13 in Section 1.2 ask you to do exactly this. There is no way that anyone could do these exercises without understanding the meaning of conditions (VS1) through (VS8). So before doing any of the exercises in a section, be certain that you understand the meaning of all of the terms defined in that section. It is also a good idea to remember a few examples of each concept that is defined.

After you have learned the new vocabulary in a section, learn its main results, which are usually found in theorems. Pay particular attention to these, being certain that you understand thoroughly everything that is being said. Then try to express the result in your own words. If you can't communicate an idea in writing, then you probably don't understand it well.

Perhaps you have heard a mathematics instructor say that mathematics is not a spectator sport. This is because one of the best ways to learn mathematics is by working exercises. The exercise sets in this book begin with true/false questions that test your understanding of important ideas

(and vocabulary!) in each section. Some problems that require only basic computations may follow. Then come exercises that ask for an explanation, justification, or conjecture. These different types of exercises will help you learn different aspects of linear algebra. Not only do the exercises help you to check your understanding of important concepts, but they also provide an opportunity to practice the vocabulary and symbolism that you are learning. For this reason, regular work on exercises is essential for success. We have provided a solution to one of the theoretical exercises in each section of the text. These exercises each have their exercise number printed within a gray colored box, and the last sentence of each of these exercises gives a short URL for its online solution. See, for example, Exercise 5 on page 6.

Here are some specific suggestions that will enable you to get the most from your study of linear algebra.

- **Carefully read each section *before* the classroom discussion**

Some students use a textbook only as a source of examples when working exercises. This approach does not yield the full benefit from either the textbook or the classroom discussions. By reading the text before a discussion occurs, you get an overview of the material to be discussed and know what you understand and what you don't.

- **Prepare regularly for each class**

You cannot expect to learn to play a musical instrument merely by attending lessons once a week—long and careful practice between lessons is necessary. So it is with linear algebra. At the least, you should study the material presented in class and work assigned exercises that deal with the new material.

Each new lesson usually introduces several important concepts or definitions that must be learned in order for subsequent sections to be understood. As a result, falling behind in your study by even a single day prevents you from understanding the material that follows. To be successful, you must learn the new material as it arises and not wait to study until you are less busy or until an exam is imminent.

- **Review often**

As you attempt to understand new material, you may become aware that you have forgotten some previous material or that you haven't understood it well enough. By relearning such material, you not only gain a deeper understanding of previous topics, but you also enable yourself to learn new ideas more quickly and more deeply. Discussing topics with classmates can be a useful way of reviewing.

We hope that your study of linear algebra is successful and that you take from the subject concepts and techniques that are useful in future courses and in your career.

*This page intentionally left blank*

# Vector Spaces

---

- 1.1 Introduction
  - 1.2 Vector Spaces
  - 1.3 Subspaces
  - 1.4 Linear Combinations and Systems of Linear Equations
  - 1.5 Linear Dependence and Linear Independence
  - 1.6 Bases and Dimension
  - 1.7\* Maximal Linearly Independent Subsets
- 

## 1.1 INTRODUCTION

Many familiar physical notions, such as forces, velocities,<sup>1</sup> and accelerations, involve both a magnitude (the amount of the force, velocity, or acceleration) and a direction. Any such entity involving both magnitude and direction is called a “vector.” A vector is represented by an arrow whose length denotes the magnitude of the vector and whose direction represents the direction of the vector. In most physical situations involving vectors, only the magnitude and direction of the vector are significant; consequently, we regard vectors with the same magnitude and direction as being equal irrespective of their positions. In this section the geometry of vectors is discussed. This geometry is derived from physical experiments that test the manner in which two vectors interact.

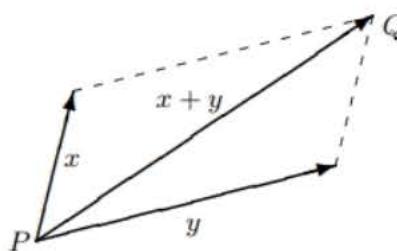
Familiar situations suggest that when two like physical quantities act simultaneously at a point, the magnitude of their effect need not equal the sum of the magnitudes of the original quantities. For example, a swimmer swimming upstream at the rate of 2 miles per hour against a current of 1 mile per hour does not progress at the rate of 3 miles per hour. For in this instance the motions of the swimmer and current oppose each other, and the rate of progress of the swimmer is only 1 mile per hour upstream. If, however,

---

<sup>1</sup>The word *velocity* is being used here in its scientific sense—as an entity having both magnitude and direction. The magnitude of a velocity (without regard for the direction of motion) is called its **speed**.

swimmer is moving downstream (with the current), then his or her rate of progress is 3 miles per hour downstream.

Experiments show that if two like quantities act together, their effect is predictable. In this case, the vectors used to represent these quantities can be combined to form a resultant vector that represents the combined effects of the original quantities. This resultant vector is called the *sum* of the original vectors, and the rule for their combination is called the *parallelogram law*. (See Figure 1.1.)



**Figure 1.1**

**Parallelogram Law for Vector Addition.** *The sum of two vectors  $x$  and  $y$  that act at the same point  $P$  is the vector beginning at  $P$  that is represented by the diagonal of parallelogram having  $x$  and  $y$  as adjacent sides.*

Since opposite sides of a parallelogram are parallel and of equal length, the endpoint  $Q$  of the arrow representing  $x + y$  can also be obtained by allowing  $x$  to act at  $P$  and then allowing  $y$  to act at the endpoint of  $x$ . Similarly, the endpoint of the vector  $x + y$  can be obtained by first permitting  $y$  to act at  $P$  and then allowing  $x$  to act at the endpoint of  $y$ . Thus two vectors  $x$  and  $y$  that both act at the point  $P$  may be added “tail-to-head”; that is, either  $x$  or  $y$  may be applied at  $P$  and a vector having the same magnitude and direction as the other may be applied to the endpoint of the first. If this is done, the endpoint of the second vector is the endpoint of  $x + y$ .

The addition of vectors can be described algebraically with the use of analytic geometry. In the plane containing  $x$  and  $y$ , introduce a coordinate system with  $P$  at the origin. Let  $(a_1, a_2)$  denote the endpoint of  $x$  and  $(b_1, b_2)$  denote the endpoint of  $y$ . Then as Figure 1.2(a) shows, the endpoint  $Q$  of  $x+y$  is  $(a_1+b_1, a_2+b_2)$ . Henceforth, when a reference is made to the coordinates of the endpoint of a vector, the vector should be assumed to emanate from the origin. Moreover, since a vector beginning at the origin is completely determined by its endpoint, we sometimes refer to *the point  $x$*  rather than *the endpoint of the vector  $x$*  if  $x$  is a vector emanating from the origin.

Besides the operation of vector addition, there is another natural operation that can be performed on vectors—the length of a vector may be magnified

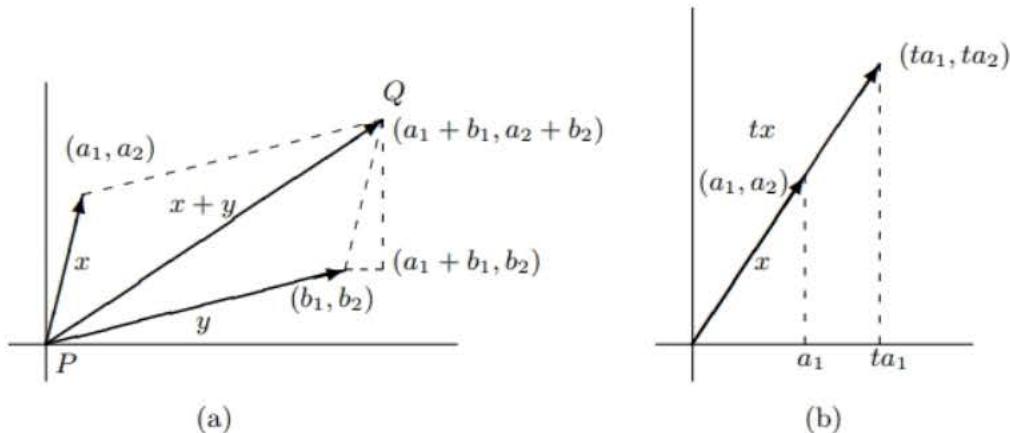


Figure 1.2

or contracted. This operation, called **scalar multiplication**, consists of multiplying the vector by a real number. If the vector  $x$  is represented by an arrow, then for any nonzero real number  $t$ , the vector  $tx$  is represented by an arrow in the same direction if  $t > 0$  and in the opposite direction if  $t < 0$ . The length of the arrow  $tx$  is  $|t|$  times the length of the arrow  $x$ . Two nonzero vectors  $x$  and  $y$  are called **parallel** if  $y = tx$  for some nonzero real number  $t$ . (Thus nonzero vectors having the same or opposite directions are parallel.)

To describe scalar multiplication algebraically, again introduce a coordinate system into a plane containing the vector  $x$  so that  $x$  emanates from the origin. If the endpoint of  $x$  has coordinates  $(a_1, a_2)$ , then the coordinates of the endpoint of  $tx$  are easily seen to be  $(ta_1, ta_2)$ . (See Figure 1.2(b).)

The algebraic descriptions of vector addition and scalar multiplication for vectors in a plane yield the following properties:

1. For all vectors  $x$  and  $y$ ,  $x + y = y + x$ .
2. For all vectors  $x$ ,  $y$ , and  $z$ ,  $(x + y) + z = x + (y + z)$ .
3. There exists a vector denoted  $\theta$  such that  $x + \theta = x$  for each vector  $x$ .
4. For each vector  $x$ , there is a vector  $y$  such that  $x + y = \theta$ .
5. For each vector  $x$ ,  $1x = x$ .
6. For each pair of real numbers  $a$  and  $b$  and each vector  $x$ ,  $(ab)x = a(bx)$ .
7. For each real number  $a$  and each pair of vectors  $x$  and  $y$ ,  $a(x + y) = ax + ay$ .
8. For each pair of real numbers  $a$  and  $b$  and each vector  $x$ ,  $(a + b)x = ax + bx$ .

Arguments similar to the preceding ones show that these eight properties, as well as the geometric interpretations of vector addition and scalar multiplication, are true also for vectors acting in space rather than in a plane. These results can be used to write equations of lines and planes in space.

Consider first the equation of a line in space that passes through two distinct points  $A$  and  $B$ . Let  $O$  denote the origin of a coordinate system in space, and let  $u$  and  $v$  denote the vectors that begin at  $O$  and end at  $A$  and  $B$ , respectively. If  $w$  denotes the vector beginning at  $A$  and ending at  $B$ , then “tail-to-head” addition shows that  $u + w = v$ , and hence  $w = v - u$ , where  $-u$  denotes the vector  $(-1)u$ . (See Figure 1.3, in which the quadrilateral  $OABC$  is a parallelogram.) Since a scalar multiple of  $w$  is parallel to  $w$  but possibly of a different length than  $w$ , any point on the line joining  $A$  and  $B$  may be obtained as the endpoint of a vector that begins at  $A$  and has the form  $tw$  for some real number  $t$ . Conversely, the endpoint of every vector of the form  $tw$  that begins at  $A$  lies on the line joining  $A$  and  $B$ . Thus an equation of the line through  $A$  and  $B$  is  $x = u + tw = u + t(v - u)$ , where  $t$  is a real number and  $x$  denotes an arbitrary point on the line. Notice also that the endpoint  $C$  of the vector  $v - u$  in Figure 1.3 has coordinates equal to the difference of the coordinates of  $B$  and  $A$ .

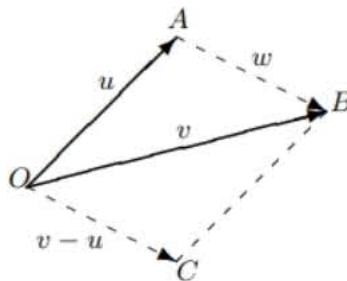


Figure 1.3

### Example 1

Let  $A$  and  $B$  be points having coordinates  $(-2, 0, 1)$  and  $(4, 5, 3)$ , respectively. The endpoint  $C$  of the vector emanating from the origin and having the same direction as the vector beginning at  $A$  and terminating at  $B$  has coordinates  $(4, 5, 3) - (-2, 0, 1) = (6, 5, 2)$ . Hence the equation of the line through  $A$  and  $B$  is

$$x = (-2, 0, 1) + t(6, 5, 2). \quad \blacklozenge$$

Now let  $A$ ,  $B$ , and  $C$  denote any three noncollinear points in space. These points determine a unique plane, and its equation can be found by use of our previous observations about vectors. Let  $u$  and  $v$  denote vectors beginning at  $A$  and ending at  $B$  and  $C$ , respectively. Observe that any point in the plane containing  $A$ ,  $B$ , and  $C$  is the endpoint  $S$  of a vector  $x$  beginning at  $A$  and having the form  $su + tv$  for some real numbers  $s$  and  $t$ . The endpoint of  $su$  is the point of intersection of the line through  $A$  and  $B$  with the line through  $S$  parallel to the line through  $A$  and  $C$ . (See Figure 1.4.) A similar procedure

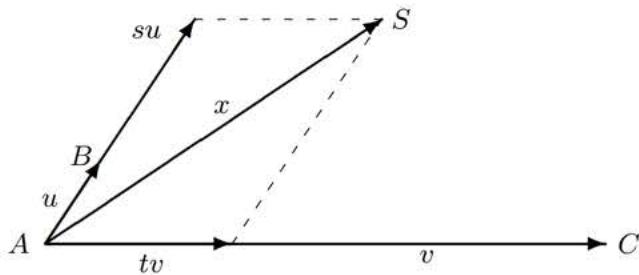


Figure 1.4

locates the endpoint of  $tv$ . Moreover, for any real numbers  $s$  and  $t$ , the vector  $su + tv$  lies in the plane containing  $A$ ,  $B$ , and  $C$ . It follows that an equation of the plane containing  $A$ ,  $B$ , and  $C$  is

$$x = A + su + tv,$$

where  $s$  and  $t$  are arbitrary real numbers and  $x$  denotes an arbitrary point in the plane.

### Example 2

Let  $A$ ,  $B$ , and  $C$  be the points having coordinates  $(1, 0, 2)$ ,  $(-3, -2, 4)$ , and  $(1, 8, -5)$ , respectively. The endpoint of the vector emanating from the origin and having the same length and direction as the vector beginning at  $A$  and terminating at  $B$  is

$$(-3, -2, 4) - (1, 0, 2) = (-4, -2, 2).$$

Similarly, the endpoint of a vector emanating from the origin and having the same length and direction as the vector beginning at  $A$  and terminating at  $C$  is  $(1, 8, -5) - (1, 0, 2) = (0, 8, -7)$ . Hence the equation of the plane containing the three given points is

$$x = (1, 0, 2) + s(-4, -2, 2) + t(0, 8, -7). \quad \blacklozenge$$

Any mathematical structure possessing the eight properties on page 3 is called a *vector space*. In the next section we formally define a vector space and consider many examples of vector spaces other than the ones mentioned above.

## EXERCISES

- Determine whether the vectors emanating from the origin and terminating at the following pairs of points are parallel.

- (a)  $(3, 1, 2)$  and  $(6, 4, 2)$   
 (b)  $(-3, 1, 7)$  and  $(9, -3, -21)$   
 (c)  $(5, -6, 7)$  and  $(-5, 6, -7)$   
 (d)  $(2, 0, -5)$  and  $(5, 0, -2)$
2. Find the equations of the lines through the following pairs of points in space.
- (a)  $(3, -2, 4)$  and  $(-5, 7, 1)$   
 (b)  $(2, 4, 0)$  and  $(-3, -6, 0)$   
 (c)  $(3, 7, 2)$  and  $(3, 7, -8)$   
 (d)  $(-2, -1, 5)$  and  $(3, 9, 7)$
3. Find the equations of the planes containing the following points in space.
- (a)  $(2, -5, -1)$ ,  $(0, 4, 6)$ , and  $(-3, 7, 1)$   
 (b)  $(3, -6, 7)$ ,  $(-2, 0, -4)$ , and  $(5, -9, -2)$   
 (c)  $(-8, 2, 0)$ ,  $(1, 3, 0)$ , and  $(6, -5, 0)$   
 (d)  $(1, 1, 1)$ ,  $(5, 5, 5)$ , and  $(-6, 4, 2)$
4. What are the coordinates of the vector  $\theta$  in the Euclidean plane that satisfies property 3 on page 3? Justify your answer.
5. Prove that if the vector  $x$  emanates from the origin of the Euclidean plane and terminates at the point with coordinates  $(a_1, a_2)$ , then the vector  $tx$  that emanates from the origin terminates at the point with coordinates  $(ta_1, ta_2)$ . Visit [goo.gl/eYTxuU](http://goo.gl/eYTxuU) for a solution.
6. Show that the midpoint of the line segment joining the points  $(a, b)$  and  $(c, d)$  is  $((a + c)/2, (b + d)/2)$ .
7. Prove that the diagonals of a parallelogram bisect each other.

## 1.2 VECTOR SPACES

In Section 1.1, we saw that with the natural definitions of vector addition and scalar multiplication, the vectors in a plane satisfy the eight properties listed on page 3. Many other familiar algebraic systems also permit definitions of addition and scalar multiplication that satisfy the same eight properties. In this section, we introduce some of these systems, but first we formally define this type of algebraic structure.

**Definitions.** A **vector space** (or **linear space**)  $V$  over a field<sup>2</sup>  $F$  consists of a set on which two operations (called **addition** and **scalar multiplication**, respectively) are defined so that for each pair of elements  $x, y$ ,

---

<sup>2</sup>Fields are discussed in Appendix C.

in  $V$  there is a unique element  $x + y$  in  $V$ , and for each element  $a$  in  $F$  and each element  $x$  in  $V$  there is a unique element  $ax$  in  $V$ , such that the following conditions hold.

- (VS 1) For all  $x, y$  in  $V$ ,  $x + y = y + x$  (commutativity of addition).
- (VS 2) For all  $x, y, z$  in  $V$ ,  $(x + y) + z = x + (y + z)$  (associativity of addition).
- (VS 3) There exists an element in  $V$  denoted by  $0$  such that  $x + 0 = x$  for each  $x$  in  $V$ .
- (VS 4) For each element  $x$  in  $V$  there exists an element  $y$  in  $V$  such that  $x + y = 0$ .
- (VS 5) For each element  $x$  in  $V$ ,  $1x = x$ .
- (VS 6) For each pair of elements  $a, b$  in  $F$  and each element  $x$  in  $V$ ,  
$$(ab)x = a(bx).$$
- (VS 7) For each element  $a$  in  $F$  and each pair of elements  $x, y$  in  $V$ ,  
$$a(x + y) = ax + ay.$$
- (VS 8) For each pair of elements  $a, b$  in  $F$  and each element  $x$  in  $V$ ,  
$$(a + b)x = ax + bx.$$

The elements  $x + y$  and  $ax$  are called the **sum** of  $x$  and  $y$  and the **product** of  $a$  and  $x$ , respectively.

The elements of the field  $F$  are called **scalars** and the elements of the vector space  $V$  are called **vectors**. The reader should not confuse this use of the word “vector” with the physical entity discussed in Section 1.1: the word “vector” is now being used to describe any element of a vector space.

A vector space is frequently discussed in the text without explicitly mentioning its field of scalars. The reader is cautioned to remember, however, that every vector space is regarded as a vector space over a given field, which is denoted by  $F$ . Occasionally we restrict our attention to the fields of real and complex numbers, which are denoted  $R$  and  $C$ , respectively. Unless otherwise noted, we assume that fields used in the examples and exercises of this book have characteristic zero (see page 549).

Observe that (VS 2) permits us to define the addition of any finite number of vectors unambiguously (without the use of parentheses).

In the remainder of this section we introduce several important examples of vector spaces that are studied throughout this text. Observe that in describing a vector space, it is necessary to specify not only the vectors but also

the operations of addition and scalar multiplication. The reader should check that each of these examples satisfies conditions (VS1) through (VS8).

An object of the form  $(a_1, a_2, \dots, a_n)$ , where the entries  $a_1, a_2, \dots, a_n$  are elements of a field  $F$ , is called an  **$n$ -tuple** with entries from  $F$ . The elements  $a_1, a_2, \dots, a_n$  are called the **entries** or **components** of the  $n$ -tuple. Two  $n$ -tuples  $(a_1, a_2, \dots, a_n)$  and  $(b_1, b_2, \dots, b_n)$  with entries from a field  $F$  are called **equal** if  $a_i = b_i$  for  $i = 1, 2, \dots, n$ .

### Example 1

The set of all  $n$ -tuples with entries from a field  $F$  is denoted by  $\mathbb{F}^n$ . This set is a vector space over  $F$  with the operations of coordinatewise addition and scalar multiplication; that is, if  $u = (a_1, a_2, \dots, a_n) \in \mathbb{F}^n$ ,  $v = (b_1, b_2, \dots, b_n) \in \mathbb{F}^n$ , and  $c \in F$ , then

$$u + v = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n) \quad \text{and} \quad cu = (ca_1, ca_2, \dots, ca_n).$$

Thus  $\mathbb{R}^3$  is a vector space over  $\mathbb{R}$ . In this vector space,

$$(3, -2, 0) + (-1, 1, 4) = (2, -1, 4) \quad \text{and} \quad -5(1, -2, 0) = (-5, 10, 0).$$

Similarly,  $\mathbb{C}^2$  is a vector space over  $\mathbb{C}$ . In this vector space,

$$(1+i, 2) + (2-3i, 4i) = (3-2i, 2+4i) \quad \text{and} \quad i(1+i, 2) = (-1+i, 2i).$$

Vectors in  $\mathbb{F}^n$  may be written as **column vectors**

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

rather than as **row vectors**  $(a_1, a_2, \dots, a_n)$ . Since a 1-tuple whose only entry is from  $F$  can be regarded as an element of  $F$ , we usually write  $F$  rather than  $\mathbb{F}^1$  for the vector space of 1-tuples with entry from  $F$ . ♦

An  $m \times n$  **matrix** with entries from a field  $F$  is a rectangular array of the form

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

where each entry  $a_{ij}$  ( $1 \leq i \leq m$ ,  $1 \leq j \leq n$ ) is an element of  $F$ . We call the entries  $a_{ij}$  with  $i = j$  the **diagonal entries** of the matrix. The entries  $a_{i1}, a_{i2}, \dots, a_{in}$  compose the  **$i$ th row** of the matrix, and the entries

$a_{1j}, a_{2j}, \dots, a_{mj}$  compose the ***j*th column** of the matrix. The rows of the preceding matrix are regarded as vectors in  $\mathbb{F}^n$ , and the columns are regarded as vectors in  $\mathbb{F}^m$ . Furthermore, we may regard a row vector in  $\mathbb{F}^n$  as a  $1 \times n$  matrix with entries from  $F$ , and we may regard a column vector in  $\mathbb{F}^m$  as an  $m \times 1$  matrix with entries from  $F$ .

The  $m \times n$  matrix in which each entry equals zero is called the **zero matrix** and is denoted by  $O$ .

In this book, we denote matrices by capital italic letters (e.g.,  $A$ ,  $B$ , and  $C$ ), and we denote the entry of a matrix  $A$  that lies in row  $i$  and column  $j$  by  $A_{ij}$ . In addition, if the number of rows and columns of a matrix are equal, the matrix is called **square**.

Two  $m \times n$  matrices  $A$  and  $B$  are called **equal** if all their corresponding entries are equal, that is, if  $A_{ij} = B_{ij}$  for  $1 \leq i \leq m$  and  $1 \leq j \leq n$ .

### Example 2

The set of all  $m \times n$  matrices with entries from a field  $F$  is a vector space, which we denote by  $M_{m \times n}(F)$ , with the following operations of **matrix addition** and **scalar multiplication**: For  $A, B \in M_{m \times n}(F)$  and  $c \in F$ ,

$$(A + B)_{ij} = A_{ij} + B_{ij} \quad \text{and} \quad (cA)_{ij} = cA_{ij}$$

for  $1 \leq i \leq m$  and  $1 \leq j \leq n$ . For instance,

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

and

$$-3 \begin{pmatrix} 1 & 0 & -2 \\ -3 & 2 & 3 \end{pmatrix} = \begin{pmatrix} -3 & 0 & 6 \\ 9 & -6 & -9 \end{pmatrix}$$

in  $M_{2 \times 3}(R)$ . ◆

Notice that the definitions of matrix addition and scalar multiplication in  $M_{m \times n}(F)$  are natural extensions of the corresponding operations in  $\mathbb{F}^n$  and  $\mathbb{F}^m$ . Thus the sum of two  $m \times n$  matrices  $A$  and  $B$  in  $M_{m \times n}(F)$  is the matrix in  $M_{m \times n}(F)$  whose  $i$ th row vector is the sum of the  $i$ th row vectors of  $A$  and  $B$ , and, for any scalar  $c$ , the matrix  $cA$  is the matrix in  $M_{m \times n}(F)$  whose  $i$ th row vector is  $c$  times the  $i$ th row vector of  $A$ . Likewise, the sum of two matrices  $A$  and  $B$  in  $M_{m \times n}(F)$  is the matrix in  $M_{m \times n}(F)$  whose  $j$ th column is the sum of the  $j$ th column vectors of  $A$  and  $B$ , and, for any scalar  $c$ , the matrix  $cA$  is the matrix in  $M_{m \times n}(F)$  whose  $j$ th column vector is  $c$  times the  $j$ th column vector of  $A$ .

### Example 3

Let  $S$  be any nonempty set and  $F$  be any field, and let  $\mathcal{F}(S, F)$  denote the set of all functions from  $S$  to  $F$ . Two functions  $f$  and  $g$  in  $\mathcal{F}(S, F)$  are called

**equal** if  $f(s) = g(s)$  for each  $s \in S$ . The set  $\mathcal{F}(S, F)$  is a vector space with the operations of addition and scalar multiplication defined for  $f, g \in \mathcal{F}(S, F)$  and  $c \in F$  by

$$(f + g)(s) = f(s) + g(s) \quad \text{and} \quad (cf)(s) = c[f(s)]$$

for each  $s \in S$ . Note that these are the familiar operations of addition and scalar multiplication for functions used in algebra and calculus. ♦

A **polynomial** with coefficients from a field  $F$  is an expression of the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where  $n$  is a nonnegative integer and each  $a_k$ , called the **coefficient** of  $x^k$ , is in  $F$ . If  $f(x) = 0$ , that is, if  $a_n = a_{n-1} = \cdots = a_0 = 0$ , then  $f(x)$  is called the **zero polynomial** and, for convenience, its degree is defined to be  $-1$ ; otherwise, the **degree** of a polynomial is defined to be the largest exponent of  $x$  that appears in the representation

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

with a nonzero coefficient. Note that the polynomials of degree zero may be written in the form  $f(x) = c$  for some nonzero scalar  $c$ . Two polynomials,

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

and

$$g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0,$$

are called **equal** if  $m = n$  and  $a_i = b_i$  for  $i = 0, 1, \dots, n$ .

When  $F$  is a field containing infinitely many scalars, we usually regard a polynomial with coefficients from  $F$  as a function from  $F$  into  $F$ . (See page 564.) In this case, the value of the function

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

at  $c \in F$  is the scalar

$$f(c) = a_n c^n + a_{n-1} c^{n-1} + \cdots + a_1 c + a_0.$$

Here either of the notations  $f$  or  $f(x)$  is used for the polynomial function

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0.$$

**Example 4**

Let

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

and

$$g(x) = b_mx^m + b_{m-1}x^{m-1} + \cdots + b_1x + b_0$$

be polynomials with coefficients from a field  $F$ . Suppose that  $m \leq n$ , and define  $b_{m+1} = b_{m+2} = \cdots = b_n = 0$ . Then  $g(x)$  can be written as

$$g(x) = b_nx^n + b_{n-1}x^{n-1} + \cdots + b_1x + b_0.$$

Define

$$f(x) + g(x) = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + (a_0 + b_0)$$

and for any  $c \in F$ , define

$$cf(x) = ca_nx^n + ca_{n-1}x^{n-1} + \cdots + ca_1x + ca_0.$$

With these operations of addition and scalar multiplication, the set of all polynomials with coefficients from  $F$  is a vector space, which we denote by  $\mathsf{P}(F)$ . ◆

We will see later that  $\mathsf{P}(F)$  is essentially the same as a subset of the vector space defined in the next example.

**Example 5**

Let  $F$  be any field. A **sequence** in  $F$  is a function  $\sigma$  from the positive integers into  $F$ . In this book, the sequence  $\sigma$  such that  $\sigma(n) = a_n$  for  $n = 1, 2, \dots$  is denoted  $(a_n)$ . Let  $\mathsf{V}$  consist of all the sequences  $(a_n)$  in  $F$ . For  $(a_n)$  and  $(b_n)$  in  $\mathsf{V}$  and  $t \in F$ , define

$$(a_n) + (b_n) = (a_n + b_n) \quad \text{and} \quad t(a_n) = (ta_n).$$

With these operations  $\mathsf{V}$  is a vector space. ◆

Our next two examples contain sets on which addition and scalar multiplication are defined, but which are *not* vector spaces.

**Example 6**

Let  $S = \{(a_1, a_2) : a_1, a_2 \in R\}$ . For  $(a_1, a_2), (b_1, b_2) \in S$  and  $c \in R$ , define

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 - b_2) \quad \text{and} \quad c(a_1, a_2) = (ca_1, ca_2).$$

Since (VS 1), (VS 2), and (VS 8) fail to hold,  $S$  is not a vector space with these operations. ◆

**Example 7**

Let  $S$  be as in Example 6. For  $(a_1, a_2), (b_1, b_2) \in S$  and  $c \in R$ , define

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, 0) \quad \text{and} \quad c(a_1, a_2) = (ca_1, 0).$$

Then  $S$  is not a vector space with these operations because (VS 3) (hence (VS 4)) and (VS 5) fail. ◆

We conclude this section with a few of the elementary consequences of the definition of a vector space.

**Theorem 1.1 (Cancellation Law for Vector Addition).** *If  $x$ ,  $y$ , and  $z$  are vectors in a vector space  $V$  such that  $x + z = y + z$ , then  $x = y$ .*

*Proof.* There exists a vector  $v$  in  $V$  such that  $z + v = 0$  (VS 4). Thus

$$\begin{aligned} x &= x + 0 = x + (z + v) = (x + z) + v \\ &= (y + z) + v = y + (z + v) = y + 0 = y \end{aligned}$$

by (VS 2) and (VS 3). ■

**Corollary 1.** *The vector  $0$  described in (VS 3) is unique.*

*Proof.* Exercise. ■

**Corollary 2.** *The vector  $y$  described in (VS 4) is unique.*

*Proof.* Exercise. ■

The vector  $0$  in (VS 3) is called the **zero vector** of  $V$ , and the vector  $y$  in (VS 4) (that is, the unique vector such that  $x + y = 0$ ) is called the **additive inverse** of  $x$  and is denoted by  $-x$ .

The next result contains some of the elementary properties of scalar multiplication.

**Theorem 1.2.** *In any vector space  $V$ , the following statements are true:*

- (a)  $0x = 0$  for each  $x \in V$ .
- (b)  $(-a)x = -(ax) = a(-x)$  for each  $a \in F$  and each  $x \in V$ .
- (c)  $a0 = 0$  for each  $a \in F$ .

*Proof.* (a) By (VS 8), (VS 3), and (VS 1), it follows that

$$0x + 0x = (0 + 0)x = 0x = 0x + 0 = 0 + 0x.$$

Hence  $0x = 0$  by Theorem 1.1.

(b) The vector  $-(ax)$  is the unique element of  $\mathbb{V}$  such that  $ax + [-(ax)] = 0$ . Thus if  $ax + (-a)x = 0$ , Corollary 2 to Theorem 1.1 implies that  $(-a)x = -(ax)$ . But by (VS 8),

$$ax + (-a)x = [a + (-a)]x = 0x = 0$$

by (a). Consequently  $(-a)x = -(ax)$ . In particular,  $(-1)x = -x$ . So, by (VS 6),

$$a(-x) = a[(-1)x] = [a(-1)]x = (-a)x.$$

The proof of (c) is similar to the proof of (a). ■

## EXERCISES

- 1.** Label the following statements as true or false.

- (a) Every vector space contains a zero vector.
- (b) A vector space may have more than one zero vector.
- (c) In any vector space,  $ax = bx$  implies that  $a = b$ .
- (d) In any vector space,  $ax = ay$  implies that  $x = y$ .
- (e) A vector in  $\mathbb{F}^n$  may be regarded as a matrix in  $M_{n \times 1}(F)$ .
- (f) An  $m \times n$  matrix has  $m$  columns and  $n$  rows.
- (g) In  $P(F)$ , only polynomials of the same degree may be added.
- (h) If  $f$  and  $g$  are polynomials of degree  $n$ , then  $f + g$  is a polynomial of degree  $n$ .
- (i) If  $f$  is a polynomial of degree  $n$  and  $c$  is a nonzero scalar, then  $cf$  is a polynomial of degree  $n$ .
- (j) A nonzero scalar of  $F$  may be considered to be a polynomial in  $P(F)$  having degree zero.
- (k) Two functions in  $\mathcal{F}(S, F)$  are equal if and only if they have the same value at each element of  $S$ .

- 2.** Write the zero vector of  $M_{3 \times 4}(F)$ .

- 3.** If

$$M = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

what are  $M_{13}$ ,  $M_{21}$ , and  $M_{22}$ ?

- 4.** Perform the indicated operations.

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

(b)  $\begin{pmatrix} -6 & 4 \\ 3 & -2 \\ 1 & 8 \end{pmatrix} + \begin{pmatrix} 7 & -5 \\ 0 & -3 \\ 2 & 0 \end{pmatrix}$

(c)  $4 \begin{pmatrix} 2 & 5 & -3 \\ 1 & 0 & 7 \end{pmatrix}$

- (d)  $-5 \begin{pmatrix} -6 & 4 \\ 3 & -2 \\ 1 & 8 \end{pmatrix}$
- (e)  $(2x^4 - 7x^3 + 4x + 3) + (8x^3 + 2x^2 - 6x + 7)$   
 (f)  $(-3x^3 + 7x^2 + 8x - 6) + (2x^3 - 8x + 10)$   
 (g)  $5(2x^7 - 6x^4 + 8x^2 - 3x)$   
 (h)  $3(x^5 - 2x^3 + 4x + 2)$

Exercises 5 and 6 show why the definitions of matrix addition and scalar multiplication (as defined in Example 2) are the appropriate ones.

5. Richard Gard (“Effects of Beaver on Trout in Sagehen Creek, California,” *J. Wildlife Management*, **25**, 221-242) reports the following number of trout having crossed beaver dams in Sagehen Creek.

#### Upstream Crossings

	Fall	Spring	Summer
Brook trout	8	3	1
Rainbow trout	3	0	0
Brown trout	3	0	0

#### Downstream Crossings

	Fall	Spring	Summer
Brook trout	9	1	4
Rainbow trout	3	0	0
Brown trout	1	1	0

Record the upstream and downstream crossings in two  $3 \times 3$  matrices, and verify that the sum of these matrices gives the total number of crossings (both upstream and downstream) categorized by trout species and season.

6. At the end of May, a furniture store had the following inventory.

	Early American	Spanish	Mediterranean	Danish
Living room suites	4	2	1	3
Bedroom suites	5	1	1	4
Dining room suites	3	1	2	6

Record these data as a  $3 \times 4$  matrix  $M$ . To prepare for its June sale, the store decided to double its inventory on each of the items listed in the preceding table. Assuming that none of the present stock is sold until the additional furniture arrives, verify that the inventory on hand after the order is filled is described by the matrix  $2M$ . If the inventory at the end of June is described by the matrix

$$A = \begin{pmatrix} 5 & 3 & 1 & 2 \\ 6 & 2 & 1 & 5 \\ 1 & 0 & 3 & 3 \end{pmatrix},$$

interpret  $2M - A$ . How many suites were sold during the June sale?

7. Let  $S = \{0, 1\}$  and  $F = R$ . In  $\mathcal{F}(S, R)$ , show that  $f = g$  and  $f + g = h$ , where  $f(t) = 2t + 1$ ,  $g(t) = 1 + 4t - 2t^2$ , and  $h(t) = 5t + 1$ .
8. In any vector space  $V$ , show that  $(a + b)(x + y) = ax + ay + bx + by$  for any  $x, y \in V$  and any  $a, b \in F$ .
9. Prove Corollaries 1 and 2 of Theorem 1.1 and Theorem 1.2(c). Visit [goo.gl/WFWgzX](http://goo.gl/WFWgzX) for a solution.
10. Let  $V$  denote the set of all differentiable real-valued functions defined on the real line. Prove that  $V$  is a vector space with the operations of addition and scalar multiplication defined in Example 3.
11. Let  $V = \{\theta\}$  consist of a single vector  $\theta$  and define  $\theta + \theta = \theta$  and  $c\theta = \theta$  for each scalar  $c$  in  $F$ . Prove that  $V$  is a vector space over  $F$ . ( $V$  is called the **zero vector space**.)
12. A real-valued function  $f$  defined on the real line is called an **even function** if  $f(-t) = f(t)$  for each real number  $t$ . Prove that the set of even functions defined on the real line with the operations of addition and scalar multiplication defined in Example 3 is a vector space.
13. Let  $V$  denote the set of ordered pairs of real numbers. If  $(a_1, a_2)$  and  $(b_1, b_2)$  are elements of  $V$  and  $c \in R$ , define

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 b_2) \quad \text{and} \quad c(a_1, a_2) = (ca_1, a_2).$$

Is  $V$  a vector space over  $R$  with these operations? Justify your answer.

14. Let  $V = \{(a_1, a_2, \dots, a_n) : a_i \in C \text{ for } i = 1, 2, \dots, n\}$ ; so  $V$  is a vector space over  $C$  by Example 1. Is  $V$  a vector space over the field of real numbers with the operations of coordinatewise addition and multiplication?

15. Let  $V = \{(a_1, a_2, \dots, a_n) : a_i \in R \text{ for } i = 1, 2, \dots, n\}$ ; so  $V$  is a vector space over  $R$  by Example 1. Is  $V$  a vector space over the field of complex numbers with the operations of coordinatewise addition and multiplication?
16. Let  $V$  denote the set of all  $m \times n$  matrices with real entries; so  $V$  is a vector space over  $R$  by Example 2. Let  $F$  be the field of rational numbers. Is  $V$  a vector space over  $F$  with the usual definitions of matrix addition and scalar multiplication?
17. Let  $V = \{(a_1, a_2) : a_1, a_2 \in F\}$ , where  $F$  is a field. Define addition of elements of  $V$  coordinatewise, and for  $c \in F$  and  $(a_1, a_2) \in V$ , define

$$c(a_1, a_2) = (a_1, 0).$$

Is  $V$  a vector space over  $F$  with these operations? Justify your answer.

18. Let  $V = \{(a_1, a_2) : a_1, a_2 \in R\}$ . For  $(a_1, a_2), (b_1, b_2) \in V$  and  $c \in R$ , define

$$(a_1, a_2) + (b_1, b_2) = (a_1 + 2b_1, a_2 + 3b_2) \quad \text{and} \quad c(a_1, a_2) = (ca_1, ca_2).$$

Is  $V$  a vector space over  $R$  with these operations? Justify your answer.

19. Let  $V = \{(a_1, a_2) : a_1, a_2 \in R\}$ . Define addition of elements of  $V$  coordinatewise, and for  $(a_1, a_2)$  in  $V$  and  $c \in R$ , define

$$c(a_1, a_2) = \begin{cases} (0, 0) & \text{if } c = 0 \\ \left(ca_1, \frac{a_2}{c}\right) & \text{if } c \neq 0. \end{cases}$$

Is  $V$  a vector space over  $R$  with these operations? Justify your answer.

20. Let  $V$  denote the set of all real-valued functions  $f$  defined on the real line such that  $f(1) = 0$ . Prove that  $V$  is a vector space with the operations of addition and scalar multiplication defined in Example 3.
21. Let  $V$  and  $W$  be vector spaces over a field  $F$ . Let

$$Z = \{(v, w) : v \in V \text{ and } w \in W\}.$$

Prove that  $Z$  is a vector space over  $F$  with the operations

$$(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2) \quad \text{and} \quad c(v_1, w_1) = (cv_1, cw_1).$$

22. How many matrices are there in the vector space  $M_{m \times n}(Z_2)$ ? (See Appendix C.)

### 1.3 SUBSPACES

In the study of any algebraic structure, it is of interest to examine subsets that possess the same structure as the set under consideration. The appropriate notion of substructure for vector spaces is introduced in this section.

**Definition.** A subset  $W$  of a vector space  $V$  over a field  $F$  is called a **subspace** of  $V$  if  $W$  is a vector space over  $F$  with the operations of addition and scalar multiplication defined on  $V$ .

In any vector space  $V$ , note that  $V$  and  $\{\theta\}$  are subspaces. The latter is called the **zero subspace** of  $V$ .

Fortunately it is not necessary to verify all of the vector space properties to prove that a subset is a subspace. Because properties (VS 1), (VS 2), (VS 5), (VS 6), (VS 7), and (VS 8) hold for all vectors in the vector space, these properties automatically hold for the vectors in any subset. Thus a subset  $W$  of a vector space  $V$  is a subspace of  $V$  if and only if the following four properties hold.

1.  $x+y \in W$  whenever  $x \in W$  and  $y \in W$ . ( $W$  is closed under addition.)
2.  $cx \in W$  whenever  $c \in F$  and  $x \in W$ . ( $W$  is closed under scalar multiplication.)
3.  $W$  has a zero vector.
4. Each vector in  $W$  has an additive inverse in  $W$ .

The next theorem shows that the zero vector of  $W$  must be the same as the zero vector of  $V$  and that property 4 is redundant.

**Theorem 1.3.** Let  $V$  be a vector space and  $W$  a subset of  $V$ . Then  $W$  is a subspace of  $V$  if and only if the following three conditions hold for the operations defined in  $V$ .

- (a)  $\theta \in W$ .
- (b)  $x+y \in W$  whenever  $x \in W$  and  $y \in W$ .
- (c)  $cx \in W$  whenever  $c \in F$  and  $x \in W$ .

*Proof.* If  $W$  is a subspace of  $V$ , then  $W$  is a vector space with the operations of addition and scalar multiplication defined on  $V$ . Hence conditions (b) and (c) hold, and there exists a vector  $\theta' \in W$  such that  $x + \theta' = x$  for each  $x \in W$ . But also  $x + \theta = x$ , and thus  $\theta' = \theta$  by Theorem 1.1 (p. 12). So condition (a) holds.

Conversely, if conditions (a), (b), and (c) hold, the discussion preceding this theorem shows that  $W$  is a subspace of  $V$  if the additive inverse of each vector in  $W$  lies in  $W$ . But if  $x \in W$ , then  $(-1)x \in W$  by condition (c), and  $-x = (-1)x$  by Theorem 1.2 (p. 12). Hence  $W$  is a subspace of  $V$ . ■

The preceding theorem provides a simple method for determining whether or not a given subset of a vector space is a subspace. Normally, it is this result that is used to prove that a subset is, in fact, a subspace.

The **transpose**  $A^t$  of an  $m \times n$  matrix  $A$  is the  $n \times m$  matrix obtained from  $A$  by interchanging the rows with the columns; that is,  $(A^t)_{ij} = A_{ji}$ . For example,

$$\begin{pmatrix} 1 & -2 & 3 \\ 0 & 5 & -1 \end{pmatrix}^t = \begin{pmatrix} 1 & 0 \\ -2 & 5 \\ 3 & -1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}^t = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}.$$

A **symmetric matrix** is a matrix  $A$  such that  $A^t = A$ . For example, the  $2 \times 2$  matrix displayed above is a symmetric matrix. Clearly, a symmetric matrix must be square. The set  $W$  of all symmetric matrices in  $M_{n \times n}(F)$  is a subspace of  $M_{n \times n}(F)$  since the conditions of Theorem 1.3 hold:

1. The zero matrix is equal to its transpose and hence belongs to  $W$ .

It is easily proved that for any matrices  $A$  and  $B$  and any scalars  $a$  and  $b$ ,  $(aA + bB)^t = aA^t + bB^t$ . (See Exercise 3.) Using this fact, we show that the set of symmetric matrices is closed under addition and scalar multiplication.

2. If  $A \in W$  and  $B \in W$ , then  $A^t = A$  and  $B^t = B$ . Thus  $(A + B)^t = A^t + B^t = A + B$ , so that  $A + B \in W$ .
3. If  $A \in W$ , then  $A^t = A$ . So for any  $a \in F$ , we have  $(aA)^t = aA^t = aA$ . Thus  $aA \in W$ .

The examples that follow provide further illustrations of the concept of a subspace. The first three are particularly important.

### Example 1

Let  $n$  be a nonnegative integer, and let  $P_n(F)$  consist of all polynomials in  $P(F)$  having degree less than or equal to  $n$ . Since the zero polynomial has degree  $-1$ , it is in  $P_n(F)$ . Moreover, the sum of two polynomials with degrees less than or equal to  $n$  is another polynomial of degree less than or equal to  $n$ , and the product of a scalar and a polynomial of degree less than or equal to  $n$  is a polynomial of degree less than or equal to  $n$ . So  $P_n(F)$  is closed under addition and scalar multiplication. It therefore follows from Theorem 1.3 that  $P_n(F)$  is a subspace of  $P(F)$ . ◆

### Example 2

Let  $C(R)$  denote the set of all continuous real-valued functions defined on  $R$ . Clearly  $C(R)$  is a subset of the vector space  $F(R, R)$  defined in Example 3 of Section 1.2. We claim that  $C(R)$  is a subspace of  $F(R, R)$ . First note

that the zero of  $\mathcal{F}(R, R)$  is the constant function defined by  $f(t) = 0$  for all  $t \in R$ . Since constant functions are continuous, we have  $f \in C(R)$ . Moreover, the sum of two continuous functions is continuous, and the product of a real number and a continuous function is continuous. So  $C(R)$  is closed under addition and scalar multiplication and hence is a subspace of  $\mathcal{F}(R, R)$  by Theorem 1.3. ♦

Two special types of matrices are frequently of interest. An  $m \times n$  matrix  $A$  is called **upper triangular** if all its entries lying below the diagonal entries are zero, that is, if  $A_{ij} = 0$  whenever  $i > j$ . An  $n \times n$  matrix  $M$  is called a **diagonal matrix** if  $M_{ij} = 0$  whenever  $i \neq j$ , that is, if all its nondiagonal entries are zero. For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 8 \end{pmatrix},$$

then  $A$  is an upper triangular  $3 \times 4$  matrix, and  $B$  is a  $3 \times 3$  diagonal matrix.

### Example 3

Clearly the zero matrix is a diagonal matrix because all of its entries are 0. Moreover, if  $A$  and  $B$  are diagonal  $n \times n$  matrices, then whenever  $i \neq j$ ,

$$(A + B)_{ij} = A_{ij} + B_{ij} = 0 + 0 = 0 \quad \text{and} \quad (cA)_{ij} = cA_{ij} = c0 = 0$$

for any scalar  $c$ . Hence  $A + B$  and  $cA$  are diagonal matrices for any scalar  $c$ . Therefore the set of diagonal matrices is a subspace of  $M_{n \times n}(F)$  by Theorem 1.3. ♦

### Example 4

The **trace** of an  $n \times n$  matrix  $M$ , denoted  $\text{tr}(M)$ , is the sum of the diagonal entries of  $M$ ; that is,

$$\text{tr}(M) = M_{11} + M_{22} + \cdots + M_{nn}.$$

It follows from Exercise 6 that the set of  $n \times n$  matrices having trace equal to zero is a subspace of  $M_{n \times n}(F)$ . ♦

### Example 5

The set of matrices in  $M_{m \times n}(R)$  having nonnegative entries is not a subspace of  $M_{m \times n}(R)$  because it is not closed under scalar multiplication (by negative scalars). ♦

The next theorem shows how to form a new subspace from other subspaces.

**Theorem 1.4.** Any intersection of subspaces of a vector space  $V$  is a subspace of  $V$ .

*Proof.* Let  $\mathcal{C}$  be a collection of subspaces of  $V$ , and let  $W$  denote the intersection of the subspaces in  $\mathcal{C}$ . Since every subspace contains the zero vector,  $0 \in W$ . Let  $a \in F$  and  $x, y \in W$ . Then  $x$  and  $y$  are contained in each subspace in  $\mathcal{C}$ . Because each subspace in  $\mathcal{C}$  is closed under addition and scalar multiplication, it follows that  $x + y$  and  $ax$  are contained in each subspace in  $\mathcal{C}$ . Hence  $x + y$  and  $ax$  are also contained in  $W$ , so that  $W$  is a subspace of  $V$  by Theorem 1.3. ■

Having shown that the intersection of subspaces of a vector space  $V$  is a subspace of  $V$ , it is natural to consider whether or not the union of subspaces of  $V$  is a subspace of  $V$ . It is easily seen that the union of subspaces must contain the zero vector and be closed under scalar multiplication, but in general the union of subspaces of  $V$  need not be closed under addition. In fact, it can be readily shown that the union of two subspaces of  $V$  is a subspace of  $V$  if and only if one of the subspaces contains the other. (See Exercise 19.) There is, however, a natural way to combine two subspaces  $W_1$  and  $W_2$  to obtain a subspace that contains both  $W_1$  and  $W_2$ . As we already have suggested, the key to finding such a subspace is to assure that it must be closed under addition. This idea is explored in Exercise 23.

## EXERCISES

1. Label the following statements as true or false.
  - (a) If  $V$  is a vector space and  $W$  is a subset of  $V$  that is a vector space, then  $W$  is a subspace of  $V$ .
  - (b) The empty set is a subspace of every vector space.
  - (c) If  $V$  is a vector space other than the zero vector space, then  $V$  contains a subspace  $W$  such that  $W \neq V$ .
  - (d) The intersection of any two subsets of  $V$  is a subspace of  $V$ .
  - (e) An  $n \times n$  diagonal matrix can never have more than  $n$  nonzero entries.
  - (f) The trace of a square matrix is the product of its diagonal entries.
  - (g) Let  $W$  be the  $xy$ -plane in  $\mathbb{R}^3$ ; that is,  $W = \{(a_1, a_2, 0) : a_1, a_2 \in \mathbb{R}\}$ . Then  $W = \mathbb{R}^2$ .
2. Determine the transpose of each of the matrices that follow. In addition, if the matrix is square, compute its trace.

$$(a) \begin{pmatrix} -4 & 2 \\ 5 & -1 \end{pmatrix}$$

$$(b) \begin{pmatrix} 0 & 8 & -6 \\ 3 & 4 & 7 \end{pmatrix}$$

(c) 
$$\begin{pmatrix} -3 & 9 \\ 0 & -2 \\ 6 & 1 \end{pmatrix}$$

(d) 
$$\begin{pmatrix} 10 & 0 & -8 \\ 2 & -4 & 3 \\ -5 & 7 & 6 \end{pmatrix}$$

(e) 
$$\begin{pmatrix} 1 & -1 & 3 & 5 \end{pmatrix}$$

(f) 
$$\begin{pmatrix} -2 & 5 & 1 & 4 \\ 7 & 0 & 1 & -6 \end{pmatrix}$$

(g) 
$$\begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix}$$

(h) 
$$\begin{pmatrix} -4 & 0 & 6 \\ 0 & 1 & -3 \\ 6 & -3 & 5 \end{pmatrix}$$

3. Prove that  $(aA + bB)^t = aA^t + bB^t$  for any  $A, B \in \mathbb{M}_{m \times n}(F)$  and any  $a, b \in F$ .
4. Prove that  $(A^t)^t = A$  for each  $A \in \mathbb{M}_{m \times n}(F)$ .
5. Prove that  $A + A^t$  is symmetric for any square matrix  $A$ .
6. Prove that  $\text{tr}(aA + bB) = a \text{tr}(A) + b \text{tr}(B)$  for any  $A, B \in \mathbb{M}_{n \times n}(F)$ .
7. Prove that diagonal matrices are symmetric matrices.
8. Determine whether the following sets are subspaces of  $\mathbb{R}^3$  under the operations of addition and scalar multiplication defined on  $\mathbb{R}^3$ . Justify your answers.
  - (a)  $W_1 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : a_1 = 3a_2 \text{ and } a_3 = -a_2\}$
  - (b)  $W_2 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : a_1 = a_3 + 2\}$
  - (c)  $W_3 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : 2a_1 - 7a_2 + a_3 = 0\}$
  - (d)  $W_4 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : a_1 - 4a_2 - a_3 = 0\}$
  - (e)  $W_5 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : a_1 + 2a_2 - 3a_3 = 1\}$
  - (f)  $W_6 = \{(a_1, a_2, a_3) \in \mathbb{R}^3 : 5a_1^2 - 3a_2^2 + 6a_3^2 = 0\}$
9. Let  $W_1$ ,  $W_3$ , and  $W_4$  be as in Exercise 8. Describe  $W_1 \cap W_3$ ,  $W_1 \cap W_4$ , and  $W_3 \cap W_4$ , and observe that each is a subspace of  $\mathbb{R}^3$ .
10. Prove that  $W_1 = \{(a_1, a_2, \dots, a_n) \in \mathbb{F}^n : a_1 + a_2 + \dots + a_n = 0\}$  is a subspace of  $\mathbb{F}^n$ , but  $W_2 = \{(a_1, a_2, \dots, a_n) \in \mathbb{F}^n : a_1 + a_2 + \dots + a_n = 1\}$  is not.
11. Is the set  $W = \{f(x) \in \mathbb{P}(F) : f(x) = 0 \text{ or } f(x) \text{ has degree } n\}$  a subspace of  $\mathbb{P}(F)$  if  $n \geq 1$ ? Justify your answer.
12. Prove that the set of  $m \times n$  upper triangular matrices is a subspace of  $\mathbb{M}_{m \times n}(F)$ .
13. Let  $S$  be a nonempty set and  $F$  a field. Prove that for any  $s_0 \in S$ ,  $\{f \in \mathcal{F}(S, F) : f(s_0) = 0\}$ , is a subspace of  $\mathcal{F}(S, F)$ .

14. Let  $S$  be a nonempty set and  $F$  a field. Let  $\mathcal{C}(S, F)$  denote the set of all functions  $f \in \mathcal{F}(S, F)$  such that  $f(s) = 0$  for all but a finite number of elements of  $S$ . Prove that  $\mathcal{C}(S, F)$  is a subspace of  $\mathcal{F}(S, F)$ .
15. Is the set of all differentiable real-valued functions defined on  $R$  a subspace of  $C(R)$ ? Justify your answer.
16. Let  $C^n(R)$  denote the set of all real-valued functions defined on the real line that have a continuous  $n$ th derivative. Prove that  $C^n(R)$  is a subspace of  $\mathcal{F}(R, R)$ .
17. Prove that a subset  $W$  of a vector space  $V$  is a subspace of  $V$  if and only if  $W \neq \emptyset$ , and, whenever  $a \in F$  and  $x, y \in W$ , then  $ax \in W$  and  $x + y \in W$ .
18. Prove that a subset  $W$  of a vector space  $V$  is a subspace of  $V$  if and only if  $0 \in W$  and  $ax + y \in W$  whenever  $a \in F$  and  $x, y \in W$ .
19. Let  $W_1$  and  $W_2$  be subspaces of a vector space  $V$ . Prove that  $W_1 \cup W_2$  is a subspace of  $V$  if and only if  $W_1 \subseteq W_2$  or  $W_2 \subseteq W_1$ .
- 20.<sup>†</sup> Prove that if  $W$  is a subspace of a vector space  $V$  and  $w_1, w_2, \dots, w_n$  are in  $W$ , then  $a_1w_1 + a_2w_2 + \dots + a_nw_n \in W$  for any scalars  $a_1, a_2, \dots, a_n$ . Visit [goo.gl/KTg35w](#) for a solution.
21. Let  $V$  denote the vector space of sequences in  $R$ , as defined in Example 5 of Section 1.2. Show that the set of convergent sequences  $(a_n)$  (that is, those for which  $\lim_{n \rightarrow \infty} a_n$  exists) is a subspace of  $V$ .
22. Let  $F_1$  and  $F_2$  be fields. A function  $g \in \mathcal{F}(F_1, F_2)$  is called an **even function** if  $g(-t) = g(t)$  for each  $t \in F_1$  and is called an **odd function** if  $g(-t) = -g(t)$  for each  $t \in F_1$ . Prove that the set of all even functions in  $\mathcal{F}(F_1, F_2)$  and the set of all odd functions in  $\mathcal{F}(F_1, F_2)$  are subspaces of  $\mathcal{F}(F_1, F_2)$ .

The following definitions are used in Exercises 23–30.

**Definition.** If  $S_1$  and  $S_2$  are nonempty subsets of a vector space  $V$ , then the **sum** of  $S_1$  and  $S_2$ , denoted  $S_1 + S_2$ , is the set  $\{x + y : x \in S_1 \text{ and } y \in S_2\}$ .

**Definition.** A vector space  $V$  is called the **direct sum** of  $W_1$  and  $W_2$  if  $W_1$  and  $W_2$  are subspaces of  $V$  such that  $W_1 \cap W_2 = \{0\}$  and  $W_1 + W_2 = V$ . We denote that  $V$  is the direct sum of  $W_1$  and  $W_2$  by writing  $V = W_1 \oplus W_2$ .

---

<sup>†</sup>A dagger means that this exercise is essential for a later section.

- 23.** Let  $W_1$  and  $W_2$  be subspaces of a vector space  $V$ .
- Prove that  $W_1 + W_2$  is a subspace of  $V$  that contains both  $W_1$  and  $W_2$ .
  - Prove that any subspace of  $V$  that contains both  $W_1$  and  $W_2$  must also contain  $W_1 + W_2$ .
- 24.** Show that  $\mathbb{F}^n$  is the direct sum of the subspaces

$$W_1 = \{(a_1, a_2, \dots, a_n) \in \mathbb{F}^n : a_n = 0\}$$

and

$$W_2 = \{(a_1, a_2, \dots, a_n) \in \mathbb{F}^n : a_1 = a_2 = \dots = a_{n-1} = 0\}.$$

- 25.** Let  $W_1$  denote the set of all polynomials  $f(x)$  in  $P(F)$  such that in the representation

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0,$$

we have  $a_i = 0$  whenever  $i$  is even. Likewise let  $W_2$  denote the set of all polynomials  $g(x)$  in  $P(F)$  such that in the representation

$$g(x) = b_mx^m + b_{m-1}x^{m-1} + \dots + b_1x + b_0,$$

we have  $b_i = 0$  whenever  $i$  is odd. Prove that  $P(F) = W_1 \oplus W_2$ .

- 26.** In  $M_{m \times n}(F)$  define  $W_1 = \{A \in M_{m \times n}(F) : A_{ij} = 0 \text{ whenever } i > j\}$  and  $W_2 = \{A \in M_{m \times n}(F) : A_{ij} = 0 \text{ whenever } i \leq j\}$ . ( $W_1$  is the set of all upper triangular matrices as defined on page 19.) Show that  $M_{m \times n}(F) = W_1 \oplus W_2$ .
- 27.** Let  $V$  denote the vector space of all upper triangular  $n \times n$  matrices (as defined on page 19), and let  $W_1$  denote the subspace of  $V$  consisting of all diagonal matrices. Define  $W_2 = \{A \in V : A_{ij} = 0 \text{ whenever } i \geq j\}$ . Show that  $V = W_1 \oplus W_2$ .
- 28.** A matrix  $M$  is called **skew-symmetric** if  $M^t = -M$ . Clearly, a skew-symmetric matrix is square. Let  $F$  be a field. Prove that the set  $W_1$  of all skew-symmetric  $n \times n$  matrices with entries from  $F$  is a subspace of  $M_{n \times n}(F)$ . Now assume that  $F$  is not of characteristic two (see page 549), and let  $W_2$  be the subspace of  $M_{n \times n}(F)$  consisting of all symmetric  $n \times n$  matrices. Prove that  $M_{n \times n}(F) = W_1 \oplus W_2$ .
- 29.** Let  $F$  be a field that is not of characteristic two. Define

$$W_1 = \{A \in M_{n \times n}(F) : A_{ij} = 0 \text{ whenever } i \leq j\}$$

and  $W_2$  to be the set of all symmetric  $n \times n$  matrices with entries from  $F$ . Both  $W_1$  and  $W_2$  are subspaces of  $M_{n \times n}(F)$ . Prove that  $M_{n \times n}(F) = W_1 \oplus W_2$ . Compare this exercise with Exercise 28.

30. Let  $W_1$  and  $W_2$  be subspaces of a vector space  $V$ . Prove that  $V$  is the direct sum of  $W_1$  and  $W_2$  if and only if each vector in  $V$  can be *uniquely* written as  $x_1 + x_2$ , where  $x_1 \in W_1$  and  $x_2 \in W_2$ .
31. Let  $W$  be a subspace of a vector space  $V$  over a field  $F$ . For any  $v \in V$  the set  $\{v\} + W = \{v + w : w \in W\}$  is called the **coset** of  $W$  containing  $v$ . It is customary to denote this coset by  $v + W$  rather than  $\{v\} + W$ .
- Prove that  $v + W$  is a subspace of  $V$  if and only if  $v \in W$ .
  - Prove that  $v_1 + W = v_2 + W$  if and only if  $v_1 - v_2 \in W$ .

Addition and scalar multiplication by scalars of  $F$  can be defined in the collection  $S = \{v + W : v \in V\}$  of all cosets of  $W$  as follows:

$$(v_1 + W) + (v_2 + W) = (v_1 + v_2) + W$$

for all  $v_1, v_2 \in V$  and

$$a(v + W) = av + W$$

for all  $v \in V$  and  $a \in F$ .

- Prove that the preceding operations are well defined; that is, show that if  $v_1 + W = v'_1 + W$  and  $v_2 + W = v'_2 + W$ , then

$$(v_1 + W) + (v_2 + W) = (v'_1 + W) + (v'_2 + W)$$

and

$$a(v_1 + W) = a(v'_1 + W)$$

for all  $a \in F$ .

- Prove that the set  $S$  is a vector space with the operations defined in (c). This vector space is called the **quotient space**  $V$  modulo  $W$  and is denoted by  $V/W$ .

## 1.4 LINEAR COMBINATIONS AND SYSTEMS OF LINEAR EQUATIONS

In Section 1.1, it was shown that the equation of the plane through three noncollinear points  $A$ ,  $B$ , and  $C$  in space is  $x = A + su + tv$ , where  $u$  and  $v$  denote the vectors beginning at  $A$  and ending at  $B$  and  $C$ , respectively, and  $s$  and  $t$  denote arbitrary real numbers. An important special case occurs when  $A$  is the origin. In this case, the equation of the plane simplifies to  $x = su + tv$ , and the set of all points in this plane is a subspace of  $\mathbb{R}^3$ . (This is proved as Theorem 1.5.) Expressions of the form  $su + tv$ , where  $s$  and  $t$  are scalars and  $u$  and  $v$  are vectors, play a central role in the theory of vector spaces. The appropriate generalization of such expressions is presented in the following definitions.

**Definitions.** Let  $V$  be a vector space and  $S$  a nonempty subset of  $V$ . A vector  $v \in V$  is called a **linear combination** of vectors of  $S$  if there exist a finite number of vectors  $u_1, u_2, \dots, u_n$  in  $S$  and scalars  $a_1, a_2, \dots, a_n$  in  $F$  such that  $v = a_1u_1 + a_2u_2 + \dots + a_nu_n$ . In this case we also say that  $v$  is a linear combination of  $u_1, u_2, \dots, u_n$  and call  $a_1, a_2, \dots, a_n$  the **coefficients** of the linear combination.

Observe that in any vector space  $V$ ,  $0v = \theta$  for each  $v \in V$ . Thus the zero vector is a linear combination of any nonempty subset of  $V$ .

TABLE 1.1 Vitamin Content of 100 Grams of Certain Foods

	A (units)	B <sub>1</sub> (mg)	B <sub>2</sub> (mg)	Niacin (mg)	C (mg)
Apple butter	0	0.01	0.02	0.2	2
Raw, unpared apples (freshly harvested)	90	0.03	0.02	0.1	4
Chocolate-coated candy with coconut center	0	0.02	0.07	0.2	0
Clams (meat only)	100	0.10	0.18	1.3	10
Cupcake from mix (dry form)	0	0.05	0.06	0.3	0
Cooked farina (unenriched)	(0) <sup>a</sup>	0.01	0.01	0.1	(0)
Jams and preserves	10	0.01	0.03	0.2	2
Coconut custard pie (baked from mix)	0	0.02	0.02	0.4	0
Raw brown rice	(0)	0.34	0.05	4.7	(0)
Soy sauce	0	0.02	0.25	0.4	0
Cooked spaghetti (unenriched)	0	0.01	0.01	0.3	0
Raw wild rice	(0)	0.45	0.63	6.2	(0)

Source: Bernice K. Watt and Annabel L. Merrill, *Composition of Foods* (Agriculture Handbook Number 8), Consumer and Food Economics Research Division, U.S. Department of Agriculture, Washington, D.C., 1963.

<sup>a</sup>Zeros in parentheses indicate that the amount of a vitamin present is either none or too small to measure.

### Example 1

Table 1.1 shows the vitamin content of 100 grams of 12 foods with respect to vitamins **A**, **B**<sub>1</sub> (thiamine), **B**<sub>2</sub> (riboflavin), niacin, and **C** (ascorbic acid).

The vitamin content of 100 grams of each food can be recorded as a column vector in  $\mathbb{R}^5$ —for example, the vitamin vector for apple butter is

$$\begin{pmatrix} 0.00 \\ 0.01 \\ 0.02 \\ 0.20 \\ 2.00 \end{pmatrix}.$$

Considering the vitamin vectors for cupcake, coconut custard pie, raw brown rice, soy sauce, and wild rice, we see that

$$\begin{pmatrix} 0.00 \\ 0.05 \\ 0.06 \\ 0.30 \\ 0.00 \end{pmatrix} + \begin{pmatrix} 0.00 \\ 0.02 \\ 0.02 \\ 0.40 \\ 0.00 \end{pmatrix} + \begin{pmatrix} 0.00 \\ 0.34 \\ 0.05 \\ 4.70 \\ 0.00 \end{pmatrix} + 2 \begin{pmatrix} 0.00 \\ 0.02 \\ 0.25 \\ 0.40 \\ 0.00 \end{pmatrix} = \begin{pmatrix} 0.00 \\ 0.45 \\ 0.63 \\ 6.20 \\ 0.00 \end{pmatrix}.$$

Thus the vitamin vector for wild rice is a linear combination of the vitamin vectors for cupcake, coconut custard pie, raw brown rice, and soy sauce. So 100 grams of cupcake, 100 grams of coconut custard pie, 100 grams of raw brown rice, and 200 grams of soy sauce provide exactly the same amounts of the five vitamins as 100 grams of raw wild rice. Similarly, since

$$2 \begin{pmatrix} 0.00 \\ 0.01 \\ 0.02 \\ 0.20 \\ 2.00 \end{pmatrix} + \begin{pmatrix} 90.00 \\ 0.03 \\ 0.02 \\ 0.10 \\ 4.00 \end{pmatrix} + \begin{pmatrix} 0.00 \\ 0.02 \\ 0.07 \\ 0.20 \\ 0.00 \end{pmatrix} + \begin{pmatrix} 0.00 \\ 0.01 \\ 0.01 \\ 0.10 \\ 0.00 \end{pmatrix} + \begin{pmatrix} 10.00 \\ 0.01 \\ 0.03 \\ 0.20 \\ 2.00 \end{pmatrix} + \begin{pmatrix} 0.00 \\ 0.01 \\ 0.01 \\ 0.30 \\ 0.00 \end{pmatrix} = \begin{pmatrix} 100.00 \\ 0.10 \\ 0.18 \\ 1.30 \\ 10.00 \end{pmatrix},$$

200 grams of apple butter, 100 grams of apples, 100 grams of chocolate candy, 100 grams of farina, 100 grams of jam, and 100 grams of spaghetti provide exactly the same amounts of the five vitamins as 100 grams of clams. ♦

Throughout Chapters 1 and 2 we encounter many different situations in which it is necessary to determine whether or not a vector can be expressed as a linear combination of other vectors, and if so, how. This question often reduces to the problem of solving a system of linear equations. In Chapter 3, we discuss a general method for using matrices to solve any system of linear equations. For now, we illustrate how to solve a system of linear equations by showing how to determine if the vector  $(2, 6, 8)$  can be expressed as a linear combination of

$$u_1 = (1, 2, 1), \quad u_2 = (-2, -4, -2), \quad u_3 = (0, 2, 3),$$

$$u_4 = (2, 0, -3), \quad \text{and} \quad u_5 = (-3, 8, 16).$$

Thus we must determine if there are scalars  $a_1, a_2, a_3, a_4$ , and  $a_5$  such that

$$\begin{aligned} (2, 6, 8) &= a_1 u_1 + a_2 u_2 + a_3 u_3 + a_4 u_4 + a_5 u_5 \\ &= a_1(1, 2, 1) + a_2(-2, -4, -2) + a_3(0, 2, 3) \\ &\quad + a_4(2, 0, -3) + a_5(-3, 8, 16) \\ &= (a_1 - 2a_2 + 2a_4 - 3a_5, 2a_1 - 4a_2 + 2a_3 + 8a_5, \\ &\quad a_1 - 2a_2 + 3a_3 - 3a_4 + 16a_5). \end{aligned}$$

Hence  $(2, 6, 8)$  can be expressed as a linear combination of  $u_1, u_2, u_3, u_4$ , and  $u_5$  if and only if there is a 5-tuple of scalars  $(a_1, a_2, a_3, a_4, a_5)$  satisfying the system of linear equations

$$\begin{aligned} a_1 - 2a_2 &+ 2a_4 - 3a_5 = 2 \\ 2a_1 - 4a_2 + 2a_3 &+ 8a_5 = 6 \\ a_1 - 2a_2 + 3a_3 - 3a_4 + 16a_5 &= 8, \end{aligned} \tag{1}$$

which is obtained by equating the corresponding coordinates in the preceding equation.

To solve system (1), we replace it by another system with the same solutions, but which is easier to solve. The procedure to be used expresses some of the unknowns in terms of others by eliminating certain unknowns from all the equations except one. To begin, we eliminate  $a_1$  from every equation except the first by adding  $-2$  times the first equation to the second and  $-1$  times the first equation to the third. The result is the following new system:

$$\begin{aligned} a_1 - 2a_2 &+ 2a_4 - 3a_5 = 2 \\ 2a_3 - 4a_4 + 14a_5 &= 2 \\ 3a_3 - 5a_4 + 19a_5 &= 6. \end{aligned} \tag{2}$$

In this case, it happened that while eliminating  $a_1$  from every equation except the first, we also eliminated  $a_2$  from every equation except the first. This need not happen in general. We now want to make the coefficient of  $a_3$  in the second equation equal to 1, and then eliminate  $a_3$  from the third equation. To do this, we first multiply the second equation by  $\frac{1}{2}$ , which produces

$$\begin{aligned} a_1 - 2a_2 &+ 2a_4 - 3a_5 = 2 \\ a_3 - 2a_4 + 7a_5 &= 1 \\ 3a_3 - 5a_4 + 19a_5 &= 6. \end{aligned}$$

Next we add  $-3$  times the second equation to the third, obtaining

$$\begin{aligned} a_1 - 2a_2 &+ 2a_4 - 3a_5 = 2 \\ a_3 - 2a_4 + 7a_5 &= 1 \\ a_4 - 2a_5 &= 3. \end{aligned} \tag{3}$$

We continue by eliminating  $a_4$  from every equation of (3) except the third. This yields

$$\begin{aligned} a_1 - 2a_2 &+ a_5 = -4 \\ a_3 &+ 3a_5 = 7 \\ a_4 - 2a_5 &= 3. \end{aligned} \tag{4}$$

System (4) is a system of the desired form: It is easy to solve for the first unknown present in each of the equations ( $a_1, a_3$ , and  $a_4$ ) in terms of the

other unknowns ( $a_2$  and  $a_5$ ). Rewriting system (4) in this form, we find that

$$\begin{aligned} a_1 &= 2a_2 - a_5 - 4 \\ a_3 &= \quad - 3a_5 + 7 \\ a_4 &= \quad 2a_5 + 3. \end{aligned}$$

Thus for any choice of scalars  $a_2$  and  $a_5$ , a vector of the form

$$(a_1, a_2, a_3, a_4, a_5) = (2a_2 - a_5 - 4, a_2, -3a_5 + 7, 2a_5 + 3, a_5)$$

is a solution to system (1). In particular, the vector  $(-4, 0, 7, 3, 0)$  obtained by setting  $a_2 = 0$  and  $a_5 = 0$  is a solution to (1). Therefore

$$(2, 6, 8) = -4u_1 + 0u_2 + 7u_3 + 3u_4 + 0u_5,$$

so that  $(2, 6, 8)$  is a linear combination of  $u_1, u_2, u_3, u_4$ , and  $u_5$ .

The procedure just illustrated uses three types of operations to simplify the original system:

1. interchanging the order of any two equations in the system;
2. multiplying any equation in the system by a nonzero constant;
3. adding a constant multiple of any equation to another equation in the system.

In Section 3.4, we prove that these operations do not change the set of solutions to the original system. Note that we employed these operations to obtain a system of equations that had the following properties:

1. The first nonzero coefficient in each equation is one.
2. If an unknown is the first unknown with a nonzero coefficient in some equation, then that unknown occurs with a zero coefficient in each of the other equations.
3. The first unknown with a nonzero coefficient in any equation has a larger subscript than the first unknown with a nonzero coefficient in any preceding equation.

To help clarify the meaning of these properties, note that none of the following systems meets these requirements.

$$\begin{array}{rcl} x_1 + 3x_2 & + & x_4 = 7 \\ 2x_3 - 5x_4 = -1 \end{array} \tag{5}$$

$$\begin{array}{rcl} x_1 - 2x_2 + 3x_3 & + & x_5 = -5 \\ x_3 & - & 2x_5 = 9 \\ x_4 + 3x_5 & = & 6 \end{array} \tag{6}$$

$$\begin{array}{rccc} x_1 & - 2x_3 & + & x_5 = 1 \\ & & x_4 - 6x_5 = 0 \\ x_2 + 5x_3 & & - 3x_5 = 2. \end{array} \quad (7)$$

Specifically, system (5) does not satisfy property 1 because the first nonzero coefficient in the second equation is 2; system (6) does not satisfy property 2 because  $x_3$ , the first unknown with a nonzero coefficient in the second equation, occurs with a nonzero coefficient in the first equation; and system (7) does not satisfy property 3 because  $x_2$ , the first unknown with a nonzero coefficient in the third equation, does not have a larger subscript than  $x_4$ , the first unknown with a nonzero coefficient in the second equation.

Once a system with properties 1, 2, and 3 has been obtained, it is easy to solve for some of the unknowns in terms of the others (as in the preceding example). *If, however, in the course of using operations 1, 2, and 3 a system containing an equation of the form  $0 = c$ , where  $c$  is nonzero, is obtained, then the original system has no solutions.* (See Example 2.)

We return to the study of systems of linear equations in Chapter 3. We discuss there the theoretical basis for this method of solving systems of linear equations and further simplify the procedure by use of matrices.

### Example 2

We claim that

$$2x^3 - 2x^2 + 12x - 6$$

is a linear combination of

$$x^3 - 2x^2 - 5x - 3 \quad \text{and} \quad 3x^3 - 5x^2 - 4x - 9$$

in  $\mathbb{P}_3(\mathbb{R})$ , but that

$$3x^3 - 2x^2 + 7x + 8$$

is not. In the first case we wish to find scalars  $a$  and  $b$  such that

$$\begin{aligned} 2x^3 - 2x^2 + 12x - 6 &= a(x^3 - 2x^2 - 5x - 3) + b(3x^3 - 5x^2 - 4x - 9) \\ &= (a + 3b)x^3 + (-2a - 5b)x^2 + (-5a - 4b)x + (-3a - 9b). \end{aligned}$$

Thus we are led to the following system of linear equations:

$$\begin{array}{rl} a + 3b &= 2 \\ -2a - 5b &= -2 \\ -5a - 4b &= 12 \\ -3a - 9b &= -6. \end{array}$$

Adding appropriate multiples of the first equation to the others in order to eliminate  $a$ , we find that

$$\begin{aligned} a + 3b &= 2 \\ b &= 2 \\ 11b &= 22 \\ 0b &= 0. \end{aligned}$$

Now adding the appropriate multiples of the second equation to the others yields

$$\begin{aligned} a &= -4 \\ b &= 2 \\ 0 &= 0 \\ 0 &= 0. \end{aligned}$$

Hence

$$2x^3 - 2x^2 + 12x - 6 = -4(x^3 - 2x^2 - 5x - 3) + 2(3x^3 - 5x^2 - 4x - 9).$$

In the second case, we wish to show that there are no scalars  $a$  and  $b$  for which

$$3x^3 - 2x^2 + 7x + 8 = a(x^3 - 2x^2 - 5x - 3) + b(3x^3 - 5x^2 - 4x - 9).$$

Using the preceding technique, we obtain a system of linear equations

$$\begin{aligned} a + 3b &= 3 \\ -2a - 5b &= -2 \\ -5a - 4b &= 7 \\ -3a - 9b &= 8. \end{aligned} \tag{8}$$

Eliminating  $a$  as before yields

$$\begin{aligned} a + 3b &= 3 \\ b &= 4 \\ 11b &= 22 \\ 0 &= 17. \end{aligned}$$

But the presence of the inconsistent equation  $0 = 17$  indicates that (8) has no solutions. Hence  $3x^3 - 2x^2 + 7x + 8$  is not a linear combination of  $x^3 - 2x^2 - 5x - 3$  and  $3x^3 - 5x^2 - 4x - 9$ . ◆

Throughout this book, we form the set of all linear combinations of some set of vectors. We now name such a set of linear combinations.

**Definition.** Let  $S$  be a nonempty subset of a vector space  $V$ . The **span** of  $S$ , denoted  $\text{span}(S)$ , is the set consisting of all linear combinations of the vectors in  $S$ . For convenience, we define  $\text{span}(\emptyset) = \{0\}$ .

In  $\mathbb{R}^3$ , for instance, the span of the set  $\{(1, 0, 0), (0, 1, 0)\}$  consists of all vectors in  $\mathbb{R}^3$  that have the form  $a(1, 0, 0) + b(0, 1, 0) = (a, b, 0)$  for some scalars  $a$  and  $b$ . Thus the span of  $\{(1, 0, 0), (0, 1, 0)\}$  contains all the points in the  $xy$ -plane. In this case, the span of the set is a subspace of  $\mathbb{R}^3$ . This fact is true in general.

**Theorem 1.5.** *The span of any subset  $S$  of a vector space  $V$  is a subspace of  $V$  that contains  $S$ . Moreover, any subspace of  $V$  that contains  $S$  must also contain the span of  $S$ .*

*Proof.* This result is immediate if  $S = \emptyset$  because  $\text{span}(\emptyset) = \{0\}$ , which is a subspace that contains  $S$  and is contained in any subspace of  $V$ .

If  $S \neq \emptyset$ , then  $S$  contains a vector  $z$ . So  $0z = 0$  is in  $\text{span}(S)$ . Let  $x, y \in \text{span}(S)$ . Then there exist vectors  $u_1, u_2, \dots, u_m, v_1, v_2, \dots, v_n$  in  $S$  and scalars  $a_1, a_2, \dots, a_m, b_1, b_2, \dots, b_n$  such that

$$x = a_1u_1 + a_2u_2 + \cdots + a_mu_m \quad \text{and} \quad y = b_1v_1 + b_2v_2 + \cdots + b_nv_n.$$

Then

$$x + y = a_1u_1 + a_2u_2 + \cdots + a_mu_m + b_1v_1 + b_2v_2 + \cdots + b_nv_n$$

and, for any scalar  $c$ ,

$$cx = (ca_1)u_1 + (ca_2)u_2 + \cdots + (ca_m)u_m$$

are clearly linear combinations of the vectors in  $S$ ; so  $x + y$  and  $cx$  are in  $\text{span}(S)$ . Thus  $\text{span}(S)$  is a subspace of  $V$ . Furthermore, if  $v \in S$ , then  $v = 1 \cdot v \in \text{span}(S)$ ; so the span of  $S$  contains  $S$ .

Now let  $W$  denote any subspace of  $V$  that contains  $S$ . If  $w \in \text{span}(S)$ , then  $w$  has the form  $w = c_1w_1 + c_2w_2 + \cdots + c_kw_k$  for some vectors  $w_1, w_2, \dots, w_k$  in  $S$  and some scalars  $c_1, c_2, \dots, c_k$ . Since  $S \subseteq W$ , we have  $w_1, w_2, \dots, w_k \in W$ . Therefore  $w = c_1w_1 + c_2w_2 + \cdots + c_kw_k$  is in  $W$  by Exercise 20 of Section 1.3. Because  $w$ , an arbitrary vector in  $\text{span}(S)$ , belongs to  $W$ , it follows that  $\text{span}(S) \subseteq W$ . ■

**Definition.** A subset  $S$  of a vector space  $V$  **generates** (or **spans**)  $V$  if  $\text{span}(S) = V$ . In this case, we also say that the vectors of  $S$  generate (or span)  $V$ .

### Example 3

The vectors  $(1, 1, 0)$ ,  $(1, 0, 1)$ , and  $(0, 1, 1)$  generate  $\mathbb{R}^3$  since an arbitrary vector  $(a_1, a_2, a_3)$  in  $\mathbb{R}^3$  is a linear combination of the three given vectors; in fact, the scalars  $r, s$ , and  $t$  for which

$$r(1, 1, 0) + s(1, 0, 1) + t(0, 1, 1) = (a_1, a_2, a_3)$$

are

$$r = \frac{1}{2}(a_1 + a_2 - a_3), \quad s = \frac{1}{2}(a_1 - a_2 + a_3), \quad \text{and} \quad t = \frac{1}{2}(-a_1 + a_2 + a_3). \quad \blacklozenge$$

### Example 4

The polynomials  $x^2 + 3x - 2$ ,  $2x^2 + 5x - 3$ , and  $-x^2 - 4x + 4$  generate  $P_2(R)$  since each of the three given polynomials belongs to  $P_2(R)$  and each polynomial  $ax^2 + bx + c$  in  $P_2(R)$  is a linear combination of these three, namely,

$$\begin{aligned} & (-8a + 5b + 3c)(x^2 + 3x - 2) + (4a - 2b - c)(2x^2 + 5x - 3) \\ & + (-a + b + c)(-x^2 - 4x + 4) = ax^2 + bx + c. \end{aligned} \quad \blacklozenge$$

### Example 5

The matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

generate  $M_{2 \times 2}(R)$  since an arbitrary matrix  $A$  in  $M_{2 \times 2}(R)$  can be expressed as a linear combination of the four given matrices as follows:

$$\begin{aligned} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} &= \left( \frac{1}{3}a_{11} + \frac{1}{3}a_{12} + \frac{1}{3}a_{21} - \frac{2}{3}a_{22} \right) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \\ &+ \left( \frac{1}{3}a_{11} + \frac{1}{3}a_{12} - \frac{2}{3}a_{21} + \frac{1}{3}a_{22} \right) \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &+ \left( \frac{1}{3}a_{11} - \frac{2}{3}a_{12} + \frac{1}{3}a_{21} + \frac{1}{3}a_{22} \right) \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \\ &+ \left( -\frac{2}{3}a_{11} + \frac{1}{3}a_{12} + \frac{1}{3}a_{21} + \frac{1}{3}a_{22} \right) \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}. \end{aligned}$$

On the other hand, the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

do not generate  $M_{2 \times 2}(R)$  because each of these matrices has equal diagonal entries. So any linear combination of these matrices has equal diagonal entries. Hence not every  $2 \times 2$  matrix is a linear combination of these three matrices.  $\blacklozenge$

At the beginning of this section we noted that the equation of a plane through three noncollinear points in space, one of which is the origin, is of the form  $x = su + tv$ , where  $u, v \in \mathbb{R}^3$  and  $s$  and  $t$  are scalars. Thus  $x \in \mathbb{R}^3$  is

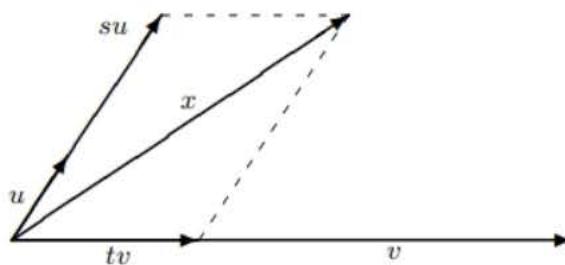


Figure 1.5

a linear combination of  $u, v \in \mathbb{R}^3$  if and only if  $x$  lies in the plane containing  $u$  and  $v$ . (See Figure 1.5.)

Usually there are many different subsets that generate a subspace  $W$ . (See Exercise 13.) It is natural to seek a subset of  $W$  that generates  $W$  and is as small as possible. In the next section we explore the circumstances under which a vector can be removed from a generating set to obtain a smaller generating set.

## EXERCISES

1. Label the following statements as true or false.
  - (a) The zero vector is a linear combination of any nonempty set of vectors.
  - (b) The span of  $\emptyset$  is  $\emptyset$ .
  - (c) If  $S$  is a subset of a vector space  $V$ , then  $\text{span}(S)$  equals the intersection of all subspaces of  $V$  that contain  $S$ .
  - (d) In solving a system of linear equations, it is permissible to multiply an equation by any constant.
  - (e) In solving a system of linear equations, it is permissible to add any multiple of one equation to another.
  - (f) Every system of linear equations has a solution.
2. Solve the following systems of linear equations by the method introduced in this section.
  - (a) 
$$\begin{array}{rcl} 2x_1 - 2x_2 - 3x_3 & = -2 \\ 3x_1 - 3x_2 - 2x_3 + 5x_4 & = 7 \\ x_1 - x_2 - 2x_3 - x_4 & = -3 \end{array}$$
  - (b) 
$$\begin{array}{rcl} 3x_1 - 7x_2 + 4x_3 & = 10 \\ x_1 - 2x_2 + x_3 & = 3 \\ 2x_1 - x_2 - 2x_3 & = 6 \end{array}$$
  - (c) 
$$\begin{array}{rcl} x_1 + 2x_2 - x_3 + x_4 & = 5 \\ x_1 + 4x_2 - 3x_3 - 3x_4 & = 6 \\ 2x_1 + 3x_2 - x_3 + 4x_4 & = 8 \end{array}$$

$$\begin{array}{lll}
 \text{(d)} & x_1 + 2x_2 + 2x_3 = 2 \\
 & x_1 + 8x_3 + 5x_4 = -6 \\
 & x_1 + x_2 + 5x_3 + 5x_4 = 3 \\
 \\ 
 \text{(e)} & x_1 + 2x_2 - 4x_3 - x_4 + x_5 = 7 \\
 & -x_1 + 10x_3 - 3x_4 - 4x_5 = -16 \\
 & 2x_1 + 5x_2 - 5x_3 - 4x_4 - x_5 = 2 \\
 & 4x_1 + 11x_2 - 7x_3 - 10x_4 - 2x_5 = 7 \\
 \\ 
 \text{(f)} & x_1 + 2x_2 + 6x_3 = -1 \\
 & 2x_1 + x_2 + x_3 = 8 \\
 & 3x_1 + x_2 - x_3 = 15 \\
 & x_1 + 3x_2 + 10x_3 = -5
 \end{array}$$

3. For each of the following lists of vectors in  $\mathbb{R}^3$ , determine whether the first vector can be expressed as a linear combination of the other two.
- (a)  $(-2, 0, 3), (1, 3, 0), (2, 4, -1)$
  - (b)  $(1, 2, -3), (-3, 2, 1), (2, -1, -1)$
  - (c)  $(3, 4, 1), (1, -2, 1), (-2, -1, 1)$
  - (d)  $(2, -1, 0), (1, 2, -3), (1, -3, 2)$
  - (e)  $(5, 1, -5), (1, -2, -3), (-2, 3, -4)$
  - (f)  $(-2, 2, 2), (1, 2, -1), (-3, -3, 3)$
4. For each list of polynomials in  $P_3(R)$ , determine whether the first polynomial can be expressed as a linear combination of the other two.
- (a)  $x^3 - 3x + 5, x^3 + 2x^2 - x + 1, x^3 + 3x^2 - 1$
  - (b)  $4x^3 + 2x^2 - 6, x^3 - 2x^2 + 4x + 1, 3x^3 - 6x^2 + x + 4$
  - (c)  $-2x^3 - 11x^2 + 3x + 2, x^3 - 2x^2 + 3x - 1, 2x^3 + x^2 + 3x - 2$
  - (d)  $x^3 + x^2 + 2x + 13, 2x^3 - 3x^2 + 4x + 1, x^3 - x^2 + 2x + 3$
  - (e)  $x^3 - 8x^2 + 4x, x^3 - 2x^2 + 3x - 1, x^3 - 2x + 3$
  - (f)  $6x^3 - 3x^2 + x + 2, x^3 - x^2 + 2x + 3, 2x^3 - 3x + 1$
5. In each part, determine whether the given vector is in the span of  $S$ .
- (a)  $(2, -1, 1), S = \{(1, 0, 2), (-1, 1, 1)\}$
  - (b)  $(-1, 2, 1), S = \{(1, 0, 2), (-1, 1, 1)\}$
  - (c)  $(-1, 1, 1, 2), S = \{(1, 0, 1, -1), (0, 1, 1, 1)\}$
  - (d)  $(2, -1, 1, -3), S = \{(1, 0, 1, -1), (0, 1, 1, 1)\}$
  - (e)  $-x^3 + 2x^2 + 3x + 3, S = \{x^3 + x^2 + x + 1, x^2 + x + 1, x + 1\}$
  - (f)  $2x^3 - x^2 + x + 3, S = \{x^3 + x^2 + x + 1, x^2 + x + 1, x + 1\}$
  - (g)  $\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix}, S = \left\{ \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right\}$
  - (h)  $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, S = \left\{ \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right\}$
6. Show that the vectors  $(1, 1, 0), (1, 0, 1)$ , and  $(0, 1, 1)$  generate  $\mathbb{F}^3$ .

7. In  $\mathbb{F}^n$ , let  $e_j$  denote the vector whose  $j$ th coordinate is 1 and whose other coordinates are 0. Prove that  $\{e_1, e_2, \dots, e_n\}$  generates  $\mathbb{F}^n$ .
8. Show that  $P_n(F)$  is generated by  $\{1, x, \dots, x^n\}$ .
9. Show that the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

generate  $M_{2 \times 2}(F)$ .

10. Show that if

$$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad M_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then the span of  $\{M_1, M_2, M_3\}$  is the set of all symmetric  $2 \times 2$  matrices.

- 11.<sup>†</sup> Prove that  $\text{span}(\{x\}) = \{ax : a \in F\}$  for any vector  $x$  in a vector space. Interpret this result geometrically in  $\mathbb{R}^3$ .
12. Show that a subset  $W$  of a vector space  $V$  is a subspace of  $V$  if and only if  $\text{span}(W) = W$ .
- 13.<sup>†</sup> Show that if  $S_1$  and  $S_2$  are subsets of a vector space  $V$  such that  $S_1 \subseteq S_2$ , then  $\text{span}(S_1) \subseteq \text{span}(S_2)$ . In particular, if  $S_1 \subseteq S_2$  and  $\text{span}(S_1) = V$ , deduce that  $\text{span}(S_2) = V$ . Visit [goo.gl/Fi8Epr](#) for a solution.
14. Show that if  $S_1$  and  $S_2$  are arbitrary subsets of a vector space  $V$ , then  $\text{span}(S_1 \cup S_2) = \text{span}(S_1) + \text{span}(S_2)$ . (The sum of two subsets is defined in the exercises of Section 1.3.)
15. Let  $S_1$  and  $S_2$  be subsets of a vector space  $V$ . Prove that  $\text{span}(S_1 \cap S_2) \subseteq \text{span}(S_1) \cap \text{span}(S_2)$ . Give an example in which  $\text{span}(S_1 \cap S_2)$  and  $\text{span}(S_1) \cap \text{span}(S_2)$  are equal and one in which they are unequal.
16. Let  $V$  be a vector space and  $S$  a subset of  $V$  with the property that whenever  $v_1, v_2, \dots, v_n \in S$  and  $a_1v_1 + a_2v_2 + \dots + a_nv_n = 0$ , then  $a_1 = a_2 = \dots = a_n = 0$ . Prove that every vector in the span of  $S$  can be *uniquely* written as a linear combination of vectors of  $S$ .
17. Let  $W$  be a subspace of a vector space  $V$ . Under what conditions are there only a finite number of distinct subsets  $S$  of  $W$  such that  $S$  generates  $W$ ?

## 1.5 LINEAR DEPENDENCE AND LINEAR INDEPENDENCE

Suppose that  $V$  is a vector space over an infinite field and that  $W$  is a subspace of  $V$ . Unless  $W$  is the zero subspace,  $W$  is an infinite set. It is desirable to find a “small” finite subset  $S$  of  $W$  that generates  $W$  because we can then describe each vector in  $W$  as a linear combination of the finite number of vectors in  $S$ . Indeed, the smaller  $S$  is, the fewer the number of computations required to represent vectors in  $W$  as such linear combinations. Consider, for example, the subspace  $W$  of  $\mathbb{R}^3$  generated by  $S = \{u_1, u_2, u_3, u_4\}$ , where  $u_1 = (2, -1, 4)$ ,  $u_2 = (1, -1, 3)$ ,  $u_3 = (1, 1, -1)$ , and  $u_4 = (1, -2, -1)$ . Let us attempt to find a proper subset of  $S$  that also generates  $W$ . The search for this subset is related to the question of whether or not some vector in  $S$  is a linear combination of the other vectors in  $S$ . Now  $u_4$  is a linear combination of the other vectors in  $S$  if and only if there are scalars  $a_1, a_2$ , and  $a_3$  such that

$$u_4 = a_1 u_1 + a_2 u_2 + a_3 u_3,$$

that is, if and only if there are scalars  $a_1, a_2$ , and  $a_3$  satisfying

$$(1, -2, -1) = (2a_1 + a_2 + a_3, -a_1 - a_2 + a_3, 4a_1 + 3a_2 - a_3).$$

Thus  $u_4$  is a linear combination of  $u_1, u_2$ , and  $u_3$  if and only if the system of linear equations

$$\begin{aligned} 2a_1 + a_2 + a_3 &= 1 \\ -a_1 - a_2 + a_3 &= -2 \\ 4a_1 + 3a_2 - a_3 &= -1 \end{aligned}$$

has a solution. The reader should verify that no such solution exists. This does not, however, answer our question of whether some vector in  $S$  is a linear combination of the other vectors in  $S$ . It can be shown, in fact, that  $u_3$  is a linear combination of  $u_1, u_2$ , and  $u_4$ , namely,  $u_3 = 2u_1 - 3u_2 + 0u_4$ .

In the preceding example, checking that some vector in  $S$  is a linear combination of the other vectors in  $S$  could require that we solve several different systems of linear equations before we determine which, if any, of  $u_1, u_2, u_3$ , and  $u_4$  is a linear combination of the others. By formulating our question differently, we can save ourselves some work. Note that since  $u_3 = 2u_1 - 3u_2 + 0u_4$ , we have

$$-2u_1 + 3u_2 + u_3 - 0u_4 = 0.$$

That is, because some vector in  $S$  is a linear combination of the others, the zero vector can be expressed as a linear combination of the vectors in  $S$  using coefficients that are not all zero. The converse of this statement is also true: If the zero vector can be written as a linear combination of the vectors in  $S$  in which not all the coefficients are zero, then some vector in  $S$  is a linear combination of the others. For instance, in the example above, the equation

$-2u_1 + 3u_2 + u_3 - 0u_4 = \theta$  can be solved for any one of  $u_1$ ,  $u_2$ , or  $u_3$  because each of these has a nonzero coefficient. Therefore any one of  $u_1$ ,  $u_2$ , or  $u_3$  can be written as a linear combination of the other three vectors. Thus, rather than asking whether some vector in  $S$  is a linear combination of the other vectors in  $S$ , it is more efficient to ask whether the zero vector can be expressed as a linear combination of the vectors in  $S$  with coefficients that are not all zero. This observation leads us to the following definition.

**Definition.** A subset  $S$  of a vector space  $V$  is called **linearly dependent** if there exist a finite number of distinct vectors  $u_1, u_2, \dots, u_n$  in  $S$  and scalars  $a_1, a_2, \dots, a_n$ , not all zero, such that

$$a_1u_1 + a_2u_2 + \cdots + a_nu_n = \theta.$$

In this case we also say that the vectors of  $S$  are linearly dependent.

For any vectors  $u_1, u_2, \dots, u_n$ , we have  $a_1u_1 + a_2u_2 + \cdots + a_nu_n = \theta$  if  $a_1 = a_2 = \cdots = a_n = 0$ . We call this the **trivial representation** of  $\theta$  as a linear combination of  $u_1, u_2, \dots, u_n$ . Thus, for a set to be linearly dependent, there must exist a nontrivial representation of  $\theta$  as a linear combination of vectors in the set. Consequently, any subset of a vector space that contains the zero vector is linearly dependent, because  $\theta = 1 \cdot \theta$  is a nontrivial representation of  $\theta$  as a linear combination of vectors in the set.

### Example 1

Consider the set

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

in  $\mathbb{R}^4$ . We show that  $S$  is linearly dependent and then express one of the vectors in  $S$  as a linear combination of the other vectors in  $S$ . To show that  $S$  is linearly dependent, we must find scalars  $a_1, a_2, a_3$ , and  $a_4$ , not all zero, such that

$$a_1(1, 3, -4, 2) + a_2(2, 2, -4, 0) + a_3(1, -3, 2, -4) + a_4(-1, 0, 1, 0) = \theta.$$

Finding such scalars amounts to finding a nonzero solution to the system of linear equations

$$\begin{aligned} a_1 + 2a_2 + a_3 - a_4 &= 0 \\ 3a_1 + 2a_2 - 3a_3 &= 0 \\ -4a_1 - 4a_2 + 2a_3 + a_4 &= 0 \\ 2a_1 - 4a_3 &= 0. \end{aligned}$$

One such solution is  $a_1 = 4$ ,  $a_2 = -3$ ,  $a_3 = 2$ , and  $a_4 = 0$ . Thus  $S$  is a linearly dependent subset of  $\mathbb{R}^4$ , and

$$4(1, 3, -4, 2) - 3(2, 2, -4, 0) + 2(1, -3, 2, -4) + 0(-1, 0, 1, 0) = \theta.$$

Hence

$$(1, 3, -4, 2) = \frac{3}{4}(2, 2, -4, 0) - \frac{1}{2}(1, -3, 2, -4) + 0(-1, 0, 1, 0). \quad \blacklozenge$$

### Example 2

In  $M_{2 \times 3}(R)$ , the set

$$\left\{ \begin{pmatrix} 1 & -3 & 2 \\ -4 & 0 & 5 \end{pmatrix}, \begin{pmatrix} -3 & 7 & 4 \\ 6 & -2 & -7 \end{pmatrix}, \begin{pmatrix} -2 & 3 & 11 \\ -1 & -3 & 2 \end{pmatrix} \right\}$$

is linearly dependent because

$$5 \begin{pmatrix} 1 & -3 & 2 \\ -4 & 0 & 5 \end{pmatrix} + 3 \begin{pmatrix} -3 & 7 & 4 \\ 6 & -2 & -7 \end{pmatrix} - 2 \begin{pmatrix} -2 & 3 & 11 \\ -1 & -3 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \blacklozenge$$

**Definition.** A subset  $S$  of a vector space that is not linearly dependent is called **linearly independent**. As before, we also say that the vectors of  $S$  are linearly independent.

The following facts about linearly independent sets are true in any vector space.

1. The empty set is linearly independent, for linearly dependent sets must be nonempty.
2. A set consisting of a single nonzero vector is linearly independent. For if  $\{u\}$  is linearly dependent, then  $au = 0$  for some nonzero scalar  $a$ . Thus

$$u = a^{-1}(au) = a^{-1}0 = 0.$$

3. A set is linearly independent if and only if the only representations of  $0$  as linear combinations of its vectors are trivial representations.

The condition in item 3 provides a useful method for determining whether a finite set is linearly independent. This technique is illustrated in the examples that follow.

### Example 3

To prove that the set

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

is linearly independent, we must show that the only linear combination of vectors in  $S$  that equals the zero vector is the one in which all the coefficients are zero. Suppose that  $a_1, a_2, a_3$ , and  $a_4$  are scalars such that

$$a_1(1, 0, 0, -1) + a_2(0, 1, 0, -1) + a_3(0, 0, 1, -1) + a_4(0, 0, 0, 1) = (0, 0, 0, 0).$$

Equating the corresponding coordinates of the vectors on the left and the right sides of this equation, we obtain the following system of linear equations.

$$\begin{array}{rl} a_1 & = 0 \\ a_2 & = 0 \\ a_3 & = 0 \\ -a_1 - a_2 - a_3 + a_4 & = 0 \end{array}$$

Clearly the only solution to this system is  $a_1 = a_2 = a_3 = a_4 = 0$ , and so  $S$  is linearly independent. ♦

#### Example 4

For  $k = 0, 1, \dots, n$  let  $p_k(x) = x^k + x^{k+1} + \dots + x^n$ . The set

$$\{p_0(x), p_1(x), \dots, p_n(x)\}$$

is linearly independent in  $P_n(F)$ . For if

$$a_0 p_0(x) + a_1 p_1(x) + \dots + a_n p_n(x) = 0$$

for some scalars  $a_0, a_1, \dots, a_n$ , then

$$a_0 + (a_0 + a_1)x + (a_0 + a_1 + a_2)x^2 + \dots + (a_0 + a_1 + \dots + a_n)x^n = 0.$$

By equating the coefficients of  $x^k$  on both sides of this equation for  $k = 1, 2, \dots, n$ , we obtain

$$\begin{array}{rl} a_0 & = 0 \\ a_0 + a_1 & = 0 \\ a_0 + a_1 + a_2 & = 0 \\ \vdots & \\ a_0 + a_1 + a_2 + \dots + a_n & = 0. \end{array}$$

Clearly the only solution to this system of linear equations is  $a_0 = a_1 = \dots = a_n = 0$ . ♦

The following important results are immediate consequences of the definitions of linear dependence and linear independence.

**Theorem 1.6.** *Let  $V$  be a vector space, and let  $S_1 \subseteq S_2 \subseteq V$ . If  $S_1$  is linearly dependent, then  $S_2$  is linearly dependent.*

*Proof.* Exercise. ■

**Corollary.** *Let  $V$  be a vector space, and let  $S_1 \subseteq S_2 \subseteq V$ . If  $S_2$  is linearly independent, then  $S_1$  is linearly independent.*

*Proof.* Exercise. ■

Earlier in this section, we remarked that the issue of whether  $S$  is a minimal generating set for its span (that is, one such that no proper subset of  $S$  is a generating set) is related to the question of whether some vector in  $S$  is a linear combination of the other vectors in  $S$ . Thus the issue of whether  $S$  is the smallest generating set for its span is related to the question of whether  $S$  is linearly dependent. To see why, consider the subset  $S = \{u_1, u_2, u_3, u_4\}$  of  $\mathbb{R}^3$ , where  $u_1 = (2, -1, 4)$ ,  $u_2 = (1, -1, 3)$ ,  $u_3 = (1, 1, -1)$ , and  $u_4 = (1, -2, -1)$ . We have previously noted that  $S$  is linearly dependent; in fact,

$$-2u_1 + 3u_2 + u_3 - 0u_4 = \theta.$$

This equation implies that  $u_3$  (or alternatively,  $u_1$  or  $u_2$ ) is a linear combination of the other vectors in  $S$ . For example,  $u_3 = 2u_1 - 3u_2 + 0u_4$ . Therefore every linear combination  $a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4$  of vectors in  $S$  can be written as a linear combination of  $u_1, u_2$ , and  $u_4$ :

$$\begin{aligned} a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4 &= a_1u_1 + a_2u_2 + a_3(2u_1 - 3u_2 + 0u_4) + a_4u_4 \\ &= (a_1 + 2a_3)u_1 + (a_2 - 3a_3)u_2 + a_4u_4. \end{aligned}$$

Thus the subset  $S' = \{u_1, u_2, u_4\}$  of  $S$  has the same span as  $S$ !

More generally, suppose that  $S$  is any linearly dependent set containing two or more vectors. Then some vector  $v \in S$  can be written as a linear combination of the other vectors in  $S$ , and the subset obtained by removing  $v$  from  $S$  has the same span as  $S$ . It follows that if no proper subset of  $S$  generates the span of  $S$ , then  $S$  must be linearly independent. Another way to view the preceding statement is given in Theorem 1.7.

**Theorem 1.7.** *Let  $S$  be a linearly independent subset of a vector space  $V$ , and let  $v$  be a vector in  $V$  that is not in  $S$ . Then  $S \cup \{v\}$  is linearly dependent if and only if  $v \in \text{span}(S)$ .*

*Proof.* If  $S \cup \{v\}$  is linearly dependent, then there are vectors  $u_1, u_2, \dots, u_n$  in  $S \cup \{v\}$  such that  $a_1u_1 + a_2u_2 + \dots + a_nu_n = \theta$  for some nonzero scalars  $a_1, a_2, \dots, a_n$ . Since  $S$  is linearly independent, one of the  $u_i$ 's, say  $u_1$ , equals  $v$ . Thus  $a_1v + a_2u_2 + \dots + a_nu_n = \theta$ , and so

$$v = a_1^{-1}(-a_2u_2 - \dots - a_nu_n) = -(a_1^{-1}a_2)u_2 - \dots - (a_1^{-1}a_n)u_n.$$

Because  $v$  is a linear combination of  $u_2, \dots, u_n$ , which are in  $S$ , we have  $v \in \text{span}(S)$ .

Conversely, let  $v \in \text{span}(S)$ . Then there exist vectors  $v_1, v_2, \dots, v_m$  in  $S$  and scalars  $b_1, b_2, \dots, b_m$  such that  $v = b_1v_1 + b_2v_2 + \dots + b_mv_m$ . Therefore

$$\theta = b_1v_1 + b_2v_2 + \dots + b_mv_m + (-1)v.$$

Note that  $v \neq v_i$  for  $i = 1, 2, \dots, m$  because  $v \notin S$ . Hence the coefficient of  $v$  in this linear combination is nonzero, and so the set  $\{v_1, v_2, \dots, v_m, v\}$  is linearly dependent. Thus  $S \cup \{v\}$  is linearly dependent by Theorem 1.6. ■

Linearly independent generating sets are investigated in detail in Section 1.6.

## EXERCISES

1. Label the following statements as true or false.
  - (a) If  $S$  is a linearly dependent set, then each vector in  $S$  is a linear combination of other vectors in  $S$ .
  - (b) Any set containing the zero vector is linearly dependent.
  - (c) The empty set is linearly dependent.
  - (d) Subsets of linearly dependent sets are linearly dependent.
  - (e) Subsets of linearly independent sets are linearly independent.
  - (f) If  $a_1x_1 + a_2x_2 + \dots + a_nx_n = \theta$  and  $x_1, x_2, \dots, x_n$  are linearly independent, then all the scalars  $a_i$  are zero.
- 2.<sup>3</sup> Determine whether the following sets are linearly dependent or linearly independent.
  - (a)  $\left\{ \begin{pmatrix} 1 & -3 \\ -2 & 4 \end{pmatrix}, \begin{pmatrix} -2 & 6 \\ 4 & -8 \end{pmatrix} \right\}$  in  $M_{2 \times 2}(R)$
  - (b)  $\left\{ \begin{pmatrix} 1 & -2 \\ -1 & 4 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ 2 & -4 \end{pmatrix} \right\}$  in  $M_{2 \times 2}(R)$
  - (c)  $\{x^3 + 2x^2, -x^2 + 3x + 1, x^3 - x^2 + 2x - 1\}$  in  $P_3(R)$
  - (d)  $\{x^3 - x, 2x^2 + 4, -2x^3 + 3x^2 + 2x + 6\}$  in  $P_3(R)$
  - (e)  $\{(1, -1, 2), (1, -2, 1), (1, 1, 4)\}$  in  $R^3$
  - (f)  $\{(1, -1, 2), (2, 0, 1), (-1, 2, -1)\}$  in  $R^3$
  - (g)  $\left\{ \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 2 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ -4 & 4 \end{pmatrix} \right\}$  in  $M_{2 \times 2}(R)$
  - (h)  $\left\{ \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 2 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 2 & -2 \end{pmatrix} \right\}$  in  $M_{2 \times 2}(R)$
  - (i)  $\{x^4 - x^3 + 5x^2 - 8x + 6, -x^4 + x^3 - 5x^2 + 5x - 3,$   
 $x^4 + 3x^2 - 3x + 5, 2x^4 + 3x^3 + 4x^2 - x + 1, x^3 - x + 2\}$  in  $P_4(R)$
  - (j)  $\{x^4 - x^3 + 5x^2 - 8x + 6, -x^4 + x^3 - 5x^2 + 5x - 3,$   
 $x^4 + 3x^2 - 3x + 5, 2x^4 + x^3 + 4x^2 + 8x\}$  in  $P_4(R)$

---

<sup>3</sup>The computations in Exercise 2(g), (h), (i), and (j) are tedious unless technology is used.

3. In  $M_{3 \times 2}(F)$ , prove that the set

$$\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \right\}$$

is linearly dependent.

4. In  $F^n$ , let  $e_j$  denote the vector whose  $j$ th coordinate is 1 and whose other coordinates are 0. Prove that  $\{e_1, e_2, \dots, e_n\}$  is linearly independent.
5. Show that the set  $\{1, x, x^2, \dots, x^n\}$  is linearly independent in  $P_n(F)$ .
6. In  $M_{m \times n}(F)$ , let  $E^{ij}$  denote the matrix whose only nonzero entry is 1 in the  $i$ th row and  $j$ th column. Prove that  $\{E^{ij} : 1 \leq i \leq m, 1 \leq j \leq n\}$  is linearly independent.
7. Recall from Example 3 in Section 1.3 that the set of diagonal matrices in  $M_{2 \times 2}(F)$  is a subspace. Find a linearly independent set that generates this subspace.
8. Let  $S = \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}$  be a subset of the vector space  $F^3$ .
  - Prove that if  $F = R$ , then  $S$  is linearly independent.
  - Prove that if  $F$  has characteristic two, then  $S$  is linearly dependent.
- 9.<sup>†</sup> Let  $u$  and  $v$  be distinct vectors in a vector space  $V$ . Show that  $\{u, v\}$  is linearly dependent if and only if  $u$  or  $v$  is a multiple of the other.
10. Give an example of three linearly dependent vectors in  $R^3$  such that none of the three is a multiple of another.
11. Let  $S = \{u_1, u_2, \dots, u_n\}$  be a linearly independent subset of a vector space  $V$  over the field  $Z_2$ . How many vectors are there in  $\text{span}(S)$ ? Justify your answer.
12. Prove Theorem 1.6 and its corollary.
13. Let  $V$  be a vector space over a field of characteristic not equal to two.
  - Let  $u$  and  $v$  be distinct vectors in  $V$ . Prove that  $\{u, v\}$  is linearly independent if and only if  $\{u + v, u - v\}$  is linearly independent.
  - Let  $u$ ,  $v$ , and  $w$  be distinct vectors in  $V$ . Prove that  $\{u, v, w\}$  is linearly independent if and only if  $\{u + v, u + w, v + w\}$  is linearly independent.
14. Prove that a set  $S$  is linearly dependent if and only if  $S = \{\mathbf{0}\}$  or there exist distinct vectors  $v, u_1, u_2, \dots, u_n$  in  $S$  such that  $v$  is a linear combination of  $u_1, u_2, \dots, u_n$ .

15. Let  $S = \{u_1, u_2, \dots, u_n\}$  be a finite set of vectors. Prove that  $S$  is linearly dependent if and only if  $u_1 = 0$  or  $u_{k+1} \in \text{span}(\{u_1, u_2, \dots, u_k\})$  for some  $k$  ( $1 \leq k < n$ ).
16. Prove that a set  $S$  of vectors is linearly independent if and only if each finite subset of  $S$  is linearly independent.
17. Let  $M$  be a square upper triangular matrix (as defined on page 19 of Section 1.3) with nonzero diagonal entries. Prove that the columns of  $M$  are linearly independent.
18. Let  $S$  be a set of nonzero polynomials in  $\mathbb{P}(F)$  such that no two have the same degree. Prove that  $S$  is linearly independent.
19. Prove that if  $\{A_1, A_2, \dots, A_k\}$  is a linearly independent subset of  $\mathbb{M}_{n \times n}(F)$ , then  $\{A_1^t, A_2^t, \dots, A_k^t\}$  is also linearly independent.
20. Let  $f, g \in \mathcal{F}(R, R)$  be the functions defined by  $f(t) = e^{rt}$  and  $g(t) = e^{st}$ , where  $r \neq s$ . Prove that  $f$  and  $g$  are linearly independent in  $\mathcal{F}(R, R)$ .
21. Let  $S_1$  and  $S_2$  be disjoint linearly independent subsets of  $V$ . Prove that  $S_1 \cup S_2$  is linearly dependent if and only if  $\text{span}(S_1) \cap \text{span}(S_2) \neq \{0\}$ . Visit [goo.gl/Fi8Epr](#) for a solution.

## 1.6 BASES AND DIMENSION

We saw in Section 1.5 that if  $S$  is a generating set for a subspace  $W$  and no proper subset of  $S$  is a generating set for  $W$ , then  $S$  must be linearly independent. A linearly independent generating set for  $W$  possesses a very useful property—every vector in  $W$  can be expressed in one and only one way as a linear combination of the vectors in the set. (This property is proved below in Theorem 1.8.) It is this property that makes linearly independent generating sets the building blocks of vector spaces.

**Definition.** A *basis*  $\beta$  for a vector space  $V$  is a linearly independent subset of  $V$  that generates  $V$ . If  $\beta$  is a basis for  $V$ , we also say that the vectors of  $\beta$  form a basis for  $V$ .

### Example 1

Recalling that  $\text{span}(\emptyset) = \{0\}$  and  $\emptyset$  is linearly independent, we see that  $\emptyset$  is a basis for the zero vector space. ♦

### Example 2

In  $\mathbb{F}^n$ , let  $e_1 = (1, 0, 0, \dots, 0), e_2 = (0, 1, 0, \dots, 0), \dots, e_n = (0, 0, \dots, 0, 1)$ ;  $\{e_1, e_2, \dots, e_n\}$  is readily seen to be a basis for  $\mathbb{F}^n$  and is called the **standard basis** for  $\mathbb{F}^n$ . ♦

**Example 3**

In  $M_{m \times n}(F)$ , let  $E^{ij}$  denote the matrix whose only nonzero entry is a 1 in the  $i$ th row and  $j$ th column. Then  $\{E^{ij} : 1 \leq i \leq m, 1 \leq j \leq n\}$  is a basis for  $M_{m \times n}(F)$ .  $\blacklozenge$

**Example 4**

In  $P_n(F)$ , the set  $\{1, x, x^2, \dots, x^n\}$  is a basis. We call this basis the **standard basis** for  $P_n(F)$ .  $\blacklozenge$

**Example 5**

In  $P(F)$ , the set  $\{1, x, x^2, \dots\}$  is a basis.  $\blacklozenge$

Observe that Example 5 shows that a basis need not be finite. In fact, later in this section it is shown that no basis for  $P(F)$  can be finite. Hence not every vector space has a finite basis.

The next theorem, which is used frequently in Chapter 2, establishes the most significant property of a basis.

**Theorem 1.8.** *Let  $V$  be a vector space and  $u_1, u_2, \dots, u_n$  be distinct vectors in  $V$ . Then  $\beta = \{u_1, u_2, \dots, u_n\}$  is a basis for  $V$  if and only if each  $v \in V$  can be uniquely expressed as a linear combination of vectors of  $\beta$ , that is, can be expressed in the form*

$$v = a_1u_1 + a_2u_2 + \cdots + a_nu_n$$

for unique scalars  $a_1, a_2, \dots, a_n$ .

*Proof.* Let  $\beta$  be a basis for  $V$ . If  $v \in V$ , then  $v \in \text{span}(\beta)$  because  $\text{span}(\beta) = V$ . Thus  $v$  is a linear combination of the vectors of  $\beta$ . Suppose that

$$v = a_1u_1 + a_2u_2 + \cdots + a_nu_n \quad \text{and} \quad v = b_1u_1 + b_2u_2 + \cdots + b_nu_n$$

are two such representations of  $v$ . Subtracting the second equation from the first gives

$$0 = (a_1 - b_1)u_1 + (a_2 - b_2)u_2 + \cdots + (a_n - b_n)u_n.$$

Since  $\beta$  is linearly independent, it follows that  $a_1 - b_1 = a_2 - b_2 = \cdots = a_n - b_n = 0$ . Hence  $a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$ , and so  $v$  is uniquely expressible as a linear combination of the vectors of  $\beta$ .  $\blacksquare$

The proof of the converse is an exercise.  $\blacksquare$

Theorem 1.8 shows that if the vectors  $u_1, u_2, \dots, u_n$  form a basis for a vector space  $V$ , then every vector in  $V$  can be uniquely expressed in the form

$$v = a_1u_1 + a_2u_2 + \cdots + a_nu_n$$

for appropriately chosen scalars  $a_1, a_2, \dots, a_n$ . Thus  $v$  determines a unique  $n$ -tuple of scalars  $(a_1, a_2, \dots, a_n)$  and, conversely, each  $n$ -tuple of scalars determines a unique vector  $v \in V$  by using the entries of the  $n$ -tuple as the coefficients of a linear combination of  $u_1, u_2, \dots, u_n$ . This fact suggests that  $V$  is like the vector space  $\mathbb{F}^n$ , where  $n$  is the number of vectors in the basis for  $V$ . We see in Section 2.4 that this is indeed the case.

In this book, we are primarily interested in vector spaces having finite bases. Theorem 1.9 identifies a large class of vector spaces of this type.

**Theorem 1.9.** *If a vector space  $V$  is generated by a finite set  $S$ , then some subset of  $S$  is a basis for  $V$ . Hence  $V$  has a finite basis.*

*Proof.* If  $S = \emptyset$  or  $S = \{0\}$ , then  $V = \{0\}$  and  $\emptyset$  is a subset of  $S$  that is a basis for  $V$ . Otherwise  $S$  contains a nonzero vector  $u_1$ . By item 2 on page 38,  $\{u_1\}$  is a linearly independent set. Continue, if possible, choosing vectors  $u_2, \dots, u_k$  in  $S$  such that  $\{u_1, u_2, \dots, u_k\}$  is a linearly independent set of  $k$  vectors. Since  $S$  is a finite set, this process must end with a linearly independent set  $\beta = \{u_1, u_2, \dots, u_n\}$ . There are two ways this could happen.

(i) The set  $\beta = S$ . In this case,  $S$  is both a linearly independent set and a generating set for  $V$ . That is,  $S$  is itself a basis for  $V$ .

(ii) The set  $\beta$  is a proper linearly independent subset of  $S$  such that adjoining to  $\beta$  any vector in  $S$  not in  $\beta$  produces a linearly dependent set. In this case, we claim that  $\beta$  is the desired subset of  $S$  that is a basis for  $V$ . Because  $\beta$  is linearly independent by construction, it suffices to show that  $\beta$  spans  $V$ . By Theorem 1.5 (p. 31), we need to show that  $S \subseteq \text{span}(\beta)$ . Let  $v \in S$ . If  $v \in \beta$ , then clearly  $v \in \text{span}(\beta)$ . Otherwise, if  $v \notin \beta$ , then the preceding construction shows that  $\beta \cup \{v\}$  is linearly dependent. So  $v \in \text{span}(\beta)$  by Theorem 1.7 (p. 40). Thus  $S \subseteq \text{span}(\beta)$ , completing the proof. ■

Because of the method by which the basis  $\beta$  was obtained in the proof of Theorem 1.9, this theorem is often remembered as saying that a *finite spanning set for  $V$  can be reduced to a basis for  $V$* . This method is illustrated in the next example.

### Example 6

Let

$$S = \{(2, -3, 5), (8, -12, 20), (1, 0, -2), (0, 2, -1), (7, 2, 0)\}.$$

It can be shown that  $S$  generates  $\mathbb{R}^3$ . We can select a basis for  $\mathbb{R}^3$  that is a subset of  $S$  by the technique used in proving Theorem 1.9. To start, select any nonzero vector in  $S$ , say  $(2, -3, 5)$ , to be a vector in the basis. Since  $4(2, -3, 5) = (8, -12, 20)$ , the set  $\{(2, -3, 5), (8, -12, 20)\}$  is linearly dependent by Exercise 9 of Section 1.5. Hence we do not include  $(8, -12, 20)$  in our basis. On the other hand,  $(1, 0, -2)$  is not a multiple of  $(2, -3, 5)$  and

vice versa, so that the set  $\{(2, -3, 5), (1, 0, -2)\}$  is linearly independent. Thus we include  $(1, 0, -2)$  as part of our basis.

Now we consider the set  $\{(2, -3, 5), (1, 0, -2), (0, 2, -1)\}$  obtained by adjoining another vector in  $S$  to the two vectors that we have already included in our basis. As before, we include  $(0, 2, -1)$  in our basis or exclude it from the basis according to whether  $\{(2, -3, 5), (1, 0, -2), (0, 2, -1)\}$  is linearly independent or linearly dependent. An easy calculation shows that this set is linearly independent, and so we include  $(0, 2, -1)$  in our basis. In a similar fashion the final vector in  $S$  is included or excluded from our basis according to whether the set

$$\{(2, -3, 5), (1, 0, -2), (0, 2, -1), (7, 2, 0)\}$$

is linearly independent or linearly dependent. Because

$$2(2, -3, 5) + 3(1, 0, -2) + 4(0, 2, -1) - (7, 2, 0) = (0, 0, 0),$$

we exclude  $(7, 2, 0)$  from our basis. We conclude that

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

is a subset of  $S$  that is a basis for  $\mathbb{R}^3$ . ◆

The corollaries of the following theorem are perhaps the most significant results in Chapter 1.

**Theorem 1.10 (Replacement Theorem).** *Let  $V$  be a vector space that is generated by a set  $G$  containing exactly  $n$  vectors, and let  $L$  be a linearly independent subset of  $V$  containing exactly  $m$  vectors. Then  $m \leq n$  and there exists a subset  $H$  of  $G$  containing exactly  $n - m$  vectors such that  $L \cup H$  generates  $V$ .*

*Proof.* The proof is by mathematical induction on  $m$ . The induction begins with  $m = 0$ ; for in this case  $L = \emptyset$ , and so taking  $H = G$  gives the desired result.

Now suppose that the theorem is true for some integer  $m \geq 0$ . We prove that the theorem is true for  $m + 1$ . Let  $L = \{v_1, v_2, \dots, v_{m+1}\}$  be a linearly independent subset of  $V$  consisting of  $m + 1$  vectors. By the corollary to Theorem 1.6 (p. 39),  $\{v_1, v_2, \dots, v_m\}$  is linearly independent, and so we may apply the induction hypothesis to conclude that  $m \leq n$  and that there is a subset  $\{u_1, u_2, \dots, u_{n-m}\}$  of  $G$  such that  $\{v_1, v_2, \dots, v_m\} \cup \{u_1, u_2, \dots, u_{n-m}\}$  generates  $V$ . Thus there exist scalars  $a_1, a_2, \dots, a_m, b_1, b_2, \dots, b_{n-m}$  such that

$$a_1 v_1 + a_2 v_2 + \cdots + a_m v_m + b_1 u_1 + b_2 u_2 + \cdots + b_{n-m} u_{n-m} = v_{m+1}. \quad (9)$$

Note that  $n - m > 0$ , lest  $v_{m+1}$  be a linear combination of  $v_1, v_2, \dots, v_m$ , which by Theorem 1.7 (p. 40) contradicts the assumption that  $L$  is linearly

independent. Hence  $n > m$ ; that is,  $n \geq m + 1$ . Moreover, some  $b_i$ , say  $b_1$ , is nonzero, for otherwise we obtain the same contradiction. Solving (9) for  $u_1$  gives

$$\begin{aligned} u_1 &= (-b_1^{-1}a_1)v_1 + (-b_1^{-1}a_2)v_2 + \cdots + (-b_1^{-1}a_m)v_m + (b_1^{-1})v_{m+1} \\ &\quad + (-b_1^{-1}b_2)u_2 + \cdots + (-b_1^{-1}b_{n-m})u_{n-m}. \end{aligned}$$

Let  $H = \{u_2, \dots, u_{n-m}\}$ . Then  $u_1 \in \text{span}(L \cup H)$ , and because  $v_1, v_2, \dots, v_m, u_2, \dots, u_{n-m}$  are clearly in  $\text{span}(L \cup H)$ , it follows that

$$\{v_1, v_2, \dots, v_m, u_1, u_2, \dots, u_{n-m}\} \subseteq \text{span}(L \cup H).$$

Because  $\{v_1, v_2, \dots, v_m, u_1, u_2, \dots, u_{n-m}\}$  generates  $V$ , Theorem 1.5 (p. 31) implies that  $\text{span}(L \cup H) = V$ . Since  $H$  is a subset of  $G$  that contains  $(n - m) - 1 = n - (m + 1)$  vectors, the theorem is true for  $m + 1$ . This completes the induction. ■

**Corollary 1.** Let  $V$  be a vector space having a finite basis. Then all bases for  $V$  are finite, and every basis for  $V$  contains the same number of vectors.

*Proof.* Suppose that  $\beta$  is a finite basis for  $V$  that contains exactly  $n$  vectors, and let  $\gamma$  be any other basis for  $V$ . If  $\gamma$  contains more than  $n$  vectors, then we can select a subset  $S$  of  $\gamma$  containing exactly  $n + 1$  vectors. Since  $S$  is linearly independent and  $\beta$  generates  $V$ , the replacement theorem implies that  $n + 1 \leq n$ , a contradiction. Therefore  $\gamma$  is finite, and the number  $m$  of vectors in  $\gamma$  satisfies  $m \leq n$ . Reversing the roles of  $\beta$  and  $\gamma$  and arguing as above, we obtain  $n \leq m$ . Hence  $m = n$ . ■

If a vector space has a finite basis, Corollary 1 asserts that the number of vectors in any basis for  $V$  is an intrinsic property of  $V$ . This fact makes possible the following important definitions.

**Definitions.** A vector space is called **finite-dimensional** if it has a basis consisting of a finite number of vectors. The unique integer  $n$  such that every basis for  $V$  contains exactly  $n$  elements is called the **dimension** of  $V$  and is denoted by  $\dim(V)$ . A vector space that is not finite-dimensional is called **infinite-dimensional**.

The following results are consequences of Examples 1 through 4.

### Example 7

The vector space  $\{0\}$  has dimension zero. ◆

### Example 8

The vector space  $F^n$  has dimension  $n$ . ◆

**Example 9**

The vector space  $M_{m \times n}(F)$  has dimension  $mn$ . ◆

**Example 10**

The vector space  $P_n(F)$  has dimension  $n + 1$ . ◆

The following examples show that the dimension of a vector space depends on its field of scalars.

**Example 11**

Over the field of complex numbers, the vector space of complex numbers has dimension 1. (A basis is  $\{1\}$ ). ◆

**Example 12**

Over the field of real numbers, the vector space of complex numbers has dimension 2. (A basis is  $\{1, i\}$ ). ◆

In the terminology of dimension, the first conclusion in the replacement theorem states that if  $V$  is a finite-dimensional vector space, then no linearly independent subset of  $V$  can contain more than  $\dim(V)$  vectors.

**Example 13**

The vector space  $P(F)$  is infinite-dimensional because, by Example 5, it has an infinite linearly independent set, namely  $\{1, x, x^2, \dots\}$ . ◆

In Example 13, the infinite linearly independent set  $\{1, x, x^2, \dots\}$  is, in fact, a basis for  $P(F)$ . Yet nothing that we have proved in this section guarantees an infinite-dimensional vector space must have a basis. In Section 1.7 it is shown, however, that every vector space has a basis.

Just as no linearly independent subset of a finite-dimensional vector space  $V$  can contain more than  $\dim(V)$  vectors, a corresponding statement can be made about the size of a generating set.

**Corollary 2.** Let  $V$  be a vector space with dimension  $n$ .

- Any finite generating set for  $V$  contains at least  $n$  vectors, and a generating set for  $V$  that contains exactly  $n$  vectors is a basis for  $V$ .
- Any linearly independent subset of  $V$  that contains exactly  $n$  vectors is a basis for  $V$ .
- Every linearly independent subset of  $V$  can be extended to a basis for  $V$ , that is, if  $L$  is a linearly independent subset of  $V$ , then there is a basis  $\beta$  of  $V$  such that  $L \subseteq \beta$ .

*Proof.* Let  $\beta$  be a basis for  $V$ .

(a) Let  $G$  be a finite generating set for  $V$ . By Theorem 1.9 some subset  $H$  of  $G$  is a basis for  $V$ . Corollary 1 implies that  $H$  contains exactly  $n$  vectors. Since a subset of  $G$  contains  $n$  vectors,  $G$  must contain at least  $n$  vectors. Moreover, if  $G$  contains exactly  $n$  vectors, then we must have  $H = G$ , so that  $G$  is a basis for  $V$ .

(b) Let  $L$  be a linearly independent subset of  $V$  containing exactly  $n$  vectors. It follows from the replacement theorem that there is a subset  $H$  of  $\beta$  containing  $n - n = 0$  vectors such that  $L \cup H$  generates  $V$ . Thus  $H = \emptyset$ , and  $L$  generates  $V$ . Since  $L$  is also linearly independent,  $L$  is a basis for  $V$ .

(c) If  $L$  is a linearly independent subset of  $V$  containing  $m$  vectors, then the replacement theorem asserts that there is a subset  $H$  of  $\beta$  containing exactly  $n - m$  vectors such that  $L \cup H$  generates  $V$ . Now  $L \cup H$  contains at most  $n$  vectors; therefore (a) implies that  $L \cup H$  contains exactly  $n$  vectors and that  $L \cup H$  is a basis for  $V$ . ■

### Example 14

It follows from Example 4 of Section 1.4 and (a) of Corollary 2 that

$$\{x^2 + 3x - 2, 2x^2 + 5x - 3, -x^2 - 4x + 4\}$$

is a basis for  $P_2(R)$ . ◆

### Example 15

It follows from Example 5 of Section 1.4 and (a) of Corollary 2 that

$$\left\{ \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \right\}$$

is a basis for  $M_{2 \times 2}(R)$ . ◆

### Example 16

It follows from Example 3 of Section 1.5 and (b) of Corollary 2 that

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

is a basis for  $R^4$ . ◆

### Example 17

For  $k = 0, 1, \dots, n$ , let  $p_k(x) = x^k + x^{k+1} + \dots + x^n$ . It follows from Example 4 of Section 1.5 and (b) of Corollary 2 that

$$\{p_0(x), p_1(x), \dots, p_n(x)\}$$

is a basis for  $P_n(F)$ . ◆

A procedure for reducing a generating set to a basis was illustrated in Example 6. In Section 3.4, when we have learned more about solving systems of linear equations, we will discover a simpler method for reducing a generating set to a basis. This procedure also can be used to extend a linearly independent set to a basis, as (c) of Corollary 2 asserts is possible.

### An Overview of Dimension and Its Consequences

Theorem 1.9 as well as the replacement theorem and its corollaries contain a wealth of information about the relationships among linearly independent sets, bases, and generating sets. For this reason, we summarize here the main results of this section in order to put them into better perspective.

A basis for a vector space  $V$  is a linearly independent subset of  $V$  that generates  $V$ . If  $V$  has a finite basis, then every basis for  $V$  contains the same number of vectors. This number is called the dimension of  $V$ , and  $V$  is said to be finite-dimensional. Thus if the dimension of  $V$  is  $n$ , every basis for  $V$  contains exactly  $n$  vectors. Moreover, every linearly independent subset of  $V$  contains *no more than*  $n$  vectors and can be extended to a basis for  $V$  by including appropriately chosen vectors. Also, each generating set for  $V$  contains *at least*  $n$  vectors and can be reduced to a basis for  $V$  by excluding appropriately chosen vectors. The Venn diagram in Figure 1.6 depicts these relationships.

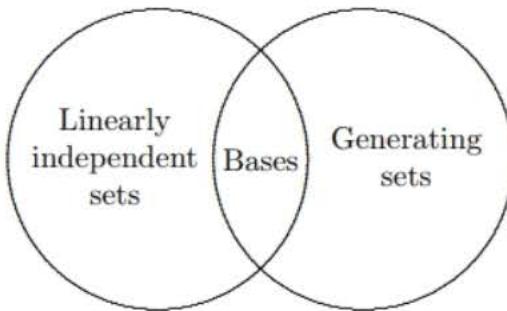


Figure 1.6

### The Dimension of Subspaces

Our next result relates the dimension of a subspace to the dimension of the vector space that contains it.

**Theorem 1.11.** *Let  $W$  be a subspace of a finite-dimensional vector space  $V$ . Then  $W$  is finite-dimensional and  $\dim(W) \leq \dim(V)$ . Moreover, if  $\dim(W) = \dim(V)$ , then  $V = W$ .*

*Proof.* Let  $\dim(V) = n$ . If  $W = \{0\}$ , then  $W$  is finite-dimensional and  $\dim(W) = 0 \leq n$ . Otherwise,  $W$  contains a nonzero vector  $x_1$ ; so  $\{x_1\}$  is a

linearly independent set. Continue choosing vectors,  $x_1, x_2, \dots, x_k$  in  $\mathbb{W}$  such that  $\{x_1, x_2, \dots, x_k\}$  is linearly independent. Since no linearly independent subset of  $\mathbb{V}$  can contain more than  $n$  vectors, this process must stop at a stage where  $k \leq n$  and  $\{x_1, x_2, \dots, x_k\}$  is linearly independent but adjoining any other vector from  $\mathbb{W}$  produces a linearly dependent set. Theorem 1.7 (p. 40) implies that  $\{x_1, x_2, \dots, x_k\}$  generates  $\mathbb{W}$ , and hence it is a basis for  $\mathbb{W}$ . Therefore  $\dim(\mathbb{W}) = k \leq n$ .

If  $\dim(\mathbb{W}) = n$ , then a basis for  $\mathbb{W}$  is a linearly independent subset of  $\mathbb{V}$  containing  $n$  vectors. But Corollary 2 of the replacement theorem implies that this basis for  $\mathbb{W}$  is also a basis for  $\mathbb{V}$ ; so  $\mathbb{W} = \mathbb{V}$ . ■

### Example 18

Let

$$\mathbb{W} = \{(a_1, a_2, a_3, a_4, a_5) \in \mathbb{F}^5 : a_1 + a_3 + a_5 = 0, a_2 = a_4\}.$$

It is easily shown that  $\mathbb{W}$  is a subspace of  $\mathbb{F}^5$  having

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

as a basis. Thus  $\dim(\mathbb{W}) = 3$ . ◆

### Example 19

The set of diagonal  $n \times n$  matrices is a subspace  $\mathbb{W}$  of  $\mathbb{M}_{n \times n}(F)$  (see Example 3 of Section 1.3). A basis for  $\mathbb{W}$  is

$$\{E^{11}, E^{22}, \dots, E^{nn}\},$$

where  $E^{ij}$  is the matrix in which the only nonzero entry is a 1 in the  $i$ th row and  $j$ th column. Thus  $\dim(\mathbb{W}) = n$ . ◆

### Example 20

We saw in Section 1.3 that the set of symmetric  $n \times n$  matrices is a subspace  $\mathbb{W}$  of  $\mathbb{M}_{n \times n}(F)$ . A basis for  $\mathbb{W}$  is

$$\{A^{ij} : 1 \leq i \leq j \leq n\},$$

where  $A^{ij}$  is the  $n \times n$  matrix having 1 in the  $i$ th row and  $j$ th column, 1 in the  $j$ th row and  $i$ th column, and 0 elsewhere. It follows that

$$\dim(\mathbb{W}) = n + (n - 1) + \dots + 1 = \frac{1}{2}n(n + 1). \quad \blacklozenge$$

**Corollary.** If  $\mathbb{W}$  is a subspace of a finite-dimensional vector space  $\mathbb{V}$ , then any basis for  $\mathbb{W}$  can be extended to a basis for  $\mathbb{V}$ .

*Proof.* Let  $S$  be a basis for  $\mathbb{W}$ . Because  $S$  is a linearly independent subset of  $\mathbb{V}$ , Corollary 2 of the replacement theorem guarantees that  $S$  can be extended to a basis for  $\mathbb{V}$ . ■

**Example 21**

The set of all polynomials of the form

$$a_{18}x^{18} + a_{16}x^{16} + \cdots + a_2x^2 + a_0,$$

where  $a_{18}, a_{16}, \dots, a_2, a_0 \in F$ , is a subspace  $W$  of  $P_{18}(F)$ . A basis for  $W$  is  $\{1, x^2, \dots, x^{16}, x^{18}\}$ , which is a subset of the standard basis for  $P_{18}(F)$ . ♦

We can apply Theorem 1.11 to determine the subspaces of  $R^2$  and  $R^3$ . Since  $R^2$  has dimension 2, subspaces of  $R^2$  can be of dimensions 0, 1, or 2 only. The only subspaces of dimension 0 or 2 are  $\{0\}$  and  $R^2$ , respectively. Any subspace of  $R^2$  having dimension 1 consists of all scalar multiples of some nonzero vector in  $R^2$  (Exercise 11 of Section 1.4).

If a point of  $R^2$  is identified in the natural way with a point in the Euclidean plane, then it is possible to describe the subspaces of  $R^2$  geometrically: A subspace of  $R^2$  having dimension 0 consists of the origin of the Euclidean plane, a subspace of  $R^2$  with dimension 1 consists of a line through the origin, and a subspace of  $R^2$  having dimension 2 is the entire Euclidean plane.

Similarly, the subspaces of  $R^3$  must have dimensions 0, 1, 2, or 3. Interpreting these possibilities geometrically, we see that a subspace of dimension zero must be the origin of Euclidean 3-space, a subspace of dimension 1 is a line through the origin, a subspace of dimension 2 is a plane through the origin, and a subspace of dimension 3 is Euclidean 3-space itself.

**The Lagrange Interpolation Formula**

In many applications, we have a collection of data obtained from experiments or samples. For example, we may know the locations of an airplane flying from New York to London at certain times and would like to be able to approximate the locations of the plane at one or more intermediate times. The process of estimating intermediate values of a variable from known values is called *interpolation*.

Corollary 2 of the replacement theorem (Theorem 1.10) can be applied to obtain a useful formula that enables us to approximate the values of an unknown function by a polynomial function. Let  $c_0, c_1, \dots, c_n$  be distinct scalars in an infinite field  $F$ . The polynomials  $f_0(x), f_1(x), \dots, f_n(x)$  defined by

$$f_i(x) = \frac{(x - c_0)(x - c_1)\cdots(x - c_{i-1})(x - c_{i+1})\cdots(x - c_n)}{(c_i - c_0)(c_i - c_1)\cdots(c_i - c_{i-1})(c_i - c_{i+1})\cdots(c_i - c_n)} = \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - c_k}{c_i - c_k}$$

are called the **Lagrange polynomials** (associated with  $c_0, c_1, \dots, c_n$ ). Note that each  $f_i(x)$  is a polynomial of degree  $n$  and hence is in  $P_n(F)$ . If we

regard  $f_i(x)$  as a polynomial function  $f_i: F \rightarrow F$ , we see that

$$f_i(c_j) = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (10)$$

This property of Lagrange polynomials can be used to show that  $\beta = \{f_0, f_1, \dots, f_n\}$  is a linearly independent subset of  $P_n(F)$ . Suppose that

$$\sum_{i=0}^n a_i f_i = 0 \quad \text{for some scalars } a_0, a_1, \dots, a_n,$$

where  $0$  denotes the zero function. Then

$$\sum_{i=0}^n a_i f_i(c_j) = 0 \quad \text{for } j = 0, 1, \dots, n.$$

But also

$$\sum_{i=0}^n a_i f_i(c_j) = a_j$$

by (10). Hence  $a_j = 0$  for  $j = 0, 1, \dots, n$ ; so  $\beta$  is linearly independent. Since the dimension of  $P_n(F)$  is  $n+1$ , it follows from Corollary 2 of the replacement theorem that  $\beta$  is a basis for  $P_n(F)$ .

Because  $\beta$  is a basis for  $P_n(F)$ , every polynomial function  $g$  in  $P_n(F)$  is a linear combination of polynomial functions of  $\beta$ , say,

$$g = \sum_{i=0}^n b_i f_i.$$

It follows that

$$g(c_j) = \sum_{i=0}^n b_i f_i(c_j) = b_j;$$

so

$$g = \sum_{i=0}^n g(c_i) f_i$$

is the unique representation of  $g$  as a linear combination of elements of  $\beta$ . This representation is called the **Lagrange interpolation formula**. Notice that the preceding argument shows that if  $b_0, b_1, \dots, b_n$  are any  $n+1$  scalars in  $F$  (not necessarily distinct), then the polynomial function

$$g = \sum_{i=0}^n b_i f_i$$

is the unique polynomial in  $P_n(F)$  such that  $g(c_j) = b_j$ . Thus we have found the unique polynomial of degree not exceeding  $n$  that has specified values  $b_j$  at given points  $c_j$  in its domain ( $j = 0, 1, \dots, n$ ). For example, let us construct the real polynomial  $g$  of degree at most 2 whose graph contains the points  $(1, 8)$ ,  $(2, 5)$ , and  $(3, -4)$ . (Thus, in the notation above,  $c_0 = 1$ ,  $c_1 = 2$ ,  $c_2 = 3$ ,  $b_0 = 8$ ,  $b_1 = 5$ , and  $b_2 = -4$ .) The Lagrange polynomials associated with  $c_0$ ,  $c_1$ , and  $c_2$  are

$$f_0(x) = \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{1}{2}(x^2 - 5x + 6),$$

$$f_1(x) = \frac{(x-1)(x-3)}{(2-1)(2-3)} = -1(x^2 - 4x + 3),$$

and

$$f_2(x) = \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{1}{2}(x^2 - 3x + 2).$$

Hence the desired polynomial is

$$\begin{aligned} g(x) &= \sum_{i=0}^2 b_i f_i(x) = 8f_0(x) + 5f_1(x) - 4f_2(x) \\ &= 4(x^2 - 5x + 6) - 5(x^2 - 4x + 3) - 2(x^2 - 3x + 2) \\ &= -3x^2 + 6x + 5. \end{aligned}$$

An important consequence of the Lagrange interpolation formula is the following result: If  $f \in P_n(F)$  and  $f(c_i) = 0$  for  $n+1$  distinct scalars  $c_0, c_1, \dots, c_n$  in  $F$ , then  $f$  is the zero function.

## EXERCISES

1. Label the following statements as true or false.
  - (a) The zero vector space has no basis.
  - (b) Every vector space that is generated by a finite set has a basis.
  - (c) Every vector space has a finite basis.
  - (d) A vector space cannot have more than one basis.
  - (e) If a vector space has a finite basis, then the number of vectors in every basis is the same.
  - (f) The dimension of  $P_n(F)$  is  $n$ .
  - (g) The dimension of  $M_{m \times n}(F)$  is  $m + n$ .
  - (h) Suppose that  $V$  is a finite-dimensional vector space, that  $S_1$  is a linearly independent subset of  $V$ , and that  $S_2$  is a subset of  $V$  that generates  $V$ . Then  $S_1$  cannot contain more vectors than  $S_2$ .

- (i) If  $S$  generates the vector space  $V$ , then every vector in  $V$  can be written as a linear combination of vectors in  $S$  in only one way.
  - (j) Every subspace of a finite-dimensional space is finite-dimensional.
  - (k) If  $V$  is a vector space having dimension  $n$ , then  $V$  has exactly one subspace with dimension 0 and exactly one subspace with dimension  $n$ .
  - (l) If  $V$  is a vector space having dimension  $n$ , and if  $S$  is a subset of  $V$  with  $n$  vectors, then  $S$  is linearly independent if and only if  $S$  spans  $V$ .
2. Determine which of the following sets are bases for  $\mathbb{R}^3$ .
- (a)  $\{(1, 0, -1), (2, 5, 1), (0, -4, 3)\}$
  - (b)  $\{(2, -4, 1), (0, 3, -1), (6, 0, -1)\}$
  - (c)  $\{(1, 2, -1), (1, 0, 2), (2, 1, 1)\}$
  - (d)  $\{(-1, 3, 1), (2, -4, -3), (-3, 8, 2)\}$
  - (e)  $\{(1, -3, -2), (-3, 1, 3), (-2, -10, -2)\}$
3. Determine which of the following sets are bases for  $P_2(R)$ .
- (a)  $\{-1 - x + 2x^2, 2 + x - 2x^2, 1 - 2x + 4x^2\}$
  - (b)  $\{1 + 2x + x^2, 3 + x^2, x + x^2\}$
  - (c)  $\{1 - 2x - 2x^2, -2 + 3x - x^2, 1 - x + 6x^2\}$
  - (d)  $\{-1 + 2x + 4x^2, 3 - 4x - 10x^2, -2 - 5x - 6x^2\}$
  - (e)  $\{1 + 2x - x^2, 4 - 2x + x^2, -1 + 18x - 9x^2\}$
4. Do the polynomials  $x^3 - 2x^2 + 1$ ,  $4x^2 - x + 3$ , and  $3x - 2$  generate  $P_3(R)$ ? Justify your answer.
5. Is  $\{(1, 4, -6), (1, 5, 8), (2, 1, 1), (0, 1, 0)\}$  a linearly independent subset of  $\mathbb{R}^3$ ? Justify your answer.
6. Give three different bases for  $F^2$  and for  $M_{2 \times 2}(F)$ .
7. The vectors  $u_1 = (2, -3, 1)$ ,  $u_2 = (1, 4, -2)$ ,  $u_3 = (-8, 12, -4)$ ,  $u_4 = (1, 37, -17)$ , and  $u_5 = (-3, -5, 8)$  generate  $\mathbb{R}^3$ . Find a subset of the set  $\{u_1, u_2, u_3, u_4, u_5\}$  that is a basis for  $\mathbb{R}^3$ .
8. Let  $W$  denote the subspace of  $\mathbb{R}^5$  consisting of all the vectors having coordinates that sum to zero. The vectors

$$\begin{aligned} u_1 &= (2, -3, 4, -5, 2), & u_2 &= (-6, 9, -12, 15, -6), \\ u_3 &= (3, -2, 7, -9, 1), & u_4 &= (2, -8, 2, -2, 6), \\ u_5 &= (-1, 1, 2, 1, -3), & u_6 &= (0, -3, -18, 9, 12), \\ u_7 &= (1, 0, -2, 3, -2), & u_8 &= (2, -1, 1, -9, 7) \end{aligned}$$

generate  $W$ . Find a subset of the set  $\{u_1, u_2, \dots, u_8\}$  that is a basis for  $W$ .

9. The vectors  $u_1 = (1, 1, 1, 1)$ ,  $u_2 = (0, 1, 1, 1)$ ,  $u_3 = (0, 0, 1, 1)$ , and  $u_4 = (0, 0, 0, 1)$  form a basis for  $\mathbb{F}^4$ . Find the unique representation of an arbitrary vector  $(a_1, a_2, a_3, a_4)$  in  $\mathbb{F}^4$  as a linear combination of  $u_1$ ,  $u_2$ ,  $u_3$ , and  $u_4$ .
10. In each part, use the Lagrange interpolation formula to construct the polynomial of smallest degree whose graph contains the following points.
- $(-2, -6), (-1, 5), (1, 3)$
  - $(-4, 24), (1, 9), (3, 3)$
  - $(-2, 3), (-1, -6), (1, 0), (3, -2)$
  - $(-3, -30), (-2, 7), (0, 15), (1, 10)$
11. Let  $u$  and  $v$  be distinct vectors of a vector space  $V$ . Show that if  $\{u, v\}$  is a basis for  $V$  and  $a$  and  $b$  are nonzero scalars, then both  $\{u + v, au\}$  and  $\{au, bv\}$  are also bases for  $V$ .
12. Let  $u$ ,  $v$ , and  $w$  be distinct vectors of a vector space  $V$ . Show that if  $\{u, v, w\}$  is a basis for  $V$ , then  $\{u + v + w, v + w, w\}$  is also a basis for  $V$ .
13. The set of solutions to the system of linear equations

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

is a subspace of  $\mathbb{R}^3$ . Find a basis for this subspace.

14. Find bases for the following subspaces of  $\mathbb{F}^5$ :

$$W_1 = \{(a_1, a_2, a_3, a_4, a_5) \in \mathbb{F}^5 : a_1 - a_3 - a_4 = 0\}$$

and

$$W_2 = \{(a_1, a_2, a_3, a_4, a_5) \in \mathbb{F}^5 : a_2 = a_3 = a_4 \text{ and } a_1 + a_5 = 0\}.$$

What are the dimensions of  $W_1$  and  $W_2$ ?

15. The set of all  $n \times n$  matrices having trace equal to zero is a subspace  $W$  of  $M_{n \times n}(F)$  (see Example 4 of Section 1.3). Find a basis for  $W$ . What is the dimension of  $W$ ?
16. The set of all upper triangular  $n \times n$  matrices is a subspace  $W$  of  $M_{n \times n}(F)$  (see Exercise 12 of Section 1.3). Find a basis for  $W$ . What is the dimension of  $W$ ?
17. The set of all skew-symmetric  $n \times n$  matrices is a subspace  $W$  of  $M_{n \times n}(F)$  (see Exercise 28 of Section 1.3). Find a basis for  $W$ . What is the dimension of  $W$ ?

18. Let  $V$  denote the vector space of all sequences in  $F$ , as defined in Example 5 of Section 1.2. Find a basis for the subspace  $W$  of  $V$  consisting of the sequences  $(a_n)$  that have only a finite number of nonzero terms  $a_n$ . Justify your answer.

19. Complete the proof of Theorem 1.8.

20.<sup>†</sup> Let  $V$  be a vector space having dimension  $n$ , and let  $S$  be a subset of  $V$  that generates  $V$ .

- (a) Prove that there is a subset of  $S$  that is a basis for  $V$ . (Be careful not to assume that  $S$  is finite.)  
(b) Prove that  $S$  contains at least  $n$  vectors.

Visit [goo.gl/wE2wwA](http://goo.gl/wE2wwA) for a solution.

21. Prove that a vector space is infinite-dimensional if and only if it contains an infinite linearly independent subset.

22. Let  $W_1$  and  $W_2$  be subspaces of a finite-dimensional vector space  $V$ . Determine necessary and sufficient conditions on  $W_1$  and  $W_2$  so that  $\dim(W_1 \cap W_2) = \dim(W_1)$ .

23. Let  $v_1, v_2, \dots, v_k, v$  be vectors in a vector space  $V$ , and define  $W_1 = \text{span}(\{v_1, v_2, \dots, v_k\})$ , and  $W_2 = \text{span}(\{v_1, v_2, \dots, v_k, v\})$ .

- (a) Find necessary and sufficient conditions on  $v$  such that  $\dim(W_1) = \dim(W_2)$ .  
(b) State and prove a relationship involving  $\dim(W_1)$  and  $\dim(W_2)$  in the case that  $\dim(W_1) \neq \dim(W_2)$ .

24. Let  $f(x)$  be a polynomial of degree  $n$  in  $P_n(R)$ . Prove that for any  $g(x) \in P_n(R)$  there exist scalars  $c_0, c_1, \dots, c_n$  such that

$$g(x) = c_0 f(x) + c_1 f'(x) + c_2 f''(x) + \cdots + c_n f^{(n)}(x),$$

where  $f^{(n)}(x)$  denotes the  $n$ th derivative of  $f(x)$ .

25. Let  $V$ ,  $W$ , and  $Z$  be as in Exercise 21 of Section 1.2. If  $V$  and  $W$  are vector spaces over  $F$  of dimensions  $m$  and  $n$ , determine the dimension of  $Z$ .

26. For a fixed  $a \in R$ , determine the dimension of the subspace of  $P_n(R)$  defined by  $\{f \in P_n(R) : f(a) = 0\}$ .

27. Let  $W_1$  and  $W_2$  be the subspaces of  $P(F)$  defined in Exercise 25 in Section 1.3. Determine the dimensions of the subspaces  $W_1 \cap P_n(F)$  and  $W_2 \cap P_n(F)$ .

- 28.** Let  $V$  be a finite-dimensional vector space over  $C$  with dimension  $n$ . Prove that if  $V$  is now regarded as a vector space over  $R$ , then  $\dim V = 2n$ . (See Examples 11 and 12.)

Exercises 29–34 require knowledge of the sum and direct sum of subspaces, as defined in the exercises of Section 1.3.

- 29.** (a) Prove that if  $W_1$  and  $W_2$  are finite-dimensional subspaces of a vector space  $V$ , then the subspace  $W_1 + W_2$  is finite-dimensional, and  $\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)$ . Hint: Start with a basis  $\{u_1, u_2, \dots, u_k\}$  for  $W_1 \cap W_2$  and extend this set to a basis  $\{u_1, u_2, \dots, u_k, v_1, v_2, \dots, v_m\}$  for  $W_1$  and to a basis  $\{u_1, u_2, \dots, u_k, w_1, w_2, \dots, w_p\}$  for  $W_2$ .
- (b) Let  $W_1$  and  $W_2$  be finite-dimensional subspaces of a vector space  $V$ , and let  $V = W_1 + W_2$ . Deduce that  $V$  is the direct sum of  $W_1$  and  $W_2$  if and only if  $\dim(V) = \dim(W_1) + \dim(W_2)$ .
- 30.** Let

$$V = M_{2 \times 2}(F), \quad W_1 = \left\{ \begin{pmatrix} a & b \\ c & a \end{pmatrix} \in V : a, b, c \in F \right\},$$

and

$$W_2 = \left\{ \begin{pmatrix} 0 & a \\ -a & b \end{pmatrix} \in V : a, b \in F \right\}.$$

Prove that  $W_1$  and  $W_2$  are subspaces of  $V$ , and find the dimensions of  $W_1$ ,  $W_2$ ,  $W_1 + W_2$ , and  $W_1 \cap W_2$ .

- 31.** Let  $W_1$  and  $W_2$  be subspaces of a vector space  $V$  having dimensions  $m$  and  $n$ , respectively, where  $m \geq n$ .
- (a) Prove that  $\dim(W_1 \cap W_2) \leq n$ .
- (b) Prove that  $\dim(W_1 + W_2) \leq m + n$ .
- 32.** Find examples of subspaces  $W_1$  and  $W_2$  of  $R^3$  such that  $\dim(W_1) > \dim(W_2) > 0$  and
- (a)  $\dim(W_1 \cap W_2) = \dim(W_2)$ ;
- (b)  $\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2)$ ;
- (c)  $\dim(W_1 + W_2) < \dim(W_1) + \dim(W_2)$ .
- 33.** (a) Let  $W_1$  and  $W_2$  be subspaces of a vector space  $V$  such that  $V = W_1 \oplus W_2$ . If  $\beta_1$  and  $\beta_2$  are bases for  $W_1$  and  $W_2$ , respectively, show that  $\beta_1 \cap \beta_2 = \emptyset$  and  $\beta_1 \cup \beta_2$  is a basis for  $V$ .
- (b) Conversely, let  $\beta_1$  and  $\beta_2$  be disjoint bases for subspaces  $W_1$  and  $W_2$ , respectively, of a vector space  $V$ . Prove that if  $\beta_1 \cup \beta_2$  is a basis for  $V$ , then  $V = W_1 \oplus W_2$ .

34. (a) Prove that if  $W_1$  is any subspace of a finite-dimensional vector space  $V$ , then there exists a subspace  $W_2$  of  $V$  such that  $V = W_1 \oplus W_2$ .
- (b) Let  $V = \mathbb{R}^2$  and  $W_1 = \{(a_1, 0) : a_1 \in \mathbb{R}\}$ . Give examples of two different subspaces  $W_2$  and  $W'_2$  such that  $V = W_1 \oplus W_2$  and  $V = W_1 \oplus W'_2$ .

The following exercise requires familiarity with Exercise 31 of Section 1.3.

35. Let  $W$  be a subspace of a finite-dimensional vector space  $V$ , and consider the basis  $\{u_1, u_2, \dots, u_k\}$  for  $W$ . Let  $\{u_1, u_2, \dots, u_k, u_{k+1}, \dots, u_n\}$  be an extension of this basis to a basis for  $V$ .
- (a) Prove that  $\{u_{k+1} + W, u_{k+2} + W, \dots, u_n + W\}$  is a basis for  $V/W$ .
- (b) Derive a formula relating  $\dim(V)$ ,  $\dim(W)$ , and  $\dim(V/W)$ .

## 1.7\* MAXIMAL LINEARLY INDEPENDENT SUBSETS

In this section, several significant results from Section 1.6 are extended to infinite-dimensional vector spaces. Our principal goal here is to prove that every vector space has a basis. This result is important in the study of infinite-dimensional vector spaces because it is often difficult to construct an explicit basis for such a space. Consider, for example, the vector space of real numbers over the field of rational numbers. There is no obvious way to construct a basis for this space, and yet it follows from the results of this section that such a basis does exist.

The difficulty that arises in extending the theorems of the preceding section to infinite-dimensional vector spaces is that the principle of mathematical induction, which played a crucial role in many of the proofs of Section 1.6, is no longer adequate. Instead, an alternate result called the *Hausdorff maximal principle* is needed. Before stating this principle, we need to introduce some terminology.

**Definition.** Let  $\mathcal{F}$  be a family of sets. A member  $M$  of  $\mathcal{F}$  is called **maximal** (with respect to set inclusion) if  $M$  is contained in no member of  $\mathcal{F}$  other than  $M$  itself.

### Example 1

Let  $\mathcal{F}$  be the family of all subsets of a nonempty set  $S$ . (This family  $\mathcal{F}$  is called the **power set** of  $S$ .) The set  $S$  is easily seen to be a maximal element of  $\mathcal{F}$ . ♦

### Example 2

Let  $S$  and  $T$  be disjoint nonempty sets, and let  $\mathcal{F}$  be the union of their power sets. Then  $S$  and  $T$  are both maximal elements of  $\mathcal{F}$ . ♦

**Example 3**

Let  $\mathcal{F}$  be the family of all finite subsets of an infinite set  $S$ . Then  $\mathcal{F}$  has no maximal element. For if  $M$  is any member of  $\mathcal{F}$  and  $s$  is any element of  $S$  that is not in  $M$ , then  $M \cup \{s\}$  is a member of  $\mathcal{F}$  that contains  $M$  as a proper subset. ♦

**Definition.** A collection of sets  $\mathcal{C}$  is called a **chain** (or **nest** or **tower**) if for each pair of sets  $A$  and  $B$  in  $\mathcal{C}$ , either  $A \subseteq B$  or  $B \subseteq A$ .

**Example 4**

For each positive integer  $n$  let  $A_n = \{1, 2, \dots, n\}$ . Then the collection of sets  $\mathcal{C} = \{A_n : n = 1, 2, 3, \dots\}$  is a chain. In fact,  $A_m \subseteq A_n$  if and only if  $m \leq n$ . ♦

With this terminology we can now state the Hausdorff maximal principle.

**Hausdorff Maximal Principle.<sup>4</sup>** Let  $\mathcal{F}$  be a family of sets. If, for each chain  $\mathcal{C} \subseteq \mathcal{F}$ , there exists a member of  $\mathcal{F}$  that contains all the members of  $\mathcal{C}$ , then  $\mathcal{F}$  contains a maximal member.

Because the Hausdorff maximal principle guarantees the existence of maximal elements in a family of sets satisfying the hypothesis above, it is useful to reformulate the definition of a basis in terms of a maximal property. In Theorem 1.12, we show that this is possible; in fact, the concept defined next is equivalent to a basis.

**Definition.** Let  $S$  be a subset of a vector space  $V$ . A **maximal linearly independent subset** of  $S$  is a subset  $B$  of  $S$  satisfying both of the following conditions.

- (a)  $B$  is linearly independent.
- (b) The only linearly independent subset of  $S$  that contains  $B$  is  $B$  itself.

**Example 5**

Example 2 of Section 1.4 shows that

$$\{x^3 - 2x^2 - 5x - 3, 3x^3 - 5x^2 - 4x - 9\}$$

is a maximal linearly independent subset of

$$S = \{2x^3 - 2x^2 + 12x - 6, x^3 - 2x^2 - 5x - 3, 3x^3 - 5x^2 - 4x - 9\}$$

---

<sup>4</sup>The Hausdorff Maximal Principle is logically equivalent to the *Axiom of Choice*, which is an assumption in most axiomatic developments of set theory. For a treatment of set theory using the Hausdorff Maximal Principle, see John L. Kelley, *General Topology*, Graduate Texts in Mathematics Series, Vol. 27, Springer-Verlag, 1991.

in  $P_3(R)$ . In this case, however, any subset of  $S$  consisting of two polynomials is easily shown to be a maximal linearly independent subset of  $S$ . Thus maximal linearly independent subsets of a set need not be unique. ♦

A basis  $\beta$  for a vector space  $V$  is a maximal linearly independent subset of  $V$ , because

1.  $\beta$  is linearly independent by definition.
2. If  $v \in V$  and  $v \notin \beta$ , then  $\beta \cup \{v\}$  is linearly dependent by Theorem 1.7 (p. 40) because  $\text{span}(\beta) = V$ .

Our next result shows that the converse of this statement is also true.

**Theorem 1.12.** *Let  $V$  be a vector space and  $S$  a subset that generates  $V$ . If  $\beta$  is a maximal linearly independent subset of  $S$ , then  $\beta$  is a basis for  $V$ .*

*Proof.* Let  $\beta$  be a maximal linearly independent subset of  $S$ . Because  $\beta$  is linearly independent, it suffices to prove that  $\beta$  generates  $V$ . We claim that  $S \subseteq \text{span}(\beta)$ , for otherwise there exists a  $v \in S$  such that  $v \notin \text{span}(\beta)$ . Since Theorem 1.7 (p. 40) implies that  $\beta \cup \{v\}$  is linearly independent, we have contradicted the maximality of  $\beta$ . Therefore  $S \subseteq \text{span}(\beta)$ . Because  $\text{span}(S) = V$ , it follows from Theorem 1.5 (p. 31) that  $\text{span}(\beta) = V$ . ■

Thus a subset of a vector space is a basis if and only if it is a maximal linearly independent subset of the vector space. Therefore we can accomplish our goal of proving that every vector space has a basis by showing that every vector space contains a maximal linearly independent subset. This result follows immediately from the next theorem.

**Theorem 1.13.** *Let  $S$  be a linearly independent subset of a vector space  $V$ . There exists a maximal linearly independent subset of  $V$  that contains  $S$ .*

*Proof.* Let  $\mathcal{F}$  denote the family of all linearly independent subsets of  $V$  containing  $S$ . To show that  $\mathcal{F}$  contains a maximal element, we show that if  $\mathcal{C}$  is a chain in  $\mathcal{F}$ , then there exists a member  $U$  of  $\mathcal{F}$  containing each member of  $\mathcal{C}$ . If  $\mathcal{C}$  is empty, take  $U = S$ . Otherwise take  $U$  equal to the union of the members of  $\mathcal{C}$ . Clearly  $U$  contains each member of  $\mathcal{C}$ , and so it suffices to prove that  $U \in \mathcal{F}$  (i.e., that  $U$  is a linearly independent subset of  $V$  that contains  $S$ ). Because each member of  $\mathcal{C}$  is a subset of  $V$  containing  $S$ , we have  $S \subseteq U \subseteq V$ . Thus we need only prove that  $U$  is linearly independent. Let  $u_1, u_2, \dots, u_n$  be in  $U$  and  $a_1, a_2, \dots, a_n$  be scalars such that  $a_1u_1 + a_2u_2 + \dots + a_nu_n = 0$ . Because  $u_i \in U$  for  $i = 1, 2, \dots, n$ , there exists a set  $A_i$  in  $\mathcal{C}$  such that  $u_i \in A_i$ . But since  $\mathcal{C}$  is a chain, one of these sets, say  $A_k$ , contains all the others. Thus  $u_i \in A_k$  for  $i = 1, 2, \dots, n$ . However,  $A_k$  is a linearly independent set; so  $a_1u_1 + a_2u_2 + \dots + a_nu_n = 0$  implies that  $a_1 = a_2 = \dots = a_n = 0$ . It follows that  $U$  is linearly independent.

The Hausdorff maximal principle implies that  $\mathcal{F}$  has a maximal element. This element is easily seen to be a maximal linearly independent subset of  $V$  that contains  $S$ . ■

**Corollary.** Every vector space has a basis.

It can be shown, analogously to Corollary 1 of the replacement theorem (p. 47), that every basis for an infinite-dimensional vector space has the same *cardinality*. (Sets have the same cardinality if there is a one-to-one and onto mapping between them.) (See, for example, N. Jacobson, *Lectures in Abstract Algebra*, vol. 2, Linear Algebra, D. Van Nostrand Company, New York, 1953, p. 240.)

Exercises 4–7 extend other results from Section 1.6 to infinite-dimensional vector spaces.

## EXERCISES

1. Label the following statements as true or false.
  - (a) Every family of sets contains a maximal element.
  - (b) Every chain contains a maximal element.
  - (c) If a family of sets has a maximal element, then that maximal element is unique.
  - (d) If a chain of sets has a maximal element, then that maximal element is unique.
  - (e) A basis for a vector space is a maximal linearly independent subset of that vector space.
  - (f) A maximal linearly independent subset of a vector space is a basis for that vector space.
2. Show that the set of convergent sequences is an infinite-dimensional subspace of the vector space of all sequences of real numbers. (See Exercise 21 in Section 1.3.)
3. Let  $V$  be the set of real numbers regarded as a vector space over the field of rational numbers. Prove that  $V$  is infinite-dimensional. *Hint:* Use the fact that  $\pi$  is transcendental, that is,  $\pi$  is not a zero of any polynomial with rational coefficients.
4. Let  $W$  be a subspace of a (not necessarily finite-dimensional) vector space  $V$ . Prove that any basis for  $W$  is a subset of a basis for  $V$ .
5. Prove the following infinite-dimensional version of Theorem 1.8 (p. 44): Let  $\beta$  be a subset of an infinite-dimensional vector space  $V$ . Then  $\beta$  is a basis for  $V$  if and only if for each nonzero vector  $v$  in  $V$ , there exist unique vectors  $u_1, u_2, \dots, u_n$  in  $\beta$  and unique nonzero scalars  $c_1, c_2, \dots, c_n$  such that  $v = c_1u_1 + c_2u_2 + \dots + c_nu_n$ . Visit [goo.gl/fNWSDM](http://goo.gl/fNWSDM) for a solution.

6. Prove the following generalization of Theorem 1.9 (p. 45): Let  $S_1$  and  $S_2$  be subsets of a vector space  $V$  such that  $S_1 \subseteq S_2$ . If  $S_1$  is linearly independent and  $S_2$  generates  $V$ , then there exists a basis  $\beta$  for  $V$  such that  $S_1 \subseteq \beta \subseteq S_2$ . Hint: Apply the Hausdorff maximal principle to the family of all linearly independent subsets of  $S_2$  that contain  $S_1$ , and proceed as in the proof of Theorem 1.13.
7. Prove the following generalization of the replacement theorem. Let  $\beta$  be a basis for a vector space  $V$ , and let  $S$  be a linearly independent subset of  $V$ . There exists a subset  $S_1$  of  $\beta$  such that  $S \cup S_1$  is a basis for  $V$ .

## INDEX OF DEFINITIONS FOR CHAPTER 1

Additive inverse	12	Scalar	7
Basis	43	Scalar multiplication	6
Cancellation law	12	Sequence	11
Column vector	8	Span of a subset	30
Chain	60	Spans	31
Degree of a polynomial	10	Square matrix	9
Diagonal entries of a matrix	8	Standard basis for $F^n$	43
Diagonal matrix	19	Standard basis for $P_n(F)$	44
Dimension	47	Subspace	17
Finite-dimensional space	47	Subspace generated by the elements of a set	31
Generates	31	Symmetric matrix	18
Infinite-dimensional space	47	Trace	19
Lagrange interpolation formula	53	Transpose	18
Lagrange polynomials	52	Upper triangular matrix	19
Linear combination	25	Trivial representation of $\theta$	37
Linearly dependent	37	Vector	7
Linearly independent	38	Vector addition	6
Matrix	8	Vector space	6
Maximal element of a family of sets	59	Zero matrix	9
Maximal linearly independent subset	60	Zero polynomial	10
$n$ -tuple	8	Zero subspace	17
Polynomial	10	Zero vector	12
Row vector	8	Zero vector space	15

# 2

## Linear Transformations and Matrices

---

- 2.1 Linear Transformations, Null spaces, and Ranges
- 2.2 The Matrix Representation of a Linear Transformation
- 2.3 Composition of Linear Transformations and Matrix Multiplication
- 2.4 Invertibility and Isomorphisms
- 2.5 The Change of Coordinate Matrix
- 2.6\* Dual Spaces
- 2.7\* Homogeneous Linear Differential Equations with Constant Coefficients

In Chapter 1, we developed the theory of abstract vector spaces in considerable detail. It is now natural to consider those functions defined on vector spaces that in some sense “preserve” the structure. These special functions are called *linear transformations*, and they abound in both pure and applied mathematics. In calculus, the operations of differentiation and integration provide us with two of the most important examples of linear transformations (see Examples 6 and 7 of Section 2.1). These two examples allow us to reformulate many of the problems in differential and integral equations in terms of linear transformations on particular vector spaces (see Sections 2.7 and 5.2).

In geometry, rotations, reflections, and projections (see Examples 2, 3, and 4 of Section 2.1) provide us with another class of linear transformations. Later we use these transformations to study rigid motions in  $\mathbb{R}^n$  (Section 6.10).

In the remaining chapters, we see further examples of linear transformations occurring in both the physical and the social sciences. Throughout this chapter, we assume that all vector spaces are over a common field  $F$ .

---

### 2.1 LINEAR TRANSFORMATIONS, NULL SPACES, AND RANGES

In this section, we consider a number of examples of linear transformations. Many of these transformations are studied in more detail in later sections. Recall that a function  $T$  with domain  $V$  and codomain  $W$  is denoted by

$T: V \rightarrow W$ . (See Appendix B.)

**Definition.** Let  $V$  and  $W$  be vector spaces over the same field  $F$ . We call a function  $T: V \rightarrow W$  a **linear transformation from  $V$  to  $W$**  if, for all  $x, y \in V$  and  $c \in F$ , we have

- (a)  $T(x + y) = T(x) + T(y)$  and
- (b)  $T(cx) = cT(x)$ .

If the underlying field  $F$  is the field of rational numbers, then (a) implies (b) (see Exercise 38), but, in general (a) and (b) are logically independent. See Exercises 39 and 40.

We often simply call  $T$  **linear**. The reader should verify the following properties of a function  $T: V \rightarrow W$ . (See Exercise 7.)

1. If  $T$  is linear, then  $T(0) = 0$ .
2.  $T$  is linear if and only if  $T(cx + y) = cT(x) + T(y)$  for all  $x, y \in V$  and  $c \in F$ .
3. If  $T$  is linear, then  $T(x - y) = T(x) - T(y)$  for all  $x, y \in V$ .
4.  $T$  is linear if and only if, for  $x_1, x_2, \dots, x_n \in V$  and  $a_1, a_2, \dots, a_n \in F$ , we have

$$T\left(\sum_{i=1}^n a_i x_i\right) = \sum_{i=1}^n a_i T(x_i).$$

We generally use property 2 to prove that a given transformation is linear.

### Example 1

Define

$$T: \mathbb{R}^2 \rightarrow \mathbb{R}^2 \text{ by } T(a_1, a_2) = (2a_1 + a_2, a_1).$$

To show that  $T$  is linear, let  $c \in \mathbb{R}$  and  $x, y \in \mathbb{R}^2$ , where  $x = (b_1, b_2)$  and  $y = (d_1, d_2)$ . Since

$$cx + y = (cb_1 + d_1, cb_2 + d_2),$$

we have

$$T(cx + y) = (2(cb_1 + d_1) + cb_2 + d_2, cb_1 + d_1).$$

Also

$$\begin{aligned} cT(x) + T(y) &= c(2b_1 + b_2, b_1) + (2d_1 + d_2, d_1) \\ &= (2cb_1 + cb_2 + 2d_1 + d_2, cb_1 + d_1) \\ &= (2(cb_1 + d_1) + cb_2 + d_2, cb_1 + d_1). \end{aligned}$$

So  $T$  is linear.  $\spadesuit$

As we will see in Chapter 6, the applications of linear algebra to geometry are wide and varied. The main reason for this is that most of the important geometrical transformations are linear. Three particular transformations that we now consider are rotation, reflection, and projection. We leave the proofs of linearity to the reader.

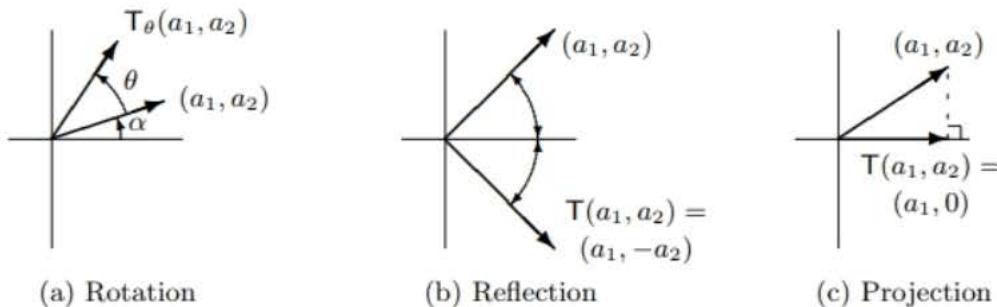


Figure 2.1

**Example 2**

For any angle  $\theta$ , define  $T_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  by the rule:  $T_\theta(a_1, a_2)$  is the vector obtained by rotating  $(a_1, a_2)$  counterclockwise by  $\theta$  if  $(a_1, a_2) \neq (0, 0)$ , and  $T_\theta(0, 0) = (0, 0)$ . Then  $T_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is a linear transformation that is called the **rotation by  $\theta$** .

We determine an explicit formula for  $T_\theta$ . Fix a nonzero vector  $(a_1, a_2) \in \mathbb{R}^2$ . Let  $\alpha$  be the angle that  $(a_1, a_2)$  makes with the positive  $x$ -axis (see Figure 2.1(a)), and let  $r = \sqrt{a_1^2 + a_2^2}$ . Then  $a_1 = r \cos \alpha$  and  $a_2 = r \sin \alpha$ . Also,  $T_\theta(a_1, a_2)$  has length  $r$  and makes an angle  $\alpha + \theta$  with the positive  $x$ -axis. It follows that

$$\begin{aligned} T_\theta(a_1, a_2) &= (r \cos(\alpha + \theta), r \sin(\alpha + \theta)) \\ &= (r \cos \alpha \cos \theta - r \sin \alpha \sin \theta, r \cos \alpha \sin \theta + r \sin \alpha \cos \theta) \\ &= (a_1 \cos \theta - a_2 \sin \theta, a_1 \sin \theta + a_2 \cos \theta). \end{aligned}$$

Finally, observe that this same formula is valid for  $(a_1, a_2) = (0, 0)$ .

It is now easy to show, as in Example 1, that  $T_\theta$  is linear. ◆

**Example 3**

Define  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  by  $T(a_1, a_2) = (a_1, -a_2)$ .  $T$  is called the **reflection about the  $x$ -axis**. (See Figure 2.1(b).) ◆

**Example 4**

Define  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  by  $T(a_1, a_2) = (a_1, 0)$ .  $T$  is called the **projection on the  $x$ -axis**. (See Figure 2.1(c).) ◆

We now look at some additional examples of linear transformations.

**Example 5**

Define  $T: M_{m \times n}(F) \rightarrow M_{n \times m}(F)$  by  $T(A) = A^t$ , where  $A^t$  is the transpose of  $A$ , defined in Section 1.3. Then  $T$  is a linear transformation by Exercise 3 of Section 1.3. ◆

**Example 6**

Let  $V$  denote the set of all real-valued functions defined on the real line that have derivatives of every order. It is easily shown that  $V$  is a vector space over  $R$ . (See Exercise 16 of Section 1.3.)

Define  $T: V \rightarrow V$  by  $T(f) = f'$ , the derivative of  $f$ . To show that  $T$  is linear, let  $g, h \in V$  and  $a \in R$ . Now

$$T(ag + h) = (ag + h)' = ag' + h' = aT(g) + T(h).$$

So by property 2 above,  $T$  is linear. ◆

**Example 7**

Let  $V = C(R)$ , the vector space of continuous real-valued functions on  $R$ . Let  $a, b \in R$ ,  $a < b$ . Define  $T: V \rightarrow R$  by

$$T(f) = \int_a^b f(t) dt$$

for all  $f \in V$ . Then  $T$  is a linear transformation because the definite integral of a linear combination of functions is the same as the linear combination of the definite integrals of the functions. ◆

Two very important examples of linear transformations that appear frequently in the remainder of the book, and therefore deserve their own notation, are the **identity** and **zero** transformations.

For vector spaces  $V$  and  $W$  (over  $F$ ), we define the **identity transformation**  $I_V: V \rightarrow V$  by  $I_V(x) = x$  for all  $x \in V$  and the **zero transformation**  $T_0: V \rightarrow W$  by  $T_0(x) = 0$  for all  $x \in V$ . It is clear that both of these transformations are linear. We often write  $I$  instead of  $I_V$ .

We now turn our attention to two very important sets associated with linear transformations: the *range* and *null space*. The determination of these sets allows us to examine more closely the intrinsic properties of a linear transformation.

**Definitions.** Let  $V$  and  $W$  be vector spaces, and let  $T: V \rightarrow W$  be linear. We define the **null space** (or **kernel**)  $N(T)$  of  $T$  to be the set of all vectors  $x$  in  $V$  such that  $T(x) = 0$ ; that is,  $N(T) = \{x \in V: T(x) = 0\}$ .

We define the **range** (or **image**)  $R(T)$  of  $T$  to be the subset of  $W$  consisting of all images (under  $T$ ) of vectors in  $V$ ; that is,  $R(T) = \{T(x): x \in V\}$ .

**Example 8**

Let  $V$  and  $W$  be vector spaces, and let  $I: V \rightarrow V$  and  $T_0: V \rightarrow W$  be the identity and zero transformations, respectively. Then  $N(I) = \{0\}$ ,  $R(I) = V$ ,  $N(T_0) = V$ , and  $R(T_0) = \{0\}$ . ◆

**Example 9**

Let  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^2$  be the linear transformation defined by

$$T(a_1, a_2, a_3) = (a_1 - a_2, 2a_3).$$

It is left as an exercise to verify that

$$N(T) = \{(a, a, 0) : a \in R\} \quad \text{and} \quad R(T) = \mathbb{R}^2. \quad \blacklozenge$$

In Examples 8 and 9, we see that the range and null space of each of the linear transformations is a subspace. The next result shows that this is true in general.

**Theorem 2.1.** *Let  $V$  and  $W$  be vector spaces and  $T: V \rightarrow W$  be linear. Then  $N(T)$  and  $R(T)$  are subspaces of  $V$  and  $W$ , respectively.*

*Proof.* To clarify the notation, we use the symbols  $\theta_V$  and  $\theta_W$  to denote the zero vectors of  $V$  and  $W$ , respectively.

Since  $T(\theta_V) = \theta_W$ , we have that  $\theta_V \in N(T)$ . Let  $x, y \in N(T)$  and  $c \in F$ . Then  $T(x+y) = T(x) + T(y) = \theta_W + \theta_W = \theta_W$ , and  $T(cx) = cT(x) = c\theta_W = \theta_W$ . Hence  $x+y \in N(T)$  and  $cx \in N(T)$ , so that  $N(T)$  is a subspace of  $V$ .

Because  $T(\theta_V) = \theta_W$ , we have that  $\theta_W \in R(T)$ . Now let  $x, y \in R(T)$  and  $c \in F$ . Then there exist  $v$  and  $w$  in  $V$  such that  $T(v) = x$  and  $T(w) = y$ . So  $T(v+w) = T(v) + T(w) = x+y$ , and  $T(cv) = cT(v) = cx$ . Thus  $x+y \in R(T)$  and  $cx \in R(T)$ , so  $R(T)$  is a subspace of  $W$ . ■

The next theorem provides a method for finding a spanning set for the range of a linear transformation. With this accomplished, a basis for the range is easy to discover using the technique of Example 6 of Section 1.6.

**Theorem 2.2.** *Let  $V$  and  $W$  be vector spaces, and let  $T: V \rightarrow W$  be linear. If  $\beta = \{v_1, v_2, \dots, v_n\}$  is a basis for  $V$ , then*

$$R(T) = \text{span}(T(\beta)) = \text{span}(\{T(v_1), T(v_2), \dots, T(v_n)\}).$$

*Proof.* Clearly  $T(v_i) \in R(T)$  for each  $i$ . Because  $R(T)$  is a subspace,  $R(T)$  contains  $\text{span}(\{T(v_1), T(v_2), \dots, T(v_n)\}) = \text{span}(T(\beta))$  by Theorem 1.5 (p. 31).

Now suppose that  $w \in R(T)$ . Then  $w = T(v)$  for some  $v \in V$ . Because  $\beta$  is a basis for  $V$ , we have

$$v = \sum_{i=1}^n a_i v_i \quad \text{for some } a_1, a_2, \dots, a_n \in F.$$

Since  $T$  is linear, it follows that

$$w = T(v) = \sum_{i=1}^n a_i T(v_i) \in \text{span}(T(\beta)).$$

So  $R(T)$  is contained in  $\text{span}(T(\beta))$ . ■

It should be noted that Theorem 2.2 is true if  $\beta$  is infinite, that is,  $R(T) = \text{span}(\{T(v) : v \in \beta\})$ . (See Exercise 34.)

The next example illustrates the usefulness of Theorem 2.2.

### Example 10

Define the linear transformation  $T: P_2(R) \rightarrow M_{2 \times 2}(R)$  by

$$T(f(x)) = \begin{pmatrix} f(1) - f(2) & 0 \\ 0 & f(0) \end{pmatrix}.$$

Since  $\beta = \{1, x, x^2\}$  is a basis for  $P_2(R)$ , we have

$$\begin{aligned} R(T) &= \text{span}(T(\beta)) = \text{span}(\{T(1), T(x), T(x^2)\}) \\ &= \text{span}\left(\left\{\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} -3 & 0 \\ 0 & 0 \end{pmatrix}\right\}\right) \\ &= \text{span}\left(\left\{\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}\right\}\right). \end{aligned}$$

Thus we have found a basis for  $R(T)$ , and so  $\dim(R(T)) = 2$ .

Now suppose that we want to find a basis for  $N(T)$ . Note that  $f(x) \in N(T)$  if and only if  $T(f(x)) = O$ , the  $2 \times 2$  zero matrix. That is,  $f(x) \in N(T)$  if and only if

$$\begin{pmatrix} f(1) - f(2) & 0 \\ 0 & f(0) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Let  $f(x) = a + bx + cx^2$ . Then

$$0 = f(1) - f(2) = (a + b + c) - (a + 2b + 4c) = -b - 3c$$

and  $0 = f(0) = a$ . Hence

$$f(x) = a + bx + cx^2 = -3cx + cx^2 = c(-3x + x^2).$$

Therefore a basis for  $N(T)$  is  $\{-3x + x^2\}$ .

Note that in this example

$$\dim(N(T)) + \dim(R(T)) = 1 + 2 = 3 = \dim(P_2(R)).$$

In Theorem 2.3, we see that a similar result is true in general. ◆

As in Chapter 1, we measure the “size” of a subspace by its dimension. The null space and range are so important that we attach special names to their respective dimensions.

**Definitions.** Let  $V$  and  $W$  be vector spaces, and let  $T: V \rightarrow W$  be linear. If  $N(T)$  and  $R(T)$  are finite-dimensional, then we define the **nullity** of  $T$ , denoted  $\text{nullity}(T)$ , and the **rank** of  $T$ , denoted  $\text{rank}(T)$ , to be the dimensions of  $N(T)$  and  $R(T)$ , respectively.

Reflecting on the action of a linear transformation, we see intuitively that the larger the nullity, the smaller the rank. In other words, the more vectors that are carried into  $0$ , the smaller the range. The same heuristic reasoning tells us that the larger the rank, the smaller the nullity. This balance between rank and nullity is made precise in the next theorem, appropriately called the *dimension theorem*.

**Theorem 2.3 (Dimension Theorem).** Let  $V$  and  $W$  be vector spaces, and let  $T: V \rightarrow W$  be linear. If  $V$  is finite-dimensional, then

$$\text{nullity}(T) + \text{rank}(T) = \dim(V).$$

*Proof.* Suppose that  $\dim(V) = n$ ,  $\dim(N(T)) = k$ , and  $\{v_1, v_2, \dots, v_k\}$  is a basis for  $N(T)$ . By the corollary to Theorem 1.11 (p. 51), we may extend  $\{v_1, v_2, \dots, v_k\}$  to a basis  $\beta = \{v_1, v_2, \dots, v_n\}$  for  $V$ . We claim that  $S = \{T(v_{k+1}), T(v_{k+2}), \dots, T(v_n)\}$  is a basis for  $R(T)$ .

First we prove that  $S$  generates  $R(T)$ . Using Theorem 2.2 and the fact that  $T(v_i) = 0$  for  $1 \leq i \leq k$ , we have

$$\begin{aligned} R(T) &= \text{span}(\{T(v_1), T(v_2), \dots, T(v_n)\}) \\ &= \text{span}(\{T(v_{k+1}), T(v_{k+2}), \dots, T(v_n)\}) = \text{span}(S). \end{aligned}$$

Now we prove that  $S$  is linearly independent. Suppose that

$$\sum_{i=k+1}^n b_i T(v_i) = 0 \quad \text{for } b_{k+1}, b_{k+2}, \dots, b_n \in F.$$

Using the fact that  $T$  is linear, we have

$$T \left( \sum_{i=k+1}^n b_i v_i \right) = 0.$$

So

$$\sum_{i=k+1}^n b_i v_i \in N(T).$$

Hence there exist  $c_1, c_2, \dots, c_k \in F$  such that

$$\sum_{i=k+1}^n b_i v_i = \sum_{i=1}^k c_i v_i \quad \text{or} \quad \sum_{i=1}^k (-c_i) v_i + \sum_{i=k+1}^n b_i v_i = 0.$$

Since  $\beta$  is a basis for  $V$ , we have  $b_i = 0$  for all  $i$ . Hence  $S$  is linearly independent. Notice that this argument also shows that  $T(v_{k+1}), T(v_{k+2}), \dots, T(v_n)$  are distinct; therefore  $\text{rank}(T) = n - k$ . ■

If we apply the dimension theorem to the linear transformation  $T$  in Example 9, we have that  $\text{nullity}(T) + 2 = 3$ , so  $\text{nullity}(T) = 1$ .

The reader should review the concepts of “one-to-one” and “onto” presented in Appendix B. Interestingly, for a linear transformation, both of these concepts are intimately connected to the rank and nullity of the transformation. This is demonstrated in the next two theorems.

**Theorem 2.4.** *Let  $V$  and  $W$  be vector spaces, and let  $T: V \rightarrow W$  be linear. Then  $T$  is one-to-one if and only if  $N(T) = \{0\}$ .*

*Proof.* Suppose that  $T$  is one-to-one and  $x \in N(T)$ . Then  $T(x) = 0 = T(0)$ . Since  $T$  is one-to-one, we have  $x = 0$ . Hence  $N(T) = \{0\}$ .

Now assume that  $N(T) = \{0\}$ , and suppose that  $T(x) = T(y)$ . Then  $0 = T(x) - T(y) = T(x - y)$  by property 3 on page 65. Therefore  $x - y \in N(T) = \{0\}$ . So  $x - y = 0$ , or  $x = y$ . This means that  $T$  is one-to-one. ■

The reader should observe that Theorem 2.4 allows us to conclude that the transformation defined in Example 9 is not one-to-one.

Surprisingly, the conditions of one-to-one and onto are equivalent in an important special case.

**Theorem 2.5.** *Let  $V$  and  $W$  be finite-dimensional vector spaces of equal dimension, and let  $T: V \rightarrow W$  be linear. Then the following are equivalent.*

- (a)  $T$  is one-to-one.
- (b)  $T$  is onto.
- (c)  $\text{rank}(T) = \dim(V)$ .

*Proof.* From the dimension theorem, we have

$$\text{nullity}(T) + \text{rank}(T) = \dim(V).$$

Now, with the use of Theorem 2.4, we have that  $T$  is one-to-one if and only if  $N(T) = \{0\}$ , if and only if  $\text{nullity}(T) = 0$ , if and only if  $\text{rank}(T) = \dim(V)$ , if and only if  $\text{rank}(T) = \dim(W)$ , and if and only if  $\dim(R(T)) = \dim(W)$ . By Theorem 1.11 (p. 50), this equality is equivalent to  $R(T) = W$ , the definition of  $T$  being onto. ■

We note that if  $V$  is not finite-dimensional and  $T: V \rightarrow V$  is linear, then it does *not* follow that one-to-one and onto are equivalent. (See Exercises 15, 16, and 21.)

The linearity of  $T$  in Theorems 2.4 and 2.5 is essential, for it is easy to construct examples of functions from  $R$  into  $R$  that are not one-to-one, but are onto, and vice versa.

The next two examples make use of the preceding theorems in determining whether a given linear transformation is one-to-one or onto.

### Example 11

Let  $T: P_2(R) \rightarrow P_3(R)$  be the linear transformation defined by

$$T(f(x)) = 2f'(x) + \int_0^x 3f(t) dt.$$

Now

$$R(T) = \text{span}(\{T(1), T(x), T(x^2)\}) = \text{span}(\{3x, 2 + \frac{3}{2}x^2, 4x + x^3\}).$$

Since  $\{3x, 2 + \frac{3}{2}x^2, 4x + x^3\}$  is linearly independent,  $\text{rank}(T) = 3$ . Since  $\dim(P_3(R)) = 4$ ,  $T$  is not onto. From the dimension theorem,  $\text{nullity}(T) + 3 = 3$ . So  $\text{nullity}(T) = 0$ , and therefore,  $N(T) = \{0\}$ . We conclude from Theorem 2.4 that  $T$  is one-to-one. ♦

### Example 12

Let  $T: F^2 \rightarrow F^2$  be the linear transformation defined by

$$T(a_1, a_2) = (a_1 + a_2, a_1).$$

It is easy to see that  $N(T) = \{0\}$ ; so  $T$  is one-to-one. Hence Theorem 2.5 tells us that  $T$  must be onto. ♦

In Exercise 14, it is stated that if  $T$  is linear and one-to-one, then a subset  $S$  is linearly independent if and only if  $T(S)$  is linearly independent. Example 13 illustrates the use of this result.

### Example 13

Let  $T: P_2(R) \rightarrow R^3$  be the linear transformation defined by

$$T(a_0 + a_1x + a_2x^2) = (a_0, a_1, a_2).$$

Clearly  $T$  is linear and one-to-one. Let  $S = \{2 - x + 3x^2, x + x^2, 1 - 2x^2\}$ . Then  $S$  is linearly independent in  $P_2(R)$  because

$$T(S) = \{(2, -1, 3), (0, 1, 1), (1, 0, -2)\}$$

is linearly independent in  $R^3$ . ♦

In Example 13, we transferred a property from the vector space of polynomials to a property in the vector space of 3-tuples. This technique is exploited more fully later.

One of the most important properties of a linear transformation is that it is completely determined by its action on a basis. This result, which follows from the next theorem and corollary, is used frequently throughout the book.

**Theorem 2.6.** Let  $V$  and  $W$  be vector spaces over  $F$ , and suppose that  $\{v_1, v_2, \dots, v_n\}$  is a basis for  $V$ . For  $w_1, w_2, \dots, w_n$  in  $W$ , there exists exactly one linear transformation  $T: V \rightarrow W$  such that  $T(v_i) = w_i$  for  $i = 1, 2, \dots, n$ .

*Proof.* Let  $x \in V$ . Then

$$x = \sum_{i=1}^n a_i v_i,$$

where  $a_1, a_2, \dots, a_n$  are unique scalars. Define

$$T: V \rightarrow W \quad \text{by} \quad T(x) = \sum_{i=1}^n a_i w_i.$$

(a)  $T$  is linear: Suppose that  $u, v \in V$  and  $d \in F$ . Then we may write

$$u = \sum_{i=1}^n b_i v_i \quad \text{and} \quad v = \sum_{i=1}^n c_i v_i$$

for some scalars  $b_1, b_2, \dots, b_n, c_1, c_2, \dots, c_n$ . Thus

$$du + v = \sum_{i=1}^n (db_i + c_i) v_i.$$

So

$$T(du + v) = \sum_{i=1}^n (db_i + c_i) w_i = d \sum_{i=1}^n b_i w_i + \sum_{i=1}^n c_i w_i = dT(u) + T(v).$$

(b) Clearly

$$T(v_i) = w_i \quad \text{for } i = 1, 2, \dots, n.$$

(c)  $T$  is unique: Suppose that  $U: V \rightarrow W$  is linear and  $U(v_i) = w_i$  for  $i = 1, 2, \dots, n$ . Then for  $x \in V$  with

$$x = \sum_{i=1}^n a_i v_i,$$

we have

$$U(x) = \sum_{i=1}^n a_i U(v_i) = \sum_{i=1}^n a_i w_i = T(x).$$

Hence  $U = T$ . ■

**Corollary.** Let  $V$  and  $W$  be vector spaces, and suppose that  $V$  has a finite basis  $\{v_1, v_2, \dots, v_n\}$ . If  $U, T: V \rightarrow W$  are linear and  $U(v_i) = T(v_i)$  for  $i = 1, 2, \dots, n$ , then  $U = T$ .

**Example 14**

Let  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be the linear transformation defined by

$$T(a_1, a_2) = (2a_2 - a_1, 3a_1),$$

and suppose that  $U: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is linear. If we know that  $U(1, 2) = (3, 3)$  and  $U(1, 1) = (1, 3)$ , then  $U = T$ . This follows from the corollary and from the fact that  $\{(1, 2), (1, 1)\}$  is a basis for  $\mathbb{R}^2$ . ♦

**EXERCISES**

1. Label the following statements as true or false. In each part,  $V$  and  $W$  are finite-dimensional vector spaces (over  $F$ ), and  $T$  is a function from  $V$  to  $W$ .
  - (a) If  $T$  is linear, then  $T$  preserves sums and scalar products.
  - (b) If  $T(x + y) = T(x) + T(y)$ , then  $T$  is linear.
  - (c)  $T$  is one-to-one if and only if the only vector  $x$  such that  $T(x) = 0$  is  $x = 0$ .
  - (d) If  $T$  is linear, then  $T(\theta_V) = \theta_W$ .
  - (e) If  $T$  is linear, then  $\text{nullity}(T) + \text{rank}(T) = \dim(W)$ .
  - (f) If  $T$  is linear, then  $T$  carries linearly independent subsets of  $V$  onto linearly independent subsets of  $W$ .
  - (g) If  $T, U: V \rightarrow W$  are both linear and agree on a basis for  $V$ , then  $T = U$ .
  - (h) Given  $x_1, x_2 \in V$  and  $y_1, y_2 \in W$ , there exists a linear transformation  $T: V \rightarrow W$  such that  $T(x_1) = y_1$  and  $T(x_2) = y_2$ .

For Exercises 2 through 6, prove that  $T$  is a linear transformation, and find bases for both  $N(T)$  and  $R(T)$ . Then compute the nullity and rank of  $T$ , and verify the dimension theorem. Finally, use the appropriate theorems in this section to determine whether  $T$  is one-to-one or onto.

2.  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^2$  defined by  $T(a_1, a_2, a_3) = (a_1 - a_2, 2a_3)$ .
3.  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by  $T(a_1, a_2) = (a_1 + a_2, 0, 2a_1 - a_2)$ .
4.  $T: M_{2 \times 3}(F) \rightarrow M_{2 \times 2}(F)$  defined by

$$T \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \begin{pmatrix} 2a_{11} - a_{12} & a_{13} + 2a_{12} \\ 0 & 0 \end{pmatrix}.$$

5.  $T: P_2(R) \rightarrow P_3(R)$  defined by  $T(f(x)) = xf(x) + f'(x)$ .

6.  $T: M_{n \times n}(F) \rightarrow F$  defined by  $T(A) = \text{tr}(A)$ . Recall (Example 4, Section 1.3) that

$$\text{tr}(A) = \sum_{i=1}^n A_{ii}.$$

7. Prove properties 1, 2, 3, and 4 on page 65.
8. Prove that the transformations in Examples 2 and 3 are linear.
9. In this exercise,  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is a function. For each of the following parts, state why  $T$  is *not* linear.
- (a)  $T(a_1, a_2) = (1, a_2)$
  - (b)  $T(a_1, a_2) = (a_1, a_1^2)$
  - (c)  $T(a_1, a_2) = (\sin a_1, 0)$
  - (d)  $T(a_1, a_2) = (|a_1|, a_2)$
  - (e)  $T(a_1, a_2) = (a_1 + 1, a_2)$
10. Suppose that  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is linear,  $T(1, 0) = (1, 4)$ , and  $T(1, 1) = (2, 5)$ . What is  $T(2, 3)$ ? Is  $T$  one-to-one?
11. Prove that there exists a linear transformation  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  such that  $T(1, 1) = (1, 0, 2)$  and  $T(2, 3) = (1, -1, 4)$ . What is  $T(8, 11)$ ?
12. Is there a linear transformation  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^2$  such that  $T(1, 0, 3) = (1, 1)$  and  $T(-2, 0, -6) = (2, 1)$ ?
13. Let  $V$  and  $W$  be vector spaces, let  $T: V \rightarrow W$  be linear, and let  $\{w_1, w_2, \dots, w_k\}$  be a linearly independent set of  $k$  vectors from  $R(T)$ . Prove that if  $S = \{v_1, v_2, \dots, v_k\}$  is chosen so that  $T(v_i) = w_i$  for  $i = 1, 2, \dots, k$ , then  $S$  is linearly independent. Visit [goo.gl/kmaQS2](http://goo.gl/kmaQS2) for a solution.
14. Let  $V$  and  $W$  be vector spaces and  $T: V \rightarrow W$  be linear.
- (a) Prove that  $T$  is one-to-one if and only if  $T$  carries linearly independent subsets of  $V$  onto linearly independent subsets of  $W$ .
  - (b) Suppose that  $T$  is one-to-one and that  $S$  is a subset of  $V$ . Prove that  $S$  is linearly independent if and only if  $T(S)$  is linearly independent.
  - (c) Suppose  $\beta = \{v_1, v_2, \dots, v_n\}$  is a basis for  $V$  and  $T$  is one-to-one and onto. Prove that  $T(\beta) = \{T(v_1), T(v_2), \dots, T(v_n)\}$  is a basis for  $W$ .
15. Recall the definition of  $P(R)$  on page 11. Define

$$T: P(R) \rightarrow P(R) \quad \text{by} \quad T(f(x)) = \int_0^x f(t) dt.$$

Prove that  $T$  linear and one-to-one, but not onto.

16. Let  $T: P(R) \rightarrow P(R)$  be defined by  $T(f(x)) = f'(x)$ . Recall that  $T$  is linear. Prove that  $T$  is onto, but not one-to-one.
17. Let  $V$  and  $W$  be finite-dimensional vector spaces and  $T: V \rightarrow W$  be linear.
- Prove that if  $\dim(V) < \dim(W)$ , then  $T$  cannot be onto.
  - Prove that if  $\dim(V) > \dim(W)$ , then  $T$  cannot be one-to-one.
18. Give an example of a linear transformation  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that  $N(T) = R(T)$ .
19. Give an example of vector spaces  $V$  and  $W$  and distinct linear transformations  $T$  and  $U$  from  $V$  to  $W$  such that  $N(T) = N(U)$  and  $R(T) = R(U)$ .
20. Let  $V$  and  $W$  be vector spaces with subspaces  $V_1$  and  $W_1$ , respectively. If  $T: V \rightarrow W$  is linear, prove that  $T(V_1)$  is a subspace of  $W$  and that  $\{x \in V: T(x) \in W_1\}$  is a subspace of  $V$ .
21. Let  $V$  be the vector space of sequences described in Example 5 of Section 1.2. Define the functions  $T, U: V \rightarrow V$  by

$$T(a_1, a_2, \dots) = (a_2, a_3, \dots) \quad \text{and} \quad U(a_1, a_2, \dots) = (0, a_1, a_2, \dots).$$

$T$  and  $U$  are called the **left shift** and **right shift** operators on  $V$ , respectively.

- Prove that  $T$  and  $U$  are linear.
  - Prove that  $T$  is onto, but not one-to-one.
  - Prove that  $U$  is one-to-one, but not onto.
22. Let  $T: \mathbb{R}^3 \rightarrow \mathbb{R}$  be linear. Show that there exist scalars  $a, b$ , and  $c$  such that  $T(x, y, z) = ax + by + cz$  for all  $(x, y, z) \in \mathbb{R}^3$ . Can you generalize this result for  $T: F^n \rightarrow F$ ? State and prove an analogous result for  $T: F^n \rightarrow F^m$ .
23. Let  $T: \mathbb{R}^3 \rightarrow \mathbb{R}$  be linear. Describe geometrically the possibilities for the null space of  $T$ . *Hint:* Use Exercise 22.
24. Let  $T: V \rightarrow W$  be linear,  $b \in W$ , and  $K = \{x \in V: T(x) = b\}$  be nonempty. Prove that if  $s \in K$ , then  $K = \{s\} + N(T)$ . (See page 22 for the definition of the sum of subsets.)

The following definition is used in Exercises 25–28 and in Exercise 31.

**Definition.** Let  $V$  be a vector space and  $W_1$  and  $W_2$  be subspaces of  $V$  such that  $V = W_1 \oplus W_2$ . (Recall the definition of direct sum given on page 22.) The function  $T: V \rightarrow V$  defined by  $T(x) = x_1$  where  $x = x_1 + x_2$  with  $x_1 \in W_1$  and  $x_2 \in W_2$ , is called the **projection of  $V$  on  $W_1$**  or the **projection on  $W_1$  along  $W_2$** .

25. Let  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ . Include figures for each of the following parts.
- Find a formula for  $T(a, b)$ , where  $T$  represents the projection on the  $y$ -axis along the  $x$ -axis.
  - Find a formula for  $T(a, b)$ , where  $T$  represents the projection on the  $y$ -axis along the line  $L = \{(s, s) : s \in \mathbb{R}\}$ .
26. Let  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ .
- If  $T(a, b, c) = (a, b, 0)$ , show that  $T$  is the projection on the  $xy$ -plane along the  $z$ -axis.
  - Find a formula for  $T(a, b, c)$ , where  $T$  represents the projection on the  $z$ -axis along the  $xy$ -plane.
  - If  $T(a, b, c) = (a - c, b, 0)$ , show that  $T$  is the projection on the  $xy$ -plane along the line  $L = \{(a, 0, a) : a \in \mathbb{R}\}$ .
27. Using the notation in the definition above, assume that  $T: V \rightarrow V$  is the projection on  $W_1$  along  $W_2$ .
- Prove that  $T$  is linear and  $W_1 = \{x \in V : T(x) = x\}$ .
  - Prove that  $W_1 = R(T)$  and  $W_2 = N(T)$ .
  - Describe  $T$  if  $W_1 = V$ .
  - Describe  $T$  if  $W_1$  is the zero subspace.
28. Suppose that  $W$  is a subspace of a finite-dimensional vector space  $V$ .
- Prove that there exists a subspace  $W'$  and a function  $T: V \rightarrow V$  such that  $T$  is a projection on  $W$  along  $W'$ .
  - Give an example of a subspace  $W$  of a vector space  $V$  such that there are two projections on  $W$  along two (distinct) subspaces.

The following definitions are used in Exercises 29–33.

**Definitions.** Let  $V$  be a vector space, and let  $T: V \rightarrow V$  be linear. A subspace  $W$  of  $V$  is said to be  **$T$ -invariant** if  $T(x) \in W$  for every  $x \in W$ , that is,  $T(W) \subseteq W$ . If  $W$  is  $T$ -invariant, we define the **restriction of  $T$  on  $W$**  to be the function  $T_W: W \rightarrow W$  defined by  $T_W(x) = T(x)$  for all  $x \in W$ .

Exercises 29–33 assume that  $W$  is a subspace of a vector space  $V$  and that  $T: V \rightarrow V$  is linear. Warning: Do not assume that  $W$  is  $T$ -invariant or that  $T$  is a projection unless explicitly stated.

- Prove that the subspaces  $\{\mathbf{0}\}$ ,  $V$ ,  $R(T)$ , and  $N(T)$  are all  $T$ -invariant.
- If  $W$  is  $T$ -invariant, prove that  $T_W$  is linear.
- Suppose that  $T$  is the projection on  $W$  along some subspace  $W'$ . Prove that  $W$  is  $T$ -invariant and that  $T_W = I_W$ .

32. Suppose that  $V = R(T) \oplus W$  and  $W$  is  $T$ -invariant. See page 22 for the definition of *direct sum*.
- Prove that  $W \subseteq N(T)$ .
  - Show that if  $V$  is finite-dimensional, then  $W = N(T)$ .
  - Show by example that the conclusion of (b) is not necessarily true if  $V$  is not finite-dimensional.
33. Suppose that  $W$  is  $T$ -invariant. Prove that  $N(T_W) = N(T) \cap W$  and  $R(T_W) = T(W)$ .
34. Prove Theorem 2.2 for the case that  $\beta$  is infinite, that is,  $R(T) = \text{span}(\{T(v) : v \in \beta\})$ .
35. Prove the following generalization of Theorem 2.6: Let  $V$  and  $W$  be vector spaces over a common field, and let  $\beta$  be a basis for  $V$ . Then for any function  $f: \beta \rightarrow W$  there exists exactly one linear transformation  $T: V \rightarrow W$  such that  $T(x) = f(x)$  for all  $x \in \beta$ .

Exercises 36 and 37 require the definition of *direct sum* given on page 22.

36. Let  $V$  be a finite-dimensional vector space and  $T: V \rightarrow V$  be linear.
- Suppose that  $V = R(T) + N(T)$ . Prove that  $V = R(T) \oplus N(T)$ .
  - Suppose that  $R(T) \cap N(T) = \{0\}$ . Prove that  $V = R(T) \oplus N(T)$ .
- Be careful to say in each part where finite-dimensionality is used.
37. Let  $V$  and  $T$  be as defined in Exercise 21.
- Prove that  $V = R(T) + N(T)$ , but  $V$  is not a direct sum of these two spaces. Thus the result of Exercise 36(a) above cannot be proved without assuming that  $V$  is finite-dimensional.
  - Find a linear operator  $T_1$  on  $V$  such that  $R(T_1) \cap N(T_1) = \{0\}$  but  $V$  is not a direct sum of  $R(T_1)$  and  $N(T_1)$ . Conclude that  $V$  being finite-dimensional is also essential in Exercise 36(b).
38. A function  $T: V \rightarrow W$  between vector spaces  $V$  and  $W$  is called **additive** if  $T(x + y) = T(x) + T(y)$  for all  $x, y \in V$ . Prove that if  $V$  and  $W$  are vector spaces over the field of rational numbers, then any additive function from  $V$  into  $W$  is a linear transformation.
39. Let  $T: C \rightarrow C$  be the function defined by  $T(z) = \bar{z}$ . Prove that  $T$  is additive (as defined in Exercise 38) but not linear.
40. Prove that there is an additive function  $T: R \rightarrow R$  (as defined in Exercise 38) that is not linear. *Hint:* Let  $V$  be the set of real numbers regarded as a vector space over the field of rational numbers. By the corollary to Theorem 1.13 (p. 61),  $V$  has a basis  $\beta$ . Let  $x$  and  $y$  be two

distinct vectors in  $\beta$ , and define  $f: \beta \rightarrow V$  by  $f(x) = y$ ,  $f(y) = x$ , and  $f(z) = z$  otherwise. By Exercise 35, there exists a linear transformation  $T: V \rightarrow V$  such that  $T(u) = f(u)$  for all  $u \in \beta$ . Then  $T$  is additive, but for  $c = y/x$ ,  $T(cx) \neq cT(x)$ .

41. Prove that Theorem 2.6 and its corollary are true when  $V$  is infinite-dimensional.

The following exercise requires familiarity with the definition of *quotient space* given in Exercise 31 of Section 1.3.

42. Let  $V$  be a vector space and  $W$  be a subspace of  $V$ . Define the mapping  $\eta: V \rightarrow V/W$  by  $\eta(v) = v + W$  for  $v \in V$ .
- (a) Prove that  $\eta$  is a linear transformation from  $V$  onto  $V/W$  and that  $N(\eta) = W$ .
  - (b) Suppose that  $V$  is finite-dimensional. Use (a) and the dimension theorem to derive a formula relating  $\dim(V)$ ,  $\dim(W)$ , and  $\dim(V/W)$ .
  - (c) Read the proof of the dimension theorem. Compare the method of solving (b) with the method of deriving the same result as outlined in Exercise 35 of Section 1.6.

## 2.2 THE MATRIX REPRESENTATION OF A LINEAR TRANSFORMATION

Until now, we have studied linear transformations by examining their ranges and null spaces. In this section, we embark on one of the most useful approaches to the analysis of a linear transformation on a finite-dimensional vector space: the representation of a linear transformation by a matrix. In fact, we develop a one-to-one correspondence between matrices and linear transformations that allows us to utilize properties of one to study properties of the other.

We first need the concept of an *ordered basis* for a vector space.

**Definition.** Let  $V$  be a finite-dimensional vector space. An **ordered basis** for  $V$  is a basis for  $V$  endowed with a specific order; that is, an ordered basis for  $V$  is a finite sequence of linearly independent vectors in  $V$  that generates  $V$ .

### Example 1

In  $F^3$ ,  $\beta = \{e_1, e_2, e_3\}$  can be considered an ordered basis. Also  $\gamma = \{e_2, e_1, e_3\}$  is an ordered basis, but  $\beta \neq \gamma$  as ordered bases. ♦

For the vector space  $F^n$ , we call  $\{e_1, e_2, \dots, e_n\}$  the **standard ordered basis** for  $F^n$ . Similarly, for the vector space  $P_n(F)$ , we call  $\{1, x, \dots, x^n\}$  the **standard ordered basis** for  $P_n(F)$ .

Now that we have the concept of ordered basis, we can identify abstract vectors in an  $n$ -dimensional vector space with  $n$ -tuples. This identification is provided through the use of *coordinate vectors*, as introduced next.

**Definition.** Let  $\beta = \{u_1, u_2, \dots, u_n\}$  be an ordered basis for a finite-dimensional vector space  $V$ . For  $x \in V$ , let  $a_1, a_2, \dots, a_n$  be the unique scalars such that

$$x = \sum_{i=1}^n a_i u_i.$$

We define the **coordinate vector of  $x$  relative to  $\beta$** , denoted  $[x]_\beta$ , by

$$[x]_\beta = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

Notice that  $[u_i]_\beta = e_i$  in the preceding definition. It is left as an exercise to show that the correspondence  $x \rightarrow [x]_\beta$  provides us with a linear transformation from  $V$  to  $F^n$ . We study this transformation in Section 2.4 in more detail.

### Example 2

Let  $V = P_2(R)$ , and let  $\beta = \{1, x, x^2\}$  be the standard ordered basis for  $V$ . If  $f(x) = 4 + 6x - 7x^2$ , then

$$[f]_\beta = \begin{pmatrix} 4 \\ 6 \\ -7 \end{pmatrix}. \quad \blacklozenge$$

Let us now proceed with the promised matrix representation of a linear transformation. Suppose that  $V$  and  $W$  are finite-dimensional vector spaces with ordered bases  $\beta = \{v_1, v_2, \dots, v_n\}$  and  $\gamma = \{w_1, w_2, \dots, w_m\}$ , respectively. Let  $T: V \rightarrow W$  be linear. Then for each  $j = 1, 2, \dots, n$ , there exist unique scalars  $a_{ij} \in F$ ,  $i = 1, 2, \dots, m$ , such that

$$T(v_j) = \sum_{i=1}^m a_{ij} w_i \quad \text{for } j = 1, 2, \dots, n.$$

**Definition.** Using the notation above, we call the  $m \times n$  matrix  $A$  defined by  $A_{ij} = a_{ij}$  the **matrix representation of  $T$  in the ordered bases  $\beta$  and  $\gamma$**  and write  $A = [T]_\beta^\gamma$ . If  $V = W$  and  $\beta = \gamma$ , then we write  $A = [T]_\beta$ .

Notice that the  $j$ th column of  $A$  is simply  $[\mathbf{T}(v_j)]_\gamma$ . Also observe that if  $\mathbf{U}: \mathbf{V} \rightarrow \mathbf{W}$  is a linear transformation such that  $[\mathbf{U}]_\beta^\gamma = [\mathbf{T}]_\beta^\gamma$ , then  $\mathbf{U} = \mathbf{T}$  by the corollary to Theorem 2.6 (p. 73).

We illustrate the computation of  $[\mathbf{T}]_\beta^\gamma$  in the next several examples.

### Example 3

Let  $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be the linear transformation defined by

$$\mathbf{T}(a_1, a_2) = (a_1 + 3a_2, 0, 2a_1 - 4a_2).$$

Let  $\beta$  and  $\gamma$  be the standard ordered bases for  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , respectively. Now

$$\mathbf{T}(1, 0) = (1, 0, 2) = 1e_1 + 0e_2 + 2e_3$$

and

$$\mathbf{T}(0, 1) = (3, 0, -4) = 3e_1 + 0e_2 - 4e_3.$$

Hence

$$[\mathbf{T}]_\beta^\gamma = \begin{pmatrix} 1 & 3 \\ 0 & 0 \\ 2 & -4 \end{pmatrix}.$$

If we let  $\gamma' = \{e_3, e_2, e_1\}$ , then

$$[\mathbf{T}]_{\beta'}^{\gamma'} = \begin{pmatrix} 2 & -4 \\ 0 & 0 \\ 1 & 3 \end{pmatrix}. \quad \blacklozenge$$

### Example 4

Let  $\mathbf{T}: \mathbb{P}_3(\mathbb{R}) \rightarrow \mathbb{P}_2(\mathbb{R})$  be the linear transformation defined by  $\mathbf{T}(f(x)) = f'(x)$ . Let  $\beta$  and  $\gamma$  be the standard ordered bases for  $\mathbb{P}_3(\mathbb{R})$  and  $\mathbb{P}_2(\mathbb{R})$ , respectively. Then

$$\begin{aligned} \mathbf{T}(1) &= 0 \cdot 1 + 0 \cdot x + 0 \cdot x^2 \\ \mathbf{T}(x) &= 1 \cdot 1 + 0 \cdot x + 0 \cdot x^2 \\ \mathbf{T}(x^2) &= 0 \cdot 1 + 2 \cdot x + 0 \cdot x^2 \\ \mathbf{T}(x^3) &= 0 \cdot 1 + 0 \cdot x + 3 \cdot x^2. \end{aligned}$$

So

$$[\mathbf{T}]_\beta^\gamma = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Note that when  $\mathbf{T}(x^j)$  is written as a linear combination of the vectors of  $\gamma$ , its coefficients give the entries of column  $j+1$  of  $[\mathbf{T}]_\beta^\gamma$ .  $\blacklozenge$

Let  $V$  and  $W$  be finite-dimensional vector spaces with ordered bases  $\beta = \{v_1, v_2, \dots, v_n\}$  and  $\gamma = \{w_1, w_2, \dots, w_m\}$ , respectively. Then

$$T_0(v_j) = 0 = 0w_1 + 0w_2 + \cdots + 0w_m.$$

Hence  $[T_0]_{\beta}^{\gamma} = O$ , the  $m \times n$  zero matrix. Also,

$$I_V(v_j) = v_j = 0v_1 + 0v_2 + \cdots + 0v_{j-1} + 1v_j + 0v_{j+1} + \cdots + 0v_n.$$

Hence the  $j$ th column of  $I_V$  is  $e_j$ , that is,

$$[I_V]_{\beta} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

The preceding matrix is called the  $n \times n$  *identity matrix* and is defined next, along with a very useful notation, the *Kronecker delta*.

**Definitions.** We define the **Kronecker delta**  $\delta_{ij}$  by  $\delta_{ij} = 1$  if  $i = j$  and  $\delta_{ij} = 0$  if  $i \neq j$ . The  $n \times n$  **identity matrix**  $I_n$  is defined by  $(I_n)_{ij} = \delta_{ij}$ . When the context is clear, we sometimes omit the subscript  $n$  from  $I_n$ .

For example,

$$I_1 = (1), \quad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus, the matrix representation of a zero transformation is a zero matrix, and the matrix representation of an identity transformation is an identity matrix.

Now that we have defined a procedure for associating matrices with linear transformations, we show in Theorem 2.8 that this association “preserves” addition and scalar multiplication. To make this more explicit, we need some preliminary discussion about the addition and scalar multiplication of linear transformations.

**Definition.** Let  $T, U: V \rightarrow W$  be arbitrary functions, where  $V$  and  $W$  are vector spaces over  $F$ , and let  $a \in F$ . We define  $T + U: V \rightarrow W$  by  $(T + U)(x) = T(x) + U(x)$  for all  $x \in V$ , and  $aT: V \rightarrow W$  by  $(aT)(x) = aT(x)$  for all  $x \in V$ .

Of course, these are just the usual definitions of addition and scalar multiplication of functions. We are fortunate, however, to have the result that both sums and scalar multiples of linear transformations are also linear.

**Theorem 2.7.** Let  $V$  and  $W$  be vector spaces over a field  $F$ , and let  $T, U: V \rightarrow W$  be linear.

(a) For all  $a \in F$ ,  $aT + U$  is linear.

(b) Using the operations of addition and scalar multiplication in the preceding definition, the collection of all linear transformations from  $V$  to  $W$  is a vector space over  $F$ .

*Proof.* (a) Let  $x, y \in V$  and  $c \in F$ . Then

$$\begin{aligned}(aT + U)(cx + y) &= aT(cx + y) + U(cx + y) \\&= a[T(cx + y)] + cU(x) + U(y) \\&= a[cT(x) + T(y)] + cU(x) + U(y) \\&= acT(x) + cU(x) + aT(y) + U(y) \\&= c(aT + U)(x) + (aT + U)(y).\end{aligned}$$

So  $aT + U$  is linear.

(b) Noting that  $T_0$ , the zero transformation, plays the role of the zero vector, it is easy to verify that the axioms of a vector space are satisfied, and hence that the collection of all linear transformations from  $V$  into  $W$  is a vector space over  $F$ . ■

**Definitions.** Let  $V$  and  $W$  be vector spaces over  $F$ . We denote the vector space of all linear transformations from  $V$  into  $W$  by  $\mathcal{L}(V, W)$ . In the case that  $V = W$ , we write  $\mathcal{L}(V)$  instead of  $\mathcal{L}(V, V)$ .

In Section 2.4, we see a complete identification of  $\mathcal{L}(V, W)$  with the vector space  $M_{m \times n}(F)$ , where  $n$  and  $m$  are the dimensions of  $V$  and  $W$ , respectively. This identification is easily established by the use of the next theorem.

**Theorem 2.8.** Let  $V$  and  $W$  be finite-dimensional vector spaces with ordered bases  $\beta$  and  $\gamma$ , respectively, and let  $T, U: V \rightarrow W$  be linear transformations. Then

(a)  $[T + U]_\beta^\gamma = [T]_\beta^\gamma + [U]_\beta^\gamma$  and

(b)  $[aT]_\beta^\gamma = a[T]_\beta^\gamma$  for all scalars  $a$ .

*Proof.* Let  $\beta = \{v_1, v_2, \dots, v_n\}$  and  $\gamma = \{w_1, w_2, \dots, w_m\}$ . There exist unique scalars  $a_{ij}$  and  $b_{ij}$  ( $1 \leq i \leq m, 1 \leq j \leq n$ ) such that

$$T(v_j) = \sum_{i=1}^m a_{ij} w_i \quad \text{and} \quad U(v_j) = \sum_{i=1}^m b_{ij} w_i \quad \text{for } 1 \leq j \leq n.$$

Hence

$$(T + U)(v_j) = \sum_{i=1}^m (a_{ij} + b_{ij}) w_i.$$

Thus

$$([\mathbf{T} + \mathbf{U}]_{\beta}^{\gamma})_{ij} = a_{ij} + b_{ij} = ([\mathbf{T}]_{\beta}^{\gamma} + [\mathbf{U}]_{\beta}^{\gamma})_{ij}.$$

So (a) is proved, and the proof of (b) is similar. ■

### Example 5

Let  $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  and  $\mathbf{U}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be the linear transformations respectively defined by

$$\mathbf{T}(a_1, a_2) = (a_1 + 3a_2, 0, 2a_1 - 4a_2) \text{ and } \mathbf{U}(a_1, a_2) = (a_1 - a_2, 2a_1, 3a_1 + 2a_2).$$

Let  $\beta$  and  $\gamma$  be the standard ordered bases of  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , respectively. Then

$$[\mathbf{T}]_{\beta}^{\gamma} = \begin{pmatrix} 1 & 3 \\ 0 & 0 \\ 2 & -4 \end{pmatrix},$$

(as computed in Example 3), and

$$[\mathbf{U}]_{\beta}^{\gamma} = \begin{pmatrix} 1 & -1 \\ 2 & 0 \\ 3 & 2 \end{pmatrix}.$$

If we compute  $\mathbf{T} + \mathbf{U}$  using the preceding definitions, we obtain

$$(\mathbf{T} + \mathbf{U})(a_1, a_2) = (2a_1 + 2a_2, 2a_1, 5a_1 - 2a_2).$$

So

$$[\mathbf{T} + \mathbf{U}]_{\beta}^{\gamma} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 5 & -2 \end{pmatrix},$$

which is simply  $[\mathbf{T}]_{\beta}^{\gamma} + [\mathbf{U}]_{\beta}^{\gamma}$ , illustrating Theorem 2.8. ◆

## EXERCISES

- Label the following statements as true or false. Assume that  $V$  and  $W$  are finite-dimensional vector spaces with ordered bases  $\beta$  and  $\gamma$ , respectively, and  $\mathbf{T}, \mathbf{U}: V \rightarrow W$  are linear transformations.
  - For any scalar  $a$ ,  $a\mathbf{T} + \mathbf{U}$  is a linear transformation from  $V$  to  $W$ .
  - $[\mathbf{T}]_{\beta}^{\gamma} = [\mathbf{U}]_{\beta}^{\gamma}$  implies that  $\mathbf{T} = \mathbf{U}$ .
  - If  $m = \dim(V)$  and  $n = \dim(W)$ , then  $[\mathbf{T}]_{\beta}^{\gamma}$  is an  $m \times n$  matrix.
  - $[\mathbf{T} + \mathbf{U}]_{\beta}^{\gamma} = [\mathbf{T}]_{\beta}^{\gamma} + [\mathbf{U}]_{\beta}^{\gamma}$ .
  - $\mathcal{L}(V, W)$  is a vector space.
  - $\mathcal{L}(V, W) = \mathcal{L}(W, V)$ .

2. Let  $\beta$  and  $\gamma$  be the standard ordered bases for  $\mathbb{R}^n$  and  $\mathbb{R}^m$ , respectively. For each linear transformation  $T: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , compute  $[T]_{\beta}^{\gamma}$ .

- (a)  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by  $T(a_1, a_2) = (2a_1 - a_2, 3a_1 + 4a_2, a_1)$ .
- (b)  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^2$  defined by  $T(a_1, a_2, a_3) = (2a_1 + 3a_2 - a_3, a_1 + a_3)$ .
- (c)  $T: \mathbb{R}^3 \rightarrow \mathbb{R}$  defined by  $T(a_1, a_2, a_3) = 2a_1 + a_2 - 3a_3$ .
- (d)  $T: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  defined by

$$T(a_1, a_2, a_3) = (2a_2 + a_3, -a_1 + 4a_2 + 5a_3, a_1 + a_3).$$

- (e)  $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$  defined by  $T(a_1, a_2, \dots, a_n) = (a_1, a_1, \dots, a_1)$ .
- (f)  $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$  defined by  $T(a_1, a_2, \dots, a_n) = (a_n, a_{n-1}, \dots, a_1)$ .
- (g)  $T: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by  $T(a_1, a_2, \dots, a_n) = a_1 + a_n$ .

3. Let  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be defined by  $T(a_1, a_2) = (a_1 - a_2, a_1, 2a_1 + a_2)$ . Let  $\beta$  be the standard ordered basis for  $\mathbb{R}^2$  and  $\gamma = \{(1, 1, 0), (0, 1, 1), (2, 2, 3)\}$ . Compute  $[T]_{\beta}^{\gamma}$ . If  $\alpha = \{(1, 2), (2, 3)\}$ , compute  $[T]_{\alpha}^{\gamma}$ .

4. Define

$$T: M_{2 \times 2}(R) \rightarrow P_2(R) \quad \text{by} \quad T \begin{pmatrix} a & b \\ c & d \end{pmatrix} = (a+b) + (2d)x + bx^2.$$

Let

$$\beta = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\} \quad \text{and} \quad \gamma = \{1, x, x^2\}.$$

Compute  $[T]_{\beta}^{\gamma}$ .

5. Let

$$\alpha = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\},$$

$$\beta = \{1, x, x^2\},$$

and

$$\gamma = \{1\}.$$

- (a) Define  $T: M_{2 \times 2}(F) \rightarrow M_{2 \times 2}(F)$  by  $T(A) = A^t$ . Compute  $[T]_{\alpha}$ .
- (b) Define

$$T: P_2(R) \rightarrow M_{2 \times 2}(R) \quad \text{by} \quad T(f(x)) = \begin{pmatrix} f'(0) & 2f(1) \\ 0 & f''(3) \end{pmatrix},$$

where  $'$  denotes differentiation. Compute  $[T]_{\beta}^{\alpha}$ .

- (c) Define  $T: M_{2 \times 2}(F) \rightarrow F$  by  $T(A) = \text{tr}(A)$ . Compute  $[T]_{\alpha}^{\gamma}$ .

- (d) Define  $T: P_2(R) \rightarrow R$  by  $T(f(x)) = f(2)$ . Compute  $[T]_{\beta}^{\gamma}$ .  
 (e) If

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

compute  $[A]_{\alpha}$ .

- (f) If  $f(x) = 3 - 6x + x^2$ , compute  $[f(x)]_{\beta}$ .  
 (g) For  $a \in F$ , compute  $[a]_{\gamma}$ .

6. Complete the proof of part (b) of Theorem 2.7.
7. Prove part (b) of Theorem 2.8.
- 8.<sup>†</sup> Let  $V$  be an  $n$ -dimensional vector space with an ordered basis  $\beta$ . Define  $T: V \rightarrow F^n$  by  $T(x) = [x]_{\beta}$ . Prove that  $T$  is linear.
9. Let  $V$  be the vector space of complex numbers over the field  $R$ . Define  $T: V \rightarrow V$  by  $T(z) = \bar{z}$ , where  $\bar{z}$  is the complex conjugate of  $z$ . Prove that  $T$  is linear, and compute  $[T]_{\beta}$ , where  $\beta = \{1, i\}$ . (Recall by Exercise 39 of Section 2.1 that  $T$  is not linear if  $V$  is regarded as a vector space over the field  $C$ .)
10. Let  $V$  be a vector space with the ordered basis  $\beta = \{v_1, v_2, \dots, v_n\}$ . Define  $v_0 = \theta$ . By Theorem 2.6 (p. 73), there exists a linear transformation  $T: V \rightarrow V$  such that  $T(v_j) = v_j + v_{j-1}$  for  $j = 1, 2, \dots, n$ . Compute  $[T]_{\beta}$ .
11. Let  $V$  be an  $n$ -dimensional vector space, and let  $T: V \rightarrow V$  be a linear transformation. Suppose that  $W$  is a  $T$ -invariant subspace of  $V$  (see the exercises of Section 2.1) having dimension  $k$ . Show that there is a basis  $\beta$  for  $V$  such that  $[T]_{\beta}$  has the form

$$\begin{pmatrix} A & B \\ O & C \end{pmatrix},$$

where  $A$  is a  $k \times k$  matrix and  $O$  is the  $(n-k) \times k$  zero matrix.

- 12.<sup>†</sup> Let  $\beta = \{v_1, v_2, \dots, v_n\}$  be a basis for a vector space  $V$  and  $T: V \rightarrow V$  be a linear transformation. Prove that  $[T]_{\beta}$  is upper triangular if and only if  $T(v_j) \in \text{span}(\{v_1, v_2, \dots, v_j\})$  for  $j = 1, 2, \dots, n$ . Visit [goo.gl/k9ZrQb](http://goo.gl/k9ZrQb) for a solution.
13. Let  $V$  be a finite-dimensional vector space and  $T$  be the projection on  $W$  along  $W'$ , where  $W$  and  $W'$  are subspaces of  $V$ . (See the definition in the exercises of Section 2.1 on page 76.) Find an ordered basis  $\beta$  for  $V$  such that  $[T]_{\beta}$  is a diagonal matrix.

14. Let  $V$  and  $W$  be vector spaces, and let  $T$  and  $U$  be nonzero linear transformations from  $V$  into  $W$ . If  $R(T) \cap R(U) = \{0\}$ , prove that  $\{T, U\}$  is a linearly independent subset of  $\mathcal{L}(V, W)$ .
15. Let  $V = P(R)$ , and for  $j \geq 1$  define  $T_j(f(x)) = f^{(j)}(x)$ , where  $f^{(j)}(x)$  is the  $j$ th derivative of  $f(x)$ . Prove that the set  $\{T_1, T_2, \dots, T_n\}$  is a linearly independent subset of  $\mathcal{L}(V)$  for any positive integer  $n$ .
16. Let  $V$  and  $W$  be vector spaces, and let  $S$  be a subset of  $V$ . Define  $S^0 = \{T \in \mathcal{L}(V, W) : T(x) = 0 \text{ for all } x \in S\}$ . Prove the following statements.
- $S^0$  is a subspace of  $\mathcal{L}(V, W)$ .
  - If  $S_1$  and  $S_2$  are subsets of  $V$  and  $S_1 \subseteq S_2$ , then  $S_2^0 \subseteq S_1^0$ .
  - If  $V_1$  and  $V_2$  are subspaces of  $V$ , then  $(V_1 \cup V_2)^0 = (V_1 + V_2)^0 = V_1^0 \cap V_2^0$ .
17. Let  $V$  and  $W$  be vector spaces such that  $\dim(V) = \dim(W)$ , and let  $T: V \rightarrow W$  be linear. Show that there exist ordered bases  $\beta$  and  $\gamma$  for  $V$  and  $W$ , respectively, such that  $[T]_{\beta}^{\gamma}$  is a diagonal matrix.

### 2.3 COMPOSITION OF LINEAR TRANSFORMATIONS AND MATRIX MULTIPLICATION

In Section 2.2, we learned how to associate a matrix with a linear transformation in such a way that both sums and scalar multiples of matrices are associated with the corresponding sums and scalar multiples of the transformations. The question now arises as to how the matrix representation of a composite of linear transformations is related to the matrix representation of each of the associated linear transformations. The attempt to answer this question leads to a definition of matrix multiplication. We use the more convenient notation of  $UT$  rather than  $U \circ T$  for the composite of linear transformations  $U$  and  $T$ . (See Appendix B.)

Our first result shows that the composite of linear transformations is linear.

**Theorem 2.9.** *Let  $V$ ,  $W$ , and  $Z$  be vector spaces over the same field  $F$ , and let  $T: V \rightarrow W$  and  $U: W \rightarrow Z$  be linear. Then  $UT: V \rightarrow Z$  is linear.*

*Proof.* Let  $x, y \in V$  and  $a \in F$ . Then

$$\begin{aligned} UT(ax + y) &= U(T(ax + y)) = U(aT(x) + T(y)) \\ &= aU(T(x)) + U(T(y)) = a(UT)(x) + UT(y). \end{aligned}$$

The following theorem lists some of the properties of the composition of linear transformations.

**Theorem 2.10.** Let  $V$  be a vector space. Let  $T, U_1, U_2 \in \mathcal{L}(V)$ . Then

- (a)  $T(U_1 + U_2) = TU_1 + TU_2$  and  $(U_1 + U_2)T = U_1T + U_2T$
- (b)  $T(U_1U_2) = (TU_1)U_2$
- (c)  $TI = IT = T$
- (d)  $a(U_1U_2) = (aU_1)U_2 = U_1(aU_2)$  for all scalars  $a$ .

*Proof.* Exercise. ■

A more general result holds for linear transformations that have domains unequal to their codomains. (See Exercise 8.)

If  $T \in \mathcal{L}(V)$ , there are circumstances where it is natural to compose  $T$  with itself one or more times. In Example 6 of Section 2.1, for instance, we considered the linear transformation  $T: V \rightarrow V$  defined by  $T(f) = f'$ , where  $V$  denotes the set of all real-valued functions on the real line that have derivatives of every order. In this context,  $TT(f) = T(f') = (f')' = f''$  is the second derivative of  $f$ , and  $TTT(f) = f'''$  is the third derivative of  $f$ . In this type of situation, the following notation is useful.

If  $T \in \mathcal{L}(V)$ , we define  $T^1 = T$ ,  $T^2 = TT$ ,  $T^3 = T^2T$ , and, in general,  $T^k = T^{k-1}T$  for  $k = 2, 3, \dots$ . For convenience, we also define  $T^0 = I_V$ .

We now turn our attention to the multiplication of matrices. Let  $V$ ,  $W$ , and  $Z$  be finite-dimensional vector spaces and  $T: V \rightarrow W$  and  $U: W \rightarrow Z$  be linear transformations. Suppose that  $A = [U]_{\beta}^{\gamma}$  and  $B = [T]_{\alpha}^{\beta}$ , where  $\alpha = \{v_1, v_2, \dots, v_n\}$ ,  $\beta = \{w_1, w_2, \dots, w_m\}$ , and  $\gamma = \{z_1, z_2, \dots, z_p\}$  are ordered bases for  $V$ ,  $W$ , and  $Z$ , respectively. We would like to define the product  $AB$  of two matrices so that  $AB = [UT]_{\alpha}^{\gamma}$ . Consider the matrix  $[UT]_{\alpha}^{\gamma}$ . For  $j = 1, 2, \dots, n$ , we have

$$\begin{aligned} (UT)(v_j) &= U(T(v_j)) = U\left(\sum_{k=1}^m B_{kj}w_k\right) = \sum_{k=1}^m B_{kj}U(w_k) \\ &= \sum_{k=1}^m B_{kj}\left(\sum_{i=1}^p A_{ik}z_i\right) = \sum_{i=1}^p \left(\sum_{k=1}^m A_{ik}B_{kj}\right)z_i \\ &= \sum_{i=1}^p C_{ij}z_i, \end{aligned}$$

where

$$C_{ij} = \sum_{k=1}^m A_{ik}B_{kj}.$$

This computation motivates the following definition of matrix multiplication.

**Definition.** Let  $A$  be an  $m \times n$  matrix and  $B$  be an  $n \times p$  matrix. We define the **product** of  $A$  and  $B$ , denoted  $AB$ , to be the  $m \times p$  matrix such that

$$(AB)_{ij} = \sum_{k=1}^n A_{ik}B_{kj} \quad \text{for } 1 \leq i \leq m, \quad 1 \leq j \leq p.$$

Note that  $(AB)_{ij}$  is the sum of products of corresponding entries from the  $i$ th row of  $A$  and the  $j$ th column of  $B$ . Some interesting applications of this definition are presented at the end of this section.

The reader should observe that in order for the product  $AB$  to be defined, there are restrictions regarding the relative sizes of  $A$  and  $B$ . The following mnemonic device is helpful: “ $(m \times n) \cdot (n \times p) = (m \times p)$ ”; that is, in order for the product  $AB$  to be defined, the two “inner” dimensions must be equal, and the two “outer” dimensions yield the size of the product.

### Example 1

We have

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 4 & -1 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \cdot 4 + 2 \cdot 2 + 1 \cdot 5 \\ 0 \cdot 4 + 4 \cdot 2 + (-1) \cdot 5 \end{pmatrix} = \begin{pmatrix} 13 \\ 3 \end{pmatrix}.$$

Notice again the symbolic relationship  $(2 \times 3) \cdot (3 \times 1) = 2 \times 1$ . ◆

As in the case with composition of functions, we have that matrix multiplication is not commutative. Consider the following two products:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Hence we see that even if both of the matrix products  $AB$  and  $BA$  are defined, it need not be true that  $AB = BA$ .

Recalling the definition of the transpose of a matrix from Section 1.3, we show that if  $A$  is an  $m \times n$  matrix and  $B$  is an  $n \times p$  matrix, then  $(AB)^t = B^t A^t$ . Since

$$(AB)_{ij}^t = (AB)_{ji} = \sum_{k=1}^n A_{jk} B_{ki}$$

and

$$(B^t A^t)_{ij} = \sum_{k=1}^n (B^t)_{ik} (A^t)_{kj} = \sum_{k=1}^n B_{ki} A_{jk},$$

we are finished. Therefore the transpose of a product is the product of the transposes *in the opposite order*.

Our definition of matrix multiplication was chosen so that the next theorem is true.

**Theorem 2.11.** *Let  $V$ ,  $W$ , and  $Z$  be finite-dimensional vector spaces with ordered bases  $\alpha$ ,  $\beta$ , and  $\gamma$ , respectively. Let  $T: V \rightarrow W$  and  $U: W \rightarrow Z$  be linear transformations. Then*

$$[UT]_\alpha^\gamma = [U]_\beta^\gamma [T]_\alpha^\beta.$$



So  $A(B + C) = AB + AC$ .

(c) We have

$$(I_m A)_{ij} = \sum_{k=1}^m (I_m)_{ik} A_{kj} = \sum_{k=1}^m \delta_{ik} A_{kj} = A_{ij}. \quad \blacksquare$$

**Corollary.** Let  $A$  be an  $m \times n$  matrix,  $B_1, B_2, \dots, B_k$  be  $n \times p$  matrices,  $C_1, C_2, \dots, C_k$  be  $q \times m$  matrices, and  $a_1, a_2, \dots, a_k$  be scalars. Then

$$A \left( \sum_{i=1}^k a_i B_i \right) = \sum_{i=1}^k a_i A B_i$$

and

$$\left( \sum_{i=1}^k a_i C_i \right) A = \sum_{i=1}^k a_i C_i A.$$

*Proof.* Exercise. ■

For an  $n \times n$  matrix  $A$ , we define  $A^1 = A$ ,  $A^2 = AA$ ,  $A^3 = A^2A$ , and, in general,  $A^k = A^{k-1}A$  for  $k = 2, 3, \dots$ . We define  $A^0 = I_n$ .

With this notation, we see that if

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

then  $A^2 = O$  (the zero matrix) even though  $A \neq O$ . Thus the cancellation property for multiplication in fields is not valid for matrices. To see why, assume that the cancellation law is valid. Then, from  $A \cdot A = A^2 = O = A \cdot O$ , we would conclude that  $A = O$ , which is false.

**Theorem 2.13.** Let  $A$  be an  $m \times n$  matrix and  $B$  be an  $n \times p$  matrix. For each  $j$ ,  $j = 1, 2, \dots, p$ , let  $u_j$  and  $v_j$  denote the  $j$ th columns of  $AB$  and  $B$ , respectively. Then

- (a)  $u_j = Av_j$
- (b)  $v_j = Be_j$ , where  $e_j$  is the  $j$ th standard vector of  $\mathbb{F}^p$ .

*Proof.* (a) We have

$$u_j = \begin{pmatrix} (AB)_{1j} \\ (AB)_{2j} \\ \vdots \\ (AB)_{mj} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n A_{1k} B_{kj} \\ \sum_{k=1}^n A_{2k} B_{kj} \\ \vdots \\ \sum_{k=1}^n A_{mk} B_{kj} \end{pmatrix} = A \begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix} = Av_j.$$

Hence (a) is proved. The proof of (b) is left as an exercise. (See Exercise 6.) ■

It follows (see Exercise 14) from Theorem 2.13 that column  $j$  of  $AB$  is a linear combination of the columns of  $A$  with the coefficients in the linear combination being the entries of column  $j$  of  $B$ . An analogous result holds for rows; that is, row  $i$  of  $AB$  is a linear combination of the rows of  $B$  with the coefficients in the linear combination being the entries of row  $i$  of  $A$ .

The next result justifies much of our past work. It utilizes both the matrix representation of a linear transformation and matrix multiplication in order to evaluate the transformation at any given vector.

**Theorem 2.14.** *Let  $V$  and  $W$  be finite-dimensional vector spaces having ordered bases  $\beta$  and  $\gamma$ , respectively, and let  $T: V \rightarrow W$  be linear. Then, for each  $u \in V$ , we have*

$$[T(u)]_\gamma = [T]_\beta^\gamma [u]_\beta.$$

*Proof.* Fix  $u \in V$ , and define the linear transformations  $f: F \rightarrow V$  by  $f(a) = au$  and  $g: F \rightarrow W$  by  $g(a) = aT(u)$  for all  $a \in F$ . Let  $\alpha = \{1\}$  be the standard ordered basis for  $F$ . Notice that  $g = Tf$ . Identifying column vectors as matrices and using Theorem 2.11, we obtain

$$[T(u)]_\gamma = [g(1)]_\gamma = [g]_\alpha^\gamma = [Tf]_\alpha^\gamma = [T]_\beta^\gamma [f]_\alpha^\beta = [T]_\beta^\gamma [f(1)]_\beta = [T]_\beta^\gamma [u]_\beta. \blacksquare$$

### Example 3

Let  $T: P_3(R) \rightarrow P_2(R)$  be the linear transformation defined by  $T(f(x)) = f'(x)$ , and let  $\beta$  and  $\gamma$  be the standard ordered bases for  $P_3(R)$  and  $P_2(R)$ , respectively. If  $A = [T]_\beta^\gamma$ , then, from Example 4 of Section 2.2, we have

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

We illustrate Theorem 2.14 by verifying that  $[T(p(x))]_\gamma = [T]_\beta^\gamma [p(x)]_\beta$ , where  $p(x) \in P_3(R)$  is the polynomial  $p(x) = 2 - 4x + x^2 + 3x^3$ . Let  $q(x) = T(p(x))$ ; then  $q(x) = p'(x) = -4 + 2x + 9x^2$ . Hence

$$[T(p(x))]_\gamma = [q(x)]_\gamma = \begin{pmatrix} -4 \\ 2 \\ 9 \end{pmatrix},$$

but also

$$[T]_\beta^\gamma [p(x)]_\beta = A[p(x)]_\beta = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ -4 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -4 \\ 2 \\ 9 \end{pmatrix}. \blacklozenge$$

We complete this section with the introduction of the *left-multiplication transformation*  $L_A$ , where  $A$  is an  $m \times n$  matrix. This transformation is probably the most important tool for transferring properties about transformations to analogous properties about matrices and vice versa. For example, we use it to prove that matrix multiplication is associative.

**Definition.** Let  $A$  be an  $m \times n$  matrix with entries from a field  $F$ . We denote by  $L_A$  the mapping  $L_A: F^n \rightarrow F^m$  defined by  $L_A(x) = Ax$  (the matrix product of  $A$  and  $x$ ) for each column vector  $x \in F^n$ . We call  $L_A$  a **left-multiplication transformation**.

#### Example 4

Let

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

Then  $A \in M_{2 \times 3}(R)$  and  $L_A: R^3 \rightarrow R^2$ . If

$$x = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix},$$

then

$$L_A(x) = Ax = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}. \quad \blacklozenge$$

We see in the next theorem that not only is  $L_A$  linear, but, in fact, it has a great many other useful properties. These properties are all quite natural and so are easy to remember.

**Theorem 2.15.** Let  $A$  be an  $m \times n$  matrix with entries from  $F$ . Then the left-multiplication transformation  $L_A: F^n \rightarrow F^m$  is linear. Furthermore, if  $B$  is any other  $m \times n$  matrix (with entries from  $F$ ) and  $\beta$  and  $\gamma$  are the standard ordered bases for  $F^n$  and  $F^m$ , respectively, then we have the following properties.

- (a)  $[L_A]_\beta^\gamma = A$ .
- (b)  $L_A = L_B$  if and only if  $A = B$ .
- (c)  $L_{A+B} = L_A + L_B$  and  $L_{aA} = aL_A$  for all  $a \in F$ .
- (d) If  $T: F^n \rightarrow F^m$  is linear, then there exists a unique  $m \times n$  matrix  $C$  such that  $T = L_C$ . In fact,  $C = [T]_\beta^\gamma$ .
- (e) If  $E$  is an  $n \times p$  matrix, then  $L_{AE} = L_A L_E$ .
- (f) If  $m = n$ , then  $L_{I_n} = I_{F^n}$ .

*Proof.* The fact that  $L_A$  is linear follows immediately from Theorem 2.12.

(a) The  $j$ th column of  $[\mathbf{L}_A]_{\beta}^{\gamma}$  is equal to  $\mathbf{L}_A(e_j)$ . However  $\mathbf{L}_A(e_j) = Ae_j$ , which is also the  $j$ th column of  $A$  by Theorem 2.13(b). So  $[\mathbf{L}_A]_{\beta}^{\gamma} = A$ .

(b) If  $\mathbf{L}_A = \mathbf{L}_B$ , then we may use (a) to write  $A = [\mathbf{L}_A]_{\beta}^{\gamma} = [\mathbf{L}_B]_{\beta}^{\gamma} = B$ . Hence  $A = B$ . The proof of the converse is trivial.

(c) The proof is left as an exercise. (See Exercise 7.)

(d) Let  $C = [\mathbf{T}]_{\beta}^{\gamma}$ . By Theorem 2.14, we have  $[\mathbf{T}(x)]_{\gamma} = [\mathbf{T}]_{\beta}^{\gamma}[x]_{\beta}$ , or  $\mathbf{T}(x) = Cx = \mathbf{L}_C(x)$  for all  $x \in F^n$ . So  $\mathbf{T} = \mathbf{L}_C$ . The uniqueness of  $C$  follows from (b).

(e) For any  $j$  ( $1 \leq j \leq p$ ), we may apply Theorem 2.13 several times to note that  $(AE)e_j$  is the  $j$ th column of  $AE$  and that the  $j$ th column of  $AE$  is also equal to  $A(Ee_j)$ . So  $(AE)e_j = A(Ee_j)$ . Thus

$$\mathbf{L}_{AE}(e_j) = (AE)e_j = A(Ee_j) = \mathbf{L}_A(Ee_j) = \mathbf{L}_A(\mathbf{L}_E(e_j)).$$

Hence  $\mathbf{L}_{AE} = \mathbf{L}_A \mathbf{L}_E$  by the corollary to Theorem 2.6 (p. 73).

(f) The proof is left as an exercise. (See Exercise 7.) ■

We now use left-multiplication transformations to establish the associativity of matrix multiplication.

**Theorem 2.16.** *Let  $A$ ,  $B$ , and  $C$  be matrices such that  $A(BC)$  is defined. Then  $(AB)C$  is also defined and  $A(BC) = (AB)C$ ; that is, matrix multiplication is associative.*

*Proof.* It is left to the reader to show that  $(AB)C$  is defined. Using (e) of Theorem 2.15 and the associativity of functional composition (see Appendix B), we have

$$\mathbf{L}_{A(BC)} = \mathbf{L}_A \mathbf{L}_{BC} = \mathbf{L}_A(\mathbf{L}_B \mathbf{L}_C) = (\mathbf{L}_A \mathbf{L}_B) \mathbf{L}_C = \mathbf{L}_{AB} \mathbf{L}_C = \mathbf{L}_{(AB)C}.$$

So from (b) of Theorem 2.15, it follows that  $A(BC) = (AB)C$ . ■

Needless to say, this theorem could be proved directly from the definition of matrix multiplication (see Exercise 19). The proof above, however, provides a prototype of many of the arguments that utilize the relationships between linear transformations and matrices.

### Applications\*

For an application of matrix multiplication to the study of population growth, visit [goo.gl/x5XDLw](http://goo.gl/x5XDLw).

A large and varied collection of interesting applications arises in connection with special matrices called *incidence matrices*. An **incidence matrix** is a square matrix in which all the entries are either zero or one and, for convenience, all the diagonal entries are zero. If we have a relationship on a

set of  $n$  objects that we denote by  $1, 2, \dots, n$ , then we define the associated incidence matrix  $A$  by  $A_{ij} = 1$  if  $i$  is related to  $j$ , and  $A_{ij} = 0$  otherwise.

To make things concrete, suppose that we have four people, each of whom owns a communication device. If the relationship on this group is “can transmit to,” then  $A_{ij} = 1$  if  $i$  can send a message to  $j$ , and  $A_{ij} = 0$  otherwise. Suppose that

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

Then since  $A_{34} = 1$  and  $A_{14} = 0$ , we see that person 3 can send to 4 but 1 cannot send to 4.

We obtain an interesting interpretation of the entries of  $A^2$ . Consider, for instance,

$$(A^2)_{31} = A_{31}A_{11} + A_{32}A_{21} + A_{33}A_{31} + A_{34}A_{41}.$$

Note that any term  $A_{3k}A_{k1}$  equals 1 if and only if both  $A_{3k}$  and  $A_{k1}$  equal 1, that is, if and only if 3 can send to  $k$  and  $k$  can send to 1. Thus  $(A^2)_{31}$  gives the number of ways in which 3 can send to 1 in two stages (namely, 3 to 2 to 1 and 3 to 4 to 1). Since

$$A^2 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix},$$

we see that there are two ways 3 can send to 1 in two stages. In general,  $(A + A^2 + \dots + A^m)_{ij}$  is the number of ways in which  $i$  can send to  $j$  in at most  $m$  stages.

A maximal collection of three or more people with the property that any two can send to each other is called a **clique**. The problem of determining cliques is difficult, but there is a simple method for determining if someone belongs to a clique. If we define a new matrix  $B$  by  $B_{ij} = 1$  if  $i$  and  $j$  can send to each other, and  $B_{ij} = 0$  otherwise, then it can be shown (see Exercise 20) that person  $i$  belongs to a clique if and only if  $(B^3)_{ii} > 0$ . For example, suppose that the incidence matrix associated with some relationship is

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

To determine which people belong to cliques, we form the matrix  $B$ , described

earlier, and compute  $B^3$ . In this case,

$$B = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B^3 = \begin{pmatrix} 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \end{pmatrix}.$$

Since all the diagonal entries of  $B^3$  are zero, we conclude that there are no cliques in this relationship.

Our final example of the use of incidence matrices is concerned with the concept of **dominance**. A relation among a group of people is called a **dominance relation** if the associated incidence matrix  $A$  has the property that for all distinct pairs  $i$  and  $j$ ,  $A_{ij} = 1$  if and only if  $A_{ji} = 0$ , that is, given any two people, exactly one of them *dominates* (or, using the terminology of our first example, can send a message to) the other. Since  $A$  is an incidence matrix,  $A_{ii} = 0$  for all  $i$ . For such a relation, it can be shown (see Exercise 22) that the matrix  $A + A^2$  has a row [column] in which each entry is positive except for the diagonal entry. In other words, there is at least one person who dominates [is dominated by] all others in one or two stages. In fact, it can be shown that any person who dominates [is dominated by] the greatest number of people in the first stage has this property. Consider, for example, the matrix

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

The reader should verify that this matrix corresponds to a dominance relation. Now

$$A + A^2 = \begin{pmatrix} 0 & 2 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 \\ 2 & 2 & 2 & 2 & 0 \end{pmatrix}.$$

Thus persons 1, 3, 4, and 5 dominate (can send messages to) all the others in at most two stages, while persons 1, 2, 3, and 4 are dominated by (can receive messages from) all the others in at most two stages.

## EXERCISES

- Label the following statements as true or false. In each part,  $V, W$ , and  $Z$  denote vector spaces with ordered (finite) bases  $\alpha, \beta$ , and  $\gamma$ , respectively;  $T: V \rightarrow W$  and  $U: W \rightarrow Z$  denote linear transformations; and  $A$  and  $B$  denote matrices.

- (a)  $[\mathbf{U}\mathbf{T}]_{\alpha}^{\gamma} = [\mathbf{T}]_{\alpha}^{\beta}[\mathbf{U}]_{\beta}^{\gamma}$ .
- (b)  $[\mathbf{T}(v)]_{\beta} = [\mathbf{T}]_{\alpha}^{\beta}[v]_{\alpha}$  for all  $v \in V$ .
- (c)  $[\mathbf{U}(w)]_{\beta} = [\mathbf{U}]_{\alpha}^{\beta}[w]_{\alpha}$  for all  $w \in W$ .
- (d)  $[\mathbf{I}_V]_{\alpha} = I$ .
- (e)  $[\mathbf{T}^2]_{\alpha}^{\beta} = ([\mathbf{T}]_{\alpha}^{\beta})^2$ .
- (f)  $A^2 = I$  implies that  $A = I$  or  $A = -I$ .
- (g)  $\mathbf{T} = \mathbf{L}_A$  for some matrix  $A$ .
- (h)  $A^2 = O$  implies that  $A = O$ , where  $O$  denotes the zero matrix.
- (i)  $\mathbf{L}_{A+B} = \mathbf{L}_A + \mathbf{L}_B$ .
- (j) If  $A$  is square and  $A_{ij} = \delta_{ij}$  for all  $i$  and  $j$ , then  $A = I$ .
2. (a) Let

$$A = \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & -3 \\ 4 & 1 & 2 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 1 & 4 \\ -1 & -2 & 0 \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}.$$

Compute  $A(2B + 3C)$ ,  $(AB)D$ , and  $A(BD)$ .

- (b) Let

$$A = \begin{pmatrix} 2 & 5 \\ -3 & 1 \\ 4 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & -2 & 0 \\ 1 & -1 & 4 \\ 5 & 5 & 3 \end{pmatrix}, \quad \text{and} \quad C = \begin{pmatrix} 4 & 0 & 3 \end{pmatrix}.$$

Compute  $A^t$ ,  $A^t B$ ,  $B C^t$ ,  $C B$ , and  $C A$ .

3. Let  $g(x) = 3 + x$ . Let  $\mathbf{T}: \mathbb{P}_2(R) \rightarrow \mathbb{P}_2(R)$  and  $\mathbf{U}: \mathbb{P}_2(R) \rightarrow \mathbb{R}^3$  be the linear transformations respectively defined by

$$\mathbf{T}(f(x)) = f'(x)g(x) + 2f(x) \quad \text{and} \quad \mathbf{U}(a + bx + cx^2) = (a + b, c, a - b).$$

Let  $\beta$  and  $\gamma$  be the standard ordered bases of  $\mathbb{P}_2(R)$  and  $\mathbb{R}^3$ , respectively.

- (a) Compute  $[\mathbf{U}]_{\beta}^{\gamma}$ ,  $[\mathbf{T}]_{\beta}$ , and  $[\mathbf{U}\mathbf{T}]_{\beta}^{\gamma}$  directly. Then use Theorem 2.11 to verify your result.
- (b) Let  $h(x) = 3 - 2x + x^2$ . Compute  $[h(x)]_{\beta}$  and  $[\mathbf{U}(h(x))]_{\gamma}$ . Then use  $[\mathbf{U}]_{\beta}^{\gamma}$  from (a) and Theorem 2.14 to verify your result.
4. For each of the following parts, let  $\mathbf{T}$  be the linear transformation defined in the corresponding part of Exercise 5 of Section 2.2. Use Theorem 2.14 to compute the following vectors:

- (a)  $[\mathbf{T}(A)]_{\alpha}$ , where  $A = \begin{pmatrix} 1 & 4 \\ -1 & 6 \end{pmatrix}$ .

- (b)  $[\mathbf{T}(f(x))]_{\alpha}$ , where  $f(x) = 4 - 6x + 3x^2$ .  
 (c)  $[\mathbf{T}(A)]_{\gamma}$ , where  $A = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$ .  
 (d)  $[\mathbf{T}(f(x))]_{\gamma}$ , where  $f(x) = 6 - x + 2x^2$ .
5. Complete the proof of Theorem 2.12 and its corollary.
6. Prove (b) of Theorem 2.13.
7. Prove (c) and (f) of Theorem 2.15.
8. Prove Theorem 2.10. Now state and prove a more general result involving linear transformations with domains unequal to their codomains.
9. Find linear transformations  $\mathbf{U}, \mathbf{T}: \mathbb{F}^2 \rightarrow \mathbb{F}^2$  such that  $\mathbf{U}\mathbf{T} = \mathbf{T}_0$  (the zero transformation) but  $\mathbf{T}\mathbf{U} \neq \mathbf{T}_0$ . Use your answer to find matrices  $A$  and  $B$  such that  $AB = O$  but  $BA \neq O$ .
10. Let  $A$  be an  $n \times n$  matrix. Prove that  $A$  is a diagonal matrix if and only if  $A_{ij} = \delta_{ij} A_{ij}$  for all  $i$  and  $j$ .
11. Let  $V$  be a vector space, and let  $\mathbf{T}: V \rightarrow V$  be linear. Prove that  $\mathbf{T}^2 = \mathbf{T}_0$  if and only if  $R(\mathbf{T}) \subseteq N(\mathbf{T})$ .
12. Let  $V$ ,  $W$ , and  $Z$  be vector spaces, and let  $\mathbf{T}: V \rightarrow W$  and  $\mathbf{U}: W \rightarrow Z$  be linear.
- (a) Prove that if  $\mathbf{U}\mathbf{T}$  is one-to-one, then  $\mathbf{T}$  is one-to-one. Must  $\mathbf{U}$  also be one-to-one?
  - (b) Prove that if  $\mathbf{U}\mathbf{T}$  is onto, then  $\mathbf{U}$  is onto. Must  $\mathbf{T}$  also be onto?
  - (c) Prove that if  $\mathbf{U}$  and  $\mathbf{T}$  are one-to-one and onto, then  $\mathbf{U}\mathbf{T}$  is also.
13. Let  $A$  and  $B$  be  $n \times n$  matrices. Recall that the trace of  $A$  is defined by

$$\text{tr}(A) = \sum_{i=1}^n A_{ii}.$$

Prove that  $\text{tr}(AB) = \text{tr}(BA)$  and  $\text{tr}(A) = \text{tr}(A^t)$ .

14. Assume the notation in Theorem 2.13.
- (a) Suppose that  $z$  is a (column) vector in  $\mathbb{F}^p$ . Use Theorem 2.13(b) to prove that  $Bz$  is a linear combination of the columns of  $B$ . In particular, if  $z = (a_1, a_2, \dots, a_p)^t$ , then show that

$$Bz = \sum_{j=1}^p a_j v_j.$$