THE UNIVERSITY OF WAIKATO
DEPARTMENT OF MATHEMATICS

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Lecture notes for

MATH102-06A ***Introduction to Algebra***

by Rua Murray and Tim Stokes

Website: http://www.math.waikato.ac.nz/~rua/102/

Contents

Contents **1**

Introductory examples **3**

0.1 A system of equations 3

0.2 Transformations and computer graphics 3

0.3 Fibonacci numbers 4

0.4 Cryptography 4

1 Linear systems **6**

1.1 Introduction to linear systems 6

1.2 Solving systems of linear equations 9

1.3 Gaussian elimination 13

1.4 Gauss–Jordan elimination 15

2 Matrices **21**

2.1 Vector and matrix algebra 21

2.2 Matrix multiplication 25

2.3 Matrix inversion 29

2.4 Homogeneous equations and the general solution to $A\,\mathbf{x} = \mathbf{b}$ 34

3 Determinants **38**

3.1 Definition of determinants 38

3.2 Determinants by row-reduction 41

Trigonometry review **48**

4 Vectors and geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$ **52**

4.1 Basic vector geometry 52

4.2 Vector products 55

4.3 Lines and planes in space 59

4.4 Linear equations and intersections of lines and planes 63

4.5 Projections in $\mathbb{R}^3$ 66

5 Induction and recursion **70**

5.1 Set theory, the natural numbers $\mathbb{N}$ and mathematical induction 70

5.2 Mathematical induction 73

5.3 Strong induction and recursion 78

6 Complex numbers **84**

6.1 Introduction to complex numbers 84

6.2 Further operations on $\mathbb{C}$ and the polar form 86

6.3 Solving equations in $\mathbb{C}$ 90

7 Elementary number theory **96**

7.1 Natural numbers and divisibility 96

7.2 Remainders and the Euclidean algorithm 99

7.3 Linear Diophantine equations 103

7.4 Modular arithmetic 106

7.5 The algebra of modular arithmetic 109

7.6 Computing remainders and solving congruences 111

8 Cryptography **116**

8.1 The shift cipher 116

8.2 The affine cipher 117

8.3 The RSA cryptosystem 119

9 Extra topics **121**

9.1 Application: Least squares model fitting 121

9.2 Matrices and linear transformations 122

9.3 Eigenvectors 125

10 Exercises **129**

Introductory examples

Here are several examples of the kinds of problems where the methods studied in this paper will be useful.

(0.1) A system of equations

- The main sources of energy in food are: carbohydrates, protein, fats and alcohol.
- The number of grams of each in 100 gram servings of four foods (and total energy content) is given in the following table:

Food	Grams per 100g serving				Energy
	Carbohydrates	Protein	Fat	Alcohol	(kcal per 100g)
Bread	47	8	2	0	227
Lean steak	0	27	12	0	218
Ice-cream	25	4	7	0	170
Red wine	0	0	0	10	68

- By assigning variables c, p, f, a to the number of kilocalories per gram of each of carbohydrates, protein, fat and alcohol we can write down a system of equations:

$$\begin{aligned} 47\,c \;+\; 8\,p \;+\; 2\,f &= 227 \\ 27\,p \;+\; 12\,f &= 218 \\ 25\,c \;+\; 4\,p \;+\; 7\,f &= 170 \\ 10\,a &= 68 \end{aligned}$$

- By solving these equations (with methods to be learned in the paper), we find that $c \approx 3.7$, $p \approx 4.3$, $f \approx 8.5$, $a = 6.8$.
- So, protein has more energy value per gram than carbohydrates, and both have significantly less than fat or alcohol!
- Most dietary advice recommends that no more than 30% of calories come from fat. How much wine is needed for a recommended diet of 2000kcal if bread is excluded, but the remaining foods are allowed?

(0.2) Transformations and computer graphics

- It is often important (for example in computer graphics programming) to get a mathematical description of how solid objects move in two or three dimensional space
- Points can be represented as *vectors* (these are really just arrays of numbers representing coordinates), and transformations like rotations can be represented by *matrices*

- For example

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{for rotation through } 45^\circ \text{ or} \quad \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{for reflection in the } y\text{–axis}$$

- We'll learn how to do algebra with such arrays in a way that represents successive applications of such transformations. For example, our rotation followed by reflection gives a reflection in a line 22.5° *clockwise* from the y–axis, whereas applying the transformations in the other order gives a reflection in a line 22.5° *anti–clockwise* from the y–axis!

(0.3) Fibonacci numbers

- A simple population model for number of rabbits R_n in n years
- Each rabbit replaces itself through breeding in two subsequent seasons so

$$R_{n+1} = R_n + R_{n-1}$$

- Start with $R_{-1} = R_0 = 1$ so that

$$\begin{array}{rcccccl} R_1 & = & R_0 + R_{-1} & = & 1+1 & = & 2, \\ R_2 & = & R_1 + R_0 & = & 2+1 & = & 3, \\ R_3 & = & R_2 + R_1 & = & 3+2 & = & 5 \ldots \end{array}$$

- The sequence is

$$\{1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots\}$$

- You'll learn techniques to prove that

$$\begin{array}{rcl} R_n & = & \frac{\sqrt{5}-3}{2\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n + \frac{\sqrt{5}+3}{2\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n \\ & \approx & (-0.1708\cdots)\,(-0.618\cdots)^n + (1.170\cdots)\,(1.618\cdots)^n \end{array}$$

(0.4) Cryptography

- We can "code" a plain-text message by converting letters into numbers:

$$A \mapsto 0,\ B \mapsto 1,\ C \mapsto 2, \ldots\ Z \mapsto 25.$$

- So "ALGEBRA" becomes "0 11 6 4 1 17 0"
- Next, we could do some arithmetic operations on each number x to get a new number y. If we look at the remainder of y after dividing by 26, the new number can be converted back into a letter; this process "encrypts" a message
- If we use $y = 5\,x + 8$ then

$$\begin{array}{rcccccccccl} A & \mapsto & 0 & \mapsto & 5\times 0+8 & = & 8 & = & 0\times 26+8 & \mapsto & 8 & \mapsto & I \\ B & \mapsto & 1 & \mapsto & 5\times 1+8 & = & 13 & = & 0\times 26+13 & \mapsto & 13 & \mapsto & N \\ \vdots & & & & & & & & & & & & \\ G & \mapsto & 6 & \mapsto & 5\times 6+8 & = & 38 & = & 1\times 26+12 & \mapsto & 12 & \mapsto & M \\ \vdots & & & & & & & & & & & & \end{array}$$

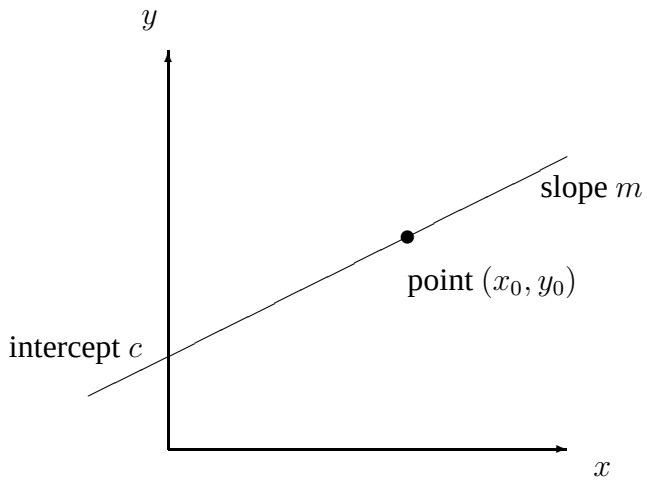
so "ALGEBRA" becomes "ILMCNPI"

- **Exercise:** "JNVWMBQ" has been encrypted using $y = 9\,x + 17$; what was the word?
- RSA public key cryptography is based on a function $y = x^a$. The idea is to choose a very large n to replace 26 and a power a so that there is a unique way to decode each message, but the "decrypt" key is impossibly hard to construct

I ○ Linear systems

(1.1) Introduction to linear systems

We will be working with systems of linear equations. We will start by looking at the most basic equations: those describing lines in the plane.



Now, points in the plane are described by an ordered pair (x, y) of *Cartesian coordinates*. For points on our line L, these coordinates satisfy some sort of equation.

Example 1. The above diagram depicts a line passing through the point $(1, 1)$, with slope $1/2$. we would like to be able to describe this line. □

Here are several formulations, of increasing sophistication:

- **Slope–intercept formula.** Using the "slope" m and "y–intercept" c of the line, the equation of the line is $y = mx + c$ and we can write

  $$L = \{(x, y) \,|y = m\, x + c\}\,.$$

  This notation reads "L is the set of the points (x, y) such that $y = m\, x + c$."

- **Point–slope formula.** More generally, if we are given the slope m and one point on the line (x_0, y_0) we can write

  $$L = \{(x, y) \,|y - y_0 = m\,(x - x_0)\}\,.$$

  This is more general than the slope–intercept formula, since we are no longer tied to knowing the y value when $x = 0$; the value of y at any given x_0 will do.

- **Algebraic formula.** The two previous approaches give a privileged position to x as the "independent variable"—the one which is allowed to vary. There is really no reason for this asymmetry between x and y, so we can git rid of it by writing[1]

$$L = \{(x, y) \,|a_1\, x + a_2\, y = b\}$$

for suitable choices of a_1, a_2, b. (The "slope" of the line is then $m = -\frac{a_1}{a_2}$ and we must have $a_1\, x_0 + a_2\, y_0 = b$.)

Example revisited. Consider the line with slope $m = \frac{1}{2}$, passing through the point $(x_0, y_0) = (1, 1)$. We can work out our three formulations. The point–slope formulation is easy (we are given the appropriate data). For the slope–intercept formula we need to work out the y–intercept. We know that $m = \frac{1}{2}$, and that $y = 1$ when $x = 1$. Thus,

$$1 = y_0 = m\, x_0 + c = \frac{1}{2}\, 1 + c = \frac{1}{2} + c$$

so $c = \frac{1}{2}$ and our three representations of the line are now:

$$\begin{aligned} y &= \frac{1}{2}x + \frac{1}{2}, \\ y - 1 &= \frac{1}{2}(x - 1), \\ x - 2\, y &= -1. \end{aligned}$$

(The last formula can be got by a bit of rearranging of either of the others.)

- **Vector formula.** All of these formulations express the relationship between x and y for points on the line. In fact, the line is simply **all** the points (x, y) which are solutions of the algebraic equation. This is the most appropriate way to think of the line in Algebra (in Calculus, the line is thought of as the graph of a function which tells you how to turn x into y). In this paper, we will learn how to find all the solutions of such equations.

Example revisited again. If we treat y as an "unknown" in the equation, then we can rearrange to obtain

$$\begin{aligned} x &= -1 + 2\, y \\ y &= 0 + 1\, y \end{aligned}$$

where $y \in \mathbb{R}$ is treated as a free variable. With some judicious bracketing (and a simple convention about pushing arithmetic operations through brackets) we can write

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

This is a *vector equation* for the line. The points (x, y) on the line are written in **vector notation** as $\binom{x}{y}$. Notice that $\binom{-1}{0}$ represents the point $(-1, 0)$, which is on the line. The final vector $\binom{2}{1}$ represents the *direction* of the line; the numbers can be thought of as meaning "2 units in the x direction moves you 1 unit in the y direction"—compatible with the original slope of $\frac{1}{2}$. There is no longer anything

[1] There is a very good reason for doing this: the slope based formulations cannot represent a vertical line!

special about y in the vector formulation; we could replace it by any other parameter. For example, if we let $y = 1 + t$ for t an arbitrary real parameter, and use sensible rules for adding vectors:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} + (1+t) \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} + t \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
$$= \begin{pmatrix} -1+2 \\ 0+1 \end{pmatrix} + t \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + t \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

showing that the original point $(x_0, y_0) = (1, 1)$ is still on the line.

Now that we have some equations for lines, we can consider *systems* of such equations, and their solution. We will try to solve them by a method of systematically eliminating variables.

Example 2. Consider the system

$$\begin{aligned} x + y &= 1 \\ 2\,x + y &= 1. \end{aligned}$$

We can solve as follows: subtract 2 times the first equation from the second equation; the second equation is now

$$-y = -1$$

so $y = 1$; substituting this into the first gives $x + 1 = 1$, so $x = 0$. The solution set for this pair of equations is simply $\{(0, 1)\}$.
Geometric interpretation: each equation describes one line in the plane, so any solution to the system of equations satisfies **both equations simultaneously** and must therefore represent a point which is on both lines. Since there is only one such point, it is the point of intersection of the two lines. □

Example 3. Consider the system

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

We can see that the second equation is a multiple of the first, so contains no new information. Let us ignore this for the moment, and use the same elimination method as above. Adding twice the first equation from the second gives the revised system

$$\begin{aligned} 2\,x + y &= 1 \\ 0 &= 0. \end{aligned}$$

We no longer have enough information to determine both x and y uniquely; in fact, the solution set is infinite, and we can write it down letting y take the value of an **arbitrary real parameter**. Effectively, we replace the missing final equation by $y = y$, and substitute this back into the first equation to get $2\,x + y = 1$ or $x = \frac{1}{2}(1 - y)$. The solution set is thus:

$$\left\{(x, y) \mid x = \frac{1-y}{2}, y \in \mathbb{R}\right\}, \text{ or } \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} + y \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}.$$

Note: This solution set is infinite; there are as many solutions as there are real numbers.
Geometric interpretation: The solution set is a line. In fact, both equations in our system describe the same line, so their common points are that same line!

Remark: We could have assigned x to be the arbitrary parameter, or even $y = t$ (or any other letter). Letting y be the **free variable** has been done for consistency with the general approach that we'll study below (many equations in many variables). □

Example 4. Consider the system

$$\begin{aligned} 2x + 3y &= 1 \\ 2x + 3y &= 2. \end{aligned}$$

What are the solutions? Whatever values x, y are given, $2x + 3y$ cannot be both 1 and 2, so there are **no solutions** to this system. This can also be seen be a systematic method: if we subtract the first equation from the second we are left with $0 = 1$! This is a ridiculous, and intolerable, situation. In order to avoid this contradiction to the rules of arithmetic, we conclude that there cannot be any values of x and y which simultaneously satisfy both equations in the system.
Geometric interpretation: The two lines described by the system do not share any common points; they are parallel. □

These examples capture three important facts about systems of linear equations: (i) their solutions can be probed by systematic methods; (ii) there could be exactly one, infinitely many, or no solutions; (iii) the solutions can be interpreted geometrically.

(1.2) Solving systems of linear equations

The three examples in the previous section were all very easy to solve; when we have more variables, and more equations in our system it is very important to follow a systematic approach.

Example. Consider the system of equations:

$$\begin{array}{rcrcrcrcl} 2\,x & - & y & - & z & + & 2\,w & = & 1 \\ 6\,x & - & 2\,y & + & z & + & 6\,w & = & 4 \\ 2\,x & & & - & z & + & 3\,w & = & 4. \end{array}$$

Does it have any solutions? how many? what are they? □

Setup

To proceed, we need a definition.

Definition. A system of m linear equations in n unknowns is one of the form:

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & & & & \vdots & & \vdots & \\ a_{m1}\,x_1 & + & a_{m2}\,x_2 & + & \cdots & + & a_{mn}\,x_n & = & b_m \end{array}$$

where each a_{ij} and b_i is a real number. □

This is sometimes called an $m \times n$ (linear) system. The **variables** (or unknowns) are $x_1, \ldots, x_n$. The numbers a_{ij} are the **coefficients** of the system, and the numbers b_i are sometimes called the **right-hand side** (or RHS). Note that none of the variables are multiplied together; this is what makes the system *linear*.

Definition. A **solution** to an $m \times n$ system is an assignment of numbers to each of $x_1, \ldots, x_n$ so that all m equations are satisfied. A system is **consistent** if it has at least one solution. Otherwise it is **inconsistent**. The **general solution** to a consistent system is the set of *all* solutions. □

Example 1. The system of equations

$$\begin{aligned} x - 2\,y &= 3 \\ 2\,x - 4\,y &= 8 \end{aligned}$$

is inconsistent, because if $x = x_0$, $y = y_0$ was a solution we would have

$$8 = 2\,x_0 - 4\,y_0 = 2\,(x_0 - 2\,y_0) = 2 \times 3 = 6,$$

which is ridiculous! □

Example 2. The system of equations

$$\begin{aligned} x + 2\,y &= 5 \\ x + \;\; y &= 3 \end{aligned}$$

is consistent, because the assignment $x = 1$, $y = 2$ is a solution. □

Systems in Echelon form

Certain linear systems are easy to solve. We will study how to solve systems with a special form, and then see later how to put an arbitrary system into a form which has this nice structure.

Definition. A system is in **Echelon Form (EF)** if the first variable in each equation is further to the right as we move down. The **leading variables** (or **pivots**) of a system in EF are those variables that are the first in one of the equations. The other variables are **free variables**. Each equation must contain a leading variable. □

Example 3. The system

$$\begin{array}{rrrrrrrr} x_1 & +2\,x_2 & +x_3 & -x_4 & & + \;\; x_6 & = & 3 \\ & x_2 & & -x_4 & +x_5 & & = & 0 \\ & & & x_4 & & +3\,x_6 & = & 3 \\ & & & & x_5 & +2\,x_6 & = & 4. \end{array}$$

is in echelon from, whereas the system

$$\begin{aligned} x + y &= 1 \\ x - y &= 7 \end{aligned}$$

is not. □

To solve a system in echelon form, we work up from the last equation expressing the value of each leading variable in terms of the free ones.

Example. In the first system of Example 3, the leading variables are x_1, x_2, x_4, x_5 and the free variables are x_3 and x_6. The last equation has x_5 as its leading variable, so we rearrange to get

$$x_5 = 4 - 2x_6. \qquad (x_5\text{–equation})$$

Similarly, rearranging the next one up to solve for x_4:

$$x_4 = 3 - 3x_6. \qquad (x_4\text{–equation})$$

Rearranging the next equation up and eliminating the non-free variables using the expressions for x_4 and x_5 gives:

$$\begin{array}{rcll} x_2 & = & x_4 - x_5 & \\ & = & (3 - 3x_6) - (4 - 2x_6) & (x_2\text{–equation}) \\ & = & -1 - x_6. & \end{array}$$

Finally, the top equation gives:

$$\begin{array}{rcll} x_1 & = & -2x_2 - x_3 + x_4 - x_6 + 3 & \\ & = & -2(-1 - x_6) - x_3 + (3 - 3x_6) - x_6 + 3 & (x_1\text{–equation}) \\ & = & 8 - x_3 - 2x_6. & \end{array}$$

we have solved for x_1, x_2, x_4, x_5 (the leading variables) in terms of x_3, x_6 (the free variables). We can collect together the various (x_i–equation)s to state the **general solution**:

$$\begin{array}{rcrcr} x_1 & = & 8 & - x_3 - & 2\,x_6 \\ x_2 & = & -1 & - & x_6 \\ x_4 & = & 3 & - & 3\,x_6 \\ x_5 & = & 4 & - & 2\,x_6 \end{array} \qquad \text{with } x_3, x_6 \in \mathbb{R}.$$

The idea is that any choice of real values for the free variables determines a solution to the system. For instance let $x_3 = 0$ and $x_6 = 1$. Then the above general solution tells us

$$x_1 = 6,\ \ x_2 = -2,\ \ x_4 = 0,\ \ x_5 = 2.$$

Finally, by adding the equations $x_3 = x_3$, $x_6 = x_6$ and applying vector notation we obtain:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 8 \\ -1 \\ 0 \\ 3 \\ 4 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_6 \begin{pmatrix} -2 \\ -1 \\ 0 \\ -3 \\ -2 \\ 1 \end{pmatrix}.$$

□

Algorithm for solving a system in Echelon Form

1. Identify the leading and free variables.
2. Assign a parameter to each free variable.
3. Work up from the last to the first equation, solving by substitution (back substitution).
4. [Optional] Write the solution in vector form.

Example 4.

$$\begin{array}{rcrcr} 3\,x & + & 2\,y & = & 1 \\ & & -y & = & 1 \end{array}$$

The leading variables are x and y. There are no free variables and the solution is

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

□

Example 5.

$$\begin{array}{rcrcrcr} x & + & y & + & z & = & 1 \\ & & & & z & = & 1 \end{array}$$

The leading variables x and z; y is a free variable, and we call it t. The solution is

$$(x, y, z) = (0, 0, 1) + t\,(-1, 1, 0),\ t \in \mathbb{R}.$$ □

Example 6.

$$\begin{array}{rcrcrcl} x & + & y & + & z & = & 1 \\ & & 2\,y & + & 2\,z & = & 2 \\ & & & & 3\,z & = & 3 \end{array}$$

The leading variables are x, y and z. There are no free variables and the solution is

$$(x, y, z) = (0, 0, 1).$$ □

Example 7.

$$\begin{array}{rcrcrcrcl} x & + & y & + & z & - & w & = & 0 \\ & & y & + & 2\,z & & & = & 0 \end{array}$$

The leading variables x and y; z and w are free variables, and we call them s and t respectively. The solution is

$$(x, y, z, w) = (0, 0, 0, 0) + s\,(1, -2, 1, 0) + t\,(1, 0, 0, 1),\ s, t \in \mathbb{R}.$$ □

Remarks on systems in echelon form

- Every system in echelon form is automatically consistent.
- The back substitution procedure can involve of a lot of arithmetic, and it is easy to make mistakes. It is therefore a good idea, when you have found the general solution, to select a particular solution (for example, by setting all the free variables to zero), and substitute it in to the system to check that it satisfies all the equations.
- It is often convenient to replace the free variables with a parameter described by a different letter. Although this has no mathematical significance, it is a good convention for the unambiguous presentation of solutions.

Remarks on general systems

- A general system of equations can have exactly 0, 1 or infinitely many solutions.
- Two systems will be called **equivalent** if they have the same solutions. That is, their solution set is the same. The basic idea that we will exploit in solving a general system is to perform some *elementary operations* on a system to obtain an equivalent system which is in EF. Since equivalent systems always have the same solutions, we can obtain the general solution to our original system by solving the equivalent EF system by back-substitution.

(1.3) Gaussian elimination

Systems of linear equations in echelon form (EF) are solved easily by back-substitution. Therefore, we'd like to know how to put a general system in EF. The idea is to systematically eliminate variables from the front of equations. We can accomplish this by a suitable sequence of **elementary operations**.

Elementary operations

There are three basic operations that we can perform on a system:

1. Interchange the order of two equations;
2. Multiply an equation by a non-zero constant;
3. Add a multiple of one equation to another equation.

Note: combining operations 2 and 3 into one operation is not an elementary operation, and can lead to the creation of "bogus" solutions.

Theorem 1.1 *Elementary operations do not change the solution set of a linear system.*

This result is important, plausible, and fairly easy to prove. The basic idea is that a solution cannot be destroyed by an elementary operation (after all, we are simply reordering, adding together or multiplying a bunch of true arithmetic statements). On the other hand, each elementary operation can be reversed by another elementary operation (again this is pretty obvious), so new solutions cannot be created (if they were, the reversed operation would have to delete them, which is not allowed!).

Theorem 1.2 *A sequence of elementary operations can be applied to any consistent linear system to put it in echelon form.*

The proof of this result is the *Gaussian elimination algorithm.* We will rehearse it a few times before writing it down precisely. The idea is to work through the system, one variable at a time, systematically eliminating the leading variable from all the other equations.

Example 1. Let us solve the following system:

$$\begin{aligned} x + 2y &= 3 \quad [1] \\ 2x + 5y &= 7. \quad [2] \end{aligned}$$

We're aiming at an echelon form. So we need to eliminate the leading variable x in the first equation [1] from the second equation [2]. Subtract twice [1] from [2]:

$$\begin{array}{rcrcl} 2x & + & 5y & = & 7 \\ 2x & + & 4y & = & 6 \\ \hline & & y & = & 1. \end{array}$$

So our new system, equivalent to the old, is:

$$\begin{aligned} x + 2y &= 3 \\ y &= 1. \end{aligned}$$

In future examples, we will record the row operation used by writing $[2] \to [2] - 2 \times [1]$. The new system is in echelon form and is easily solved: $y = 1$, so $x = 3 - 2 = 1$. There are no free variables, so $x = y = 1$ is the solution. (Check!) □

Example 2. We apply the elimination procedure to the following system:

$$\begin{array}{rcrcrcl} & & y & + & z & = & 1 \\ x & + & y & + & z & = & 1 \\ 2\,x & + & 2\,y & + & 3\,z & = & 1 \end{array} \qquad [1] \to [2]$$
$$[2] \to [1]$$

$$\begin{array}{rcrcrcl} x & + & y & + & z & = & 1 \\ & & y & + & z & = & 1 \\ 2\,x & + & 2\,y & + & 3\,z & = & 1 \end{array} \qquad [3] \to [3] - 2 \times [1]$$

$$\begin{array}{rcrcrcl} x & + & y & + & z & = & 1 \\ & & y & + & z & = & 1 \\ & & & & z & = & -1 \end{array}$$

The solution can now be found by back substitution: $x = 0$, $y = 2$, $z = -1$. □

Formal statement of the Gaussian elimination algorithm

The main idea in Gaussian elimination is to use elementary operations to systematically eliminate variables. The basic procedure is to work through the leading variables one at a time, removing the terms from lower equations:

Gaussian elimination algorithm for m equations

0. Identify the first leading variable. Call it X, set $i := 1$.

1. If equation $[i]$ does not start with an X then swap with another equation which does.

2. [Optional] Divide equation $[i]$ by the coefficient of X to turn the first coefficient into a 1.

3. Eliminate X from each of equations $[i+1]$–$[m]$ by subtracting an appropriate multiple of equation $[i]$.

4. If any equations "$0 = 0$" are obtained, remove them from the system; if any equations "$0 = b$" (where $b \neq 0$) are obtained, then **STOP**, as the system is inconsistent.

5. If there are no more equations, then **STOP**. Otherwise, identify the next leading variable which appears to the right of X in any of the equations $[i+1]$–$[m]$. Call it X. Set $i := i + 1$ and go to Step 1.

Note: This procedure, if followed correctly, will produce a system in echelon form which is equivalent to the original one. The only way that it can fail is if the original system is inconsistent. In this case, the algorithm with terminate. □

Using Gaussian elimination to solve a word problem

Example 3. A traveller recorded the following data about $ spent in several categories on a brief European

tour. Determine how much money she spent in each country.

\$ / day	UK	France	Spain	TOTAL \$ spent
Accommodation	50	20	20	340
Food	20	30	20	320
Sundry	10	10	10	140

First of all, let x, y, z represent days spent in the UK, France and Spain respectively. Then, the data can be written as several equations, which we proceed to solve by Gaussian elimination:

$$\begin{array}{rcrcrcll} 50\,x & + & 20\,y & + & 20\,z & = & 340 & \\ 20\,x & + & 30\,y & + & 20\,z & = & 320 & [2] \to [2] - \frac{20}{50}[1] \\ 10\,x & + & 10\,y & + & 10\,z & = & 140 & [3] \to [3] - \frac{10}{50}[1] \end{array}$$

$$\begin{array}{rcrcrcll} 50\,x & + & 20\,y & + & 20\,z & = & 340 & \\ & & 22\,y & + & 12\,z & = & 184 & \\ & & 6\,y & + & 6\,z & = & 72 & [3] \to [3] - \frac{6}{22}[2] \end{array}$$

$$\begin{array}{rcrcrcl} 50\,x & + & 20\,y & + & 20\,z & = & 340 \\ & & 22\,y & + & 12\,z & = & 184 \\ & & & & \frac{30}{11}\,z & = & \frac{240}{11} \end{array}$$

The system is now in Echelon Form, and we can solve it by back substitution to obtain: $x = 2$, $y = 4$, $z = 8$. So, with 2 days in the UK, her total expenditure there was $2 \times (\$50 + \$20 + \$10) = \$160$. Similar calculations reveal that she spent \$240 in France and \$400 in Spain. □

Notice that in this example the arithmetic could have been simplified by interchanging the second and third equations before the final step. Once you get familiar with the process of Gaussian elimination, you can introduce these extra manouvres to simplify your calculations.

(1.4) Gauss–Jordan elimination

We can improve our Gaussian elimination procedure, reduce the amount of writing required and simplify the back substitution step by employing **matrix notation** and performing some extra elementary row operations .

Augmented matrix notation

In Gaussian elimination, there is considerable redundancy in notation by having to constantly rewrite the whole system of equations, variables and all. Since the variables never change (and are implicitly identified by their positions), there is really no need to keep writing them down. In fact, the elementary operations change only the *coeffi cients* of this system, so we need only keep track of these numbers.

Example 1. We represent the system

$$\begin{array}{rcrcrcl} x & + & 2\,y & + & 3\,z & = & 4 \\ 3\,x & + & 4\,y & + & 5\,z & = & 2 \\ 2\,x & + & 5\,y & + & z & = & 3 \end{array}$$

as the *augmented matrix*

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 3 & 4 & 5 & 2 \\ 2 & 5 & 1 & 3 \end{array}\right).$$

Now we can apply the Gaussian elimination approach to the **rows** of this matrix to put it in EF, and convert back to a linear system at the end. The only difference is that we will record our steps using a slightly different notation. First, we want to eliminate the coefficients of x in the second and third equations. This corresponds to transforming the first entries of the second and third rows of the augmented matrix into 0s. This is done by *elementary row operations*, and they exactly mirror the elementary operations used in Gaussian elimination. Now, our first move is to subtract 3 times the first row from the second row, in order to put a 0 in the first column of the second row. We denote this by $R_2 \to R_2 - 3\,R_1$; the similar move $R_3 \to R_3 - 2\,R_1$ is applied to the second row. We then obtain:

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -2 & -4 & -10 \\ 0 & 1 & -5 & -5 \end{array}\right).$$

We then apply $R_3 \to R_3 + \frac{1}{2}\,R_2$ to get

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -2 & -4 & -10 \\ 0 & 0 & -7 & -10 \end{array}\right),$$

which corresponds to the system of equations:

$$\begin{array}{rcrcrcl} x & + & 2\,y & + & 3\,z & = & 4 \\ & & -2\,y & - & 4\,z & = & -10 \\ & & & & -7\,z & = & -10. \end{array}$$

This can be solved by back substitution to get $x = -\frac{32}{7}$, $y = \frac{15}{7}$, $z = \frac{10}{7}$. □

Notice that when we performed a "row operation", we simply worked along the row, one column at a time, doing the (same) indicated arithmetic operation to each entry.

Definition. The system of equations:

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & & & & & \vdots & & \vdots \\ a_{m1}\,x_1 & + & a_{m2}\,x_2 & + & \cdots & + & a_{mn}\,x_n & = & b_m \end{array}$$

can be written in **augmented matrix notation** as:

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right) \quad \text{or} \quad (A|\mathbf{b})$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Then A is the **matrix of coefficients** and $\mathbf{b}$ is the **right hand side** (RHS). □

Definition. The **elementary row operations** are:

1. Interchange two rows $\begin{pmatrix} R_i \to R_j \\ R_j \to R_i \end{pmatrix}$

2. Multiply a row by a non-zero constant $(R_i \to a\,R_i)$

3. Add a multiple of one row to another $(R_i \to R_i + a\,R_j)$

They do not change the solution, and are equivalent to the elementary operations for a systems of equations (each equation is associated with a **row** of the matrix). □

In Example 1 we used elementary row operations to put the matrix in Echelon Form. It turns out that we can continue doing further row operations to make the matrix as simple as possible. Basically, we'll use row operations to do most of the work from the back-substitution step.

Example 2. Let us pick up the augmented matrix from the previous example, and do some more row operations. First of all, we'll turn the first non-zero coefficient of each row into a one:

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -2 & -4 & -10 \\ 0 & 0 & -7 & -10 \end{array}\right) \begin{array}{l} \\ R_2 \to \frac{-1}{2} R_2 \\ R_3 \to \frac{-1}{7} R_3 \end{array}$$

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 5 \\ 0 & 0 & 1 & \frac{10}{7} \end{array}\right) \begin{array}{l} R_1 \to R_1 - 3\,R_3 \\ R_2 \to R_2 - 2\ R_3 \\ \\ \end{array}$$

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & -\frac{2}{7} \\ 0 & 1 & 0 & \frac{15}{7} \\ 0 & 0 & 1 & \frac{10}{7} \end{array}\right) \begin{array}{l} R_1 \to R_1 - 2\,R_2 \\ \\ \\ \end{array}$$

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & -\frac{32}{7} \\ 0 & 1 & 0 & \frac{15}{7} \\ 0 & 0 & 1 & \frac{10}{7} \end{array}\right) \qquad .$$

The extra operations were motivated by trying to get as many 0s and 1s in the matrix as possible. Note that the system of equations now turns out to be

$$\begin{aligned} x &= -\tfrac{32}{7} \\ y &= \tfrac{15}{7} \\ z &= \tfrac{10}{7} \end{aligned}$$

so the system is effectively solved. □

We now introduce some terminology to describe what we have done.

Definition. The first non-zero entry of a matrix row is called a **pivot**. A **leading 1** is a pivot which has the value 1. A matrix is in **Reduced Row Echelon Form (RREF)** if every pivot is a leading 1, and each column containing a leading 1 has 0s above and below the leading 1. □

We can make some enhancements to our methods to put the augmented matrix in RREF. This involves putting as many 0s and 1s as possible in the augmented matrix, and is accomplished by further row operations. At the end of this procedure, back–substitution is no longer necessary, since the solution can be written down immediately.

Gauss–Jordan elimination

Example 3. Use augmented matrices and row operations to solve the following system of equations via a matrix in RREF:

$$\begin{array}{rcrcrcl} 2\,x & + & 4\,y & + & 2\,z & = & 6 \\ & & y & + & z & = & 1 \\ x & + & 3\,y & + & 3\,z & = & 5. \end{array}$$

We will write down the augmented matrix and proceed by row operations:

$$\left(\begin{array}{ccc|c} 2 & 4 & 2 & 6 \\ 0 & 1 & 1 & 1 \\ 1 & 3 & 3 & 5 \end{array}\right) \quad R_3 \to R_3 - \tfrac{1}{2}\,R_1$$

$$\left(\begin{array}{ccc|c} 2 & 4 & 2 & 6 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 \end{array}\right) \quad R_3 \to R_3 - R_2$$

$$\left(\begin{array}{ccc|c} 2 & 4 & 2 & 6 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right).$$

This is an EF, which could be solved by back-substitution. Instead, we scale the first row to make the top-left pivot a leading 1. This requires the operation $R_1 \to \frac{1}{2}\,R_1$. We then continue with further operations to get the RREF:

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right) \quad \begin{array}{l} R_1 \to R_1 - \quad R_2 \\ R_2 \to R_2 - \quad R_3 \end{array}$$

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right) \quad R_1 \to R_1 - \quad R_2$$

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right)$$

from which we immediately recover $x = 2, y = 0, z = 1$. □

Gauss–Jordan elimination procedure

0. Write the system in augmented matrix form.

1. Use row operations to obtain the echelon form.

2. Divide each row by its leading coefficient (so all the pivots are leading 1s).

3. Eliminate all non-zero entries above each leading 1 by subtracting appropriate multiples of its row.

4. Write down the solution.

Note: One can clear each column with a leading variable in succession instead, without getting to echelon form along the way. The answer will be the same.

Theorem 1.3 *The RREF for a linear system is unique, and always can be found by Gauss–Jordan elimination.*

Example 4. Use Gauss–Jordan elimination to solve the system:

$$\begin{array}{rcrcrcl} x & + & y & + & z & = & 1 \\ 2\,x & - & y & + & 3\,z & = & 2\ . \\ 4\,x & + & y & + & 5\,z & = & 5 \end{array}$$

By the Gauss–Jordan algorithm, we write:

$$\left(\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 2 & -1 & 3 & 2 \\ 4 & 1 & 5 & 5 \end{array}\right) \quad \begin{array}{l} R_2 \to R_2 - 2\,R_1 \\ R_3 \to R_3 - 4\,R_1 \end{array}$$

$$\left(\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & -3 & 1 & 0 \\ 0 & -3 & 1 & 1 \end{array}\right) \quad R_3 \to R_3 - R_2$$

$$\left(\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

The last row indicates an inconsistency, so we stop after step 1, concluding that there is **no solution**. □

Example 5. Use Gauss–Jordan elimination to solve the following system:

$$\begin{array}{rcrcrcrcl} x_1 & & & + & x_3 & + & 4\,x_4 & = & -1 \\ 2\,x_1 & - & x_2 & + & x_3 & + & 7\,x_4 & = & -2 \\ -2\,x_1 & + & x_2 & & & - & 6\,x_4 & = & 2 \\ x_1 & + & x_2 & + & x_3 & + & 4\,x_4 & = & -1. \end{array}$$

Solution:

$$\left(\begin{array}{rrrr|r} 1 & 0 & 1 & 4 & -1 \\ 2 & -1 & 1 & 7 & -2 \\ -2 & 1 & 0 & -6 & 2 \\ 1 & 1 & 1 & 4 & -1 \end{array}\right) \quad \begin{array}{l} R_2 \to R_2 - 2\,R_1 \\ R_3 \to R_3 + 2\,R_1 \\ R_4 \to R_4 - R_1 \end{array}$$

$$\left(\begin{array}{rrrr|r} 1 & 0 & 1 & 4 & -1 \\ 0 & -1 & -1 & -1 & 0 \\ 0 & 1 & 2 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{array}\right) \quad \begin{array}{l} R_2 \to R_4 \\ \\ R_4 \to R_2 \end{array}$$

$$\left(\begin{array}{rrrr|r} 1 & 0 & 1 & 4 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 2 & 0 \\ 0 & -1 & -1 & -1 & 0 \end{array}\right) \quad \begin{array}{l} \\ R_3 \to R_3 - R_2 \\ R_4 \to R_4 + R_2 \end{array}$$

$$\left(\begin{array}{rrrr|r} 1 & 0 & 1 & 4 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 \end{array}\right) \quad \begin{array}{l} \\ R_3 \to \frac{1}{2}R_3 \\ R_4 \to (-1)R_4 \end{array}$$

$$\left(\begin{array}{cccc|c} 1 & 0 & 1 & 4 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{array}\right) \begin{array}{l} R_1 \to R_1 - R_3 \\ \\ \\ R_4 \to R_4 - R_3 \end{array}$$

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 3 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

The solution can be read off simply: the positions of the leading–1s tell us that x_1, x_2, x_3 are leading variables, and the lack of leading–1 in the fourth column tells us that x_4 is a free variable. Thus, we set $x_4 = t$ and recover: $x_1 = -1 - 3\,t$, $x_2 = 0$, $x_3 = -t$. □

Remarks

- If you are a little confused by the final step in Example 5 ("reading off the solution"), try writing out the equations from the RREF and then solve!
- If a system of equations has no solutions, then the coefficient matrix of the RREF will have a row of zeros, with the corresponding element in the RHS column being non-zero.

The idea of an equation $A\mathbf{x} = \mathbf{b}$

In our augmented matrix notation, we have replaced the system of equations with an augmented matrix $(A|\mathbf{b})$. It is tempting to think of the augmented matrix $(A|\mathbf{b})$ as a short-hand notation for a "matrix equation": $A \times \mathbf{x} = \mathbf{b}$, (where $\mathbf{x}$ is a vector of the unknown variables $x_1, \ldots, x_n$). It turns out that this is a completely reasonable point of view, and the trick in making this precise is to get the correct the idea of "multiplication" for matrices. We will do this in the next few lectures, by studying *matrix algebra*. Although we can't solve the equation by writing $\mathbf{x} = \mathbf{b} \div A$ (matrix division certainly doesn't make sense), we will see that there is sometimes another matrix A^{-1} such that $\mathbf{x} = A^{-1}\mathbf{b}$. So, we will next start looking at matrices as objects in their own right, and study their algebraic properties. Once we understand these basic properties, we can come back to the connection with linear equations.

II ∘ Matrices

(2.1) Vector and matrix algebra

We have worked with augmented matrices for solving linear systems, and have employed "vector notation" to describe solutions of these equations. It is now time to give these objects status in their own right, and study their algebraic properties.

Basically, a matrix is an array of (usually real) numbers. Such things have many uses:

- solving systems of linear equations (already seen, but more to come)
- operations research applications (linear programming etc.)
- computer graphics
- all sorts of scientific computations
- statistical analysis

Definition. An $m \times n$ matrix

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

is an array of numbers with m rows and n columns. We will often write $A = (a_{ij})$ (where a_{ij} is the *entry* (or element or component) in the ith row and jth column of A. The **size** or **dimension** of the matrix is $m \times n$. If $B = (b_{ij})$ is another $m \times n$ matrix, then $A = B$ if and only if $a_{ij} = b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. □

Example 1. The following matrix is a 3×2 matrix: $A = \begin{pmatrix} 2 & 4 \\ 3 & 6 \\ 4 & 0 \end{pmatrix}$. If we employ the notation $A = (a_{ij})$ then, $a_{11} = 2,\ a_{12} = 4,\ a_{21} = 3,\ a_{22} = 6,\ a_{31} = 4,\ a_{32} = 0.$ □

Example 2. A necessary (but not sufficient) condition for equality of matrices is that they have the same size:

$$\begin{pmatrix} -2 & 0 \\ 1 & 3 \end{pmatrix} \neq \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

□

Special types of matrices

There are some special types of matrices that are important:

- **column vector:** $n \times 1$ matrix, essentially just an element of $\mathbb{R}^n$;

- **row vector:** $1 \times m$ matrix, an element of $\mathbb{R}^m$ written in a different way;
- **zero matrix:** all entries are zero (often denoted $\mathbf{0}$, with its size clear from context);
- **square matrix:** an $n \times n$ matrix;
- **triangular matrix:** $a_{ij} = 0$ when $i > j$ (upper triangular), *or* $a_{ij} = 0$ when $i < j$ (lower triangular). Triangular matrices are especially nice since they represent systems of equations which are in EF, so are easily solved by back-substitution;
- **diagonal matrix:** square, and the only non-zero entries are on the main diagonal (entries of the form a_{ii}). These are even better than triangular matrices, since they represent systems of equations of the form

$$\begin{array}{rcccl} a_{11}\,x_1 & & & = & b_1 \\ & a_{22}\,x_2 & & = & b_2 \\ & & \ddots & \vdots & \vdots \\ & & a_{nn}\,x_n & = & b_n \end{array}$$

  which are solved by ordinary division;
- **identity matrix:** diagonal matrix with all main diagonal entries 1 (denoted I or I_n). These are the target coefficient matrices with Gauss–Jordan elimination, since the RHS with such an augmented system is the solution;
- **block matrix:** these are built out of smaller matrices. Augmented matrices are block matrices. Here is another example, if

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 0 \end{pmatrix},\ B = \begin{pmatrix} 1 & 5 \\ 0 & 7 \end{pmatrix} \text{ and } C = \begin{pmatrix} 8 & 9 \end{pmatrix}$$

  then

$$\left(\begin{array}{c|c} A & \begin{array}{c} B \\ \hline C \end{array} \end{array}\right) = \begin{pmatrix} 1 & 2 & 1 & 5 \\ 3 & 4 & 0 & 7 \\ 1 & 0 & 8 & 9 \end{pmatrix}.$$

  Many configurations are possible, subject to consistency in number of rows and columns.

You can think of an $m \times n$ matrix as an array of m rows, each of which is an n–dimensional row vector, or as an array of n columns, each of which is an m–dimensional column vector. Sometimes this will be useful!

Example 3. Let $\mathbf{r}_1 = (2\ 4), \mathbf{r}_2 = (3\ 6), \mathbf{r}_3 = (4\ 0), \mathbf{c}_1 = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}, \mathbf{c}_2 = \begin{pmatrix} 4 \\ 6 \\ 0 \end{pmatrix}$. Then, with A as in Example 1, $A = \left(\begin{array}{c} \mathbf{r}_1 \\ \hline \mathbf{r}_2 \\ \hline \mathbf{r}_3 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{c}_1 & \mathbf{c}_2 \end{array}\right)$. □

Remarks about vectors

Row and column vectors have a number of special properties that are not shared by all matrices:

- they provide a natural description of points in $\mathbb{R}^n$ (the components represent "number of units in a coordinate direction"), and perform very well as the basic objects in geometry;
- they have a natural notion of *length* and *direction* (we have had a taste of this with the vector description of lines);
- their algebra (*addition* and *scalar multiplication*) has a geometric interpretation;
- they can be transformed by matrices.

In view of their special importance, vectors are will usually be denoted by bold-face, lower-case letters. For example, $\mathbf{v}$. (When writing by hand, it is conventional to underline vectors: $\underline{v}$, or even place an arrow above them: $\vec{v}$.)

Matrix algebra

We can **add** matrices of the same size together simply by adding **all** their respective entries. For example,

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

Matrices can also be **scaled** by multiplying all entries by the same constant number. For example,

$$3 \begin{pmatrix} 1 & 2 \\ 3 & -4 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 9 & -12 \end{pmatrix}.$$

We also define: $-A = (-1)\, A$.

Note: Any matrix can be multiplied by a given scalar, but only matrices of the same size can be added together.

These operations behave in much the same way as ordinary arithmetic (they are, after all, ordinary arithmetic performed "entry–by–entry"). The rules can be summarized formally. Let A, B and C be any $m \times n$ matrices and let α, β be scalars. Then:

1. $A + B = B + A$;

2. $A + \mathbf{0} = A$;

3. $(A + B) + C = A + (B + C)$;

4. $\alpha\,(A + B) = \alpha\, A + \alpha\, B$;

5. $(\alpha + \beta)\, A = \alpha\, A + \beta\, A$

6. $A + (-A) = \mathbf{0}$

7. $\alpha\,(\beta A) = (\alpha\beta)\, A$;

8. $1\, A = A$.

Transpose of a matrix

The final basic operation that we will sometimes perform on a matrix is to "flip it around", exchanging rows for columns.

Definition. The **transpose** A^T of an $m \times n$ matrix A is the $n \times m$ matrix obtained by interchanging rows and columns:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix}.$$

A square matrix is called **symmetric** if $A = A^T$. □

Remark. One can easily check that $(A^T)^T = A$ and $(A + B)^T = A^T + B^T$.

Examples

1. $\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}^T = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \neq \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}$ so this matrix is not symmetric.
2. $\begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}^T = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$ so this matrix is symmetric.
3. $\begin{pmatrix} 2 & 3 & 1 \\ -1 & 4 & 6 \end{pmatrix}^T = \begin{pmatrix} 2 & -1 \\ 3 & 4 \\ 1 & 6 \end{pmatrix}$.
4. $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$. □

(2.2) Matrix multiplication

It is desirable to have a notion of *multiplication* for certain matrices. This will make matrices much more useful, and provide a richer algebraic structure for us to study.

Row–column products

We seek to write the system of equations

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ & & \vdots & & & & \vdots & & \vdots \\ a_{m1}\,x_1 & + & a_{m2}\,x_2 & + & \cdots & + & a_{mn}\,x_n & = & b_m \end{array}$$

as a **matrix equation**: $A\,\mathbf{x} = \mathbf{b}$, for certain choices of matrices. By convention, we put:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix};$$

we are hoping to be able say $A \times \mathbf{x} = \mathbf{b}$. If we pick off the first row of A and $\mathbf{b}$ we are aiming for:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = b_1$$

as a representation of the first equation in our system:

$$a_{11}\,x_1 + a_{12}\,x_2 + \cdots + a_{1n}\,x_n = b_1.$$

So, we make the following definition.

Definition. Let $\mathbf{r} = (r_1\ r_2\ \cdots\ r_n)$ be a $1 \times n$ *row vector* and let $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$ be an $n \times 1$ *column vector*.

Then

$$\mathbf{r}\,\mathbf{c} = r_1\,c_1 + r_2\,c_2 + \cdots + r_n\,c_n = \sum_{j=1}^{n} r_j\,c_j$$

is the **row–column product** of $\mathbf{r}$ and $\mathbf{c}$. □

Example. We compute

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

□

Note: the row–column product always produces a **number**; the dimensions of the vectors must match; the row vector must always be written first. □

General matrix products

With a simple generalization, we can define matrix products.

Definition. Let A be an $m \times n$ matrix and B an $n \times p$ matrix. Then, their **product** AB is the $m \times p$ matrix whose ikth entry is the row–column product of the ith row of A with the kth column of B. □

Example 1. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 5 & 1 \\ -1 & 3 \end{pmatrix}$. Then,

$$\begin{aligned} (AB)_{11} &= \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 5 \\ -1 \end{pmatrix} &= 5 - 2 = 3 \\ (AB)_{12} &= \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} &= 1 + 6 = 7 \\ (AB)_{21} &= \begin{pmatrix} 3 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ -1 \end{pmatrix} &= 15 - 4 = 11 \\ (AB)_{22} &= \begin{pmatrix} 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} &= 3 + 12 = 15 \end{aligned}$$

so $A\,B = \begin{pmatrix} 3 & 7 \\ 11 & 15 \end{pmatrix}$. On the other hand,

$$B\,A = \begin{pmatrix} 5 & 1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 5+3 & 10+4 \\ -1+9 & -2+12 \end{pmatrix} = \begin{pmatrix} 8 & 14 \\ 8 & 10 \end{pmatrix}.$$

□

Example 2. Let

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 1 & 4 \end{pmatrix}, \; B = \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then A is a 2×3 matrix, and B is a 2×2 matrix. Since the dimensions do not agree in the required way, the product AB does not exist. □

Example 3. Let A and B be as in the previous example. If we let $m = 2$, $n = 2$ and $p = 3$ then B is an $m \times n$ matrix and A in an $n \times p$ matrix, so we can define the product $B\,A$, and we expect an $n \times p$ matrix (2×3) for the product:

$$\begin{aligned} B\,A &= \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 3 & 2 \\ 2 & 1 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 3+2 & 9+1 & 6+4 \\ -1 & -3 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 5 & 10 & 10 \\ -1 & -3 & -2 \end{pmatrix}. \end{aligned}$$

□

Example 4. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 2 & 3 \end{pmatrix}$ and $C = \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix}$. Then

(i) $AB = \begin{pmatrix} 1 & 3 & 7 \\ 3 & 5 & 15 \end{pmatrix}$;

(ii) AC is not defined, due to incompatible dimensions;

(iii) $BC = \begin{pmatrix} 0 \\ 17 \end{pmatrix}$;

(iv) $ABC = \begin{pmatrix} 34 \\ 68 \end{pmatrix}$;

(v) BA is not defined, due to incompatible dimensions;

(vi) $B^T A = \begin{pmatrix} 1 & 2 \\ 5 & 6 \\ 10 & 14 \end{pmatrix}$.

□

Remark. The product formula can be written reasonably compactly: if $A = (a_{ij})$ is $m \times n$ and $B = (b_{jk})$ is $n \times p$ (so number of columns of A = number of rows of B) then we define

$$AB = (a_{ij})(b_{jk}) \ = \ (c_{ik}) \text{ where } c_{ik} \ = \ \sum_{j=1}^{n} a_{ij} b_{jk}.$$

Properties of matrix multiplication

Certain properties of ordinary algebra carry over to matrix multiplication.

- We have the *associative* and *distributive* laws: assume the indicated operations can be performed on matrices A, B, C then:

  1. $(AB)C = A(BC)$

  2. $A(B + C) = AB + AC$

  3. $(A + B)C = AC + BC$

  4. $\alpha(AB) = (\alpha A)B = A(\alpha B)$, $\alpha \in \mathbb{R}$.

- Some identities of ordinary algebra carry over too. For example, if A, B are square of the same size, then

$$\begin{aligned} (A + B)^2 &= (A + B)(A + B) \\ &= A(A + B) + B(A + B) \\ &= A^2 + AB + BA + B^2. \end{aligned}$$

  (We cannot assume $AB = BA$ though, recall Example 1 above)

- Because of associativity of multiplication, brackets are not needed in products, and for a square matrix A, we write

$$A^k \ = \ A \times A \times \cdots \times A \ (k \text{ times}).$$

- If A and B are such that their product is defined, then

$$(A\,B)^T = B^T\,A^T.$$

- For every integer $n > 0$, the $n \times n$ *identity matrix* I_n "fixes" a matrix under multiplication. (Recall, I_n is diagonal and has all its diagonal entries equal to 1.) For example,

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

  Basic property: if A is $m \times n$, then

$$A\, I_n \;=\; I_m\, A \;=\; A.$$

Some odd things happen with matrix multiplication.

- Not all pairs of matrices can be multiplied together; the dimensions must be compatible.
- Generally, $AB \neq BA$, even if both products are defined and have the same sizes (for example if A, B are both $n \times n$). We saw an example of this above. Here is another one:

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

  whereas

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

- Two non-zero matrices can multiply to give a zero matrix. For example,

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

Remarks

- We have seen that a system of linear equations

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & & & & \vdots & & \vdots & \\ a_{m1}\,x_1 & + & a_{m2}\,x_2 & + & \cdots & + & a_{mn}\,x_n & = & b_m \end{array}$$

  can be written as a matrix equation: $A\,\mathbf{x} = \mathbf{b}$ (A is the $m \times n$ coefficient matrix, and $\mathbf{x}$ is the $n \times 1$ matrix of variables). However, the temptation to replace the solution of the underlying system with the algebraic manouvre $\mathbf{x} = \mathbf{b} \div A$ should be **resisted at all costs.** We have said **nothing** about division of matrices; we have had enough trouble with matrix multiplication—matrix division does not make sense in general. We will see shortly how the idea of an *inverse matrix* can be employed to make some sense of "undoing matrix multiplication".

- You may find it interesting to think of a matrix as inducing a "function on vectors", via multiplication. If A is an $m \times n$ matrix and $\mathbf{x}$ is an n–dimensional column vector, then $\mathbf{y} = A\,\mathbf{x}$ is an m–dimensional column vector. This is especially important for square matrices, since they can be used in this way to describe *transformations* on $\mathbb{R}^n$ (the points in $\mathbb{R}^n$ are identified with vectors).

(2.3) Matrix inversion

Suppose that $\mathbf{b}$ is an n–dimensional column vector. Then the action of the *identity matrix* I_n on $\mathbf{b}$ is just like multiplying by 1:

$$I_n\,\mathbf{b} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{b}.$$

Thus, if we could find a matrix B such that $B\,A = I_n$ we could solve the system $A\,\mathbf{x} = \mathbf{b}$:

$$\mathbf{x} = I_n\,\mathbf{x} = (B\,A)\,\mathbf{x} = B\,(A\,\mathbf{x}) = B\,\mathbf{b},$$

so that $\mathbf{x}$ can be found by matrix multiplication.

Definition. An $n \times n$ matrix A is **invertible** (or non-singular) if there is another $n \times n$ matrix B such that

$$A\,B = I_n = B\,A.$$

The matrix B is called the **inverse** of A and is written A^{-1}. If no such matrix B exists then A is non-invertible or **singular**. □

Remarks

- Technically, it is only necessary to require $AB = I_n$ in the definition, since one can then prove that also $BA = I_n$ (see Theorem 1.6.3 on page 61 of Anton, or Theorem 1.8.7 on page 112 of Grossman.).
- We have also glossed over the fact that an inverse is not guaranteed to be unique in the definition. The proof of uniqueness is a classic in algebra: suppose that $AB = BA = I_n = AC = CA$ for matrices B and C. Then
$$C = C\,I_n = C\,(AB) = (CA)\,B = I_n\,B = B.$$

Example 1. Let $A = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}$ and $B = \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix}$. Then

$$A\,B = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2 = \cdots = B\,A$$

so $\begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix}$. Now, suppose we wish to solve the system:

$$\begin{array}{rcrcl} 3\,x & + & 2\,y & = & 5 \\ 4\,x & + & 3\,y & = & 2 \end{array}.$$

We can write this as:

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

so

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$
$$= \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$
$$= \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix} \begin{pmatrix} 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 11 \\ -14 \end{pmatrix};$$

that is, $x = 11$ and $y = -14$. □

Of course, being invertible is a rather special property, and not all square matrices have it.

Example 2. The matrix $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is singular. *Proof:* if A is to be invertible, we must be able to find a matrix B such that $I_2 = AB$. Let $B = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$, so writing out the equation:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} x+z & y+w \\ x+z & y+w \end{pmatrix}.$$

But looking at the 11 and 21 components of this equation we get

$$1 = x + z \text{ from the 11 entry and } 0 = x + z \text{ from the 21 entry.}$$

This is inconsistent, so there can be no such matrix B. □

Theorem 2.1 *Let A and B be invertible square matrices. Then:*

1. *A^{-1} is invertible and $(A^{-1})^{-1} = A$;*
2. *AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$;*
3. *A^T is invertible and $(A^T)^{-1} = (A^{-1})^T$.*

These are all straight-forward applications of the definition of the inverse.

Finding the inverse

Example 3. Let us find the inverse of the matrix $\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$. We seek an unknown matrix $B = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$ such that $AB = I_2$; that is:

$$\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This is really **two** systems of equations, one for each column:

$$\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

It turns out the same sequence of row operations works for solving both systems, so we do them simultaneously:

$$\left(\begin{array}{cc|c} 2 & 1 & 1 \\ 5 & 3 & 0 \end{array}\right) \qquad \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 5 & 3 & 1 \end{array}\right)$$

$$R_2 \to R_2 - \tfrac{5}{2}R_1$$

$$\left(\begin{array}{cc|c} 2 & 1 & 1 \\ 0 & \frac{1}{2} & -\frac{5}{2} \end{array}\right) \qquad \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{array}\right)$$

$$R_2 \to 2R_2$$

$$\left(\begin{array}{cc|c} 2 & 1 & 1 \\ 0 & 1 & -5 \end{array}\right) \qquad \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 1 & 2 \end{array}\right)$$

$$R_1 \to R_1 - R_2$$

$$\left(\begin{array}{cc|c} 2 & 0 & 6 \\ 0 & 1 & -5 \end{array}\right) \qquad \left(\begin{array}{cc|c} 2 & 0 & -2 \\ 0 & 1 & 2 \end{array}\right)$$

$$R_1 \to \tfrac{1}{2}R_2$$

$$\left(\begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & -5 \end{array}\right) \qquad \left(\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & 2 \end{array}\right).$$

Reading off the solution, $x = 3, z = -5, y = -1, w = 2$, so that the inverse is $\begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$. □

The procedure we used in this example turns out to have general applicability: essentially, we construct a large augmented matrix, with a family of RHSs for the columns of the identity matrix, and solve simultaneously for all the columns of A^{-1}.

Algorithm for calculating A^{-1}

1. Form the $n \times 2n$ augmented matrix:

$$(A|I_n).$$

2. Perform Gauss–Jordan elimination to obtain the RREF.

3. If row operations cease with an augmented system

$$(I_n|B)$$

then $B = A^{-1}$;

otherwise, a row of 0s appears in the left-hand matrix and A is *singular*.

Theorem 2.2 *The augmented matrix method for a square matrix either produces the inverse or proves the matrix is non-invertible.*

The augmented matrix method works in practice, and this theorem tells us that it always will. The theorem is immediate, once we have established a few more facts about inverses. Convince yourself of this after reading Theorem 2.3 below.

Examples

1. Find $\begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 2 \\ 9 & -2 & 5 \end{pmatrix}^{-1}$.

Solution:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 3 & 1 & 2 & 0 & 1 & 0 \\ 9 & -2 & 5 & 0 & 0 & 1 \end{array}\right) \quad R_3 \to R_3 - R_2$$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 3 & 1 & 2 & 0 & 1 & 0 \\ 0 & -5 & -1 & 0 & -3 & 1 \end{array}\right) \quad \begin{array}{l} R_2 \to R_2 - 3\,R_1 \\ R_3 \to -R_3 \end{array}$$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -5 & -1 & -3 & 1 & 0 \\ 0 & 5 & 1 & 0 & 3 & -1 \end{array}\right) \quad R_3 \to R_3 + R_2$$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -5 & -1 & -3 & 1 & 0 \\ 0 & 0 & 0 & -3 & 4 & -1 \end{array}\right) \quad \text{Stop! No Inverse.}$$

2. Find $\left(\begin{array}{ccc} 2 & 1 & -1 \\ 0 & 2 & 1 \\ 5 & 2 & -3 \end{array}\right)^{-1}$.

Solution:

$$\left(\begin{array}{ccc|ccc} 2 & 1 & -1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 5 & 2 & -3 & 0 & 0 & 1 \end{array}\right) \quad R_3 \to R_3 - 2R_1$$

$$\left(\begin{array}{ccc|ccc} 2 & 1 & -1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & -2 & 0 & 1 \end{array}\right) \quad \begin{array}{l} R_1 \to R_3 \\ \\ R_3 \to R_1 \end{array}$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -2 & 0 & 1 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 2 & 1 & -1 & 1 & 0 & 0 \end{array}\right) \quad R_3 \to R_3 - 2R_1$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -2 & 0 & 1 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 5 & 0 & -2 \end{array}\right) \quad R_2 \to R_2 - R_3$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -2 & 0 & 1 \\ 0 & 1 & 0 & -5 & 1 & 2 \\ 0 & 1 & 1 & 5 & 0 & -2 \end{array}\right) \quad R_3 \to R_3 - R_2$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -2 & 0 & 1 \\ 0 & 1 & 0 & -5 & 1 & 2 \\ 0 & 0 & 1 & 10 & -1 & -4 \end{array}\right) \quad R_1 \to R_1 + R_3$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 8 & -1 & -3 \\ 0 & 1 & 0 & -5 & 1 & 2 \\ 0 & 0 & 1 & 10 & -1 & -4 \end{array}\right)$$

So the inverse is: $\left(\begin{array}{ccc} 8 & -1 & -3 \\ -5 & 1 & 2 \\ 10 & -1 & -4 \end{array}\right)$.

3. Find $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}^{-1}$.

Solution:

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right) \quad R_4 \to R_4 - R_3$$

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 \end{array}\right) \quad R_3 \to R_3 - R_2$$

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 \end{array}\right) \quad R_2 \to R_2 - R_1$$

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 \end{array}\right)$$

So the inverse is: $\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$.

Invertibility and linear equations

We have seen (in Example 1) how to use the inverse to solve linear equations. We will finish this section by recording the importance of invertibility in a theorem.

Theorem 2.3 *Let A be an $n \times n$ matrix. The following are equivalent:*
(1) A is invertible;
(2) every equation $A\,\mathbf{x} = \mathbf{b}$ has a unique solution ($\mathbf{x} = A^{-1}\,\mathbf{b}$);
(3) if $A\,\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$;
(4) no echelon form of A has a row of 0s;
(5) the RREF of A is I_n.

Proof of Theorem 2.3 (Optional)

Proof: (1)⇒(2) First of all, $\mathbf{x} = A^{-1}\,\mathbf{b}$ is a solution, since $A(A^{-1}\mathbf{b}) = A\,A^{-1}\mathbf{b} = I_n\mathbf{b} = \mathbf{b}$. If $\mathbf{x}'$ is another solution then $A^{-1}\mathbf{b} = A^{-1}(A\,\mathbf{x}') = \mathbf{x}'$, so the solution is unique.
(2)⇒(3) Clearly, $A\,\mathbf{0} = \mathbf{0}$, so $\mathbf{x} = \mathbf{0}$ follows by uniqueness.
(3)⇒(4) If an echelon form E of A had a row of zeros, then E would have at most $n-1$ leading variables, and thus at least one free variable. Therefore, (by inspecting the row reduced augmented matrix $(E|\mathbf{0})$),

there is a solution to the equation $A\,\mathbf{x} = \mathbf{0}$ of the form $\mathbf{x} = t\,\mathbf{v}$ for a non-zero vector $\mathbf{v}$ (it has a 1 in the row corresponding to the free variable).
(4)⇒(5) The RREF has no row of zeros, therefore, every row contains a leading 1, and since there are n rows and n columns, there can be no free variables. Thus, the Gauss-Jordan elimination procedure can row reduce A to I_n.
(5)⇒(1) Since the RREF of A is I_n, the equation $AB = I_n$ can be solved "column-by-column" by the augmented matrix procedure outlined above. □

(2.4) Homogeneous equations and the general solution to $A\,\mathbf{x} = \mathbf{b}$

We have seen above that a square matrix A is invertible precisely when every equation $A\,\mathbf{x} = \mathbf{b}$ has a unique solution for each choice of $\mathbf{b}$. In this case, $\mathbf{x} = A^{-1}\mathbf{b}$. For non-square matrices, such a characterization makes no sense: a non-square matrix cannot be invertible. However, we can say a little more about the solutions to $A\,\mathbf{x} = \mathbf{b}$ in the general case. The essential possibilities are that the equation may have 0, 1 or infinitely many solutions, depending on A and $\mathbf{b}$. The Gauss–Jordan algorithm is a very powerful method for finding all of these solutions, and we'll now spend some time having a closer look at what exactly it is finding! The basic questions about $A\,\mathbf{x} = \mathbf{b}$ which are resolved by Gauss–Jordan elimination are: *is the system consistent?* and *what are all the solutions?*

Solution spaces, homogeneous and particular solutions

We need a couple of pieces of terminology.

Definition. The *solution space* is the general solution to $A\,\mathbf{x} = \mathbf{b}$ (*ie.* the set of vectors $\mathbf{x}$ which satisfy the equation). The *homogeneous equation* (associated with A) is

$$A\,\mathbf{x} = \mathbf{0}.$$

The general solution to the homogenous equation is called the *homogeneous solution*. A solution to $A\,\mathbf{x} = \mathbf{b}$ is called a *particular solution*. □

Similar terminology will occur when we solve linear diophantine equations later on.

Suppose $\mathbf{x}_p, \mathbf{y}_p$ are both particular solutions. Then

$$A\,\mathbf{x}_p = \mathbf{b} \text{ and } A\,\mathbf{y}_p = \mathbf{b}$$

so

$$A\,(\mathbf{x}_p - \mathbf{y}_p) = A\,\mathbf{x}_p - A\,\mathbf{y}_p = \mathbf{b} - \mathbf{b} = \mathbf{0};$$

that is, $\mathbf{x}_p - \mathbf{y}_p$ belongs to the homogeneous solution. Conversely, if $\mathbf{x}_h$ is the homogeneous solution, then if $\mathbf{x}_p$ is a particular solution,

$$A\,(\mathbf{x}_p + \mathbf{x}_h) = A\,\mathbf{x}_p + A\,\mathbf{x}_h = \mathbf{b} + \mathbf{0} = \mathbf{b}.$$

Thus, the general solution can be expressed as the sum of a particular solution and the homogenous solution.

In fact, the Gaussian elimination algorithm automatically finds both homogeneous and particular solutions. We will investigate this by revisiting several of our earlier examples.

Example 1. Let $A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 4 & 1 & 5 \end{pmatrix}$ and $\mathbf{b} = (-2, 5, 1)^T$. The RREF is

$$\left(\begin{array}{ccc|c} 1 & 0 & \frac{4}{3} & 1 \\ 0 & 1 & -\frac{1}{3} & -3 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

First of all, from the bottom row, the system is consistent. Next, we find the homogeneous solution by setting the RHS to zero, assigning a free parameter to each free variable (in this case, only z, corresponding to the third column), and applying back substitution. That is, we solve

$$\begin{array}{rcrcl} x & & + & \frac{4}{3}\,z & = & 0 \\ & y & - & \frac{1}{3}\,z & = & 0 \\ & & & z & = & z \end{array}$$

to get $(x, y, z) = (-\frac{4}{3}\,z, \frac{1}{3}\,z, z)$. It is more convenient to write this in terms of a parameter $t = 3\,z$ to get

$$\mathbf{x}_h = t \begin{pmatrix} -4 \\ 1 \\ 3 \end{pmatrix}, \quad t \in \mathbb{R}.$$

Next, in finding $\mathbf{x}_p$ we need only **one** particular solution, since any particular solutions differ by a homogeneous solution. This means that in finding $\mathbf{x}_p$ we can ignore any homogeneous contributions and so set the corresponding parameters to 0. Since $\mathbf{x}_h$ was generated by the third column (without a leading 1), we can ignore this column, and solve

$$\left(\begin{array}{ccc|c} 1 & 0 & \times & 1 \\ 0 & 1 & \times & -3 \\ 0 & 0 & \times & 0 \end{array}\right).$$

Clearly, $\mathbf{x}_p = (1, -3, 0)^T$ and we write the general solution as

$$\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix} + t \begin{pmatrix} -4 \\ 1 \\ 3 \end{pmatrix}.$$

□

Notice that the free parameters are used up by solving the homogeneous equation, leaving the leading 1s for use in finding a particular solution.

Example 2. Let $A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 4 & 1 & 5 \end{pmatrix}$ and $\mathbf{b} = (1, 2, 5)^T$. The RREF is

$$\left(\begin{array}{ccc|c} 1 & 0 & \frac{4}{3} & 1 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{array}\right).$$

The system is **inconsistent** so has no solution space. However, the homogeneous system has RREF

$$\left(\begin{array}{ccc|c} 1 & 0 & \frac{4}{3} & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right),$$

and the homogeneous solution is the same as in the previous example. □

Example 3. Let $A = \begin{pmatrix} 2 & 4 & 2 \\ 0 & 1 & 1 \\ 1 & 3 & 3 \end{pmatrix}$ and $\mathbf{b} = (6, 1, 5)^T$. The RREF is

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right).$$

In this case, $\mathbf{x}_h = (0,0,0)^T$ is the homogeneous solution (there are no free parameters), and the particular solution is $\mathbf{x}_p = (2,0,1)^T$. Since $\mathbf{x}_h = \{\mathbf{0}\}$, there is only one particular solution in this case. □

Remark: Notice that the homogeneous equation is the same for every choice of $\mathbf{b}$ ($\mathbf{b}$ is a particular RHS, and in the homogeneous equation the RHS is $\mathbf{0}$). Also, since $A\,\mathbf{0} = \mathbf{0}$, the homogeneous equation is **always consistent**, and the homogeneous solution always includes $\mathbf{0}$. □

Some special "vector spaces"

Let A be an $m \times n$ matrix. There are some special collections of vectors associated with every such matrix. These are intimately connected with the "size of the solution space" of $A\,\mathbf{x} = \mathbf{b}$. The first special space specifies "the possible $\mathbf{b}$s for which $A\,\mathbf{x} = \mathbf{b}$ is consistent", and the second space determines the dimension of the solution space when the system is consistent.

Definition. The *column space* of A col(A) consists of the vectors in $\mathbb{R}^m$ that can be written as $A\,\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$. This is sometimes called the *range* of A. The *null space* (or *kernel*) of A, denoted null(A), is the set of vectors $\mathbf{x}$ in $\mathbb{R}^n$ for which $A\,\mathbf{x} = \mathbf{0}$. □

Note: If $A = (\mathbf{c}_1|\mathbf{c}_2|\cdots|\mathbf{c}_n)$ (where the m–dimensional vectors $\mathbf{c}_i$ are the *columns* of A), then

$$A\,\mathbf{x} = x_1\,\mathbf{c}_1 + x_2\,\mathbf{c}_2 + \cdots + x_n\,\mathbf{c}_n$$

(where $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$). So every vector $A\,\mathbf{x}$ is a sum of scalar multiples of the columns of A. The columns of A thus "span" the space col(A) (explaining the name). Note also that null(A) is the set of solutions to the *homogeneous equation*. □

Example 4. Let $A = \begin{pmatrix} 1 & -2 \\ 3 & -6 \end{pmatrix}$. If $\mathbf{x} = \begin{pmatrix} s \\ t \end{pmatrix}$ then

$$A\,\mathbf{x} = \begin{pmatrix} s - 2\,t \\ 3s - 6t \end{pmatrix} = (s - 2t)\begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

This means that **every** vector of the form $A\,\mathbf{x}$ can be written as some multiple of the vector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$. Thus

$$\text{col}(A) = \left\{ \lambda \begin{pmatrix} 1 \\ 3 \end{pmatrix} \middle| \, \lambda \in \mathbb{R} \right\}.$$

In fact, our calculation above also shows how to find null(A). If $A\,\mathbf{x} = \mathbf{0}$, then we must have $s - 2\,t = 0$ so $s = 2\,t$ and $\mathbf{x} = \begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} 2\,t \\ t \end{pmatrix} = t \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Thus,

$$\text{null}(A) = \left\{ t \begin{pmatrix} 2 \\ 1 \end{pmatrix} \middle| \, t \in \mathbb{R} \right\}.$$

□

In this example, finding col(A) and null(A) was quite easy; in general, we might need to use Gauss–Jordan elimination. The solution space to the homogeneous system is null(A). It turns out that col(A) is spanned by the columns of A which contain a leading 1 in the RREF. The proof of this is beyond the scope of MATH102. The point of all this is the characterization:

$$A\,\mathbf{x} = \mathbf{b} \text{ is consistent } \Leftrightarrow \mathbf{b} \in \text{col}(A).$$

Example 5. Let $A = \begin{pmatrix} 1 & 0 & 1 & 4 \\ 2 & -1 & 1 & 7 \\ 1 & -2 & -1 & 2 \end{pmatrix}$. The RREF of A is

$$\begin{pmatrix} 1 & 0 & 1 & 4 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

from which we can read off the homogeneous solution so

$$\text{null}(A) = \left\{ s \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -4 \\ -1 \\ 0 \\ 1 \end{pmatrix} \middle| \; s, t \in \mathbb{R} \right\}.$$

Suppose now that $\mathbf{b} \in \mathbb{R}^3$, and that the augmented matrix $(A|\mathbf{b})$ has RREF

$$\left(\begin{array}{cccc|c} 1 & 0 & 1 & 4 & u \\ 0 & 1 & 1 & 1 & v \\ 0 & 0 & 0 & 0 & w \end{array}\right).$$

If there exists a particular solution, we must have $w = 0$ (otherwise the system $A\,\mathbf{x} = \mathbf{b}$ is inconsistent). From our work above, we know that **any** particular solutions differ by a homogeneous solution, so all we need to do is find **one** particular solution. Since the homogeneous solution is "generated" by the columns without leading 1s (the third and fourth columns), these columns can be ignored safely and a particular solution can be found by looking at

$$\left(\begin{array}{cccc|c} 1 & 0 & \times & \times & u \\ 0 & 1 & \times & \times & v \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Clearly, $\mathbf{x}_p = (u, v, 0, 0)^T$ is a particular solution. Notice that this also tells us how to identify col(A): if $\mathbf{b} \in \text{col}(A)$ then $A\,\mathbf{x} = \mathbf{b}$ has solutions of the form $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h$. Then, using the notation above,

$$\mathbf{b} = A\,\mathbf{x} = A\,(\mathbf{x}_p + \mathbf{x}_h) = A\,\mathbf{x}_p + A\,\mathbf{x}_h = \begin{pmatrix} 1 & 0 & 1 & 4 \\ 2 & -1 & 1 & 7 \\ 1 & -2 & -1 & 2 \end{pmatrix} \begin{pmatrix} u \\ v \\ 0 \\ 0 \end{pmatrix} + \mathbf{0} = u \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + v \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix}.$$

The method of this example can be applied in general to find the null and columns spaces of matrices. □

Aside: canonical decomposition with respect to A

There are two additional spaces associated with A. These are the null space and columns space of A^T. The latter space is called the *row space*, since the columns of A^T are the rows of A. One can prove that every vector in null(A^T) is orthogonal to every vector in col(A), and every vector in null(A) is orthogonal to every vector in row(A). In fact, these spaces provide nice *decompositions*:

$$\mathbb{R}^n = \text{null}(A) \oplus \text{row}(A) \text{ and } \mathbb{R}^m = \text{null}(A^T) \oplus \text{col}(A).$$

Moreover, the dimensions of row(A) and col(A) are the same, and are the *rank* of A. The rank is the number of leading ones in the matrix. These facts are beyond the scope of MATH102, but lead to a very deep and applicable understanding of the geometry of linear transformations.

III ∘ Determinants

(3.1) Definition of determinants

Recall that invertible matrices correspond to systems of equations with unique solutions. This makes detection of non-invertibility important. We will develop a numerical criterion for assessing invertibility of square matrices.

2×2 determinants

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. A careful reading of Theorem 2.3 (equivalence of (1) and (4)) reveals that A is singular if and only if it has an echelon form containing a row of zeros. So, let's try to find the REF of A. By performing the row operation $R_2 \to R_2 - \frac{c}{a}R_1$ we get

$$\begin{pmatrix} a & b \\ 0 & d - \frac{bc}{a} \end{pmatrix}.$$

Multiplying R_2 by a, we see that the REF has a row of zeros only if $a\,d - b\,c = 0$. Thus:

Theorem 3.1 *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *and define the* **determinant** *of* A *to be* $\det(A) = ad - bc$. *Then* A *is singular if and only if* $\det(A) = 0$.

Notation: Sometimes, instead of writing $\det(A)$ we will write $|A|$.

Example 1. Let $A = \begin{pmatrix} 2 & 1 \\ -3 & 3 \end{pmatrix}$. Then

$$\det(A) = \begin{vmatrix} 2 & 1 \\ -3 & 3 \end{vmatrix} = 2 \times 3 - 1 \times (-3) = 6 + 3 = 9.$$

Thus, A is invertible. □

3×3 determinants

Similar ideas can be extended to 3×3 matrices, but the formulas are more complicated. Let

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

and let M_{1j} be the **minor** obtained by deleting the first row and jth column from A. Thus,

$$M_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}, M_{12} = \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} \text{ and } M_{13} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}.$$

Definition. The determinant of a 3×3 matrix $A = (a_{ij})$ is

$$\det(A) = a_{11}|M_{11}| - a_{12}|M_{12}| + a_{13}|M_{13}|.$$

□

This determinant is computed recursively (it is in written in terms of 2×2 determinants), and also has the property of determining whether a matrix is singular or not.

Example 2. Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}$. Then the minors are

$$M_{11} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}, M_{12} = \begin{pmatrix} 0 & 3 \\ 1 & 1 \end{pmatrix}, M_{13} = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}.$$

Hence,

$$\begin{aligned} \det(A) &= 1 \times \begin{vmatrix} 2 & 3 \\ 1 & 1 \end{vmatrix} - 1 \times \begin{vmatrix} 0 & 3 \\ 1 & 1 \end{vmatrix} + 0 \times \begin{vmatrix} 0 & 2 \\ 1 & 1 \end{vmatrix} \\ &= (2 \times 1 - 3 \times 1) - (0 \times 1 - 3 \times 1) + 0 \\ &= -1 - (-3) = 2. \end{aligned}$$

Since $\det(A) \neq 0$, A is invertible. □

$n \times n$ determinants

The idea of a determinant proves useful for square matrices of any size. The general definition of an $n \times n$ determinant can be made "recursively".

For an $n \times n$ matrix A, with

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

let M_{ij} be the matrix obtained from A by deleting the ith row and jth column. Each such matrix is an $(n-1) \times (n-1)$ **minor** of A.

Then we define $\det(A)$ to be

$$a_{11} \det(M_{11}) - a_{12} \det(M_{12}) + \cdots \pm a_{1n} \det(M_{1n}).$$

The sign of the last term depends on whether n is even or odd: it will alternate $+, -, +, -, \ldots$.

Example 3. Let

$$A = \begin{pmatrix} 1 & 0 & 7 & 2 \\ 0 & 3 & 0 & 3 \\ 2 & 4 & 2 & 0 \\ 1 & 0 & 0 & 4 \end{pmatrix}.$$

Then

$$\det(A) = 1 \times \begin{vmatrix} 3 & 0 & 3 \\ 4 & 2 & 0 \\ 0 & 0 & 4 \end{vmatrix} - 0 \times (\cdots) + 7 \times \begin{vmatrix} 0 & 3 & 3 \\ 2 & 4 & 0 \\ 1 & 0 & 4 \end{vmatrix} - 2 \times \begin{vmatrix} 0 & 3 & 0 \\ 2 & 4 & 2 \\ 1 & 0 & 0 \end{vmatrix}.$$

For the first 3×3 determinant above, we can continue the process.

$$\begin{aligned}\begin{vmatrix} 3 & 0 & 3 \\ 4 & 2 & 0 \\ 0 & 0 & 4 \end{vmatrix} &= 3\begin{vmatrix} 2 & 0 \\ 0 & 4 \end{vmatrix} - 0 \times (\cdots) + 3\begin{vmatrix} 4 & 2 \\ 0 & 0 \end{vmatrix} \\ &= 3\left(2 \times 4 - 0 \times 0\right) - 0 + 3\left(0 \times 0\right) \\ &= 24.\end{aligned}$$

The second 3×3 determinant is irrelevant since it is multiplied by zero anyway. The third and fourth come out to be -36 and 6 respectively. So

$$\det(A) = 1 \times 24 + 7 \times (-36) - 2 \times 6 = -240.$$

□

This way of computing the determinant is called a **cofactor expansion along the first row**. The cofactors are the (signed) determinants of the minors, so the cofactor of a_{ij} is $(-1)^{i+j}|M_{ij}|$. Somewhat surprisingly, this process produces a notion of the determinant which behaves in exactly the right way.

Theorem 3.2 *A is invertible if and only if* $\det(A) \neq 0$.

We will defer the proof of this theorem until we've collected a few more facts about the way that determinants behave.

Theorem 3.3 *Let A and B be square matrices. Then*

1. *The determinant can be computed by a cofactor expansion along any row:*

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(M_{ij}), \quad (i = 1, \cdots, n);$$

2. $\det\left(A^T\right) = \det(A)$*;*
3. $\det(AB) = \det(A)\det(B)$.

The proof of Theorem 3.3 is uninspiring, and technical. Therefore, it is omitted (see Theorems 1,4,5 in Section 2.2 of Grossman, if you are really keen, or Theorems in Sections 2.3—2.4 of Anton!)

The general formula for expanding along the ith row is in terms of the minors M_{ij} of A, obtained by deleting the ith row and the jth column. Thus $\det(A)$ equals

$$(-1)^{i+1} a_{i1} \det(M_{i1}) + (-1)^{i+2} a_{i2} \det(M_{i2}) + \cdots + (-1)^{i+n} a_{in} \det(M_{in}).$$

The $(-1)^{i+j}$ terms tell us which sign to use: if $i + j$ is even, it keeps the sign of a_{ij} as it is, while if $i + j$ is odd, it changes the sign.

Taken together, Theorem 3.3 (1) and Theorem 3.3 (2) show that we can expand by cofactors along *any* row or column. The best idea is to choose the one with the most 0s in it. In fact, if a matrix has an entire row or column of zeros, then its determinant must be zero!

Example 4. As in the previous example, let

$$A = \begin{pmatrix} 1 & 0 & 7 & 2 \\ 0 & 3 & 0 & 3 \\ 2 & 4 & 2 & 0 \\ 1 & 0 & 0 & 4 \end{pmatrix}.$$

We are free to compute the determinant by expanding in cofactors along any row or column. Since the final row has two zeros, we could usefully use this row:

$$\det(A) = (-1)^{4+1} \times 1 \times \begin{vmatrix} 0 & 7 & 2 \\ 3 & 0 & 3 \\ 4 & 2 & 0 \end{vmatrix} + 0 + 0 + (-1)^{4+4} \times 4 \times \begin{vmatrix} 1 & 0 & 7 \\ 0 & 3 & 0 \\ 2 & 4 & 2 \end{vmatrix}.$$

Evaluating the two 3×3 determinants gives 96 and -36 respectively, so that

$$\det(A) = -1 \times 96 + 4 \times (-36) = -240,$$

in agreement with the earlier answer. □

(3.2) Determinants by row-reduction

The recursive formula for calculating the determinant requires a significant amount of arithmetic, and remembering to put in the appropriate power of (-1) adds more room for error. We'll therefore look at an alternative method of computation.

Determinant of a triangular matrix

The central observation is that determinants of triangular matrices are easy to evaluate.

Theorem 3.4 *Let A be a triangular $n \times n$ matrix. Then* $\det(A) = a_{11}a_{22}\cdots a_{nn}$.

Proof: Assume that A is upper triangular; the lower triangular case is similar. Then, by expanding about the first column,

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & & a_{2n} \\ 0 & 0 & a_{33} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & a_{nn} \end{vmatrix} &= a_{11} \begin{vmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & a_{33} & a_{34} & \ddots \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & a_{nn} \end{vmatrix} - 0 \times \cdots \\ &= a_{11}\, a_{22} \begin{vmatrix} a_{33} & a_{34} & \cdots \\ 0 & \ddots & \vdots \\ & 0 & a_{nn} \end{vmatrix} + a_{11} \times 0 \times \cdots \\ &= \cdots \\ &= a_{11}\, a_{22}\, a_{33} \cdots a_{nn}. \end{aligned}$$

□

Example 1. Let $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 3 \\ 0 & 0 & 1 \end{pmatrix}$. One can expand the determinant by cofactors, obtaining:

$$\det(A) = 1 \times \begin{vmatrix} 2 & 3 \\ 0 & 1 \end{vmatrix} - 0 \times \begin{vmatrix} 0 & 3 \\ 0 & 1 \end{vmatrix} + 1 \times \begin{vmatrix} 0 & 2 \\ 0 & 0 \end{vmatrix} = (2-0) + 0 + (0-0) = 2.$$

Alternatively, one can simply use the formula, giving $\det(A) = 1 \times 2 \times 1 = 2$. □

Example 2. Let $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 1 & 3 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$. Then $\det(A) = 1 \times 2 \times 3 \times 2 = 12$. □

Effect of row operations on $\det(A)$

The idea is to use row operations to put an arbitrary square matrix in triangular form, and then compute the determinant via Theorem 3.4.

Theorem 3.5 *Elementary row operations have the following effects on the determinant:*
(1) interchanging two rows multiplies the determinant by (-1);
(2) multiplying a row by a constant a multiplies the determinant by the same constant a;
(3) adding a constant multiple of one row to another row leaves the determinant unchanged.

The proof of Theorem 3.5 is a little bit technical; see Properties 2,4 and 7 in Section 2.2 of Grossman, or Theorem 2.2.4 in Anton for details. Our main interest is in using the result.

Theorem 3.6 *A square matrix A which has two rows the same has zero determinant.*

Proof: If two rows of A are the same, then adding (-1) times one of these rows to the other one will create a row of zeros. By Theorem 3.5, the determinant of the new matrix will be the same as $\det(A)$. But, by expanding in cofactors along the row of zeros, it is clear that the determinant of the new matrix is 0. Thus $\det(A) = 0$. □

Computations of $\det(A)$

We will perform a suitable sequence of row operations, keeping track of the effect of our operations on the determinant.

Example 3. Let $A = \begin{pmatrix} 2 & 4 & 0 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{pmatrix}$. Then, by row-reduction and Theorem 3.5,

$$\begin{aligned}
\det(A) &= \begin{vmatrix} 2 & 4 & 0 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{vmatrix} && R_1 \to \tfrac{1}{2}R_1 \\
&= 2 \times \begin{vmatrix} 1 & 2 & 0 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{vmatrix} && \\
&&& R_2 \to R_2 - R_1 \\
&&& R_3 \to R_3 - R_1 \\
&= 2 \begin{vmatrix} 1 & 2 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 1 \end{vmatrix} && R_2 \to R_3 \\
&&& R_3 \to R_2 \\
&= (-1) \times 2 \begin{vmatrix} 1 & 2 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{vmatrix} && \\
&= -2 \times 1 \times -1 \times -1 = -2,
\end{aligned}$$

using Theorem 3.4. Notice that when the first row was multiplied by $r = \frac{1}{2}$, the overall effect on the determinant was also to be multiplied by $\frac{1}{2}$. Thus, in order to maintain equality with $\det(A)$, a factor of $2 = \frac{1}{r}$ was inserted. □

Note: if an $n \times n$ matrix cannot be converted into upper triangular form in this way, it must be singular and its determinant will be zero. (See the reasoning in the proof of Theorem 3.2 below.) This will become obvious during the calculation since a row of zeros will occur.

Determinant calculations grow exponentially as n increases, but grow roughly in proportion to n^3 if row reductions are used. So for large n, the row reduction method is much preferred. *Even in the* 3×3 *case, the row operation method is generally faster and easier to use.*

Example 4. Finally, let us revisit the example of the previous section, using our new method:

$$\det\begin{pmatrix} 1 & 0 & 7 & 2 \\ 0 & 3 & 0 & 3 \\ 2 & 4 & 2 & 0 \\ 1 & 0 & 0 & 4 \end{pmatrix} = \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 3 & 0 & 3 \\ 2 & 4 & 2 & 0 \\ 1 & 0 & 0 & 4 \end{vmatrix} \quad \begin{matrix} \\ \\ R_3 \to R_3 - 2\,R_1 \\ R_4 \to R_4 - R_1 \end{matrix}$$

$$= \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 3 & 0 & 3 \\ 0 & 4 & -12 & -4 \\ 0 & 0 & -7 & 2 \end{vmatrix} \quad R_3 \to \tfrac{1}{3}\,R_3$$

$$= 3 \times \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 4 & -12 & -4 \\ 0 & 0 & -7 & 2 \end{vmatrix} \quad R_3 \to R_3 - 4\,R_2$$

$$= 3 \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -12 & -8 \\ 0 & 0 & -7 & 2 \end{vmatrix} \quad R_3 \to -\tfrac{1}{12}R_3$$

$$= -12 \times 3 \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & \frac{2}{3} \\ 0 & 0 & -7 & 2 \end{vmatrix} \quad R_4 \to R_4 + 7\,R_3$$

$$= -36 \begin{vmatrix} 1 & 0 & 7 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & \frac{2}{3} \\ 0 & 0 & 0 & \frac{20}{3} \end{vmatrix} .$$

By multiplying down the diagonal we again recover the determinant: $-36 \times \frac{20}{3} = -240$. □

Proof of Theorem 3.2

From Theorem 2.3, A is singular if and only if there is a echelon form E for A containing a row of zeros. In fact, there is no loss of generality in assuming that the last row of E is zeros. Since E was obtained from A by a sequence of elementary row operations, there is a number $r \neq 0$ such that $\det(A) = r \det(E)$ (this follows from Theorem 3.5). Moreover, since E is an EF, it is necessarily upper triangular, so by Theorem 3.4,

$$\det(A) = r \det(E) = r E_{11}\, E_{22} \cdots E_{nn} = 0$$

since the last row of E (which includes the entry E_{nn}) is $\mathbf{0}$. □

Aside: Geometry and determinants

Determinants have a geometric meaning, in that they can be used to determine the volumes and areas of certain geometric objects.

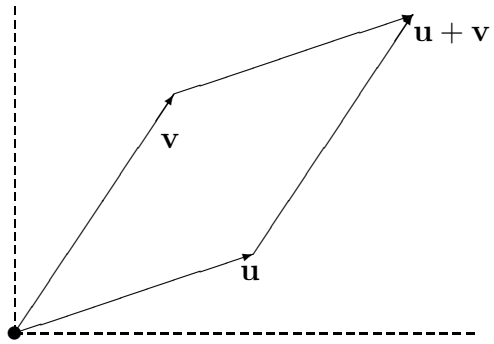
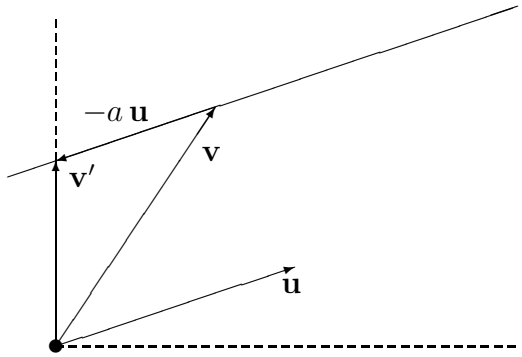


Figure 3.1: (a) Parallelogram generated by $\mathbf{u}$ and $\mathbf{v}$



(b) The row operation $R_2 \to R_2 - a\,R_1$

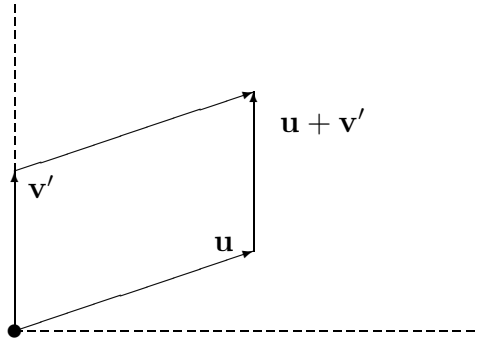
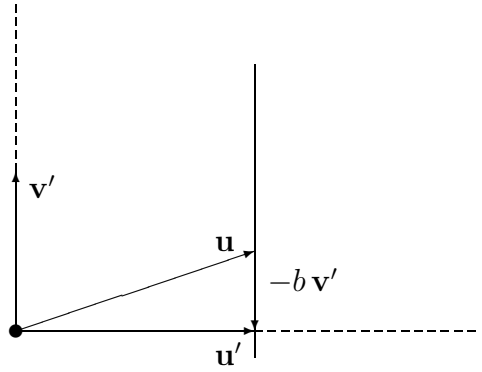


Figure 3.2: (a) Parallelogram generated by $\mathbf{u}$ and $\mathbf{v}'$



(b) The row operation $R_1 \to R_1 - b\,R_2$

2×2 determinants and parallelograms

Let $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ be two vectors in $\mathbb{R}^2$. These vectors determine a parallelogram, by letting $\mathbf{u}$ and $\mathbf{v}$ be the directions of the two pairs of parallel sides. The corners of the parallelogram have the coordinates of $\mathbf{0}$, $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u}+\mathbf{v}$, see Figure 3.1 (a). It turns out the area can be computed via determinants!

First of all, form the matrix $A = \begin{pmatrix} u_1 & u_2 \\ v_1 & v_2 \end{pmatrix}$. Then, one aims to compute the determinant via row operations. First of all, a suitable multiple $a = \frac{v_1}{u_1}$ of R_1 is subtracted from R_2.

Algebraically, this puts a 0 in the lower left hand corner of the matrix, so the transformed matrix is of the form $\begin{pmatrix} u_1 & u_2 \\ 0 & v_2' \end{pmatrix}$. We also know that we haven't changed the determinant, so

$$\det(A) = \begin{vmatrix} u_1 & u_2 \\ 0 & v_2' \end{vmatrix}.$$

Since the rows of A were the vectors $\mathbf{u}$ and $\mathbf{v}$, the geometric interpretation of the row operation is that we have replaced $\mathbf{v}$ with a vector $\mathbf{v}' = \mathbf{v} - a\,\mathbf{u}$; this corresponds to "sliding" the vector $\mathbf{v}$ down the line parallel to $\mathbf{u}$, and is depicted in Figure 3.1 (b). This induces a "shear" on the parallelogram, which preserves area; the revised parallelogram (with sides $\mathbf{u}$ and $\mathbf{v}'$ and the same area as the original) is depicted in Figure 3.2 (a).

Finally, a further row operation $R_1 \to R_1 - b\,R_2$ reduces the matrix to diagonal form, without changing the determinant. The geometric effect of this is to replace the first vector $\mathbf{u}$ by the vector $\mathbf{u}' = \mathbf{u} - b\,\mathbf{v}'$, obtained

by sliding along the line through $\mathbf{u}$, parallel to $\mathbf{v}'$. This is depicted in Figure 3.2 (b). The new parallelogram has the same area as the original, but now it is easy to compute. The vectors $\mathbf{u}'$ and $\mathbf{v}'$ are aligned parallel to the x and y axes, so take the form $\mathbf{u}' = (u_1', 0)$ and $\mathbf{v}' = (0, v_2)$. Since this parallelogram is a rectangle, its area is simply base $\times$ height $= u_1' v_2'$. Combining all this together, we have:

$$\det(A) = \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} = \begin{vmatrix} u_1' & 0 \\ 0 & v_2' \end{vmatrix} = u_1'\, v_2' = \text{area of new parallelogram} = \text{area of original parallelogram.}$$

By the calculation $\det(A) = u_1\, v_2 - u_2\, v_1$, we have proved:

Theorem 3.7 *The area of the parallelogram generated by vectors* $\mathbf{u} = (u_1, u_2)$ *and* $\mathbf{v} = (v_1, v_2)$ *is* $|u_1\, v_2 - u_2\, v_2|$.

Note: The absolute value signs are to ensure that the area is positive. □

Example 1. The area of the parallelogram with sides parallel to the vectors $(2, 1)$ and $(1, 2)$ is $2\times2-1\times1 = 3$. □

Areas of triangles and convex bodies in $\mathbb{R}^2$

Another neat application of 2×2 determinants is to the computation of areas of triangles and other convex bodies in the plane.

Theorem 3.8 *Consider the triangle with corners* $\mathbf{p}$, $\mathbf{q}$ *and* $\mathbf{r}$ *(written as row vectors). Let* A *be the matrix whose rows are* $\mathbf{p} - \mathbf{r}$ *and* $\mathbf{q} - \mathbf{r}$. *Then the area of the triangle is* $\frac{|\det(A)|}{2}$.

Proof: Certainly, the area of the triangle is unchanged if a constant vector is subtracted from all the corners (this corresponds to shifting the triangle about in the plane. So, the area is the same as the area of the triangle with corners $\mathbf{p} - \mathbf{r}, \mathbf{q} - \mathbf{r}, \mathbf{r} - \mathbf{r} = \mathbf{0}$. However, this area is precisely half the area of the parallelogram generated by the vectors $\mathbf{p} - \mathbf{r}$ and $\mathbf{q} - \mathbf{r}$ (draw a diagram). The result now follows by Theorem 3.7. □

Example 2. What is the area of the triangle with corners $\mathbf{p} = (3, 2)$, $\mathbf{q} = (5, 3)$ and $\mathbf{r} = (4, 4)$? *Solution:* The area of this triangle is

$$\frac{\begin{vmatrix} \mathbf{p} - \mathbf{r} \\ \mathbf{q} - \mathbf{r} \end{vmatrix}}{2} = \frac{\begin{vmatrix} -1 & -2 \\ 1 & -1 \end{vmatrix}}{2} = \frac{|(-1)(-1) - (-2)1|}{2} = \frac{3}{2}.$$

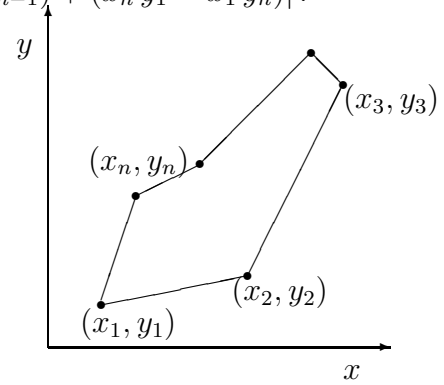
□

In fact, this approach to computing areas can be extended.

Theorem 3.9 *The* n*–sided polygon in the plane with corners* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *has area*

$$\frac{1}{2}\left|(x_1\, y_2 - x_2\, y_1) + (x_2\, y_3 - x_3\, y_2) + \cdots + (x_{n-1}\, y_n - x_n\, y_{n-1}) + (x_n\, y_1 - x_1\, y_n)\right|.$$

Remarks on proof: This formula is proved in several stages. First of all, assume that the polygon is convex, and contains the origin. Then, it can be thought of as the union of triangles with corners $(0, 0)$, (x_i, y_i) and (x_{i+1}, y_{i+1}). The area formula can be proved by adding up the areas of all these triangles. Next, one can prove that shifting the origin to any other $\mathbf{x}$ inside the polygon has no effect on the formula. Finally, one can show that non–convexity is not a problem either. □



Volume of a parallelepiped

In 3–dimensions, the determinant also has a geometric meaning. A *parallelepiped* is a 3–dimensional "box", which rather than being rectangular, is a kind of pushed over prism with 3 pairs of parallel faces. (Each face is a section of a 2–dimensional plane in $\mathbb{R}^3$, and we will soon see that these can be specified by a pair of directions.) All of the edges of the parallelepiped are generated by 3 vectors, $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$. The corners of the parallelepiped are the points represented by $\mathbf{0}, \mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{w}, \mathbf{v}+\mathbf{w}, \mathbf{u}+\mathbf{v}+\mathbf{w}$. A similar argument to the proof of Theorem 3.7 gives:

Theorem 3.10 *Consider the parallelogram generated by the (row) vectors* $\mathbf{u}$, $\mathbf{v}$ *and* $\mathbf{w}$, *and let* A *be the matrix with rows* $\mathbf{u}, \mathbf{v}, \mathbf{w}$. *Then the volume of the parallelepiped is* $|\det(A)|$.

Trigonometry review

You will need a mastery of the basic facts of trigonometry. These include the definitions of the functions $\sin$ and $\cos$, their values for a few common angles, and some idea of how to work out other angles. The main point to get to grips with is that angles are measured as a "proportion of a circle". Thus, if working in degrees, 360° is a whole circle, so 90° is $\frac{90}{360} = \frac{1}{4}$ of a circle. The trig functions are a convenient way of writing down the coordinates of the points of a circle. If you think of the circle[1] as being drawn in the xy–plane, with radius 1 (centred at $(0,0)$) then we associate 0° (or 0rad) with the point $(1,0)$ (on the x–axis). Angles are then measured anti-clockwise from the positive x–axis. Thus, a 90° angle is $\frac{1}{4}$ of the way around the circle, so lies on the y–axis; it corresponds to the point $(0,1)$ in the xy–plane. Angles which differ by a multiple of 360° are considered to be the same (an exact number of additional "winds" around the circle[2]).

The trigonometric functions are defined by saying that the (x,y) coordinates of the angle θ are $(\cos\theta, \sin\theta)$.

$\cos\theta$—x coordinate $\qquad\qquad$ $\sin\theta$—y–coordinate

Try drawing a picture of a circle with an embedded triangle (one vertex at $(0,0)$, one at the point on the circle with angle θ, and one side along the x–axis) to convince yourself that this definition is the same as the "SOHCAHTOA" taught in school ("SOH" stands for $\sin = \frac{\text{opposite}}{\text{hypotenuse}}$, "CAH" stands for $\cos = \frac{\text{adjacent}}{\text{hypotenuse}}$).

Radians

You may be used to working with *degrees*, wherein 360° is a "whole circle". Most mathematicians prefer to use *radians* wherein $2\,\pi$rad comprise the whole circle. The reason for this is simple: if you measure the circumference of a circle of radius 1, it is precisely $2\,\pi$; so an angle of θrad means that you have travelled a distance θ along the circumference of the circle while subtending[3] the angle. Thought of this way, angle is a dimensionless quantity, expressing the ratio of arclength to radius.

For quoting an angle, it doesn't much matter whether you use degrees or radians, so long as it is clear which you are using. Angles quoted without a clear statement of angular measure will be assumed to be in radians, since this is the natural, dimensionless angular measure. Another good reason for using radians is that the basic trigonometric functions have very nice expressions in terms of *power series* when written in radians, for example,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots .$$

(But this formula is FALSE if x is expressed in degrees!) In summary, you can use degrees to express angles, but radians are better! In other papers (especially calculus), you must use radians.

[1] In calculus, this is the curve of points satisfying the *implicit* formula $x^2 + y^2 = 1$.

[2] Or think of the fact that a clock depicting 2o'clock looks exactly the same today as it does it 2o'clock tomorrow, despite the fact that the hands have undergone two complete rotations around the clock-face in between.

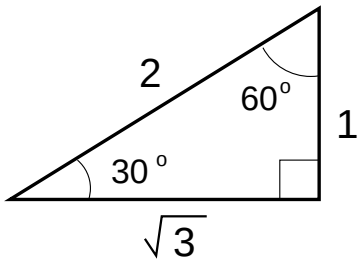
[3] To "subtend" is what angles do!

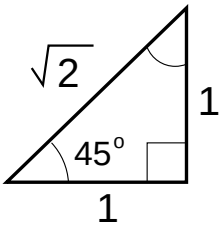
Key ingredients in working out sine and cosine

- 2π radians $\equiv 360^\circ$ so

$$\frac{\pi}{6}\text{rad} \equiv 30^\circ, \frac{\pi}{4}\text{rad} \equiv 45^\circ, \frac{\pi}{3}\text{rad} \equiv 60^\circ, \frac{\pi}{2}\text{rad} \equiv 90^\circ, \ldots$$

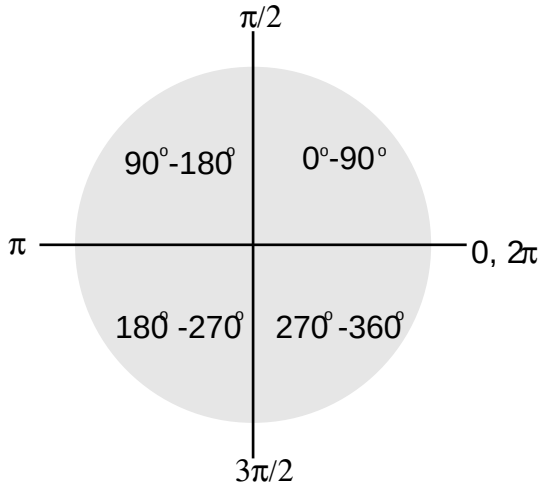
- The following basic triangles can be used to remember some standard values of sine and cosine. (Note that the sidelengths of the triangles satisfy Pythagoras' Theorem.)





so
$$\begin{aligned} \sin(30) = \cos(60) &= \tfrac{1}{2}, \\ \sin(60) = \cos(30) &= \tfrac{\sqrt{3}}{2}, \\ \sin(45) = \cos(45) &= \tfrac{1}{\sqrt{2}}. \end{aligned}$$

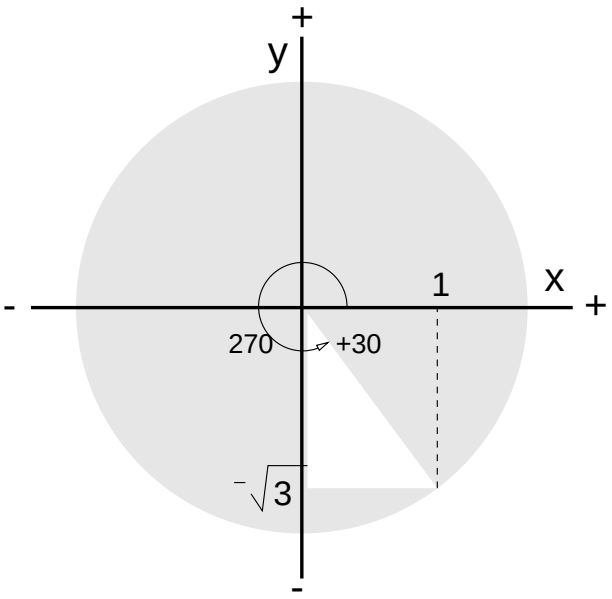
- Quadrants of circle; angles anti-clockwise from +ve x–axis



- $\cos\theta$—x coordinate, $\sin\theta$—y–coordinate

Example 1. Calculate sine and cosine of $\frac{10\,\pi}{6}$.

- First, $\frac{10\,\pi}{6}$rad is $\frac{360}{2\,\pi}\,\frac{10\,\pi}{6} = 300^\circ$
- Notice that $300 = 270 + 30$, so the angle corresponds to the final quadrant, being 30° anti-clockwise from the negative y–axis
- Now take the appropriate triangle (in this case the 1–$\sqrt{3}$–2 triangle) and arrange it with the 30° angle at the origin, and the adjacent edge along the negative y–axis.



- The coordinates of the opposite corner allow us to read off the $\cos$ and $\sin$ values. From the diagram,

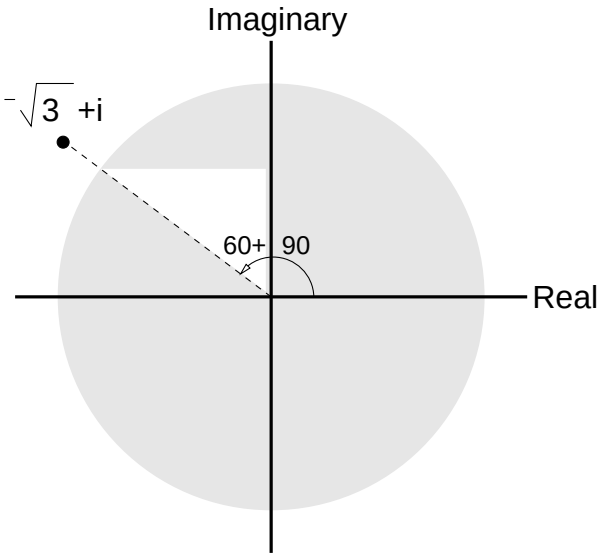
$$\sin(300) = \frac{-\sqrt{3}}{2} \text{ and } \cos(300) = \frac{1}{2}$$

[since the triangle has "adjacent" side of $-\sqrt{3}$, "opposite" side of length 1, and "hypotenuse" (or radius) 2].

□

Example 2. (Requires some knowledge of complex numbers) Figure out $\arg(-\sqrt{3}+i)$.

- plot the point $-\sqrt{3}+i$ in the complex plane
- note that it lies in the **second** quadrant
- inscribe the appropriate triangle, with its adjacent side along the positive y–axis [since the +ve y–axis is the boundary of the first and second quadrants]



- the triangle which fits is the one with a 60° angle, so the total angle from the x-axis is

$$90^\circ + 60^\circ = 150^\circ = \frac{5\,\pi}{6} = \arg(-\sqrt{3} + i)$$

□

Some useful tables

Some important values of $\sin$ and $\cos$ are summarized in the following tables. The right-hand table gives some "translation rules", wherein if an angle $\phi = \theta + x$ for x a multiple of $\frac{\pi}{2}$, then the values of $\sin\phi$ and $\cos\phi$ can be recovered from the values of $\sin\theta$ and $\cos\theta$ (these may be listed in the left-hand table).

θ	$\sin\theta$	$\cos\theta$
0	0	1
$\frac{\pi}{6}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$
$\frac{\pi}{4}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$
$\frac{\pi}{3}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$
$\frac{\pi}{2}$	1	0

ϕ	$\sin\phi$	$\cos\phi$
$-\theta$	$-\sin\theta$	$\cos\theta$
$\theta + \frac{\pi}{2}$	$\cos\theta$	$-\sin\theta$
$\theta + \pi$	$-\sin\theta$	$-\cos\theta$
$\theta - \pi$	$-\sin\theta$	$-\cos\theta$
$\theta + 2\,\pi$	$\sin\theta$	$\cos\theta$

Table 3.1: Common values of trigonometric functions (left) and addition rules (right).

Think about the fact that the angle $-\frac{\pi}{2}$ is the same as the angle $\frac{3\,\pi}{2}$; why is this?

Example 3. We can use these tables to compute other commonly occurring values. For example, to calculate $\sin(\frac{2\,\pi}{3})$, note that $\frac{2\,\pi}{3} = \frac{\pi}{6} + \frac{\pi}{2}$. Using $\theta = \frac{\pi}{6}$ (and $\phi = \frac{2\,\pi}{3}$) in the right-hand table we see

$$\sin\frac{2\,\pi}{3} = \sin\left(\frac{\pi}{6} + \frac{\pi}{2}\right) = \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2},$$

(where the value of $\cos\frac{\pi}{6}$ is obtained from the second row the left-hand table). □

Remark: The values in the right-hand table are recovered from the *trigonometric addition formulae*:

$$\sin(\theta + \alpha) = \sin\theta\,\cos\alpha + \sin\alpha\,\cos\theta,$$

$$\cos(\theta + \alpha) = \cos\theta\,\cos\alpha - \sin\alpha\,\sin\theta.$$

The other important basic identity is: $(\sin\theta)^2 + (\cos\theta)^2 = 1$.

Exercise III.1 Can you see how to derive the values of $\sin -\theta$ and $\cos -\theta$ from these identities? [Hint: $\sin 0 = 0$, $\cos 0 = 1$.]

IV ∘ Vectors and geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$

(4.1) Basic vector geometry

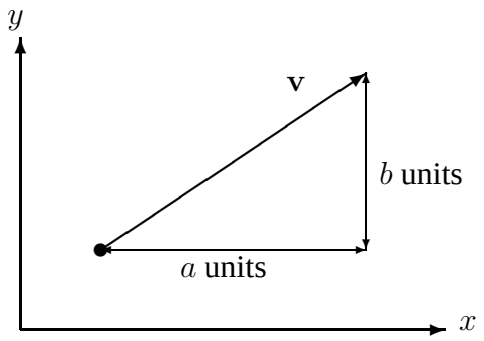
We have seen vectors crop up in several places so far: as a useful notation for describing points and lines in the plane; for writing down solutions to linear equations; and as special kinds of matrices. Now, we would like to think of vectors as "generalized numbers" which can describe things happening in multi-dimensional space (this is important in geometry, physics, economics, biology, ...). We will start our geometric study with vectors in the plane.

Geometry of vectors in $\mathbb{R}^2$

We will think of a vector in 2–space is a 2×1 matrix, and will often write it as an ordered pair of real numbers:

$$\mathbf{v} = (a, b) \equiv \begin{pmatrix} a \\ b \end{pmatrix}.$$

To draw the vector (a, b), we can choose Cartesian coordinates, and draw it as a "directed arrow": we put its "tail" at any point we like, and its "head" a units to the right (if a is positive, left if it's negative) and b units above (if b is positive, below if it's negative) its tail.



A vector $\mathbf{v} = (a, b)$ with its tail at the origin[1] is a *position vector*, because it specifies a particular point: the point with coordinates a and b. We call a the x-coordinate and b the y-coordinate[2].

Algebra of vectors

Just like matrices, two vectors are *equal* if their coordinates agree. In $\mathbb{R}^2$:

$$(a, b) = (c, d) \text{ if and only if } a = c \text{ and } b = d.$$

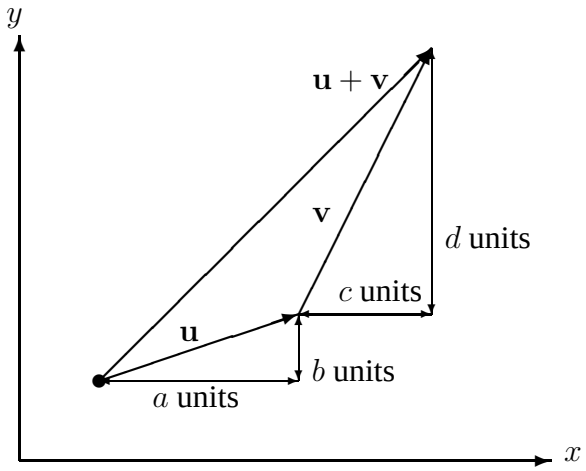
[1] The origin is where the x and y axes meet.

[2] Formally, the *vector* $\mathbf{v}$ is not to be confused with the point with coordinates (a, b). The reason for making the distinction is that we want to be able to do algebra with vectors (add them, multiply by scalars, multiply them by matrices, and so on), but "algebra with points" doesn't make much sense!

Recall that vectors can be *scaled* by constants α, and added together:

$$\alpha\,(a,b) = (\alpha\,a, \alpha\,b) \text{ and } (a,b)+(c,d) = (a+c, b+d).$$

These operations satisfy all the normal rules of algebra (given in Section 2.1), and have a sensible geometric interpretation. In particular, if $\mathbf{u} = (a,b)$ and $\mathbf{v} = (c,d)$ then $\mathbf{u}+\mathbf{v}$ is the vector obtained by putting the tail of $\mathbf{v}$ at the head of $\mathbf{u}$:



Geometrically, $\alpha\,\mathbf{v}$ has a similar "direction" to $\mathbf{v}$, but the "length" is scaled by α (if $\alpha < 0$ then the direction is opposite to the direction of $\mathbf{v}$).

Length and direction

Definition. The **length** of $\mathbf{v} = (a,b)$ is

$$\|\mathbf{v}\| = \sqrt{a^2+b^2}.$$

Let $\theta \in [0, 2\pi)$ be such that $\frac{a}{\sqrt{a^2+b^2}} = \cos\theta$ and $\frac{b}{\sqrt{a^2+b^2}} = \sin\theta$. Then the **direction** of $\mathbf{v}$ is the vector

$$(\cos\theta, \sin\theta)$$

(θ is the angle measured in radians anti-clockwise from the x–axis). Then, $\mathbf{v}$ can be written as: $\mathbf{v} = \|\mathbf{v}\|\,(\cos\theta, \sin\theta)$. □

Two vectors are called **parallel** if one is a scalar multiple of the other. If two vectors are parallel they point in equal or opposite directions.

Example. $\mathbf{u} = (1,2)$ and $\mathbf{v} = (-2,-4)$ are parallel since $\mathbf{v} = -2\mathbf{u}$. Their direction vectors are

$$\frac{1}{\sqrt{1^2+2^2}}(1,2) = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right) \text{ and } \frac{1}{\sqrt{(-2)^2+(-4)^2}}(-2,-4) = \left(\frac{-1}{\sqrt{5}}, \frac{-2}{\sqrt{5}}\right)$$

respectively, so their directions are opposite. □

Note: if $\mathbf{u}, \mathbf{v}$ are non-zero, and $\mathbf{u} = \alpha\mathbf{v}$, we must have $\alpha \neq 0$, so also $\mathbf{v} = \beta\mathbf{u}$, where $\beta = \frac{1}{\alpha}$. Thus, it doesn't matter which way round we do things. □

Example. Are $\mathbf{u} = (1,2)$ and $\mathbf{v} = (2,5)$ parallel? *Answer:* No, because if $\mathbf{v} = \alpha\mathbf{u}$ for some real α, then we must have $(2,5) = \alpha(1,2) = (\alpha, 2\alpha)$, so

$$2 = \alpha,\ \ 5 = 2\alpha$$

must hold simultaneously! □

Example. For what value of the unknown a is $(-2, a)$ parallel to $(3, 6)$? *Answer:* If $(3, 6) = \alpha(-2, a)$, then

$$3 = -2\alpha, \;\; 6 = a\,\alpha.$$

So $\alpha = -\frac{3}{2}$ from the first equality, and from the second,

$$a = \frac{6}{\alpha} = 6\left(-\frac{2}{3}\right) = -4.$$

So $(-2, -4)$ is parallel to $(3, 6)$. □

Generalization of vector properties beyond $\mathbb{R}^2$

Just like in $\mathbb{R}^2$, higher dimensional vectors have a meaning as objects on their own. We associate vectors in $\mathbb{R}^n$ with $n \times 1$ matrices, work with them algebraically, and think about them as geometric objects.

In 3-space $\mathbb{R}^3$, we have a third z–coordinate as well (often drawn coming out of the page). Then, a vector (a, b, c) represents an arrow whose head has moved from its tail: a units in the x–direction, b units in the y–direction and c units in the z–direction. The idea generalizes formally to any number of dimensions, giving the concept of vectors $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ (n-space)—although pictures become problematic!

Vector arithmetic is simply the arithmetic associated with matrices (two vectors are equal if their components are equal, they can be added component by component, and so on), but we have the supplementary concepts of length and direction.

Length and direction in $\mathbb{R}^n$

The **length** of a vector $\mathbf{v} = (x_1, x_2, \ldots, x_n)$ is

$$\|\mathbf{v}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \;\; \left(= \sqrt{\mathbf{v}^T\mathbf{v}}\right).$$

If $\mathbf{u}$ and $\mathbf{v}$ are position vectors of points, then the distance between the points they represent is $\|\mathbf{u} - \mathbf{v}\|$.

Example 1. Find the distance between points in $\mathbb{R}^3$ with coordinates $(1, -3, 2)$ and $(-1, 0, -4)$.
Answer: Working with the corresponding position vectors, the distance is

$$\|(1, -3, 2) - (-1, 0, -4)\| = \|(2, -3, 6)\| = \sqrt{4 + 9 + 36} = \sqrt{49} = 7.$$

□

A **unit vector** is a vector of length 1. For any non-zero $\mathbf{v} \in \mathbb{R}^n$, $\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$ is the unit vector in the direction of $\mathbf{v}$, and we can write

$$\mathbf{v} = \|\mathbf{v}\|\,\hat{\mathbf{v}}.$$

Three unit vectors in $\mathbb{R}^3$ are especially important, namely

$$\mathbf{i} = (1, 0, 0),\; \mathbf{j} = (0, 1, 0),\; \mathbf{k} = (0, 0, 1).$$

Note that $(a, b, c) = a\,\mathbf{i} + b\,\mathbf{j} + c\,\mathbf{k}$ for any a, b, c. The set $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ is often called the *standard basis* of $\mathbb{R}^3$: it has the property that every vector can be uniquely expressed as a "linear combination" of its elements (as just shown). (There is something similar for each $\mathbb{R}^n$.)

(4.2) Vector products

We will now study two "products" on vectors which are fundamentally important for the geometry of $\mathbb{R}^n$. The basic idea is that since $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are $n \times 1$ matrices, we can't simply multiply them together with matrix multiplication; something else is required.

Scalar product

The **scalar product** (or **dot product** or **inner product**) of two vectors $\mathbf{u},\ \mathbf{v} \in \mathbb{R}^n$ is the real number (1×1 matrix):

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\, \mathbf{v}.$$

Thus, in $\mathbb{R}^2$, if $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ then

$$\mathbf{u} \cdot \mathbf{v} = \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = u_1\, v_1 + u_2\, v_2.$$

In $\mathbb{R}^3$, if $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ then:

$$\mathbf{u} \cdot \mathbf{v} = u_1\, v_1 + u_2\, v_2 + u_3\, v_3.$$

The idea can be generalized in an obvious way to any number of dimensions.

Example 1. In $\mathbb{R}^3$,

$$(2, 3, -1) \cdot (1, 4, 1) \ = \ (2)(1) + (3)(4) + (-1)(1) = 13.$$

□

Note: The scalar product of two vectors is **always a number, not another vector.** □

The scalar product has some sensible algebraic properties: For vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and scalars $\alpha \in \mathbb{R}$,

1. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
2. $\mathbf{u} \cdot \mathbf{u} = ||\mathbf{u}||^2$
3. $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$
4. $\alpha\, (\mathbf{u} \cdot \mathbf{v}) = (\alpha\, \mathbf{u}) \cdot \mathbf{v}$.

Other facts now follow easily: for example,

$$\mathbf{u} \cdot (-\mathbf{v}) \ = \ \mathbf{u} \cdot ((-1)\mathbf{v}) \ = \ (-1)\, (\mathbf{u} \cdot \mathbf{v}) \ = \ -(\mathbf{u} \cdot \mathbf{v}).$$

Angle

Definition. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors in $\mathbb{R}^n$, and let θ be such that

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\, \|\mathbf{v}\|}.$$

Then θ is the **angle between $\mathbf{u}$ and $\mathbf{v}$**. We say that the two vectors are **orthogonal** if the angle between them is a right angle; that is $\cos\theta = \cos\frac{\pi}{2} = 0$ so $\mathbf{u} \cdot \mathbf{v} = 0$. □

This is justified in $\mathbb{R}^2$ and $\mathbb{R}^3$ by the following:

Theorem 4.1 *If* $\mathbf{u}, \mathbf{v}$ *are non-zero vectors in* $\mathbb{R}^2$ *or* $\mathbb{R}^3$*, then*

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \, \cos\theta,$$

where θ *is the angle between* $\mathbf{u}$ *and* $\mathbf{v}$ *when their tails are at the same point.*

Note: $\cos\theta = \cos(2\,\pi - \theta)$, so it doesn't matter if we go clockwise or anti-clockwise when measuring this angle! □

Proof: From the cosine rule of geometry,

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\,\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta.$$

But

$$\begin{aligned}
\|\mathbf{u} - \mathbf{v}\|^2 &= (\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v}) \\
&= (\mathbf{u} - \mathbf{v}) \cdot \mathbf{u} + (\mathbf{u} - \mathbf{v}) \cdot (-\mathbf{v}) \\
&= \mathbf{u} \cdot \mathbf{u} + (-\mathbf{v}) \cdot \mathbf{u} + \mathbf{u} \cdot (-\mathbf{v}) + (-\mathbf{v}) \cdot (-\mathbf{v}) \\
&= \mathbf{u} \cdot \mathbf{u} - \mathbf{u} \cdot \mathbf{v} - \mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{v} \\
&= \|\mathbf{u}\|^2 - 2\,(\mathbf{u} \cdot \mathbf{v}) + \|\mathbf{v}\|^2.
\end{aligned}$$

Comparing with the cosine rule, we see that $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta$. □

Example 2. Find the angle between the vectors $(1, 1, 1)$ and $(1, 1, -1)$. *Answer:* If the angle is θ, then

$$\cos\theta \;=\; \frac{1 + 1 - 1}{\sqrt{3}\sqrt{3}} \;=\; \frac{1}{3}.$$

So $\theta = \arccos(\frac{1}{3}) = \cos^{-1}(\frac{1}{3}) \approx 71$ degrees. □

Example 3. In 2-space, let points A, B, C have position vectors $\mathbf{u} = (-4, 3)$, $\mathbf{v} = (1, 0)$ and $\mathbf{w} = (0, -2)$ respectively. Find the angle at vertex A. *Answer:* The angle at vertex A in triangle ABC has cosine

$$\frac{(\mathbf{v} - \mathbf{u}) \cdot (\mathbf{w} - \mathbf{u})}{\|\mathbf{v} - \mathbf{u}\| \, \|\mathbf{w} - \mathbf{u}\|} = \frac{(5, -3) \cdot (4, -5)}{\|(5, -3)\| \, \|(4, -5)\|} = \frac{35}{\sqrt{34}\sqrt{41}} \approx 0.9374,$$

so $\theta = \arccos(0.9374) = \cos^{-1}(0.9374) \approx 20.38$ degrees. □

Vector cross product

We will study a special vector "product" of vectors on $\mathbb{R}^3$ which produces another vector. It is important in 3–dimensional geometry, and is defined via determinants. This product is very different to matrix multiplication (we can't multiply two (3×1) matrices together), and makes sense only in 3–space.

Definition. Let $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$ and $\mathbf{k} = (0, 0, 1)$ be **coordinate vectors** in $\mathbb{R}^3$. For any pair $\mathbf{u}, \mathbf{v} \in \mathbb{R}_3$ define the **cross product**:

$$\begin{aligned}
\mathbf{u} \times \mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \\
&= (u_2\,v_3 - u_3\,v_2)\,\mathbf{i} - (u_1\,v_3 - u_3\,v_1)\,\mathbf{j} + (u_1\,v_2 - u_2\,v_1)\,\mathbf{k} \\
&= (u_2\,v_3 - u_3\,v_2, 0, 0) - (0, u_1\,v_3 - u_3\,v_1, 0) + (0, 0, u_1\,v_2 - u_2\,v_1) \\
&= (u_2\,v_3 - u_3\,v_2, u_3\,v_1 - u_1\,v_3, u_1\,v_2 - u_2\,v_1).
\end{aligned}$$

□

Remarks on the cross product

- From the definition, the matrix whose determinant gives $\mathbf{u} \times \mathbf{u}$ has two rows the same (the second and third rows are both the row vector $\mathbf{u}$). Consequently, the determinant is zero, so $\mathbf{u} \times \mathbf{u} = \mathbf{0}$.
- Comparing the definitions of $\mathbf{u} \times \mathbf{v}$ and $\mathbf{v} \times \mathbf{u}$, the only difference is that the second and third rows have been interchanged. Since this multiplies determinants by (-1), we have $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$.
- The definition involving determinants is really a mnemonic; it doesn't make sense to talk about a matrix where some of the entries are vectors (the **i**, **j** and **k** in the first row). Nonetheless, it is a useful trick for remembering the formula (the last line of the definition), if the symbols **i**, **j** and **k** are given a purely formal interpretation.
- The vector cross product has the unusual properties listed above ($\mathbf{u} \times \mathbf{u} = \mathbf{0}$ and $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$), and is unique to 3–dimensions. Thus, we will reserve the symbol $\times$ for the vector cross product, and not use it elsewhere from now on.

Example 4. Let $\mathbf{u} = (2, -1, 2)$ and $\mathbf{v} = (4, -1, -4)$. Calculate $\mathbf{u} \times \mathbf{v}$. *Solution:* From the formula:

$$\mathbf{u} \times \mathbf{v} = ((-1)(-4) - 2(-1), 2(4) - 2(-4), 2(-1) - (-1)(4)) = (6, 16, 2).$$

□

Scalar triple product

In $\mathbb{R}^3$, there is a special ternary operation. This is a combination of vector products called the **scalar triple product**:

Theorem 4.2 *Let* $\mathbf{u}, \mathbf{v}, \mathbf{w}$ *be vectors in* $\mathbb{R}^3$*, and let* A *be the matrix whose rows are* $\mathbf{u}$*,* $\mathbf{v}$ *and* $\mathbf{w}$ *respectively. Then*

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \det(A) = (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}.$$

Proof: The first equality follows by a cofactor expansion of

$$\det(A) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

along the first row (use the definitions of $\cdot$ and $\mathbf{v} \times \mathbf{w}$.) For the other equality, note that

$$(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w} = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}),$$

so this product is the same as the determinant of the matrix with rows $\mathbf{w}$, $\mathbf{u}$ and $\mathbf{v}$ respectively. By Theorem 3.5 this matrix has the same determinant as A, since it can be obtained from A by the row operations $R_1 \leftrightarrow R_2$ then $R_3 \leftrightarrow R_1$ (each of which changes the sign of $|A|$ once). □

We can use this result to compute volumes of parallelepipeds (Theorem 3.10), or to establish further properties of the cross product.

Theorem 4.3 *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$*. Then* $\mathbf{u} \times \mathbf{v}$ *is orthogonal to both* $\mathbf{u}$ *and* $\mathbf{v}$.

Proof: By Theorems 4.2 and 3.6 (a matrix with two rows the same has zero determinant),

$$\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = 0$$

This shows that $\mathbf{u}$ is orthogonal to $\mathbf{u} \times \mathbf{v}$. The argument for $\mathbf{v}$ is similar. □

Further properties of the cross product

Lemma. *If* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ *then* $\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - (\mathbf{u} \cdot \mathbf{v})^2$.

Proof: Let $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$. The result follows by expanding both sides. □

Theorem 4.4 *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ *be non-zero vectors. Then*
(1) $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| |\sin\theta|$ *where* θ *is the angle between* $\mathbf{u}$ *and* $\mathbf{v}$*;*
(2) $\|\mathbf{u} \times \mathbf{v}\|$ *is the area of the parallelogram defined by* $\mathbf{u}$ *and* $\mathbf{v}$*;*
(3) $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ *if and only if* $\mathbf{u}$ *and* $\mathbf{v}$ *are parallel.*

Proof: From the lemma,

$$\begin{aligned} \|\mathbf{u} \times \mathbf{v}\|^2 &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - (\mathbf{u} \cdot \mathbf{v})^2 \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - (\|\mathbf{u}\| \|\mathbf{v}\| \cos\theta)^2 \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 (1 - (\cos\theta)^2) \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 (\sin\theta)^2. \end{aligned}$$

Part (1) follows by taking square roots of both sides of the equality. For part (2), notice that the parallelogram generated by $\mathbf{u}$ and $\mathbf{v}$ is comprised of two triangles of side length $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$, separated by an angle θ (draw a diagram). If $\mathbf{u}$ is the base of the triangle, its height is $\|\mathbf{v}\| |\sin\theta|$, so the area of each triangle is $\frac{1}{2}\|\mathbf{u}\| \|\mathbf{v}\| |\sin\theta|$. Since the parallelogram has twice the area of such a triangle, the result follows from part (1). For part (3), note that

$$\mathbf{u} \times \mathbf{v} = \mathbf{0} \Leftrightarrow \|\mathbf{u} \times \mathbf{v}\| = 0 \Leftrightarrow \sin\theta = 0 \Leftrightarrow \theta = 0 \text{ or } \pi.$$

But the angle between two vectors being 0 or π means they have the same or opposite directions, so they are parallel. □

Example 5. Find the area of the parallelogram whose sides are the vectors $(1, 3, 0)$ and $(2, 1, 2)$. *Answer:* the area is

$$\|(1, 3, 0) \times (2, 1, 2)\| = \|(6, -2, -5)\| = \sqrt{36 + 4 + 25} = \sqrt{65}.$$

Note also that if the angle between them is α, then

$$\sin\alpha = \frac{\sqrt{65}}{\sqrt{10}\sqrt{9}} = \frac{1}{3}\sqrt{\frac{65}{10}} \approx 0.85.$$

□

Example 6. Find the area of the triangle with vertices at $(1, 2, 1), (3, 3, 3), (2, 1, 2)$. *Answer:* Let

$$\mathbf{u} = (3, 3, 3) - (1, 2, 1) = (2, 1, 2),$$
$$\mathbf{v} = (2, 1, 2) - (1, 2, 1) = (1, -1, 1).$$

Then we seek half the area of the parallelogram determined by these two vectors. Hence the area is

$$\begin{aligned} \frac{1}{2}\|\mathbf{u} \times \mathbf{v}\| &= \frac{1}{2}\|(2, 1, 2) \times (1, -1, 1)\| \\ &= \frac{1}{2}\|(3, 0, -3)\| \\ &= \frac{1}{2}\sqrt{9 + 9} \\ &= \frac{3}{\sqrt{2}} \\ &\approx 2.121. \end{aligned}$$

□

(4.3) Lines and planes in space

We can use vector notation to describe interesting geometric objects.

Vector form of a line

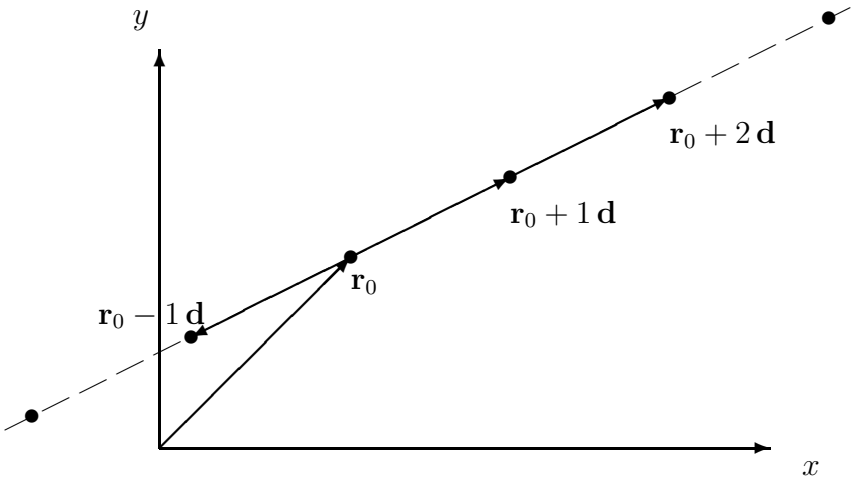
Vectors provide a very convenient way of representing lines; let us recall this. In $\mathbb{R}^2$, a line is specified by **two** vectors:

1. a position vector $\mathbf{r}_0$ for one point on the line; and
2. a direction vector $\mathbf{d}$ for the line.

Given $\mathbf{r}_0$ and $\mathbf{d}$, every point on the line has position vector

$$\mathbf{r}_0 + t\,\mathbf{d}, \quad t \in \mathbb{R}.$$

We call the real variable t a *parameter*.



Given two points on a line with position vectors $\mathbf{p}, \mathbf{q}$, we can define the line as follows: let

$$\mathbf{r}_0 = \mathbf{p} \text{ and } \mathbf{d} = \mathbf{q} - \mathbf{p}.$$

Then the line is

$$\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d} = \mathbf{p} + t\,(\mathbf{q} - \mathbf{p}) = \mathbf{p} + t\,\mathbf{q} - t\,\mathbf{p} = (1-t)\,\mathbf{p} + t\,\mathbf{q}, \qquad t \in \mathbb{R}.$$

This makes sense in any number of dimensions (including 3); we will do some examples later.

Just like in $\mathbb{R}^2$, a line has the general equation

$$\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d}, \qquad t \in \mathbb{R}$$

where $\mathbf{r}$ is a position vector for points on the line, $\mathbf{r}_0$ is the position vector of a given point on the line, and $\mathbf{d}$ is the direction vector for the line.

Example 1. Find the line through $(1, 2, 1)$ and $(2, -1, -1)$ in $\mathbb{R}^3$. Let $\mathbf{r}_0 = (1, 2, 1)$,

$$\mathbf{d} = (2, -1, -1) - (1, 2, 1) = (1, -3, -2).$$

The line has parametric vector equation $\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d}$, so

$$\mathbf{r} = (1, 2, 1) + t\,(1, -3, -2) = (1 + t, 2 - 3t, 1 - 2t),$$

$t \in \mathbb{R}$. Letting t vary gives all points on the line. In coordinate-based parametric form:

$$x = 1 + t,\ y = 2 - 3t,\ z = 1 - 2t.$$

□

Planes in $\mathbb{R}^3$

We have seen how lines are idealized geometric objects that arise naturally as solutions of systems of linear equations, and that they have a "parametric" description via vectors. *Planes*—another kind of geometric idealization—also arise in this way.

Example 2. Consider the equation in $\mathbb{R}^3$: $x + 2\,y + 3\,z = 1$. This can be thought of as a system in echelon form, so that the solution can be written down as: $y = s$, $z = t$, $x = 1 - 2\,s - 3\,t$. In vector notation, this is simply

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix}, \qquad s, t \in \mathbb{R}.$$

□

The set of solutions to this system of equations is a 2–dimensional object (it has two free parameters). It makes a great deal of sense to think of this as a flat two–dimensional sheet (of infinite extent in all directions) sitting in $\mathbb{R}^3$. It is a plane.

Parametric form of a plane in space

Definition. Let $\mathbf{r}_0 \in \mathbb{R}^3$ and let $\mathbf{d}_1, \mathbf{d}_2 \in \mathbb{R}^3$ be two non-parallel[3] vectors in $\mathbb{R}^3$. Then the set of all position vectors

$$\mathbf{r} = \mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2, \qquad t_1, t_2 \in \mathbb{R}$$

describes the **plane through $\mathbf{r}_0$, parallel to $\mathbf{d}_1$ and $\mathbf{d}_2$.** □

Note: if $\mathbf{d}_1$ and $\mathbf{d}_2$ were parallel, then the parametric form above is simply describing a line[4]. □

So, in Example 2 above, the set of solutions to the single equation given is a plane. The solution has been written in terms of two parameters, and this is the **vector parametric form of the plane**. Planes can be given several different mathematical descriptions. The vector parametric form is one of these. It turns out that a single equation (like the one in Example 2) is also a legitimate description.

Plane through three points

Just as two points can be used to define a line, three points in $\mathbb{R}^3$ are sufficient to describe a plane (provided they do not all lie on a common line).

Example 3. We will find an equation for the plane containing the points with position vectors

$$\mathbf{u} = (0, -4, 1), \mathbf{v} = (3, 0, 2), \mathbf{w} = (2, -1, -3).$$

[3] That is, there is no $\alpha \in \mathbb{R}$ such that either $\mathbf{d}_1 = \alpha\,\mathbf{d}_2$ or $\mathbf{d}_2 = \alpha\,\mathbf{d}_1$.

[4] If $\alpha\,\mathbf{d}_1 = \mathbf{d}_2$ then $\mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2 = \mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\alpha\,\mathbf{d}_1 = \mathbf{r}_0 + s\,\mathbf{d}_1$ where $(t_1 + \alpha\,t_2) = s \in \mathbb{R}$ is arbitrary.

One point in the plane is $\mathbf{r}_0 = \mathbf{u}$, while two non-parallel vectors lying in the plane are $\mathbf{d}_1 = \mathbf{v} - \mathbf{u}$ and $\mathbf{d}_2 = \mathbf{w} - \mathbf{u}$. So the plane is

$$\begin{aligned} \mathbf{r} &= \mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2 \\ &= \mathbf{u} + t_1\,(\mathbf{v} - \mathbf{u}) + t_2\,(\mathbf{w} - \mathbf{u}) \\ &= (0, -4, 1) + t_1\,(3, 4, 1) + t_2\,(2, 3, -4). \end{aligned}$$

As for lines, we can write down a parametric form:

$$(x, y, z) = (0 + 3\,t_1 + 2\,t_2, -4 + 4\,t_1 + 3\,t_2, 1 + t_1 - 4\,t_2),$$

so

$$x = 3t_1 + 2t_2,\ y = -4 + 4t_1 + 3t_2,\ z = 1 + t_1 - 4t_2,$$

$t_1, t_2 \in \mathbb{R}$. □

From the method of this example, we can see that in general, if $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are the position vectors of three points in a plane (which are not on a common line) then the plane has vector parametric form:

$$\mathbf{r} = (1 - t_1 - t_2)\,\mathbf{u} + t_1\,\mathbf{v} + t_2\,\mathbf{w}, \qquad t_1, t_2 \in \mathbb{R}.$$

Cartesian form of a plane

Example 0 shows that we can go from three points on a plane to the vector parametric description. It is easy to see how to go in the other direction: if $\mathbf{r} = \mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2$ then choosing any three different values of (t_1, t_2) will give three different points. For example, $(t_1, t_2) = (0, 0), (1, 0), (0, 1)$ gives $\mathbf{r}_0, \mathbf{r}_0 + \mathbf{d}_1, \mathbf{r}_0 + \mathbf{d}_2$ respectively—the position vectors of three points on the plane. It is natural to wonder whether we can move freely back and forth between the parametric form, and a single equation like in Example 2.

Example 4. Consider the parametric form obtained in the previous example. We can eliminate t_1, t_2 from the three equations in the parametric form to find a single equation for the plane.

We treat t_1, t_2 as unknowns, and solve for them in terms of the variables x, y, z. The system is:

$$\begin{aligned} 3\,t_1 + 2\,t_2 &= x \\ 4\,t_1 + 3\,t_2 &= y + 4 \\ t_1 - 4\,t_2 &= z - 1. \end{aligned}$$

Using Gaussian elimination to solve the first two equations we get

$$t_1 = 3\,x - 2\,y - 8 \text{ and } t_2 = 3\,y - 4\,x + 12.$$

Putting these into the third equation in our system, we get

$$(3\,x - 2\,y - 8) - 4\,(3\,y - 4\,x + 12) = z - 1.$$

After rearranging:

$$19\,x - 14\,y - z = 55$$

is the single Cartesian equation for the plane. □

Definition. Let $a, b, c, d \in \mathbb{R}$. Provided at least one of a, b, c is non–zero, the set of all points (x, y, z) satisfying

$$a\,x + b\,y + c\,z = d$$

is a plane. The formula is called the **Cartesian form of the plane**. □

The Cartesian form has a very nice geometric interpretation. Let $\mathbf{n} = (a, b, c)$ and $\mathbf{r} = (x, y, z)$ be the position vector of a point on the plane. Then the Cartesian form can be written as

$$\mathbf{n} \cdot \mathbf{r} = d.$$

Now suppose that $\mathbf{r}_0$ is the position vector of a point on the plane, and the vector $\mathbf{v}$ is parallel to the plane, then $\mathbf{r}_0 + \mathbf{v}$ is also a point on the plane, so we have

$$\mathbf{n} \cdot \mathbf{r}_0 = d = \mathbf{n} \cdot (\mathbf{r}_0 + \mathbf{v}) = \mathbf{n} \cdot \mathbf{r}_0 + \mathbf{n} \cdot \mathbf{v}.$$

From this we conclude that $\mathbf{n} \cdot \mathbf{v} = 0$. That is, **$\mathbf{n}$ is orthogonal to any vector which is parallel to the plane**. In fact, if $\mathbf{r}$ is the position vector of an arbitrary point on the plane, then $\mathbf{r} - \mathbf{r}_0$ is parallel to the plane, and we have:

Definition. The general Cartesian form of a plane is

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0,$$

where $\mathbf{r} = (x, y, z)$, $\mathbf{r}_0$ is the position vector of a point on the plane, and $\mathbf{n}$ is the **normal vector** for the plane. This is sometimes called the **implicit** form of the plane. □

Example 5. Consider the plane in $\mathbb{R}^3$ containing the point $(1, 2, -1)$ and perpendicular to the line with parametric form

$$x = 2t - 1,\ y = t + 2,\ z = -t,\ t \in \mathbb{R}.$$

We can find a Cartesian form of this plane as follows. First, we find the direction of the given line, via the vector form:

$$(x, y, z) \ = \ (2\,t - 1, t + 2, -t) \ = \ (-1, 2, 0) + t\,(2, 1, -1).$$

So the direction vector is $(2, 1, -1)$. Then, we want a plane perpendicular to this vector which contains $(1, 2, -1)$. Hence, the plane is:

$$\begin{aligned} 0 &= (2, 1, -1) \cdot ((x, y, z) - (1, 2, -1)) \\ &= 2\,(x - 1) + (y - 2) - (z + 1) \\ &= 2\,x + y - z - 2 - 2 - 1 = 2\,x + y - z - 5, \end{aligned}$$

that is, $2\,x + y - z = 5$ is the Cartesian form. □

The final piece in our study of the description of planes is to see how to move efficiently from a parametric form, to a Cartesian form. Let the parametric form be

$$\mathbf{r} = \mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2.$$

We need the normal vector $\mathbf{n}$ to be orthogonal to both $\mathbf{d}_1$ and $\mathbf{d}_2$. From our study of the cross product, this is easy to accomplish: simply set $\mathbf{n} = \mathbf{d}_1 \times \mathbf{d}_2$.

(4.4) Linear equations and intersections of lines and planes

We can consolidate our understanding of lines and planes by working with these objects some more.

Cartesian form of a line

Let us suppose that we have a *pair* of planes P_1 and P_2. Each of these will be described by a Cartesian equation:

$$\begin{aligned} P_1: \quad \mathbf{n}_1 \cdot \mathbf{r} &= k_1 \\ P_2: \quad \mathbf{n}_2 \cdot \mathbf{r} &= k_2 \end{aligned}$$

for normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ and constants $k_1, k_2 \in \mathbb{R}$. We would like to study the set of points common to both planes. Written out in coordinates, we have a pair of Cartesian equations in three variables. If we were to solve these equations via Gaussian elimination, we will obtain one of the following possibilities:

- there are no solutions because the system is inconsistent;
- there is one free variable, leading to a one-parameter general solution;
- there are two free variables, and hence a two parameter general solution.

The first case will occur when the planes are parallel, but do not intersect. In the second case, the solution is a line, and in the final case, the solution is a plane—the same plane as both P_1 and P_2. In summary, we have: *the intersection of two planes is usually a line, although it could happen that the planes are coincident or parallel.*

The representation of a line as the intersection of two planes is called the **Cartesian representation of the line**. It is important to understand what is going on geometrically here, but it is not so important to be able to find the Cartesian representation of a given line—it is non-unique. Of course, finding a parametric representation, given a Cartesian representation is easy: you simply solve the system of equations.

Some problems involving lines and planes

Example 1. Let us use vectors to test whether the points $(-1, 2, 0)$, $(3, 3, 2)$ and $(11, 5, 6)$ lie on a single line. First of all, we'll write down the equation of the line through the first two points:

$$(x, y, z) = (1 - t)\,(-1, 2, 0) + t\,(3, 3, 2) = (4\,t - 1, 2 + t, 2\,t).$$

If the points are co-linear, then there will be a value of t which allows $(11, 5, 6)$ to be written in this way. The equations (for x, y and z respectively) would then be:

$$11 = 4\,t - 1, \qquad 5 = 2 + t, \qquad 6 = 2\,t.$$

These are all solved by $t = 3$, so the points are indeed co-linear. We can even write $(11, 5, 6) = -2\,(-1, 2, 0) + 3\,(3, 3, 2)$. □

Example 2. Do the points $(1, 3, 2)$, $(-2, 3, 5)$, $(-1, 0, 1)$ and $(8, 3, -5)$ lie on a common plane? The first thing to do is to find the plane containing the first three points, and then test whether the fourth point is on it. We'll take $\mathbf{r}_0 = (1, 3, 2)$ and let the two direction vectors be $(-2, 3, 5) - (1, 3, 2) = (-3, 0, 3)$ and $(-1, 0, 1) - (1, 3, 2) = (-2, -3, -1)$. A suitable normal vector is thus

$$\mathbf{n} = (-3, 0, 3) \times (-2, -3, -1) = (9, -9, 9).$$

The plane containing the first three points thus has equation:

$$(9,-9,9)\cdot(x,y,z)=\mathbf{n}\cdot\mathbf{r}=\mathbf{n}\cdot\mathbf{r}_0=(9,-9,9)\cdot(1,3,2)=0.$$

Dividing by 9, this becomes $x-y+z=0$. It is easy to see that first three points satisfy this equation (as we would expect). But also:

$$8-3+(-5)=0,$$

so the fourth point is on the plane. That is, the four points are coplanar. □

Remarks

- The points represented by $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are co-linear if and only if $\mathbf{v}-\mathbf{u}$ and $\mathbf{w}-\mathbf{u}$ are parallel. But, by Theorem 4.4 (3) this happens if and only $(\mathbf{v}-\mathbf{u})\times(\mathbf{w}-\mathbf{u})=\mathbf{0}$. Then, we'll also have $\mathbf{u}\cdot((\mathbf{v}-\mathbf{u})\times(\mathbf{w}-\mathbf{u}))=\mathbf{u}\cdot\mathbf{0}=0$, so by regarding $\mathbf{u},\mathbf{v},\mathbf{w}$ as row vectors, by Theorem 4.2

$$0=\mathbf{u}\cdot((\mathbf{v}-\mathbf{u})\times(\mathbf{w}-\mathbf{u}))=\begin{vmatrix}\mathbf{u}\\ \mathbf{v}-\mathbf{u}\\ \mathbf{w}-\mathbf{u}\end{vmatrix}=\begin{vmatrix}\mathbf{u}\\ \mathbf{v}\\ \mathbf{w}\end{vmatrix}$$

  (the last equality is by Theorem 3.5, since the second matrix is obtained from the first by elementary row operations). Therefore, three points in $\mathbb{R}^3$ are co-linear if and only if their scalar triple product is 0.

- Two planes intersect in a common line if and only if their normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ are non-parallel. In this case, the direction vector for the line is $\mathbf{n}_1\times\mathbf{n}_2$.

- Two (distinct) lines with parallel direction vectors always lie in a common plane.

- If two lines $\mathbf{r}_i+t_i\,\mathbf{d}_i$ $(i=1,2)$ lie in a common plane then the normal vector of the plane is $\mathbf{n}=\mathbf{d}_1\times\mathbf{d}_2$ and we would need $\mathbf{r}_1\cdot\mathbf{n}=\mathbf{r}_2\cdot\mathbf{n}$. That is, $(\mathbf{r}_1-\mathbf{r}_2)\cdot(\mathbf{d}_1\times\mathbf{d}_2)=0$.

Intersections of lines and planes

Suppose we are given a line and plane in three–dimensional space. We might expect these to have some common points. There are several ways we could go about finding these points, depending on the way the objects are presented to us:

- If the line is given as a pair of Cartesian equations, and the plane is given as a single Cartesian equation, then any point in their intersection must satisfy all three equations. Consequently, we could: *solve the system of three equations for* x,y,z.

- If the line and plane are both given in parametric form, then we are looking for a point which simultaneously satisfies

$$\mathbf{r}=\mathbf{r}_0+t\,\mathbf{d}\text{ and }\mathbf{r}=\mathbf{r}'_0+t_1\,\mathbf{d}_1+t_2\,\mathbf{d}_2$$

  for the line and plane respectively (the position vectors are $\mathbf{r}_0,\mathbf{r}'_0$, $\mathbf{d},\mathbf{d}_1,\mathbf{d}_2$ are direction vectors, $t,t_1,t_2\in\mathbb{R}$ are real parameters). By equating the two equations, and reorganizing, we then need to solve

$$t_1\,\mathbf{d}_1+t_2\,\mathbf{d}_2-t\,\mathbf{d}=\mathbf{r}_0-\mathbf{r}'_0.$$

This can be written as a matrix equation

$$\left(\begin{array}{c|c|c} \mathbf{d}_1 & \mathbf{d}_2 & -\mathbf{d} \end{array}\right) \begin{pmatrix} t_1 \\ t_2 \\ t \end{pmatrix} = \mathbf{r}_0 - \mathbf{r}'_0$$

for the variables t, t_1, t_2. The intersection can be found either by substituting the value of t into the equation for the line, or by substituting t_1, t_2 into the formula for the plane; both should give the same answer!

- If the line is given as a pair of equations, and the plane is given in parametric form, then substituting the plane equation into the pair of equations for the line will give two equations for t_1 and t_2. These can be solved, and the values substituted back into the parametric form for the plane to give the required points.
- If the line is given in parametric form, and the plane is given in Cartesian form, then the parametric equation for the line should be substituted into the Cartesian equation for the plane. This gives one equation for the one parameter t describing the line. This can be solved easily, and the resulting value of t substituted into the parametric equation for the line.

It turns out that the last of these is the easiest to use.

Example 3. Find the point of intersection of the line $(x, y, z) = (1, 2, 1) + t\,(2, -3, 1)$ and the plane $2\,x + y - 3\,z = 5$. To solve this, substitute

$$x = 1 + 2\,t, y = 2 - 3\,t, z = 1 + t \text{ into } 2\,x + y - 3\,z = 5.$$

Thus,

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

This is solved by $t = -2$, so the point of intersection is $(x, y, z) = (1, 2, 1) + (-2)\,(2, -3, 1) = (-3, 8, -1)$. Substituting these numbers back into the equation for the plane reveals that $2\,x + y - 3\,z = 2\,(-3) + 8 - 3\,(-1) = 5$, providing a useful check for our calculations. □

The method of this example can be written down as an algorithm.

Finding the intersection of a line and plane

1. Write down the line in the form $\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d}$ (if the line is given as a pair of Cartesian equations, solve the equations);
2. write down the plane in Cartesian form: $\mathbf{r} \cdot \mathbf{n} = d$ (if the plane is given as $\mathbf{r} = \mathbf{r}'_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2$ then put $\mathbf{n} = \mathbf{d}_1 \times \mathbf{d}_2$ and $d = \mathbf{r}'_0 \cdot \mathbf{n}$);
3. substitute the equation for the line into the equation for the plane to get a single equation for t;
4. solve for t;
5. the intersection point(s) are obtained by substituting the solution(s) for t into the equation for the line;

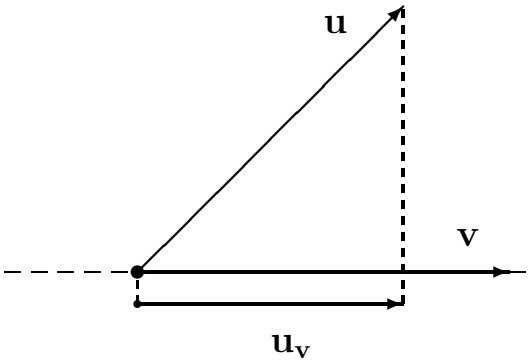
   if there are **no solutions** for t, then the plane and the line do not intersect;
6. check the points satisfy the equation of the plane.

(4.5) Projections in $\mathbb{R}^3$

We have now seen how to describe lines and planes in space via parametric or Cartesian equations. We can use the scalar product and vector algebra to work out solutions to some geometric problems. For example: *given a point P in $\mathbb{R}^n$ and a line L, what is the distance from P to L?* To answer this kind of question we will need the concept of a **projection**. We will introduce some important concepts, and solve some geometric problems.

Projection onto a vector

We will be begin by describing the process of "projection onto a vector". Let $\mathbf{u}$ be the position vector of a given point; we would like to know $\mathbf{u_v}$—the *component* of $\mathbf{u}$ in a given direction $\mathbf{v}$.



To calculate such components, we need some additional notation. We would like $\mathbf{u_v}$ to be a vector *parallel* to $\mathbf{v}$, such that $\mathbf{u} - \mathbf{u_v}$ is *orthogonal* to $\mathbf{v}$. This gives us a pair of equations to solve:

$$\begin{array}{rcll} \mathbf{u_v} & = & \alpha\,\mathbf{v} & \text{since } \mathbf{u_v} \text{ is parallel to } \mathbf{v}, \\ 0 & = & (\mathbf{u} - \mathbf{u_v}) \cdot \mathbf{v} & \text{since } \mathbf{u} - \mathbf{u_v} \text{ is orthogonal to } \mathbf{v}. \end{array}$$

Thus

$$0 = (\mathbf{u} - \alpha\,\mathbf{v}) \cdot \mathbf{v} = \mathbf{u} \cdot \mathbf{v} - \alpha\,\mathbf{v} \cdot \mathbf{v}$$

and hence $\alpha = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}$. Thus:

Definition. The **component of $\mathbf{u}$ in the direction of $\mathbf{v}$** is

$$\mathbf{u_v} = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}\,\mathbf{v}.$$

□

Example 1. To find the component of the vector $(2, 0, 1)$ in the direction of the vector $(1, -1, 0)$, let $\mathbf{u} = (2, 0, 1)$ and $\mathbf{v} = (1, -1, 0)$. Then

$$\begin{array}{rcl} \mathbf{u_v} & = & \dfrac{(2,0,1)\cdot(1,-1,0)}{(1,-1,0)\cdot(1,-1,0)}(1,-1,0) \\ & = & \dfrac{2+0+0}{(1+1+0)}(1,-1,0) \\ & = & (1,-1,0). \end{array}$$

So the component is $\mathbf{v}$ itself in this case. We can check that the orthogonality of $\mathbf{u} - \mathbf{u_v}$ with $\mathbf{v}$:

$$(\mathbf{u} - \mathbf{u_v}) \cdot \mathbf{v} = ((2,0,1) - (1,-1,0)) \cdot (1,-1,0) = (1,1,1) \cdot (1,-1,0) = 0,$$

so the method has worked! □

Projections and the method of projection

The component of a vector in a certain direction is a special case of a general kind of operation:

Definition. Let V be a collection of vectors. Suppose that a vector $\mathbf{u}$ can be written as

$$\mathbf{u} = \mathbf{u}_{\parallel} + \mathbf{u}_{\perp}$$

where $\mathbf{u}_{\parallel} \in V$ and $\mathbf{u}_{\perp}$ is orthogonal to every $\mathbf{v}$ which is parallel to V. Then $\mathbf{u}_{\parallel}$ is the **projection** of $\mathbf{u}$ onto V, and we write $\mathbf{u}_{\parallel} = \text{proj}_V \mathbf{u}$. □

Projection principle. $\mathbf{u}_{\parallel}$ *represents the closest point to* $\mathbf{u}$ *in* V*. The distance from* $\mathbf{u}$ *to* V *is*

$$\|\mathbf{u} - \mathbf{u}_{\parallel}\| = \|\mathbf{u}_{\perp}\|.$$

Method of Projection. *To compute the distance from a point with position vector* $\mathbf{u}$ *to a collection* V *of vectors, first compute* $\text{proj}_V \mathbf{u}$*, and then*

$$\|\mathbf{u} - \text{proj}_V \mathbf{u}\| = \|\mathbf{u}_{\perp}\|$$

is the distance from $\mathbf{u}$ *to* V*.* □

Example 2. By the projection principle, the position vector $\mathbf{u}_{\mathbf{v}}$ represents the closest point to $\mathbf{u}$ on the line with direction vector $\mathbf{v}$. □

Distance from a point to a line

We can apply the definition of projections to calculate the distance from a point to an arbitrary line. Let $\mathbf{u}$ be a position vector of a point, and let the line L have parametric representation

$$\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d}.$$

Then, $\text{proj}_L \mathbf{u}$ must be on the line, and $\mathbf{u}_{\perp} = \mathbf{u} - \text{proj}_L \mathbf{u}$ must be perpendicular to any vector parallel to L—that is, any vector parallel to $\mathbf{d}$. As in projecting onto a vector, we have

$$\begin{array}{rcll} \text{proj}_L \mathbf{u} & = & \mathbf{r}_0 + \alpha\,\mathbf{d} & \text{since proj}_L\mathbf{u} \text{ is on } L, \\ 0 & = & [\mathbf{u} - \text{proj}_L \mathbf{u}] \cdot \mathbf{d} & \text{since } \mathbf{u}_{\perp} \text{ is orthogonal to } \mathbf{d}. \end{array}$$

So, we need to find an α such that

$$0 = (\mathbf{u} - (\mathbf{r}_0 + \alpha\,\mathbf{d})) \cdot \mathbf{d} = (\mathbf{u} - \mathbf{r}_0) \cdot \mathbf{d} - \alpha\,\mathbf{d} \cdot \mathbf{d}.$$

Thus,

$$\alpha = \frac{(\mathbf{u} - \mathbf{r}_0) \cdot \mathbf{d}}{\mathbf{d} \cdot \mathbf{d}}$$

and the method of projection gives an algorithm:

Finding the distance from a point to a line

1. Write the line as $\mathbf{r} = \mathbf{r}_0 + t\,\mathbf{d}$, let the point be represented by $\mathbf{u}$;
2. find
$$\text{proj}_L \mathbf{u} = \mathbf{r}_0 + \alpha\,\mathbf{d} \text{ where } [\mathbf{u} - \text{proj}_L \mathbf{u}] \cdot \mathbf{d} = 0;$$
3. calculate $\|\mathbf{u}_{\perp}\| = \|\mathbf{u} - \text{proj}_L \mathbf{u}\|$.

Example 3. Use the method of projection to find the distance between the point $P = (2, 2, 1)$ and the line $(x, y, z) = (-1 + 2\,t, 1 + 3\,t, -3 + 2\,t), t \in \mathbb{R}$.
Solution: The line can be written as $(-1, 1, -3) + t\,(2, 3, 2)$, so the important vectors are $\mathbf{u} = (2, 2, 1), \mathbf{r}_0 = (-1, 1, -3)$ and $\mathbf{d} = (2, 3, 2)$. We need to solve

$$\begin{array}{rcll} \text{proj}_L\mathbf{u} & = & \mathbf{r}_0 + \alpha\,\mathbf{d} & \text{since proj}_L\mathbf{u} \text{ is on } L, \\ 0 & = & [\mathbf{u} - (\mathbf{r}_0 + \alpha\,\mathbf{d})] \cdot \mathbf{d} & \text{since } \mathbf{u}_\perp \text{ is orthogonal to } \mathbf{d}. \end{array}$$

Since $\mathbf{u} - \mathbf{r}_0 = (3, 1, 4)$, we need:

$$[(3, 1, 4) - \alpha\,(2, 3, 2)] \cdot (2, 3, 2) = 0$$

or

$$(3, 1, 4) \cdot (2, 3, 2) = \alpha(2, 3, 2) \cdot (2, 3, 2).$$

This says $17 = \alpha\,17$,so $\alpha = 1$ and

$$\text{proj}_L\mathbf{u} = \mathbf{r}_0 + \alpha\,\mathbf{d} = (-1, 1, -3) + 1\,(2, 3, 2) = (1, 4, -1).$$

Finally,

$$\mathbf{u}_\perp = \mathbf{u} - \text{proj}_L\mathbf{u} = (2, 2, 1) - (1, 4, -1) = (1, -2, 2)$$

so the distance from P to L is $\|\mathbf{u}_\perp\| = \|(1, -2, 2)\| = \sqrt{1 + (-2)^2 = (2)^2} = 3$. □

Distance from a point to a plane

By the projection principle, the distance from a point to a plane can be obtained via projections.

Example 4. Let P be the plane $\mathbf{r} = (1, 2, 1) + t_1\,(1, 3, 4) + t_2\,(-1, 0, 2)$. Find the projection of $(-1, 8, -1)$ onto P.
Solution: We need to solve the projection equations. Here, $\mathbf{r}_0 = (1, 2, 1)$, and the two direction vectors are $\mathbf{d}_1 = (1, 3, 4)$ and $\mathbf{d}_2 = (-1, 0, 2)$. We need $\mathbf{u}_\perp = \mathbf{u} - \text{proj}_P\mathbf{u}$ to be orthogonal to the directions $\mathbf{d}_1, \mathbf{d}_2$, so we solve:

$$\begin{array}{rcll} \text{proj}_P\mathbf{u} & = & \mathbf{r}_0 + \alpha\,\mathbf{d}_1 + \beta\,\mathbf{d}_2 & \text{since proj}_P\mathbf{u} \text{ is on } P, \\ 0 & = & [\mathbf{u} - (\mathbf{r}_0 + \alpha\,\mathbf{d}_1 + \beta\,\mathbf{d}_2)] \cdot \mathbf{d}_1 & \text{since } \mathbf{u}_\perp \text{ is orthogonal to } \mathbf{d}_1 \\ 0 & = & [\mathbf{u} - (\mathbf{r}_0 + \alpha\,\mathbf{d}_1 + \beta\,\mathbf{d}_2)] \cdot \mathbf{d}_2 & \text{since } \mathbf{u}_\perp \text{ is orthogonal to } \mathbf{d}_2. \end{array}$$

After substituting in the vectors:

$$\begin{array}{rcl} 0 & = & (-2, 6, -2) \cdot (1, 3, 4) - \alpha\,(1, 3, 4) \cdot (1, 3, 4) - \beta\,(-1, 0, 2) \cdot (1, 3, 4) \\ 0 & = & (-2, 6, -2) \cdot (-1, 0, 2) - \alpha\,(1, 3, 4) \cdot (-1, 0, 2) - \beta\,(-1, 0, 2) \cdot (-1, 0, 2) \end{array}$$

or

$$\begin{array}{rcrcr} 26\,\alpha & + & 7\,\beta & = & 8 \\ 7\,\alpha & + & 5\,\beta & = & -2. \end{array}$$

This system has solution: $\alpha = \frac{2}{3}, \beta = \frac{-4}{3}$ so

$$\text{proj}_P\mathbf{u} = (1, 2, 1) + \frac{2}{3}\,(1, 3, 4) - \frac{4}{3}\,(-1, 0, 2) = (3, 4, 1).$$

□

We apply the projection principle to find the distance from a point to a plane: the least distance from a point $\mathbf{u}$ to a plane P is $\|\mathbf{u} - \text{proj}_P\mathbf{u}\|$:

Finding the distance from a point to a plane

1. Express the plane in parametric form: $\mathbf{r}_0 + t_1\,\mathbf{d}_1 + t_2\,\mathbf{d}_2$
   (if necessary, solve a Cartesian equation to get this form);
2. find $\text{proj}_P\mathbf{u} = \mathbf{r}_0 + \alpha\,\mathbf{d}_1 + \beta\,\mathbf{d}_2$, so that $[\mathbf{u} - \text{proj}_P\mathbf{u}] \cdot \mathbf{d}_i = 0$ $(i = 1, 2)$;
3. calculate $\|\mathbf{u} - \text{proj}_P\mathbf{u}\|$.

Example 5. Find the shortest distance between the point with position vector $\mathbf{u} = (-1, 1, 3)$ and the plane P with Cartesian equation $2\,x - y - 3\,z = 2$.
Solution: The Cartesian equation can be solved to give the plane as

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + t_1 \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}.$$

We therefore solve the pair of equations:

$$\begin{aligned} 0 &= [(-1,1,3) - (1,0,0) - \alpha\,(1,2,0) - \beta\,(3,0,2)] \cdot (1,2,0) \\ 0 &= [(-1,1,3) - (1,0,0) - \alpha\,(1,2,0) - \beta\,(3,0,2)] \cdot (3,0,2); \end{aligned}$$

that is

$$\begin{array}{rcrcl} 5\,\alpha & + & 3\,\beta & = & 0 \\ 3\,\alpha & + & 11\,\beta & = & 0. \end{array}$$

This system has solution $\alpha = \beta = 0$ (which was lucky, and not the usual situation), so

$$\text{proj}_P\mathbf{u} = (1,0,0) + 0\,(1,2,0) - 0\,(3,0,2) = (1,0,0).$$

The shortest distance is now

$$\|\mathbf{u} - \text{proj}_P\mathbf{u}\| = \|(-1,1,3) - (1,0,0)\| = \|(-2,1,3)\| = \sqrt{4+1+9} = \sqrt{14}.$$

□

Remark. If the plane P is given implicitly $\mathbf{r} \cdot \mathbf{n} = d$, there is an alternative approach. The vector $\mathbf{u}_\perp$ must be orthogonal to all directions in P, so is in fact parallel to $\mathbf{n}$ (the normal to the plane). We thus solve the plane equations for $\mathbf{u}_\perp = \gamma\,\mathbf{n}$. Since the projection of $\mathbf{u}$ onto P must satisfy the plane equation, we know

$$\text{proj}_P\mathbf{u} \cdot \mathbf{n} = d \text{ and } \text{proj}_P\mathbf{u} = \mathbf{u} - \mathbf{u}_\perp = \mathbf{u} - \gamma\,\mathbf{n}.$$

Combining these equations gives

$$(\mathbf{u} - \gamma\,\mathbf{n}) \cdot \mathbf{n} = d,$$

so $\gamma = \frac{d - \mathbf{u}\cdot\mathbf{n}}{\mathbf{n}\cdot\mathbf{n}}$. This allows us to write down $\mathbf{u}_\perp$, and thus recover $\text{proj}_P\mathbf{u} = \mathbf{u} - \mathbf{u}_\perp$ and the distance from $\mathbf{u}$ to P. (In this method, we are really looking for where the line through $\mathbf{u}$ with direction $\mathbf{n}$ intersects P.) This strategy suggests an alternative algorithm for finding the distance from a point to a plane when P is given parametrically: find a Cartesian representation using $\mathbf{n} = \mathbf{d}_1 \times \mathbf{d}_2$ and proceed. □

Proof of the projection principle

Proof: Suppose that $\mathbf{v} \in V$ and write $\mathbf{v} = \mathbf{u}_\parallel + \mathbf{v}'$ where $\mathbf{v}' = \mathbf{v} - \mathbf{u}_\parallel$ is parallel to V. Then,

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}_\parallel + \mathbf{u}_\perp - (\mathbf{u}_\parallel + \mathbf{v}')\|^2 = \|\mathbf{u}_\perp - \mathbf{v}'\|^2 = (\mathbf{u}_\perp - \mathbf{v}') \cdot (\mathbf{u}_\perp - \mathbf{v}') = \mathbf{u}_\perp \cdot \mathbf{u}_\perp - 2\mathbf{v}' \cdot \mathbf{u}_\perp + \mathbf{v}' \cdot \mathbf{v}'.$$

However, since $\mathbf{v}'$ is parallel to V, $\mathbf{u}_\perp \cdot \mathbf{v}' = 0$. Moreover, $\mathbf{v}' \cdot \mathbf{v}' = \|\mathbf{v}'|^2 \geq 0$. This shows that

$$\|\mathbf{u} - \mathbf{v}\|^2 = \mathbf{u}_\perp \cdot \mathbf{u}_\perp + \mathbf{v}' \cdot \mathbf{v}' \geq \mathbf{u}_\perp \cdot \mathbf{u}_\perp = \|\mathbf{u}_\perp\|^2.$$

By taking square roots, this shows that $\mathbf{u}$ is at least a distance $\|\mathbf{u}_\perp\|$ from $\mathbf{w} \in V$, so the projection principle is true. □

V ○ Induction and recursion

So far we have worked with vectors of real numbers without much regard for the underlying structure of $\mathbb{R}$. In this part of the paper, we will take a tour into the foundations of mathematics, to explore the properties of number systems, and discuss how the numbers we work with are "built". We will begin with the most basic of systems.

(5.1) Set theory, the natural numbers $\mathbb{N}$ and mathematical induction

Numbers are a mathematical abstraction. For millennia (we don't know how many millennia), humans have been counting things. Thus, we have the notion of *natural numbers:*

$$\mathbb{N} = 1, 2, 3, 4, \cdots .$$

We can think of these numbers as being universal names for "how many" objects, but it is very difficult to be more explicit than that. *Do these numbers really exist?* Well, the answer is "yes", so long as you are prepared to accept some axioms from "Set theory".

Set theory and the history of mathematics

There are many axioms of set theory, and they are all quite plausible. Things like "there exists a set", "given any pair of sets, there is another set that contains them both", and so on. Taken together, and worked with cleverly, these axioms are enough to construct most of mathematics! Most of set theory was soundly established about 100 years ago. There are some traps however, one of the most famous is "the set of all sets which are not members of themselves". Think about this set. Does it contain itself? This is *Russell's paradox*, after Bertrand Russell (1872–1970). In the early 20th century, great mathematicians and philosophers (Cantor, Frege, Russell and others) tried to put all of mathematics on rigorous, axiomatic foundations, and Russell's paradox was a devastating blow[1]. It turned out that Russell's approach could be fixed up (with some more sophisticated notions), but *every* axiomatic system suffers from a certain weakness; this is the content of *Gödel's incompleteness theorem*, which states that any formal axiomatic system will contain valid statements which are neither provably true, nor provably false. Partly due to Gödel's result, and partly due to the whims of mathematical fashion, most contemporary mathematicians do not worry too often about these foundations. Indeed, like students, working mathematicians take much of what they learn on trust, and put off checking all the details until really needed. In fact, you may notice that the "deeper" we probe into the structure of mathematics, the more recently the foundations have been secured. For example, Gauss (1777–1855) was working masterfully with complex numbers in the late 18th century, but the foundations of the natural numbers were not really sorted out until the 20th century.

Set theory and the natural numbers

Clever use of the set theory axioms lets us construct the "empty set": $\emptyset$. Basically, you take any set you can find (remember, there is an axiom that says "there exists a set"[2]), and throw away all its members; you are

[1] At least to Russell!

[2] If this axiom is still a bit much to swallow, what do you think about Descartes' cornerstone epistemological existence theorem: "Cogito Ergo Sum"?

left with $\emptyset$. We will associate $\emptyset$ with the number 0. Next, we can define an operation on sets, the "successor" operation.

Definition. The **successor** of a set A is the set $\text{succ}(A)$ containing A, and $\emptyset$. A set B is called an **inductive set** if $\text{succ}(x) \in B$ whenever $x \in B$. □

The idea is to build 1 as the successor of 0, 2 as the successor of 1, and so on. So,

$$\begin{aligned} \text{succ}(\emptyset) &= \{\emptyset\} \\ \text{succ}(\text{succ}(\emptyset)) &= \{\emptyset, \{\emptyset\}\} \\ \text{succ}(\text{succ}(\text{succ}(\emptyset))) &= \{\emptyset, \{\emptyset, \{\emptyset\}\}\} \\ &\vdots \end{aligned}$$

specifies the numbers $1, 2, 3$. This construction allows any natural number to be constructed, by iterative application of the successor operation. Roughly speaking, $\mathbb{N}$ is defined as the smallest non-empty inductive set.

Note: Quite a lot is still swept under the carpet with this construction of $\mathbb{N}$. For example, *are there any inductive sets*? □

Basically, we observe that, $1 \in \mathbb{N}$ and $n + 1 \in \mathbb{N}$ whenever $n \in \mathbb{N}$. We can build basic arithmetic with definitions like "$A + \text{succ}(\emptyset) = \text{succ}(A)$" (which just says that $n+1$ means what we already think it means)! Further clever use of set theory axioms lets us construct multiplication on $\mathbb{N}$.

Principle of induction

With our construction of $\mathbb{N}$, we have:

Principle of Induction: Let S be a subset of the set of the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$. Suppose the following properties hold for S.

- $1 \in S$; and
- $k \in S$ implies $k + 1 \in S$.

Then $S = \mathbb{N}$. □

This principle has a simply analogy: climbing a ladder. Suppose that, starting from the ground,

- we can climb to the lowest rung; and
- if we can climb to any given rung, we can always climb to the next one.

Then we can climb to any rung we like, however high up the ladder!

Note: We defined $\mathbb{N}$ to be a non–empty inductive set. This is why the principle of induction requires us to check that $1 \in S$. □

A proof by induction

We will now see how to use induction to establish formulae which depend on n.

Example 1. We would like to show that

$$1^2 + 2^2 + \cdots + n^2 = \frac{n\,(n+1)\,(2\,n+1)}{6}.$$

For *all* n. This formula depends on an integer in that n is a "variable". □

The motivation is simple: if we let S be the set of $n \in \mathbb{N}$ for which a given statement is true, then we try to establish that S is an inductive set. That is, that $1 \in S$, and that $k+1 = \text{succ}(k) \in S$ whenever $k \in S$.

Proof of Example 1. We would like to show that

$$1^2 + 2^2 + \cdots + n^2 = \frac{n\,(n+1)\,(2\,n+1)}{6}.$$

This formula depends on an integer in that n is a "variable", and we will think of it as indexing an inductive set. Let

$$S = \left\{ n \,\middle|\, 1^2 + 2^2 + \cdots + n^2 \;=\; \frac{n(n+1)(2n+1)}{6} \text{ is true} \right\}.$$

Thus we must show:

(B) Base case: the formula is true for $n = 1$; and

(R) Recursive step: for any $k \geq 1$, the formula being true for $n = k$ implies it is true for $n = k+1$.

(Note that (B) really says $1 \in S$, and (R) says that $k \in S$ implies $k+1 \in S$.) Once (B) and (R) are shown, the principle of induction gives that $S = \mathbb{N}$, so the formula is true for all $n \in \mathbb{N}$.

When $k = 1$, we have

$$1^2 \;=\; 1 \times 2 \times 3/6,$$

which is clearly TRUE. So (B) is OK.

To work on (R), assume that for *some particular* $k \geq 1$:

$$1^2 + 2^2 + \cdots + k^2 \;=\; \frac{k(k+1)(2k+1)}{6}.$$

We must show, assuming only this, that the formula holds for $n = k+1$. Namely:

$$1^2 + 2^2 + \cdots + (k+1)^2 = \frac{(k+1)((k+1)+1)(2(k+1)+1)}{6}.$$

Let's start with the more complicated side of $P(k+1)$ and try to simplify it.

$$\begin{aligned}
\text{LHS} &= 1^2 + 2^2 + \cdots + (k+1)^2 \\
&= (1^2 + 2^2 + \cdots + k^2) + (k+1)^2 \\
&= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \quad \text{(by assumption for } n = k\text{)} \\
&= \frac{k(k+1)(2k+1) + 6(k+1)^2}{6} \\
&= \frac{(k+1)(k(2k+1) + 6(k+1))}{6} \\
&= \frac{(k+1)(2k^2 + k + 6k + 6)}{6} \\
&= \frac{(k+1)(2k^2 + 7k + 6)}{6}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\text{RHS} &= \frac{(k+1)((k+1)+1)(2(k+1)+1)}{6} \\ &= \frac{(k+1)(k+2)(2k+3)}{6} \\ &= \frac{(k+1)(2k^2+4k+3k+6)}{6} \\ &= \frac{(k+1)(2k^2+7k+6)}{6} \\ &= \text{LHS}.\end{aligned}$$

This establishes (R), the recursive step. It now follows by the Principle of Mathematical Induction that the equation is true for all $n = 1, 2, 3, \ldots$. □

Final remark: The rigorous mathematical foundations of the method of induction were only completed with the development of modern set theory in the early–20th century. Despite this, Blaise Pascal (1623–1662) used the method of induction competently and correctly in the mid–17th century!

Justification of induction

We have already seen how the natural numbers are an "inductive set" constructed from set theory. It is also possible to derive the Principle of Induction from other, more concrete, assertions. It is easy to define an *order* in $\mathbb{N}$ (to accommodate statements like $2 < 4$), and this gives rise to a swag of rigorously constructed (but familiar) concepts. Amongst these is:

Least integer principle. Every non-empty subset S of $\mathbb{N}$ has a least element[3]. □

This unremarkable looking fact gets us a really long way when it comes to proving things about the integers! It turns out to be equivalent to the principle of induction, although we'll prove only one direction here.

Theorem 5.1 *The least integer principle implies the principle of induction.*

Proof: Let $S \subseteq \mathbb{N}$ be such that $1 \in S$ and $n \in S$ implies $n+1 \in S$. Let $\bar{S}$ be the complement of S in $\mathbb{N}$, *ie.* all the natural numbers **not** in S. We want to show this set is empty.

By the least integer principle, if $\bar{S}$ is not empty, it has a least element m. Now,

- $m \in \bar{S}$, so by definition, $m \notin S$; and
- since m is the **least** element of $\bar{S}$, we have $k \notin \bar{S}$ whenever $k < m$; that is, $k \in S$.

From the first observation, we conclude that $m \neq 1$ (since $1 \in S$). This implies that $m > 1$ (since m is a natural number, it cannot be negative), so $m - 1 > 0$; that is, $m - 1 \in \mathbb{N}$. But, by the second observation above, since $m-1 < m$, we must have $m-1 \in S$. Since S is the set of natural numbers where the recursive step holds, we must have $m = (m-1)+1 \in S$. This is a contradiction. The only possible conclusion is that $\bar{S}$ cannot have any elements, so $S = \mathbb{N}$. □

(5.2) Mathematical induction

In this section we will be a bit more formal about the methodology of induction proofs. To start with, we will clarify what sort of things can be proved by induction.

[3]Note that the least integer principle fails in $\mathbb{Z}$, where negative numbers are allowed.

Propositions

A valid statement for us will be one that can be expressed in plain, simple language. We will refer to **propositions** P; such a P is a statement with a definite truth value.

Example 1. "Auckland is the capital of New Zealand." Here the proposition has truth value FALSE. □

Generally, we don't allow imprecise statements such as "Hamilton is more than you expect." However, a statement such as "The average annual rainfall in Hamilton is more than twice that of Christchurch" is allowable (and FALSE in this case!).

Some propositions are functions, whose truth depends on a variable. For example, let $P(x)$ be the statement that the person in office G3.0x is a lecturer. When $x = 6$, this is TRUE (G3.06 is a lecturer's office); when $x = 5$ it is FALSE (the current occupant of G3.05 is a postdoc). Obviously the value of x determines whether this statement is true or false.

Some statements are true for all possible values of their variables, such as when we let $P(x)$ be the statement "$x^2 \geq 0$", where x is a real number. This is true for all real numbers.

We are interested in propositions which depend on an integer value. Let $P(n)$ be a statement which is either true or false for any given n. For example, $P(7)$ may be true while $P(191)$ may be false[4]. Our strategy for showing that $P(n)$ is true for *all* n is to apply the principle of induction to the set of integers

$$S = \{n \mid P(n) \text{ is true}\}.$$

Example 2. For each positive integer n, let $P(n)$ be the proposition that

$$1^2 + 2^2 + \cdots + n^2 \;=\; \frac{n(n+1)(2n+1)}{6}.$$

(If $n = 1$, this just says $1^2 = \frac{1\times 2\times(2+1)}{6}$.) We proved above that $P(n)$ is true for all n. More precisely, we proved that the set of n for which $P(n)$ is true is $\mathbb{N}$. □

Many such facts in mathematics and computer science (*eg.* whether an algorithm works, or how fast it is) are difficult to prove directly but can be proved with *induction.*

Method of induction

The basic idea is to show the $n = 1$ case works, and then to show that if it works for some given fixed $k \geq 1$, (*ie.* if $P(k)$ is true for some k), then it must also work for $n = k + 1$ (i.e. $P(k + 1)$ is true too). From there we will conclude that $P(n)$ is true for *all* $n \geq 1$.

Example 3. Prove that for all $n \in \mathbb{N}$, $2^n \geq n + 1$. We will solve this problem with ideal setting out. (Always try to stick to this, even if you cannot solve a problem in full.)
Proof: For $n \geq 1$, let $P(n)$ be the proposition that $2^n \geq n + 1$.

(B) $P(1)$ says that $2^1 \geq 1 + 1$, that is, $2 \geq 2$, which is true.

(R) Assume $P(k)$ is true for some particular $k \geq 1$, that is,

$$2^k \geq k + 1.$$

[4]Showing that a proposition $P(n)$ is true for every single value of n is obviously not practical, and usually it will be too hard to derive the formula. In induction, we start with the formula, and prove that it is correct.

We show that $P(k+1)$ is true, that is,

$$2^{k+1} \geq (k+1)+1.$$

$$\begin{aligned} \text{LHS} &= 2^{k+1} \\ &= 2 \times 2^k \\ &\geq 2 \times (k+1) \text{ by } P(k) \\ &= 2k+2 \\ &\geq k+2 \\ &= \text{RHS}. \end{aligned}$$

So LHS$\geq$RHS, and $P(k+1)$ is true. Thus, $P(k)$ implies $P(k+1)$.

Hence by the Principle of Induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Note the strategy was the same as before: take the more complicated side and try to simplify it using the $P(k)$ assumption. In this case we make the LHS smaller or equal at each step, rather than just equal. Usually, the key step is seeing how to write $P(k+1)$ in a form where you can use $P(k)$.

Writing a proof by induction

1. Make sure you understand what the proposition $P(n)$ means, and write down the algebraic description of $P(n)$: "Let $P(n)$ be the proposition that $\cdots$."
2. Write down the base case, $P(1)$, and check that it is true.
3. Write: "We assume $P(k)$ is true. That is, $\cdots$." where the $\cdots$ are replaced by the algebraic description of $P(k)$.
4. Use mathematical reasoning to deduce that $P(k+1)$ is true. It may be necessary to write down what $P(k+1)$ actually says, and work from the more complicated side to the simpler side, using a substitution of $P(k)$ where appropriate. Then, write: "$P(k)$ implies $P(k+1)$."
5. Write: "Hence, by the Principle of Induction, $P(n)$ is true for all $n \cdots$."

Example 4. Prove that $5^n - 1$ is divisible by 4 for all $n \geq 0$. First of all, note:

- "m is divisible by n" is saying that if m, n are integers and $m = nq$, then q also an integer. So 12 is divisible by 4 (since $12 = 4 \times 3$), but not divisible by 5 (since $12 = 5 \times (2\,\frac{2}{5})$ and $q = 2\,\frac{2}{5}$ is not an integer).
- Now $n = 0$ is the base case! The principle of induction can be modified to allow any starting point, be it positive, negative or zero.

Now we can get on with the proof. For each $n \geq 0$, let $P(n)$ be the proposition that $5^n - 1$ is divisible by 4.

(B) $P(0)$ says that $5^0 - 1$ is divisible by 4. But $5^0 - 1 = 0 = 0 \times 4$, so this is TRUE.

(R) Assume $k \geq 0$ is such that $P(k)$ is true, that is, $5^k - 1$ is divisible by 4. This means

$$\text{there is an integer } q \text{ such that } 5^k - 1 \;=\; 4q.$$

We must show that $P(k+1)$ is true, that is, that $5^{k+1} - 1$ is divisible by 4. But

$$\begin{aligned} 5^{k+1} - 1 &= 5 \times 5^k - 1 \\ &= 5(4q+1) - 1 \text{ by } P(k) \\ &\quad \text{(we assume } 5^k - 1 = 4q) \\ &= 20q + 5 - 1 \\ &= 20q + 4 \\ &= 4(5q+1). \end{aligned}$$

But $5\,q+1$ is an integer since q is, so by definition, $5^{k+1}-1$ is divisible by 4, that is, $P(k+1)$ is true. Hence $P(k)$ implies $P(k+1)$.

Thus, by the Principle of Induction, $P(n)$ is true for all $n \geq 0$. □

Example 5. Prove that for all $n \in \mathbb{N}$,

$$\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^n = \begin{pmatrix} 1+2n & 2n \\ -2n & 1-2n \end{pmatrix}.$$

Proof: For each $n = 1, 2, 3, \ldots$, let $P(n)$ be the proposition that

$$\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^n = \begin{pmatrix} 1+2n & 2n \\ -2n & 1-2n \end{pmatrix}.$$

(B) $P(1)$ says that

$$\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^1 = \begin{pmatrix} 1+2 & 2 \\ -2 & 1-2 \end{pmatrix},$$

which is TRUE.

(R) Assume $k \in \mathbb{N}$ is such that $P(k)$ is true. That is,

$$\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^k = \begin{pmatrix} 1+2k & 2k \\ -2k & 1-2k \end{pmatrix}.$$

We show $P(k+1)$ must therefore be true, *ie.*

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

$$\begin{aligned}
\text{LHS} &= \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^{k+1} \\
&= \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}^{k} \\
&= \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}\begin{pmatrix} 1+2k & 2k \\ -2k & 1-2k \end{pmatrix} \quad \text{by } P(k) \\
&= \begin{pmatrix} 3(1+2k)+2(-2k) & 3(2k)+2(1-2k) \\ -2(1+2k)-1(-2k) & -2(2k)-1(1-2k) \end{pmatrix} \\
&= \begin{pmatrix} 3+6k-4k & 6k+2-4k \\ -2-4k+2k & -4k-1+2k \end{pmatrix} \\
&= \begin{pmatrix} 3+2k & 2k+2 \\ -2-2k & -1-2k \end{pmatrix},
\end{aligned}$$

whereas

$$\begin{aligned}
\text{RHS} &= \begin{pmatrix} 1+2k+2 & 2k+2 \\ -2k-2 & 1-2\,(k+1) \end{pmatrix} \\
&= \begin{pmatrix} 3+2k & 2k+2 \\ -2-2k & -1-2k \end{pmatrix} \\
&= \text{LHS}.
\end{aligned}$$

So $P(k+1)$ is true. That is, $P(k)$ implies $P(k+1)$.

Hence by the Principle of Induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Now we'll attempt a different kind of example.

How many subsets does a set with n elements have? Any set with one element has two subsets: the empty set and itself. A two-elements set has four subsets, and a three-element set has eight.

Example 6. We can use induction to prove that an n-element set has 2^n subsets.

For $n \geq 1$, let $P(n)$ be the proposition that any set with n elements has 2^n subsets.

(B) $P(1)$ we have already considered: it was true.

(R) Assume $P(k)$ is true for some $k \geq 1$: any set with k elements has 2^k subsets. We will show $P(k+1)$: any set with $k+1$ elements has 2^{k+1} subsets.

Let S be any set with $k+1$ elements. Suppose $a \in S$, and let S' be S with a removed, so S' has k elements. So S' has 2^k subsets by $P(k)$.

Each subset of S' is a subset of S as well. How many other subsets of S are there? All the subsets which **do** contain a, and there are exactly as many of these as there are subsets that *don't* contain a, namely 2^k of each. Thus

$$\begin{aligned}
\text{total number of subsets of } S &= 2^k + 2^k \\
&= 2 \times 2^k \\
&= 2^{k+1}.
\end{aligned}$$

So since S was arbitrary, we have shown every set with $k+1$ elements has 2^{k+1} subsets, so $P(k+1)$ is true and we have established that $P(k)$ implies $P(k+1)$.

Hence by the Principle of Induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Final remark

Your answers to induction problems will get around half marks if you get all the steps right except the recursive step (the only possibly hard step). But conversely, if your setting out is poor, even if you get the recursive step to work, you may not get much more than half the marks!

(5.3) Strong induction and recursion

In mathematics and computing, an important class of propositions are those which are defined *recursively*. A recursively defined sequence of propositions is a collection in which the proposition is defined in terms of the values of previous propositions. Given this setup, it is very natural to try to prove the truth of recursively defined sequences of propositions by induction.

Example 1. Let $x_1 = 0$ and for each $n > 1$ let

$$x_n = \frac{1}{2}\, x_{n-1} + 2.$$

Then $x_2 = \frac{1}{2}\, x_1 + 2 = \frac{1}{2}\, 0 + 2 = 2$, $x_3 = \frac{1}{2}\, x_2 + 2 = \frac{1}{2}\, 2 + 2 = 3$, $x_4 = \frac{1}{2}\, x_3 + 2 = \frac{1}{2}\, 3 + 2 = 3\,\frac{1}{2}$, and so on. In fact, we can use induction to prove

$$P(n) : x_n = 4 - 8\left(\frac{1}{2}\right)^n.$$

(B) When $n = 1$, $P(n)$ states $x_1 = 4 - 8\,\frac{1}{2} = 0$, which is TRUE.

(R) Suppose that $P(n)$ holds for $n = k$, so that $x_k = 4 - 8\,(\frac{1}{2})^k$. Then,

$$x_{k+1} = \frac{1}{2}\, x_k + 2 = \frac{1}{2}\,\left(4 - 8\,(\tfrac{1}{2})^k\right) + 2 = 2 - 8\,(\tfrac{1}{2})^{k+1} + 2 = 4 - 8\,(\tfrac{1}{2})^{k+1}.$$

Thus, $P(k+1)$ is true.

Since $P(k)$ implies $P(k+1)$, the proposition $P(n)$ is true for all $n \in \mathbb{N}$ by the principle of mathematical induction. □

Of course, most recursively defined sequences are far more complicated than the one above. Although we will get nowhere near the level of complexity required for some computer science applications, we can certainly equip ourselves for handling more complicated propositions than the one above. We will start with a "second order recursion".

Example 2. Recall the "Fibonacci" sequence from the first lecture:

$$R_0 = 1, R_1 = 2, R_{n+1} = R_n + R_{n-1}.$$

The first few terms of this sequence are

$$1, 2, 3, 5, 8, 13, 21, \ldots.$$

There is actually a formula for the nth term of the sequence! Let $P(n)$ be the statement:

$$R_n = \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

We would like to use induction to prove this statement. Unfortunately, our basic method of induction is insufficient: in order to evaluate the formula for R_{k+1}, we need to use the formulae for both R_k and R_{k-1}. Thus, the best we will be able to prove is that

$$P(k-1)\&P(k) \;\Rightarrow\; P(k+1),$$

which is logically weaker than what is needed by the principle of induction. □

Fortunately, there is an equivalent (although apparently stronger) version of induction which saves the day.

Strong induction

Strong principle of induction: Let S be a subset of the set of the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$. Suppose the following properties hold for S:

- $1 \in S$; and
- $1, 2, 3, \ldots, k \in S$ implies $k + 1 \in S$.

Then $S = \mathbb{N}$. □

When proving a proposition $P(n)$ is true for all n, using strong induction means that in the recursive step (R), we can assume not just that $P(k)$ is true, but *all* $P(i)$ for all $i \leq k$. Note that the statement "$P(1)\&P(2)\&\cdots\&P(k) \;\Rightarrow\; P(k+1)$" is a logically weaker statement than "$P(k) \;\Rightarrow\; P(k+1)$", so the sufficiency of this weaker recursive step suggests that the strong principle of induction is actually stronger than ordinary induction. However, the two principles are equivalent (as can be shown in the exercises).

Example 2 revisited. We will use Strong induction to prove that

$$R_n = \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

Let S be the subset of $\mathbb{N}$ for which the formula is true.

(B) If $n = 1$ then the RHS of the formula is:

$$\begin{aligned}
\text{RHS} &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^1 + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^1 \\
&= \frac{(\sqrt{5}-3)(1-\sqrt{5}) + (\sqrt{5}+3)(1+\sqrt{5})}{(2\sqrt{5})\,2} \\
&= \frac{(\sqrt{5}-3-(\sqrt{5})^2+3\sqrt{5}) + (\sqrt{5}+3+(\sqrt{5})^2+3\sqrt{5})}{4\sqrt{5}} \\
&= \frac{8\sqrt{5}}{4\sqrt{5}} = 2,
\end{aligned}$$

so the formula holds for $k = 1$. A similar calculation proves the formula for $k = 2$, so we have a "base case" of the formula holding for all $k \leq 2$.

(R) We will assume now, for a given k, that the formula holds for all $i \leq k$. In particular

$$\begin{aligned} R_{k-1} &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{k-1} + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k-1}, \\ R_k &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{k} + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k} \\ &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{k-1}\left(\frac{1-\sqrt{5}}{2}\right) + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k-1}\left(\frac{1+\sqrt{5}}{2}\right). \end{aligned}$$

We now have:

$$R_{k+1} = R_{k-1} + R_k = \tfrac{\sqrt{5}-3}{2\sqrt{5}}\left(\tfrac{1-\sqrt{5}}{2}\right)^{k-1}\left(1+\tfrac{1-\sqrt{5}}{2}\right) + \tfrac{\sqrt{5}+3}{2\sqrt{5}}\left(\tfrac{1+\sqrt{5}}{2}\right)^{k-1}\left(1+\tfrac{1+\sqrt{5}}{2}\right).$$

However,

$$\begin{aligned} \left(1+\frac{1-\sqrt{5}}{2}\right) &= \frac{3-\sqrt{5}}{2} \\ &= \frac{6-2\sqrt{5}}{4} \\ &= \frac{1-2\sqrt{5}+5}{4} \\ &= \frac{1-2\sqrt{5}+(\sqrt{5})^2}{4} \\ &= \left(\frac{1-\sqrt{5}}{2}\right)^2. \end{aligned}$$

A similar calculation shows that

$$\left(1+\frac{1+\sqrt{5}}{2}\right) = \left(\frac{1+\sqrt{5}}{2}\right)^2.$$

Putting these formulas back in the expression for R_{k+1} gives:

$$\begin{aligned} R_{k+1} &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{k-1}\left(\frac{1-\sqrt{5}}{2}\right)^2 + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k-1}\left(\frac{1+\sqrt{5}}{2}\right)^2 \\ &= \frac{\sqrt{5}-3}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{k+1} + \frac{\sqrt{5}+3}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}. \end{aligned}$$

This shows that the proposition $P(k+1)$ holds.

By the Strong Principle of Induction, $P(k)$ holds for every $k \geq 1$ and the formula is correct. □

Example 3. Let a sequence a_n be generated by $a_{n+1} = -2\,a_n + 3\,a_{n-1}$ where $a_0 = 1$ and $a_1 = 13$ are given (this is a *second order recurrence*). Prove by induction that

$$a_n = 4 + (-3)^n.$$

Solution: We will use strong induction.

(B) For a base case, notice that $P(1)$ and $P(2)$ are true, simply by use of the formula.

(R) Let $k \geq 2$ and assume now that $P(1), \ldots, P(k)$ all hold. In fact, we need only the formulae in $P(k-1)$ and $P(k)$, since these are the terms needed in the recursion. The two formulae are:

$$a_{k-1} = 4 + (-3)^{k-1} \text{ and } a_k = 4 + (-3)^k.$$

Now, from the recurrence,

$$\begin{aligned} a_{k+1} = -2\,a_k + 3\,a_{k-1} &= -2\,(4 + (-3)^k) + 3\,(4 + (-3)^{k-1}) \\ &= -8 - 2\,(-3)^k + 12 + 3\,(-3)^{k-1} \\ &= 4 - 2\,(-3)^k - (-3)\,(-3)^{k-1} \\ &= 4 - 2\,(-3)^k - (-3)^k \\ &= 4 - (2+1)\,(-3)^k \\ &= 4 + (-3)\,(-3)^k = 4 + (-3)^{k+1}. \end{aligned}$$

This establishes $P(k+1)$. We have thus established that "$P(k-1) \& P(k) \Rightarrow P(k+1)$, so certainly the weaker condition "$P(1) \& \cdots \& P(k-1) \& P(k) \Rightarrow P(k+1)$ holds.

Thus $P(n)$ is true for all $n \geq 1$ by the Strong Principle of Induction. □

Higher order linear recurrences

The last two examples are called second order linear recurrences, since the formula for the nth term depends on scalar multiples of the previous two terms.

Definition. A *kth order linear recurrence* is a sequence $\{a_n\}$ defined by

$$a_n = b_1\,a_{n-1} + b_2\,a_{n-2} + \cdots + b_k\,a_{n-k}$$

where $b_1, b_2, \ldots, b_k$ are fixed constants and $a_1, a_2, \ldots, a_k$ are given "initial values". □

To solve for the general term of a kth order linear recurrence, one needs do some algebra. Let

$$p(x) = x^k - b_1\,x^{k-1} - b_2\,x^{k-2} - \cdots - b_{k-1}\,x - b_k.$$

If $\lambda_1, \cdots, \lambda_k$ are k distinct *roots* of the polynomial equation $p(x) = 0$, then the general term of the sequence has the form

$$a_n = c_1\,\lambda_1^n + c_2\,\lambda_2^n + \cdots + c_k\,\lambda_k^n.$$

So, to solve an kth order linear recurrence one first has to find these roots, and then solve a system of equations (using the values of $a_1, \cdots, a_k$) to find $c_1, \cdots, c_k$. Once this is done, strong induction can be used to prove that your formula is correct.

Notes

- In general the above procedure can be fairly involved.
- The constants $c_1, \ldots, c_k$ can be found using linear algebra, and the work done here can be re-used when doing the "base case" in strong induction (the base case will involve checking your formula for all $a_1, \ldots, a_k$).
- If the polynomial doesn't have distinct roots, then things get more complicated.

We will do just one example.

Example 4. Let a sequence a_n be generated by $a_n = -2\,a_{n-1} + 3\,a_{n-2}$ where a_0 and a_1 are given (this is a *second order recurrence*). Prove by induction that

$$a_n = c_1\,\alpha^n + c_2\,\beta^n$$

where c_1, c_2, α, β solve the equations:

$$\alpha^2 + 2\,\alpha - 3 = \beta^2 + 2\,\beta - 3 = 0$$

$$a_0 = c_1 + c_2 \text{ and } a_1 = c_1 - 3\,c_2.$$

Solution: We will use strong induction. First of all, notice α and β are solutions to the equation $p(x) = x^2 + 2\,x - 3 = 0$ (in the notation of the general case we have $b_1 = -2$ and $b_2 = 3$). Thus, we can take $\alpha = 1$, $\beta = -3$. Let $P(n)$ be the proposition that

$$a_n = c_1\,1^n + c_2\,(-3)^n = c_1 + c_2\,(-3)^n.$$

(B) For a base case, notice that with this choice of α, β, $P(0)$ and $P(1)$ will be true if

$$\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}.$$

This equation is the same as the given one for c_1, c_2, fixing a choice of constants for which $P(0)$ and $P(1)$ hold.

(R) We will assume now that $P(k-1)$ and $P(k)$ both hold. Then

$$a_{k-1} = c_1 + c_2\,(-3)^{k-1} \text{ and } a_k = c_1 + c_2\,(-3)^k.$$

Now, from the recurrence,

$$\begin{aligned} a_{k+1} = -2\,a_k + 3\,a_{k-1} &= -2\,(c_1 + c_2\,(-3)^k) + 3\,(c_1 + c_2\,(-3)^{k-1}) \\ &= c_1 - 2\,c_2\,(-3)^k + 3\,c_2\,(-3)^{k-1} \\ &= c_1 - 2\,c_2\,(-3)^k - (-3)\,c_2\,(-3)^{k-1} \\ &= c_1 - 2\,c_2\,(-3)^k - c_2\,(-3)^k \\ &= c_1 - (2+1)\,c_2\,(-3)^k \\ &= c_1 + (-3)\,c_2\,(-3)^k = c_1 + c_2\,(-3)^{k+1}. \end{aligned}$$

This establishes $P(k+1)$.

Thus $P(n)$ is true for all $n \geq 0$ by the Strong Principle of Induction. □

Prime factorization

You may have encountered the idea of a **prime number** (you certainly will in a couple of weeks time when we study elementary number theory). A prime is a positive integer **greater than** 1 which has no factors other than itself and 1. The first few primes are:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43.$$

Primes are important because every (positive) integer can be written as a unique product of prime factors. For example:

$$819 = 3 \times 3 \times 7 \times 13.$$

The only time a number cannot be expressed as a product of several primes is if it actually *is* a prime.

Example 5. Let us prove that all non-primes can be factored in this way. The proof involves strong induction.

For $n \geq 2$, let $P(n)$ be the proposition that n is either prime or is a product of primes:

$$n = p_1\, p_2 \cdots p_r, \text{ all } p_i \text{ prime.}$$

(B) $P(2)$ is the base case, but 2 is a prime, so $P(2)$ is true.

(R) Assume $P(i)$ is true for all $i \leq k$, for some $k \geq 2$. We shall show $P(k+1)$ is true. There are two possibilities:

(i) If $k+1$ is prime, then $P(k+1)$ is true.

(ii) If $k+1$ is not prime, then by definition there must exist positive integers p, q, both greater than 1, for which $k+1 = p\,q$. But both p, q are less than $k+1$ so both $P(p)$ and $P(q)$ are true by the strong induction assumption above. So we can write

$$p = p_1\, p_2 \cdots p_r, \;\; q = q_1\, q_2 \cdots q_s,$$

where $p_1, p_2, \ldots, p_r, q_1, q_2, \ldots, q_s$ are primes. So

$$k+1 \;=\; pq \;=\; p_1 p_2 \cdots p_r q_1 q_2 \cdots q_s,$$

which is a product of primes. So $P(k+1)$ is true.

So whether or not $k+1$ is prime, $P(k+1)$ is true. Hence by the Strong Principle of Induction, $P(n)$ is true for all $n \geq 2$. □

VI ∘ Complex numbers

(6.1) Introduction to complex numbers

The *complex numbers* are an enlargement of the reals that allow arbitrary polynomial equations to be solved. (For example, $x^2 + 1 = 0$ has no real solution.) This turns out to be a very useful extension.

History of number systems

The first number system "invented" was the natural numbers $1, 2, 3, \ldots$, or $\mathbb{N}$. We have discussed the modern, set theoretic, approach to their structure, but the basic idea has been around since at least Babylonian times (c. 2000BC). The operations of addition and multiplication can be defined on $\mathbb{N}$, so that equations like

$$x_1 + x_2 = y \text{ or } x_1\, x_2 = y$$

made sense, and could be solved for y using arithmetic, given x_1, x_2.

It made sense to extend this by allowing negative numbers and zero as well, so that subtraction would work in general. This allowed the solution of equations like:

$$x_1 + y = x_2.$$

Remember, we take subtraction for granted, but it really does mean "undoing addition", or "solving an equation". Since it is a "reverse" operation, it may be more difficult than a "forward" operation, like addition or multiplication. And sometimes, it can't be done without an "extension" of the system[1]. To make subtraction work properly, we need the integers $\mathbb{Z}$. The other operations on $\mathbb{N}$ are easily extended, but facts such as

$$(-2) \times (-3) \;=\; 6$$

are needed, to ensure the usual number laws still work (distributivity of multiplication mainly).

Now, to "undo" multiplication by division, we need to be able to solve equations like

$$2\, x = 5.$$

This cannot be done in $\mathbb{Z}$, so fractions or rational numbers $\mathbb{Q}$ proved necessary. Linear equations with integer coefficients can be solved in $\mathbb{Q}$. The rational numbers were well understood in the times of the ancient Greeks, but $\cdots$

$\cdots$ eventually, it was realized that the real numbers were needed so that the "gaps" on the number line are filled, and equations like $x^2 = 2$ could be solved. Pythagoras (c. 500 BC) and his followers knew that the rational numbers didn't solve everything: the length of the hypotenuse of a right-angle triangle with its other sides of length 1 was irrational (*ie.* $\sqrt{2}$ is irrational).

Strange as it seems nowadays, the existence of irrational numbers was amongst the secret knowledge of the Pythagorean sect! Amazingly, the intricacies of "filling in the gaps" in the real number system weren't properly worked out until the late–19th century, when Dedekind's (1831–1916) work in the foundations of analysis and calculus provided the modern construction of the real number system $\mathbb{R}$.

[1] Try solving $3 + x = 2$ in $\mathbb{N}$.

But mathematicians didn't stop there. Many more number systems have been considered. The cross product operation on vectors in $\mathbb{R}^3$ defines a kind of number system. Square matrices of fixed size is another example.

The complex numbers are much closer to the real numbers.

Complex numbers and the basic operations

Complex numbers allow all algebraic equations with real coefficients to have solutions. Most of their properties were well understood by Gauss (1777–1855), and the complex numbers have a beautiful geometric interpretation.

Definition. A **complex number** is a "number" of the form

$$z = a + bi,$$

where a, b are real and i is assumed to satisfy $i^2 = -1$. Then a is the "real part" of z and b is the "imaginary part". The set of all complex numbers is denoted by $\mathbb{C}$. □

We simply define the operations of addition, multiplication and subtraction by using all the usual laws of algebra and replacing i^2 by -1 wherever it arises. So we assume associativity and commutativity of *both* addition and multiplication, as well as the distributive law.

Examples

1. $(2 + 3i) + (1 - 5i) \;=\; 2 + 1 + 3i - 5i \;=\; 3 - 2i$

2.

$$\begin{aligned}(2+3i)(1-5i) &= 2(1-5i) + 3i(1-5i)\\ &= 2 - 10i + 3i - 15i^2\\ &= 2 - 7i - 15(-1)\\ &= 17 - 7i.\end{aligned}$$

3. $-(-2 + 6i) \;=\; 2 - 6i$

In a similar way, we can deduce the general formulas for the sum or product of two complex numbers.
Sum:

$$(a + bi) + (c + di) \;=\; (a + c) + (b + d)i$$

Product:

$$\begin{aligned}(a+bi)(c+di) &= a(c+di) + bi(c+di)\\ &= ac + adi + bci + bdi^2\\ &= (ac - bd) + (ad + bc)i\end{aligned}$$

Negatives:

$$-(a + bi) \;=\; -a - bi$$

Actually, these are usually taken to be the *definitions* of the complex number operations, and the usual laws can be deduced from these. The numbers themselves can be viewed as ordered pairs of reals, *ie.* elements of $\mathbb{R}^2$:

$$a + bi \;\leftrightarrow\; (a, b)$$

with the operations defined in terms of them. Addition is then just ordinary vector addition in $\mathbb{R}^2$:

$$(a + b\,i) + (c + d\,i) \leftrightarrow (a, b) + (c, d) = (a + c, b + d) \leftrightarrow (a + c) + (b + d)\,i$$

and something similar for negatives. Multiplying one complex number by another with imaginary part zero is really just scalar multiplication in this view:

$$c\,(a + b\,i) \leftrightarrow c\,(a, b) = (c\,a, c\,b) \leftrightarrow c\,a + c\,b\,i.$$

The only weird one is multiplication by complex numbers with non–zero real part:

$$(a, b)\,(c, d) \;=\; (a\,c - b\,d, a\,d + b\,c),$$

although even this can be represented by matrix multiplication[2].

The view of complex numbers as elements of $\mathbb{R}^2$ is useful for visualizing complex numbers. Representing $z = a + b\,i$ as (a, b) allows us to draw z in the **Argand plane**. This is just the usual $\mathbb{R}^2$, with the x–axis for real parts, and the y-axis for imaginary parts. While complex addition and scalar multiplication are easy to visualize as vector operations, multiplication is more tricky—we'll see more of this later.

(6.2) Further operations on $\mathbb{C}$ and the polar form

There are several special operations with the complex numbers.

Conjugation, modulus and inversion

Definition. The **conjugate** of $z = a + b\,i \in \mathbb{C}$ is $\bar{z} = a - b\,i$. In the Argand plane, the conjugate of z is obtained by reflecting about the x-axis. □

Example 1. If $z = 3 + 4\,i$ then $\bar{z} = 3 - 4\,i$. □

Conjugation has several reasonable properties (all are easy to prove):

1. $\bar{\bar{z}} = z$
2. $\overline{z_1 + z_2} = \bar{z_1} + \bar{z_2}$
3. $\overline{z_1 z_2} = \bar{z_1}\bar{z_2}$

Multiplying z by its conjugate gives a non-negative real number with a geometric interpretation: Write $z = a + bi$, so that

$$\begin{aligned} z\bar{z} &= (a + b\,i)(a - b\,i) \\ &= a^2 + b\,a\,i - a\,bi - (b\,i)^2 \\ &= a^2 + (b\,a - a\,b)\,i - (i^2)\,b^2 \\ &= a^2 + 0 - (-1)\,b^2 \\ &= a^2 + b^2. \end{aligned}$$

This is just the square of the length of the vector (a, b) in the Argand plane!

[2]The idea is to identify the complex number $a + b\,i$ with the matrix $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$. Then complex addition and multiplication are equivalent to ordinary matrix addition and multiplication, but visualization is more tricky.

Definition. The **modulus** of z is

$$|z| = \sqrt{z\,\bar{z}}.$$

The modulus is the length of z, and it is a positive real number, providing $z \neq 0$. □

Example 2. If $z = 3 + 4\,i$ then $|z| = \sqrt{z\bar{z}} = \sqrt{(3 - 4\,i)(3 + 4\,i)} = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$. □

Properties of the modulus:

1. $|-z| = |z|$
2. $|z_1 + z_2| \leq |z_1| + |z_2|$ (triangle inequality)
3. $|z_1 z_2| = |z_1| \cdot |z_2|$

Proof: The first of these is easy to prove. The second is really a fact about lengths of vectors in $\mathbb{R}^2$ and is obvious from a picture. The third follows from the properties of conjugation:

$$\begin{aligned} |z_1\,z_2|^2 &= (z_1\,z_2)(\overline{z_1\,z_2}) \\ &= z_1\,z_2\,\bar{z_1}\,\bar{z_2} \\ &= (z_1\,\bar{z_1})\,(z_2\,\bar{z_2}) \\ &= |z_1|^2\,|z_2|^2. \end{aligned}$$

Taking square roots of both sides gives us what we want. □

Definition. The inverse of a non-zero complex number z is:

$$z^{-1} = \frac{1}{z} = \frac{1}{z}\,\frac{\bar{z}}{\bar{z}} = \frac{\bar{z}}{z\,\bar{z}} = \frac{\bar{z}}{|z|^2}.$$

Clearly, $z^{-1}\,z = 1$. □

Example 3. The inverse of $4 - 3\,i$ is

$$\begin{aligned} \frac{1}{|4 - 3\,i|^2}\overline{4 - 3\,i} &= \frac{1}{4^2 + 3^2}(4 + 3\,i) \\ &= \frac{1}{16 + 9}(4 + 3\,i) \\ &= \frac{1}{25}(4 + 3\,i) \\ &= \frac{4}{25} + \frac{3}{25}\,i. \end{aligned}$$

We can check this by multiplying out! □

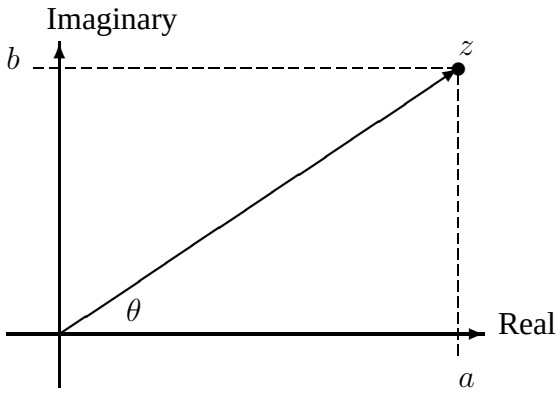
Finally, we can define **complex division**: For $z_1, z_2 \in \mathbb{C}$ with $z_2 \neq 0$, we now define

$$z_1/z_2 = z_1\,z_2^{-1}.$$

Polar form

A complex number written as $a + bi$ is in *rectangular form*. There is also the *polar form*; this is the representation on the *Argand plane*.

Any complex number (or indeed position vector in $\mathbb{R}^2$) is completely determined by knowing the angle it makes with the positive x–axis θ (as shown, with anti-clockwise positive and clockwise negative, as usual), and its length.



Suppose the length of $z = a + bi$ is r. Then by definition,

$$\cos\ \theta = a/r, \text{ and } \sin\ \theta = b/r.$$

So

$$z = a + b\,i = (r\ \cos\ \theta) + (r\ \sin\ \theta)\,i = r\,(\cos\ \theta + i\ \sin\ \theta).$$

Definition. Write cis $\theta = \cos\ \theta + i\ \sin\ \theta$. We call $z = r\,\text{cis}\ \theta$ the **polar form** of z. Here $r = |z|$, and θ is the **argument** of z, $\arg(z)$. □

Note that $|\text{cis}\ \theta| = |\cos\theta + i\ \sin\theta| = \sqrt{\cos^2\theta + \sin^2\theta} = \sqrt{1} = 1$.

Example 4. Suppose $z = \sqrt{2}\,\text{cis}\ \frac{\pi}{4}$. Express z in rectangular form $z = a + bi$ for a, b real. We'll be using radians from now on unless stated otherwise.

$$\begin{aligned} z &= \sqrt{2}\,\text{cis}\ \frac{\pi}{4} \\ &= \sqrt{2}\,(\cos\ \frac{\pi}{4} + i\ \sin\ \frac{\pi}{4}) \\ &= \sqrt{2}\,(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\,i) \\ &= 1 + i. \end{aligned}$$

(Recall that $\pi/4$ radians is $\frac{180}{\pi} \cdot \pi/4 = 45$ degrees.) □

Example 5. Suppose $|z| = 2$ and $\arg(z) = \frac{5\pi}{6}$. To put z in rectangular form, note that $\frac{5\pi}{6} = \pi - \frac{\pi}{6}$, so

$$\cos\frac{5\,\pi}{6} = \cos\left(\pi + \left(-\frac{\pi}{6}\right)\right) = -\cos\left(-\frac{\pi}{6}\right) = -\cos\ \frac{\pi}{6} = -\frac{\sqrt{3}}{2},$$

and likewise,

$$\sin\left(\pi + (-\frac{\pi}{6})\right) = -\sin\left(-\frac{\pi}{6}\right) = \sin\ \frac{\pi}{6} = \frac{1}{2}.$$

So cis $\frac{5\pi}{6} = -\frac{\sqrt{3}}{2} + \frac{1}{2}\,i$ and hence

$$z = 2\,\text{cis}\,\frac{5\pi}{6} = 2\left(-\frac{\sqrt{3}}{2} + \frac{1}{2}\,i\right) = -\sqrt{3} + i.$$

□

Converting from polar to rectangular form

1. Let $z = r\,\text{cis}\,\theta$.
2. Calculate $\cos\theta$ and $\sin\theta$.
3. Let $a = r\,\cos\theta$ and $b = r\,\sin\theta$.

Now let's convert from rectangular to polar.

Example 6. Let $z = 1 + i$. To find the polar form $z = r\,\text{cis}\,\theta$, we calculate $r = |z| = \sqrt{1^2 + 1^2} = \sqrt{1+1} = \sqrt{2}$, and write z as

$$z = \sqrt{2}\left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\,i\right).$$

We must have

$$\text{cis}\;\theta = \cos\,\theta + i\,\sin\,\theta = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\,i,$$

so

$$\cos\,\theta = \frac{1}{\sqrt{2}},\;\text{and}\;\sin\,\theta = \frac{1}{\sqrt{2}}.$$

Now $\cos\phi = \frac{1}{\sqrt{2}}$ if $\phi = \pm\frac{\pi}{4}$, but $\sin\frac{\pi}{4} = \frac{1}{\sqrt{2}}$ and $\sin\left(-\frac{\pi}{4}\right) = -\frac{1}{\sqrt{2}}$ so the only solution to both equations is $\theta = \pi/4$ (or 45 degrees). Then

$$z = \sqrt{2}\,\text{cis}\;\frac{\pi}{4}.$$

□

Example 7. Convert $z = -\frac{1}{2} + \frac{\sqrt{3}}{2}\,i$ to polar form. First of all, $|z| = \sqrt{\frac{1}{4} + \frac{3}{4}} = 1$, so cis $\theta = -\frac{1}{2} + \frac{\sqrt{3}}{2}\,i$. Then

$$\cos\,\theta = -\frac{1}{2},\;\text{and}\;\;\sin\,\theta = \frac{\sqrt{3}}{2}.$$

Solving the cosine equation gives $\theta = \pm\frac{2\pi}{3}$, and from the sine equation, $\theta = \frac{\pi}{3}$ or $\frac{2\pi}{3}$. The only common solution is $\theta = \frac{2\pi}{3}$, so

$$z\;= \text{cis}\;\frac{2\pi}{3}.$$

□

Converting from rectangular to polar form

1. Let $z = a + i\,b$. Let $r = \sqrt{a^2 + b^2}$.
2. Find the two values of $\phi = \cos^{-1}\frac{a}{r}$.
3. The correct value of θ will be the ϕ for which $\sin\phi = \frac{b}{r}$ (the other one is $-\frac{b}{r}$).
4. The polar form is $z = r\,\text{cis}\,\theta$.

Interpretation of complex multiplication

We can use the polar representation to help understand what happens when we multiply together two complex numbers. Let $z_1 = r_1 \operatorname{cis}\ \theta_1$, $z_2 = r_2 \operatorname{cis}\ \theta_2$. Then

$$\begin{aligned} z_1\, z_2 &= (r_1 \operatorname{cis}\ \theta_1)\,(r_2 \operatorname{cis}\ \theta_2) \\ &= r_1\, r_2 \operatorname{cis}\ \theta_1 \operatorname{cis}\ \theta_2. \end{aligned}$$

With the help of some trig identities, one can prove that

$$\operatorname{cis}\ \theta_1 \operatorname{cis}\ \theta_2 = \operatorname{cis}\,(\theta_1 + \theta_2),$$

so that

$$z_1\, z_2 = r_1\, r_2 \operatorname{cis}\,(\theta_1 + \theta_2).$$

This shows the fact mentioned earlier, that

$$|z_1\, z_2| = r_1\, r_2 = |z_1|\,|z_2|,$$

and that

$$\arg(z_1\, z_2) = \arg(z_1) + \arg(z_2).$$

Sometimes, instead of cis θ the notation $e^{i\theta}$ is used. The above formula involving cis then has the more intuitive form:

$$e^{i\,\theta_1}\, e^{i\,\theta_2} = e^{i\,(\theta_1+\theta_2)}.$$

One can show how to generalize exponentials to allow complex arguments in a way consistent with this usage, but we do not pursue this further now.

(6.3) Solving equations in $\mathbb{C}$

Solving linear equations with complex numbers

Using only complex division, we can solve basic equations in $\mathbb{C}$. Suppose that

$$z_1\, w = z_2.$$

Then

$$w = z_1^{-1}\, z_1\, w = z_1^{-1}\, z_2 = \frac{z_2}{z_1} = \frac{z_2\, \bar{z_1}}{z_1\, \bar{z_1}}.$$

Example 1. Compute $(-1 + 7\,i)/(2 + i)$.

$$\begin{aligned} \frac{-1+7\,i}{2+i} &= \left(\frac{-1+7\,i}{2+i}\right)\left(\frac{2-i}{2-i}\right) \\ &= \frac{(-1+7\,i)\,(2-i)}{(2+i)\,(2-i)} \\ &= \frac{-1(2-i)+7\,i\,(2-i)}{2^2+1^2} \\ &= \frac{-2+i+14\,i+7}{5} \\ &= \frac{5+15\,i}{5} \\ &= 1+3\,i. \end{aligned}$$

□

We can even use this method to find the inverse of a complex number efficiently:

Example 2. Let $z = 2 - i$. Then,

$$\begin{aligned} \frac{1}{2-i} &= \left(\frac{1}{2-i}\right)\left(\frac{2+i}{2+i}\right) \\ &= \frac{2+i}{2^2+1^2} \\ &= \frac{2+i}{5} \\ &= \frac{2}{5} + \frac{1}{5}\,i. \end{aligned}$$

□

We can solve more complicated equations too: if u, v, w are complex we can solve $uz + v = w$ for z:

$$z = (w - v)/u = (w - v)\,u^{-1}.$$

In fact, all the techniques of linear algebra work for matrices with complex coefficients: you might like to test out a few!

Equations of the form $z^n = w$

From facts established above, we can use induction to prove *De Moivre's formula*:

$$(r \text{ cis } \theta)^n \;=\; r^n \text{ cis } (n\theta).$$

This will allow us to do a lot of calculations involving powers of complex numbers.

Example 3. Find $(-1 + \sqrt{3}\,i)^8$, in rectangular form.

Commentary: A key useful point here is that in general

$$\text{cis } \theta_1 \;=\; \text{cis } \theta_2$$

means both the sines and cosines of the two angles are equal, so θ_1 and θ_2 differ by a multiple of 2π, as we saw earlier. □

Now, let $z = -1 + \sqrt{3}\,i$. Then $|z| = \sqrt{1+3} = \sqrt{4} = 2$, so

$$z \;=\; 2\left(-\frac{1}{2} + \frac{\sqrt{3}}{2}\,i\right),$$

so $\cos\ \theta = -\frac{1}{2}$, $\sin\ \theta = \frac{\sqrt{3}}{2}$, and hence $\theta = \pi - \pi/3 = 2\pi/3$ giving $z = 2\,\text{cis}\ \frac{2\,\pi}{3}$. Thus,

$$\begin{aligned}
(-1+\sqrt{3}\,i)^8 &= \left(2\,\text{cis}\ \frac{2\,\pi}{3}\right)^8 \\
&= 2^8\,\text{cis}\ 8\left(\frac{2\,\pi}{3}\right) \\
&= 2^8\,\text{cis}\ \frac{16\,\pi}{3} \\
&= 2^8\,\text{cis}\ \left(4\,\pi + \frac{4\,\pi}{3}\right) \\
&= 2^8\,\text{cis}\ \frac{4\,\pi}{3} \\
&= 2^8\,\text{cis}\ \left(\pi + \frac{\pi}{3}\right) \\
&= 2^8\,\left(\cos\left(\pi+\frac{\pi}{3}\right) + i\,\sin\left(\pi+\frac{\pi}{3}\right)\right) \\
&= 2^8\,\left(-\cos\ \frac{\pi}{3} - i\,\sin\ \frac{\pi}{3}\right) \\
&= 2^8\,\left(-\frac{1}{2} - \frac{\sqrt{3}}{2}\,i\right) \\
&= 2^7\,(-1-\sqrt{3}\,i) \\
&= -128 - 128\sqrt{3}\,i.
\end{aligned}$$

□

Finding nth powers of $z \in \mathbb{C}$

1. Write z in polar form: $z = r\,\text{cis}\,\theta$;
2. Calculate r^n and $n\,\theta$. The nth power is $r^n\text{cis}\,n\theta$.
3. Write the results of step 2 in rectangular form.

Equations of the form $w^n = z$

We can use the same idea to find n-th roots of complex numbers. In general, this means that for a given $w \in \mathbb{C}$ we want to solve the equation $z^n = w$.

Let us suppose that $w = |w|\,\text{cis}\ \theta$, and write $z = r\,\text{cis}\ \phi$, where both r and ϕ are to be found. Then, by De Moivre's formula,

$$\begin{aligned}
|w|\,\text{cis}\ \theta &= (r\,\text{cis}\ \phi)^n \\
&= r^n\,\text{cis}\,(n\,\phi),
\end{aligned}$$

so $r^n = |w|$ and hence $r = |w|^{\frac{1}{n}}$. But we also know that $\text{cis}\,(n\,\phi) = \text{cis}\ \theta$, so $n\phi = \theta + 2\,k\,\pi$ for any integer k.

Commentary: This step is really critical, since we want to find **all** the roots of z. It turns out that there are exactly n of them, and their arguments are evenly spaced around the unit circle in $\mathbb{C}$. □

We can now write $\phi = \frac{\theta}{n} + \frac{2\,k\,\pi}{n}$, so

$$\begin{aligned} z &= r\,\text{cis}\ \phi \\ &= |w|^{\frac{1}{n}}\,\text{cis}\left(\frac{\theta}{n} + \frac{2\,k\,\pi}{n}\right) \end{aligned}$$

for any integer k. In practice, we need to consider only $k = 0, 1, 2, \ldots, n-1$, since for $k = n$ we are just adding 2π to the $k = 0$ case and we start repeating ourselves[3]. Except for 0, there are exactly n n-th roots of a complex number.

Example 4. Solve $z^3 = 8\,i$.
Solution: First, write $w = 8\,i$ in polar form: $8\,i = 8\,(0 + i)$. So $\cos\theta = 0$, $\sin\theta = 1$, and hence $\theta = \pi/2$. Thus,

$$z^3 = 8\,i = 8\,\text{cis}\,\frac{\pi}{2} = 8\,\text{cis}\,\left(2\,k\,\pi + \frac{\pi}{2}\right),$$

$k = 0, 1, 2, \ldots$. Hence

$$\begin{aligned} z &= 8^{\frac{1}{3}}\,\text{cis}\,\frac{2\,k\,\pi + \pi/2}{3} \\ &= 2\,\text{cis}\,\frac{4\,k\,\pi + \pi}{6}, k = 0, 1, 2 \\ &= 2\,\text{cis}\,\frac{\pi}{6},\ 2\,\text{cis}\,\frac{5\pi}{6},\ 2\,\text{cis}\,\frac{9\pi}{6} \\ &= 2\,(\sqrt{3}/2 + i/2),\ 2\,(-\sqrt{3}/2 + i/2),\ 2\,(-i) \\ &= \sqrt{3} + i,\ -\sqrt{3} + i,\ -2i. \end{aligned}$$

Note: $k = 3$ would give $2\,\text{cis}\ \frac{13\pi}{6} = 2\,\text{cis}\ \frac{\pi}{6}$—nothing new. Also, the cube roots are equally spaced around the circle $|z| = 2$, centred on the origin of the Argand plane and having radius 2. □

Finding nth roots of $z \in \mathbb{C}$

1. Write z in polar form: $z = r\,\text{cis}\,\theta$;
2. the nth roots are $w = r^{1/n}\text{cis}\ \left(\frac{\theta + 2\,k\,\pi}{n}\right)$, $k = 0, 1, \ldots, n-1$;
3. write the results of step 2 in rectangular form.

Example 5. Solve $z^2 = 2 + 2\sqrt{3}\,i$.
Solution: $|2 + 2\sqrt{3}\,i| = \sqrt{4 + 12} = 4$, so

$$2 + 2\sqrt{3}\,i\ =\ 4\left(\frac{2}{4} + \frac{2\sqrt{3}}{4}\,i\right)\ =\ 4\left(\frac{1}{2} + \frac{\sqrt{3}}{2}\,i\right).$$

If $\arg(w) = \theta$, then cis $\theta = \frac{1}{2} + \frac{\sqrt{3}}{2}\,i$, so

$$\cos\,\theta\ =\ \frac{1}{2},\ \sin\,\theta\ =\ \frac{\sqrt{3}}{2},$$

so we're in the first quadrant and our table tells us that $\theta = \pi/3$. So

$$z^2\ =\ 2 + 2\sqrt{3}\,i = 4\,\text{cis}\,\frac{\pi}{3} = 4\,\text{cis}\,\left(2\,k\,\pi + \frac{\pi}{3}\right),$$

[3]The same consideration lets us exclude all $k < 0$.

for $k = 0, 1$. Then

$$\begin{aligned} z &= 4^{\frac{1}{2}} \operatorname{cis}\left(\frac{2k\pi + \pi/3}{2}\right) \\ &= 2 \operatorname{cis}(k\,\pi + \pi/6), \quad k = 0, 1 \\ &= 2 \operatorname{cis}\frac{\pi}{6}, \quad 2 \operatorname{cis}\frac{7\,\pi}{6}. \end{aligned}$$

But

$$\operatorname{cis}\frac{\pi}{6} = \cos\frac{\pi}{6} + \sin\frac{\pi}{6}\,i = \frac{\sqrt{3}}{2} + \frac{1}{2}\,i,$$

and

$$\operatorname{cis}\frac{7\pi}{6} = \operatorname{cis}\left(\pi + \frac{\pi}{6}\right) = -\operatorname{cis}\frac{\pi}{6} = -\frac{\sqrt{3}}{2} - \frac{1}{2}\,i.$$

Thus $z = \sqrt{3} + i,\ -\sqrt{3} - i$.

Note: If $z^2 = w$, then also $(-z)^2 = z^2 = w$, so the two solutions are $z, -z$, just as with real numbers (though now there is no concept of "positive" or "negative"). □

Solving quadratic equations

To solve the general quadratic equation

$$a\,z^2 + b\,z + c \;=\; 0,\ a, b, c \in \mathbb{C},$$

the method is as for the real numbers. First complete the square:

$$\begin{aligned} a\,z^2 + b\,z + c &= a\left(z^2 + \frac{b}{a}\,z + \frac{c}{a}\right) \\ &= a\left(\left(z + \frac{b}{2\,a}\right)^2 - \frac{b^2}{4\,a^2} + \frac{c}{a}\right) \end{aligned}$$

which is zero exactly when

$$\left(z + \frac{b}{2\,a}\right)^2 = \frac{b^2}{4\,a^2} - \frac{c}{a} = \frac{b^2 - 4\,a\,c}{4\,a^2}.$$

So taking square roots:

$$z + \frac{b}{2a} \;=\; \pm\frac{\sqrt{b^2 - 4\,a\,c}}{2a},$$

so

$$\begin{aligned} z &= -\frac{b}{2\,a} \pm \frac{\sqrt{b^2 - 4\,a\,c}}{2\,a} \\ &= \frac{-b \pm \sqrt{b^2 - 4\,a\,c}}{2\,a}. \end{aligned}$$

Example 6. Solve $z^2 + 2\,z + 3 = 0$.

$$\begin{aligned} z &= \frac{-2 \pm \sqrt{4 - 12}}{2} \\ &= -1 \pm \frac{\sqrt{-8}}{2} \\ &= -1 \pm \frac{2\sqrt{2}\,i}{2} \\ &= -1 + \sqrt{2}\,i,\ -1 - \sqrt{2}\,i. \end{aligned}$$

□

Example 7. Solve $z^2 + 2z - \sqrt{3}\,i \;=\; 0$.

$$\begin{aligned} z &= \frac{-2 \pm \sqrt{4 + 4\sqrt{3}\,i}}{2} \\ &= \frac{-2 \pm 2\sqrt{1 + \sqrt{3}\,i}}{2} \\ &= -1 \pm \sqrt{1 + \sqrt{3}\,i}. \end{aligned}$$

To simplify the writing down, we'll work on $w = \sqrt{1 + \sqrt{3}\,i}$. Then

$$\begin{aligned} w^2 &= 1 + \sqrt{3}\,i \\ &= 2\left(\frac{1}{2} + \frac{\sqrt{3}}{2}\,i\right) \\ &= 2\,\mathrm{cis}\,\frac{\pi}{3} \end{aligned}$$

so

$$\begin{aligned} w &= \pm\sqrt{2}\,\mathrm{cis}\,\frac{\pi}{6} \\ &= \pm\sqrt{2}\left(\frac{\sqrt{3}}{2} + \frac{i}{2}\right) \\ &= \pm\left(\frac{\sqrt{3}}{\sqrt{2}} + \frac{i}{\sqrt{2}}\right) \\ &= \pm\left(\sqrt{\frac{3}{2}} + \frac{1}{\sqrt{2}}\,i\right), \end{aligned}$$

and so

$$z \;=\; \left(-1 + \sqrt{\frac{3}{2}}\right) + \frac{1}{\sqrt{2}}\,i, \text{ or } \left(-1 - \sqrt{\frac{3}{2}}\right) - \frac{1}{\sqrt{2}}\,i.$$

□

VII ○ Elementary number theory

(7.1) Natural numbers and divisibility

Number theory is the study of properties of the natural numbers $\mathbb{N}$, especially relating to addition and multiplication. It turns out that these basic arithmetic operations lead to an amazing amount of structure on $\mathbb{N}$, including the notion of prime numbers, and their apparently random pattern. This area has many modern applications, *eg.* coding theory and cryptography—we'll see some of these applications later.

Divisibility and remainders

Definition. An integer b **divides** an integer a if $\frac{a}{b}$ is an integer, that is, if $a = k\,b$ for some integer k. *Notation:* $b|a$, and we say b is a *divisor* of a. The *divisor list* for a is the set of all positive divisors of a. □

Properties: For all integers a, b, c:

1. $a|a$
2. if $a|b$ and $b|a$ then $a = \pm b$
3. if $a|b$ and $b|c$ then $a|c$

Sometimes it is easy to tell if a base ten integer is divisible by a certain natural number. Obviously $2|n$ if and only if the final digit of n is even, but more interesting are the following:

- $4|n$ iff the integer given by the final two digits is divisible by 4
- $3|n$ iff sum of digits is divisible by 3
- $9|n$ iff sum of digits is divisible by 9
- $5|n$ iff last digit is 0 or 5

Example 1. The divisors of 15 are $\{1, 3, 5, 15\}$.
The divisors of 72 are $\{1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, 72\}$.
The divisor list for -5 is $\{1, 5\}$. □

Prime numbers

Prime numbers are the "least divisible" numbers. There are infinitely many of them, their distribution obeys a remarkably regular law[1], but there is no known formula for generating them. Indeed, the distribution was derived by Gauss (1777–1855) when he was a teenager, but a full proof did not arrive until 1896! Primes are amongst the most beguiling objects in mathematics: so simple to define, yet the set of all them is so complicated that mathematicians have struggled to understand it from ancient times to the present day.

[1] The "Prime number theorem" asserts that the number of primes less than or equal to n is approximately $\frac{n}{\log n}$. A brief discussion, including some history, can be found on the WWW at: http://mathworld.wolfram.com/PrimeNumberTheorem.html

Definition. If a positive integer $p > 1$ has only 1 and p as divisors, we call it **prime**. □

The first few primes are:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \ldots$$

There are lots of primes, a fact known to Euclid (c. 300 B.C.):

Theorem 7.1 *There are infinitely many primes.*

Proof: Suppose on the contrary that there are only finitely many, say

$$p_1, p_2, \ldots, p_k.$$

Let $n = 1 + p_1 \times p_2 \times \cdots \times p_k$. By assumption, this is not prime since it is bigger than any of the p_i. So it must have divisors, and can in fact be written as a product of primes (as we showed in the induction section). So it is certainly divisible by at least one of the p_j. So $n = k\,p_j$ for some integer k. Now let P be the product of all the primes except p_j. Obviously $p_1 \times p_2 \times \cdots \times p_k = P\,p_j$ so we can write

$$\begin{aligned} 1 &= n - p_1 \times p_2 \times \cdots \times p_k \\ &= k\,p_j - P\,p_j \\ &= (k - P)\,p_j. \end{aligned}$$

This suggests that 1 is divisible by p_j which is obviously false. We conclude that there cannot be finitely many primes! □

A fundamental fact of algebra is that any non-prime n can be written as a product of primes **in only one way**. So not only can we write

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r},$$

where the p_i are primes with $p_1 < p_2 < \cdots < p_r$, and the α_i are naturals, but n cannot be written as any *other* such product.

The divisors of such an n all have the form

$$p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}$$

with $0 \le m_i \le \alpha_i$ for each i. There are thus $\alpha_i + 1$ possible choices of each m_i and hence

$$(\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_r + 1)$$

distinct divisors.

Example 2. Write 504 as a product of primes. How many divisors has it have? *Solution:* Successively divide by each prime as many times as you can and continue dividing until you get 1:

2	504
2	252
2	126
3	63
3	21
7	7
	1

So $504 = 2^3 \times 3^2 \times 7$, and the number of divisors is: $(3 + 1)(2 + 1)(1 + 1) = 4 \times 3 \times 2 = 24$. □

Greatest common divisor

Definition. For $a, b \in \mathbb{N}$, the **greatest common divisor** of a and b is the largest $d \in \mathbb{N}$ with the property that $d|a$ and $d|b$. *Notation:* $\gcd(a, b)$. □

Example 3. 30 and 75 have divisor lists $\{1, 2, 3, 5, 6, 10, 15, 30\}$ and $\{1, 3, 5, 15, 25, 75\}$ (respectively). Of these, the following are common divisors: $1, 3, 5, 15$. The largest of these is 15 so $\gcd(30, 75) = 15$.

We can find this gcd by writing each of $30, 75$ as unique products of primes. then:

$$30 \;=\; 2 \times 3 \times 5,\; 75 \;=\; 3 \times 5^2.$$

(To obtain these, successively divide by the smallest prime you can, as described earlier.) We can read off the gcd by taking the highest power of each prime present in *both* numbers and multiplying them together. (In this case, we get $3 \times 5 = 15$.) □

For large a, b this method of obtaining the gcd (used in the above example) is very inefficient; we'll shortly see a much better way.

Least common multiple

This is an important "dual" concept to the gcd.

Definition. The **least common multiple** $n = \operatorname{lcm}(a, b)$ is the smallest natural number n divisible by both a and b (*ie.* such that $a|n$ and $b|n$). □

Example 4. 30 and 75 have as their first few multiples

$$30, 60, 90, 120, 150, 180, \ldots \text{ and } 75, 150, 225, 300, \ldots$$

respectively. The smallest of these common to both is obviously 150, so $\operatorname{lcm}(30, 75) = 150$. □

We can find lcms as for gcds, by first rewriting the two numbers as their unique products of primes and taking the highest (rather than lowest) power of each prime that occurs in *either* (rather than both), and multiplying them together. Thus

$$30 \;=\; 2 \times 3 \times 5,\; 75 \;=\; 3 \times 5^2$$

so

$$\operatorname{lcm}(30, 75) \;=\; 2 \times 3 \times 5^2 \;=\; 150.$$

However, as noted above, this method is rather inefficient, finding the prime factorization can be very difficult **and** there are some neat tricks that will lead to a much better way.

Theorem 7.2 *We have* $a\,b = \gcd(a, b)\operatorname{lcm}(a, b)$.

Example 5. We will soon see that $\gcd(10362, 12397) = 11$. Therefore,

$$\operatorname{lcm}(10362, 12397) = \frac{10362 \times 12397}{\gcd(10362, 12397)} = \frac{128457714}{11} = 11677974,$$

a calculation that might be infeasible otherwise! □

(7.2) Remainders and the Euclidean algorithm

Remainders

Most of our study of divisibility is built on the following result.

Theorem 7.3 (Remainder Theorem.) *For $a, b \in \mathbb{Z}$ with $b > 0$, there are* unique *integers $q, r \geq 0$ such that*

$$a \;=\; bq + r, \;\; 0 \leq r < b.$$

Notation: We call q the **quotient** and r the **remainder** of a on division by b and write: $\text{rem}(a, b) = r$. □

Example 1. If $a = 548$ and $b = 24$, then $\frac{a}{b} = 22 + \frac{20}{24}$ (we can find with the help of a calculator). So

$$548 \;=\; 22 \times 24 + 20, \text{ and } 0 \leq 20 < 24,$$

so the quotient is 22 and remainder is 20. That is, $\text{rem}(548, 24) = 20$. □

The ideas work just as well if a is negative. The divisor b must be positive, but a often is not.

Example 2. If $a = -548$ and $b = 24$, then $\frac{a}{b} = -23 + \frac{4}{24}$:

$$-548 \;=\; -23 \times 24 + 4, \text{ and } 0 \leq 4 < 24,$$

so quotient is -23 and remainder is 4. □

The Euclidean algorithm

We now consider a very efficient method of computing $\text{gcd}(a, b)$ which does not involve factorizing each of a, b into its product of primes representation. First, a useful fact, which allows us to simplify gcd calculations.

Theorem 7.4 *For $a > b > 0$, $\text{gcd}(a, b) = \text{gcd}(\text{rem}(a, b), b)$.*

To use this result to compute $\text{gcd}(a, b)$, with $a > b$, we first replace a by its remainder on division by b—the gcd will be the same. Let $a' = \text{rem}(a, b)$, we have $\text{gcd}(a', b) = \text{gcd}(a, b)$, with $b > a'$. Notice that the new pair of numbers (b, a') is "smaller" than the original pair (a, b). We can repeat this step by computing $b' = \text{rem}(b, a')$, and our new pair will be (a', b') with $\text{gcd}(a', b') = \text{gcd}(a', b) = \text{gcd}(a, b)$. By continuing in this way, at each step replacing the larger one by its remainder on division by the smaller one, eventually we get a remainder of zero[2]. In the final case, we have $\text{gcd}(c, 0)$ to compute, and the answer is c.

Example 3. To compute $\text{gcd}(10362, 12397)$ we divide the smaller into the bigger initially and replace the bigger by its remainder:

$$12397 \;=\; 10362 \times 1 + 2035.$$

Replace 12397 by 2035 (since $\text{gcd}(12397, 10362) = \text{gcd}(2035, 10362)$) and repeat. Next, we calculate:

$$10362 \;=\; 2035 \times 5 + 187.$$

So $\text{gcd}(2035, 10362) = \text{gcd}(2035, 187)$. Again,

$$2035 \;=\; 187 \times 10 + 165.$$

[2]This follows from the least integer principle since the remainders are going down each time, and are never negative.

So $\gcd(2035, 187) = \gcd(165, 187)$. And again:

$$187 \;=\; 165 \times 1 + 22.$$

So $\gcd(165, 187) = \gcd(165, 22)$. Once more:

$$165 \;=\; 22 \times 7 + 11.$$

So $\gcd(165, 22) = \gcd(11, 22)$. And again:

$$22 \;=\; 11 \times 2 + 0.$$

Now, $\gcd(11, 22) = \gcd(11, 0) = 11$, but stringing together all of our equalities, each gcd in this list is equal to the others, so the original gcd must be 11. Thus

$$\gcd(10362, 12397) = 11.$$

In general, the setting out can be streamlined as follows:

$$\begin{aligned} 12397 &= 10362 \times 1 + 2035 \\ 10362 &= 2035 \times 5 + 187 \\ 2035 &= 187 \times 10 + 165 \\ 187 &= 165 \times 1 + 22 \\ 165 &= 22 \times 7 + 11 \\ 22 &= 11 \times 2 + 0 \end{aligned}$$

and we simply read off the gcd from the line immediately before the zero remainder: 11. □

The method of this example is easily generalized, and can be written down formally:

Euclidean algorithm

1. Let $a > b$ and set $r_0 = a$, $r_1 = b$; let $n = 1$.
2. Find numbers q_n and $r_{n+1} < r_n$ such that
$$r_{n-1} = r_n\, q_n + r_{n+1}.$$
3. **If** $r_{n+1} = 0$ then $\gcd(a, b) = r_n$; **otherwise** increment $n := n + 1$ and return to Step 2,

Note, at the first iteration, one is simply writing $a = b\,q + r$, with $a = r_0$, $b = r_1$ and $r = r_2$; the iterative steps are justified by the remainder theorem, and keep repeating until the algorithm stops. In essence, Theorem 7.4 says that $\gcd(r_{n-1}, r_n) = \gcd(r_n, r_{n+1})$, so the algorithm is mathematically correct.

Perhaps more extraordinary is that we can rewrite the entire computation in the Euclidean algorithm to express $\gcd(a, b)$ as a certain integer combination of a and b. This will let us solve special kinds of linear equations: *linear Diophantine equations*.

Example 4. Let's rewrite the equations in our computation of $\gcd(12397, 10362)$ to have the remainders on the left:

$$\begin{aligned} 2035 &= 12397 - 10362 \times 1 \\ 187 &= 10362 - 2035 \times 5 \\ 165 &= 2035 - 187 \times 10 \\ 22 &= 187 - 165 \times 1 \\ 11 &= 165 - 22 \times 7 \end{aligned}$$

The final equation expresses 11 in terms of 22 and 165. But the fourth equation tells us how to write 22 in terms of 165 and 187, so we can eliminate 22 from the final equation to get 11 in terms of 165 and 187. Then, we can use the third equation to eliminate 165, resulting in an expression for 11 in terms of 187 and 2035. Continuing with second and then first equations we will end up with integers x, y such that

$$11 = 12397\,x + 10362\,y.$$

Let's do this in full. We have,

$$\begin{aligned}
11 &= 165 - 22 \times 7 \\
&= 165 - (187 - 165 \times 1)7 \text{ from previous line} \\
&= 165 \times 8 - 187 \times 7 \\
&= (2035 - 187 \times 10) \times 8 - 187 \times 7 \\
&= 2035 \times 8 - 187 \times 87 \\
&= 2035 \times 8 - (10362 - 2035 \times 5)87 \\
&= 2035 \times 443 - 10362 \times 87 \\
&= (12397 - 10362 \times 1)443 - 10362 \times 87 \\
&= 12397 \times 443 - 10362 \times 530.
\end{aligned}$$

So that in this case we've shown

$$11 = 12397 \times x + 10362 \times y$$

where $x = 443$ and $y = -530$. □

Because the Euclidean algorithm followed by the back-substitution method just discussed can be performed for any a, b, the general method proves:

Theorem 7.5 (Bezout's Theorem) *For any $a, b \in \mathbb{N}$ there are integers x and y such that*

$$a\,x + b\,y = \text{gcd}(a, b).$$

We will use Bezout's theorem to solve linear Diophantine equations.

Proof of the remainder theorem

To prove Theorem 7.3 we need to dust–off the *least integer principle*. Recall that this asserts the plausible fact that every non-empty set of positive integers has a smallest element.

Now, we can get on with the proof, although some care needs to be taken to deal with all cases correctly!

First of all, suppose that $a = 0$, then the theorem holds with $q = r = 0$. If $b|a$ then $a = b\,q$ for some q, and the theorem obviously holds.

Next suppose that b does not divide a and $a > 0$. Then $a - bq \neq 0$ for **all** $q \in \mathbb{Z}$. We show the least element of the following set R of natural numbers contains a remainder r with the desired properties:

$$R = \{a - bq \mid q \geq 0 \text{ and } a - b\,q > 0\}.$$

Clearly $a \in R$ (use $q = 0$) so R is not empty. By the least integer principle, R has a smallest element r. Clearly, $r > 0$. Since $r \in R$, there is $q_0 \geq 0$ such that $r = a - b\,q_0$, so $a = b\,q_0 + r$. We need to show that $r < b$, and that q and r are unique.

If $r = b$ then $b|a$, and we have dealt with this case already[3]. If $r > b$, then $r = b + r'$ for some $r' > 0$, and then $a = b\,q_0 + b + r' = b\,(q_0 + 1) + r'$. We'd then have $r' = a - b(q_0 + 1) > 0$, so $r' \in R$ and $r' < r$, contradicting the minimality of r, so $r < b$.

The proof of uniqueness is similar: suppose that also $a = b\,q' + r'$ with $0 \leq r' < b$. Then $r' \in R$, so necessarily

$$0 \leq r' - r < b$$

(remember that r is the minimal member of R). However, we also know that

$$r' - r = (a - b\,q') - (a - b\,q) = b\,(q - q').$$

Putting these facts together, we have

$$0 \leq b\,(q - q') < b,$$

so (upon division by b) the only possible value of $q - q'$ is 0. Thus $q = q'$ and $r = r'$.

Finally, we consider the case $a < 0$ and b does not divide a. Then, let $a' = -a$, so $a' > 0$ and by our work above there are unique q', r' such that $a' = b\,q' + r'$ and $0 < r' < b$. Then,

$$a = -a' = -(b\,q' + r') = b\,(-q') - r' = b\,(-1 - q') + (b - r').$$

We take $q = -1 - q'$ and $r = b - r'$. This is the last case, and the proof is complete. □

Proofs of the gcd/lcm theorems

First of all, we need a technical result.

Theorem 7.6 *Suppose that $a|m$ and $b|m$. Then* $\text{lcm}(a, b)|m$.

Proof: Since m is a multiple of both a and b, it follows that $\text{lcm}(a, b) \leq m$. Therefore, there are positive integers q, r with $r < \text{lcm}(a, b)$ such that

$$m = q\,\text{lcm}(a, b) + r.$$

If $r = 0$, there is no more work to do. We will assume that $r > 0$, and derive a contradiction. Now, since $a|m$ and $a|\text{lcm}(a, b)$ there are integers p_1, p_2 such that

$$m = p_1\,a \text{ and } \text{lcm}(a, b) = p_2\,a.$$

Thus,

$$r = m - q\,\text{lcm}(a, b) = p_1\,a - q\,(p_2\,a) = (p_1 - q\,p_2)\,a,$$

so $a|r$. A similar argument shows that $b|r$. Thus, r is a common multiple of a and b, so $r \geq \text{lcm}(a, b)$. This is a contradiction. Therefore, the only possibility is that $r = 0$, so that

$$m = q\,\text{lcm}(a, b) + 0 = q\,\text{lcm}(a, b);$$

that is $\text{lcm}(a, b)|m$. □

Now:

[3] You should check that you understand why $b = r$ implies that $b|a$.

Proof of Theorem 7.2: We prove this in two steps. First of all, since $\gcd(a,b)$ divides both a and b, there are p_1, p_2 such that $a = p_1 \gcd(a,b)$ and $b = p_2 \gcd(a,b)$. Then

$$p_1\, b = p_1\, p_2\, \gcd(a,b) = p_2\, p_1\, \gcd(a,b) = p_2\, a,$$

so $p_1\, p_2\, \gcd(a,b)$ is a multiple of both a and b. It follows that $\operatorname{lcm}(a,b) \le p_1\, p_2\, \gcd(a,b)$ and hence

$$\gcd(a,b)\, \operatorname{lcm}(a,b) \le \gcd(a,b)\, p_1\, p_2\, \gcd(a,b) = (p_1\, \gcd(a,b))\, (p_2\, \gcd(a,b)) = a\, b.$$

The proof will be complete if we can also derive the opposite inequality. Clearly, $a\, b$ is a multiple of both a and b, so by Theorem 7.6 there is an integer d such that

$$d\, \operatorname{lcm}(a,b) = a\, b.$$

In particular, both $q_1 = \frac{\operatorname{lcm}(a,b)}{a}$ and $q_2 = \frac{\operatorname{lcm}(a,b)}{b}$ are integers, so $a = d\, q_2$ and $b = d\, q_1$. This shows that d divides both a and b, so $\gcd(a,b) \ge d$. In particular,

$$\gcd(a,b)\, \operatorname{lcm}(a,b) \ge d\, \operatorname{lcm}(a,b) = a\, b.$$

Since we have proved inequalities in both directions, the theorem is true. □

Proof of Theorem 7.4: Let $d = \gcd(\operatorname{rem}(a,b), b)$, $r = \operatorname{rem}(a,b)$, and let $a = b\, q + r$. Clearly, $\gcd(a,b)$ divides both a and b, so there are integers p_1, p_2 such that

$$a = p_1\, \gcd(a,b) \text{ and } b = p_2\, \gcd(a,b).$$

Then,

$$r = a - b\, q = p_1\, \gcd(a,b) - p_2\, \gcd(a,b)\, q = (p_1 - p_2\, q)\, \gcd(a,b)$$

so also $\gcd(a,b) | r$. Since $\gcd(a,b)$ divides both b and r, we have

$$\gcd(a,b) \le \gcd(r,b) = \gcd(\operatorname{rem}(a,b), b) = d.$$

On the other hand, $d|r$ and $d|b$ so d is also a divisor of $b\, q + r = a$. Consequently,

$$d \le \gcd(a,b).$$

Taken together, we have proved that $d = \gcd(a,b)$. □

(7.3) Linear Diophantine equations

Definition. A **Diophantine equation** is an equation in which the unknowns and coefficients are all integers. □

The most famous examples have the form

$$x^n + y^n \;=\; z^n,$$

where n is a fixed natural number and $x, y, z > 0$ are integers. It was recently shown that this equation has no solutions x, y, z if $n > 2$. (If $n = 2$, there are many, e.g. $x = 3, y = 4, z = 5$.) This is Fermat's Last Theorem (although Fermat probably didn't prove it).

We are interested in linear Diophantine equations in two variables:

$$a\, x + b\, y \;=\; k,$$

where a, b, k are given and we solve for x, y.

Example 1. Does $21\,x + 35\,y \;=\; 12$ have any integer solutions? Let us suppose that it does, and observe that 7 divides both 21 and 35. So if there is a solution to this equation x_0, y_0, then because $7|(21\,x_0 + 35\,y_0)$, it must also be a factor of 12. It is not, so no solution to this equation can exist. □

We can generalize this reasoning to prove:

Theorem 7.7 *$a\,x_0 + b\,y_0 = k$ has integer solutions if and only if* $\gcd(a, b)|k$.

Proof: ($\Rightarrow$) Suppose that $a\,x_0 + b\,y_0 = k$. Since $\gcd(a, b)|a$ and $\gcd(a, b)|b$ we can write

$$a = p_1 \gcd(a, b) \text{ and } b = p_2 \gcd(a, b).$$

Thus,

$$k = a\,x_0 + b\,y_0 = p_1 \gcd(a, b)\,x_0 + p_2 \gcd(a, b)\,y_0 = (p_1\,x_0 + p_2\,y_0) \gcd(a, b),$$

so $\gcd(a, b)|k$.
($\Leftarrow$) Suppose that $k = d \gcd(a, b)$. By Bezout's Theorem, there are integers x and y such that

$$a\,x + b\,y = \gcd(a, b).$$

Then, using $x_p = x\,d$ and $y_p = y\,d$ we have

$$a\,x_p + b\,y_p = a\,x\,d + b\,y\,d = (a\,x + b\,y)\,d = \gcd(a, b)\,d = k,$$

so the theorem is proved. □

The argument in the proof also shows us how to construct solutions; let us see how it works:

Example 2. Find integers x, y such that

$$10362\,x + 12397\,y \;=\; 33.$$

We saw earlier using the Euclidean algorithm that $\gcd(10362, 12397) = 11$, and $11|33$, so solutions will exist. In fact we also found using back-substitution that

$$10362 \times (-530) + 12397 \times 443 \;=\; 11.$$

Multiplying -530 and 443 by $3 = \frac{33}{11}$ gives a solution to the given equation:

$$x = -1590, \;\; y = 1329.$$

□

How do we find **all** solutions to a linear Diophantine equation? Notice that if x_0, y_0 is one solution and x_1, y_1 is a second solution, then by subtraction,

$$\begin{array}{rrcrcl} & a\,x_1 & + & b\,y_1 & = & k \\ - & a\,x_0 & + & b\,y_0 & = & k \\ \hline & a\,(x_1 - x_0) & + & b\,(y_1 - y_0) & = & 0 \end{array}$$

so

$$a(x_1 - x_0) = -b(y_1 - y_0).$$

Thus, our problem is equivalent to finding all solutions to

$$a\,x + b\,y = 0$$

since the full solution can then be recovered as $x_0 + x, y_0 + y$. The problem is solved by the following theorem[4]:

[4]The proof shows why we need the concept of lcm.

Theorem 7.8 *Let $a, b > 0$. All integer solution to $a\,x + b\,y = 0$ have the form*

$$x = t\,\frac{b}{\gcd(a,b)},\quad y = -t\,\frac{a}{\gcd(a,b)},\quad t \in \mathbb{Z}.$$

Proof: Assume that $a\,x + b\,y = 0$. If $a\,x = 0$, then the result follows by letting $t = 0$. We will assume that[5] $a\,x > 0$. Put $m = a\,x$. Then m is a multiple of a, and since $m = a\,x = -b\,y$, m is also a multiple of y. By Theorem 7.6, $\text{lcm}(a,b)|m$. Thus, $m = t\,\text{lcm}(a,b)$ for an integer t. That is,

$$x = \frac{a\,x}{a} = \frac{m}{a} = t\,\frac{\text{lcm}(a,b)}{a} = t\,\frac{a\,b}{\gcd(a,b)}\,\frac{1}{a} = t\,\frac{b}{\gcd(a,b)}$$

(by Theorem 7.2). The expression for y follows immediately, since $y = \frac{-a\,x}{b}$. □

Now, we can write down an algorithm.

Solving linear Diophantine equations

Let the equation be $a\,x + b\,y = k$.

1. Use the Euclidean algorithm to find $\gcd(a,b)$.
2. **If** $\gcd(a,b)|k$ then use back-substitution through the working of the Euclidean algorithm to find integers x_0, y_0 such that
$$a\,x_0 + b\,y_0 = \gcd(a,b);$$
otherwise, the equation has no integer solutions, so **STOP.**
3. The general solution is
$$x = x_0\,\frac{k}{\gcd(a,b)} + t\,\frac{b}{\gcd(a,b)},\quad y = y_0\,\frac{k}{\gcd(a,b)} - t\,\frac{a}{\gcd(a,b)}$$
for any $t \in \mathbb{Z}$.

This says that any choice of $t \in \mathbb{Z}$ gives a solution to the original equation, and also that every possible solution arises in this way. The situation is very like what happens with the general solution of a system of linear equations where there are free variables.

Example 3. In solving

$$10362\,x + 12397\,y \;=\; 33,$$

we already have the particular solution

$$x_p \;=\; -1590,\;\; y_p \;=\; 1329.$$

We can now write down the general solution.

$$\begin{aligned} x &= x_p + t \cdot \frac{b}{\gcd(a,b)} \\ &= -1590 + \frac{12397}{11}\,t \\ &= -1590 + 1127\,t \end{aligned}$$

[5] Otherwise, let $m = b\,y > 0$, and a similar proof goes through.

and

$$\begin{aligned} y &= y_p - t \cdot \frac{a}{\gcd(a,b)} \\ &= 1329 - \frac{10362}{11}\, t \\ &= 1329 - 942\, t. \end{aligned}$$

So general solution is

$$x = -1590 + 1127\, t, \;\; y = 1329 - 942\, t, \;\; t \in \mathbb{Z}.$$

□

Further applications of divisibility

Bezout's Theorem is a very useful fact. It allows us to show things like the following.

Theorem 7.9 *Suppose p is a prime and $a, b \in \mathbb{N}$. If $p|(a\,b)$, then either $p|a$ or $p|b$.*

Proof: Suppose $p|(a\,b)$. Of course $\gcd(a,p)|p$, so since p is prime, either (i) $\gcd(a,p) = p$ or (ii) $\gcd(a,p) = 1$.

(i) If $\gcd(a,p) = p$, then $p|a$ by definition.

(ii) If $\gcd(a,p) = 1$ then by Bezout's Theorem, $a\,x + p\,y = 1$ for some integers x, y. So multiplying both sides by b, $a\,b\,x + b\,p\,y = b$. But $p|(a\,b)$, so $a\,b = p\,q$ for some natural q. Hence

$$b = (a\,b)\,x + b\,p\,y = (p\,q)\,x + b\,p\,y = p\,(q\,x + b\,y)$$

and so $p|b$. □

This theorem can be used to prove that the prime factorization of an integer is unique.

Let's finish this subsection with an elegant proof (very like the original ancient Greek proof) that $\sqrt{2}$ is irrational, that is, not a fraction.

Proof: Suppose instead that $\sqrt{2}$ is rational, with $\sqrt{2} = \frac{p}{q}$. Assume that this fraction is in reduced form, so that any common factors in the top and bottom lines have been cancelled out. (This can always be done.) Then squaring both sides gives $2 = p^2/q^2$.

Hence $p^2 = 2\,q^2$, and so $2|p^2$. But 2 is prime, $2|p$ or $2|p$ by the last theorem. So $2|p$. Hence $p = 2\,k$ for some natural number k. So $2\,q^2 = p^2 = (2\,k)^2 = 4\,k^2$, so $q^2 = 2\,k^2$. Hence $2|q^2$ and so $2|q$ for the same reason that $2|p$.

So 2 is a common factor of p, q, contradicting our assumption that their common factors had been cancelled out. So $\sqrt{2}$ cannot be rational. □

(7.4) Modular arithmetic

In the Euclidean algorithm, we used the fact that replacing one number by its remainder when divided by the other gave the same answer, but also simplified the gcd calculation considerably. It also turns out that computing remainders of complicated expressions can be made much easier by replacing the numbers involved by their remainders.

Congruence modulo n

Definition. Let n be a natural number. We say integers a, b are **congruent modulo** n if they differ by a multiple of n. *Notation*:

$$a \equiv b \text{ (mod } n).$$

□

This is what happened with angles in the complex numbers section: two angles are essentially equal if they are "congruent modulo $2\,\pi$", or in degrees, congruent modulo 360.

The remainder theorem shows us that any integer is congruent to its remainder modulo n: since $a = n\,q + r$, $a - r = nq$, so by definition, $a \equiv r$ (mod n). The next fact tells us we can decide if two things are congruent by comparing their remainders.

Theorem 7.10 *For integers a, b and a natural number n, $a \equiv b$ (mod n) if and only if*

$$\text{rem}(a, n) = \text{rem}(b, n).$$

Proof: Dividing both a, b by n using the remainder theorem gives

$$a \;=\; q_1\, n + r_1, \;\; b \;=\; q_2\, n + r_2, \;\; 0 \le r_1, r_2 < n.$$

Assume $r_1 \ge r_2$ (w.l.o.g.!). Then

$$\begin{aligned} a - b &= (q_1 n + r_1) - (q_2 n + r_2) \\ &= (q_1 n - q_2 n) + (r_1 - r_2) \\ &= (q_1 - q_2) n + (r_1 - r_2). \end{aligned}$$

Here, $0 \le r_1 - r_2 < n$, so $r_1 - r_2$ must be *the* remainder when $a - b$ is divided by n. So $n|(a - b)$ if and only if $r_1 - r_2 = 0$; that is, $r_1 = r_2$. □

Shorthand notation: We'll sometimes write "a mod n" for rem(a, n) or "the remainder when a is divided by n". □

Example 1. Are 423 and 321 congruent modulo 17? Calculate 423 mod 17 and 321 mod 17 to confirm the theorem. Then

$$423 - 321 \;=\; 102 \;=\; 17 \times 6,$$

so by definition, $423 \equiv 321$ (mod 17). Note that

$$423 = 17 \times 24 + 15, \;\; 321 = 17 \times 18 + 15,$$

so rem$(423, 17) =$ rem$(321, 17)$, confirming the theorem in this case. □

The notion of equivalence modulo n has some convenient properties:

Theorem 7.11 *For any $n \in \mathbb{N}$ and $a, b, c \in \mathbb{Z}$,*

1. $a \equiv a$ *(mod n);*
2. $a \equiv b$ *(mod n) implies $b \equiv a$ (mod n);*
3. $a \equiv b$ *(mod n) and $b \equiv c$ (mod n) imply $a \equiv c$ (mod n).*

The proofs of these are pretty straightforward, in view of Theorem 7.10: two integers are congruent if and only if their remainders are the same. Together these conditions show that congruence modulo n is an *equivalence relation*: every integer is in a unique **congruence class**, defined by what its remainder is. Each such congruence class consists of all things congruent to that remainder (and hence to each other!)

The possible remainders modulo n are

$$0, 1, 2, \ldots, n-1$$

so there are n distinct such classes.

Modular arithmetic

It turns out to be quite a reasonable proposition to do arithmetic within equivalence classes. First of all, it is interesting to note that the basic operations of addition and multiplication are well-behaved.

Theorem 7.12 *For any $n \in \mathbb{N}$ and $a_1, a_2, b_1, b_2 \in \mathbb{Z}$, if $a_1 \equiv a_2$ (mod n) and $b_1 \equiv b_2$ (mod n), then*

1. $a_1 + b_1 \equiv a_2 + b_2$ *(mod n);*
2. $a_1\, b_1 \equiv a_2\, b_2$ *(mod n);*
3. $-a_1 \equiv -a_2$ *(mod n);*
4. $a_1^k \equiv a_2^k$ *(mod n) for any natural k.*

Proof: We show the first of these, leaving the others as exercises. If $a_1 \equiv a_2$ (mod n) and $b_1 \equiv b_2$ (mod n), then $a_1 - a_2 = k_1\, n$ and $b_1 - b_2 = k_2\, n$ for some integers k_1, k_2. So

$$\begin{aligned}(a_1 + b_1) - (a_2 + b_2) &= (a_1 - a_2) + (b_1 - b_2) \\ &= k_1 n + k_2 n \\ &= (k_1 + k_2) n,\end{aligned}$$

so by definition $a_1 + b_1 \equiv a_2 + b_2$ (mod n). □

These rules tell us that congruence modulo n behaves like equality: if we have some expression built out of integers and the operations on them, we can replace things by other things they're congruent to (*e.g.* their remainders) without changing what the overall expression is congruent to.

We can now show one of the rules for checking divisibility of 3 we gave earlier: **a (positive) integer n is divisible by 3 if and only if the sum of its digits is 3.**

Example 2. Is 23985 divisible by 3? Work out $2 + 3 + 9 + 8 + 5 = 27$. This is divisible by 3 (since $2 + 7 = 9$ is!). So 23985 is also. □

Proof of criterion for divisibility by 3: Write $n = n_1\, n_2 \cdots n_k$, where the n_i are the digits in the decimal representation. We note that $10 \equiv 1$ (mod 3). Then

$$\begin{aligned}n &= 10^{k-1} \times n_1 + 10^{k-2} \times n_2 + \cdots + 10^{k-k} \times n_k \\ &\equiv 1^{k-1} \times n_1 + 1^{k-2} \times n_2 + \cdots \\ &\quad + 1^{k-k} \times n_k \text{ (mod } 3) \\ &= n_1 + n_2 + \cdots n_k,\end{aligned}$$

so $n \equiv (n_1+n_2+\cdots+n_k) \pmod 3$. Therefore, $3|n$ if and only if $3|(n_1+n_2+\cdots n_k)$ (since both remainders will equal zero together). □

We can prove lots of other familiar facts quite easily also. For example: **the product of two odd numbers is odd**.

Proof: Let m, n be odd, so $m \equiv 1 \pmod 2$ and $n \equiv 1 \pmod 2$. Then

$$m\,n \;\equiv\; 1 \times 1 \pmod 2 \;=\; 1,$$

and so $m\,n$ is odd. □

Some of our earlier induction proofs about divisibility can now be superseded.

Example 3. We show $6^n - 1$ is divisible by 5 :

$$6^n - 1 \;\equiv\; 1^n - 1 \pmod 5 \;=\; 0,$$

so $5|(6^n - 1)$. □

Example 4. What is 13^{511} mod 7? To solve this problem, note that $13-2\times 7 = -1$, so $13 \equiv (-1) \pmod 7$. Thus,

$$13^{511} \equiv (-1)^{511} = (-1)^{2\times 255+1} = ((-1)^2)^{255}(-1) = (1)^{255} \times (-1) = (-1) \equiv 6 \pmod 7.$$

□

(7.5) The algebra of modular arithmetic

We know that we can replace the two numbers in a sum or product by their remainders and the result will be congruent to what we started with, and hence congruent to the *remainder* of what we started with. This suggests there is an underlying "algebra of modular arithmetic".

Example 1. Consider the following simple congruence:

$$2 \times 4 \;=\; 8 \;\equiv\; 3 \pmod 5.$$

It reflects the following fact: *any* integer with remainder of $2 \pmod 5$, when multiplied by *any* integer with remainder $4 \pmod 5$ gives an integer with remainder $3 \pmod 5$. Now, -8 has remainder 2 on division by 5 and 14 has remainder 4. So their product should have remainder 3. Sure enough:

$$(-8) \times 14 \;=\; -112 = (-23) \times 5 + 3 \;\equiv\; 3 \pmod 5.$$

□

This works the same way for all three operations $+, \times, -$, and for any n.

Definition. For an integer a, let $\overline{a}$ denote the **equivalence class of** a **modulo** n. That is,

$$\overline{a} = \{x \in \mathbb{Z} | a \equiv x \pmod n\}.$$

We write

$$\mathbb{Z}_n \;=\; \{\bar{0}, \bar{1}, \ldots, \overline{n-1}\}$$

and call $\mathbb{Z}_n$ the **integers modulo** n. (We use the bar notation to emphasize that elements of $\mathbb{Z}_n$ are not ordinary numbers, because the operations are not the ordinary ones!) □

Example 2. The equivalence class of 17 modulo 5 is

$$\overline{17} = \{x \in \mathbb{Z} | x \equiv 17 \text{ (mod } 5)\} = \{\ldots, -8, -3, 2, 7, 12, 17, 22, 27, 32, \ldots\}.$$

Its representative in $\mathbb{Z}_5 = \{\bar{0}, \bar{1}, \bar{2}, \bar{3}, \bar{4}\}$ is $\bar{2}$. □

The arithmetic operations on $\mathbb{Z}_n$ are defined as follows:

$$\begin{aligned} \bar{a} + \bar{b} &= \overline{\text{rem}(a+b, n)} \\ \bar{a}\bar{b} &= \overline{\text{rem}(ab, n)} \\ -\bar{a} &= \overline{\text{rem}(-a, n)}. \end{aligned}$$

They simply duplicate the ordinary properties of congruences:

- $a + b \equiv c$ (mod n) if and only if $\bar{a} + \bar{b} = \bar{c}$ in $\mathbb{Z}_n$, and
- $ab \equiv c$ (mod n) if and only if $\bar{a}\bar{b} = \bar{c}$ in $\mathbb{Z}_n$.

Example 3. The congruence

$$2 \times 4 \equiv 3 \text{ (mod } 5)$$

corresponds to the fact that in $\mathbb{Z}_5$, $\bar{2} \times \bar{4} = \bar{3}$. □

It turns out that all the usual identities of algebra work for $\mathbb{Z}_n$: addition and multiplication are both commutative and associative, and the distributive law works. Also, $\bar{0}$ acts like the number zero and we call it the zero element, and each element has an additive inverse. Also, $\bar{1}$ is an identity element for multiplication.

Invertibility in $\mathbb{Z}_n$

We would like to know when it is possible to solve equations like

$$a\,x \equiv b \text{ (mod } n).$$

If n is small enough, we can solve by replacing a, b by their remainders modulo n, computing all answers in $\mathbb{Z}_n$ for x, and then adding $t\,n$ (t an integer) to the result to get the general solution. However, it turns out that there is a much neater way. Motivated by the situation for matrices, the ideal would be to have

$$x \equiv a^{-1}\,b \text{ (mod } n)$$

for an appropriately defined a^{-1}.

Definition. We will say that $\overline{a} \in \mathbb{Z}_n$ is **invertible in** $\mathbb{Z}_n$ if there is an element $\overline{y} \in \mathbb{Z}_n$ such that

$$a\,y = y\,a \equiv 1 \text{ (mod } n).$$

Such a $\overline{y}$ (if it exists) is unique, and is the **inverse** of $\overline{a}$, so we write $\overline{y} = \overline{a}^{-1}$. □

Amazingly, the existence of inverses is probed via the gcd.

Definition. Numbers $a, b \in \mathbb{N}$ are called **relatively prime** if $\gcd(a, b) = 1$. □

For example, 51 and 5 are relatively prime, and any natural number is relatively prime to 1.

Theorem 7.13 *A non-zero $\overline{a} \in \mathbb{Z}_n$ is invertible in $\mathbb{Z}_n$ if and only if a and n are relatively prime.*

Proof: First of all, notice that $a\,y \equiv 1 \pmod{n}$ if and only if there is an integer $x \in \mathbb{Z}$ such that

$$n\,x + a\,y = 1.$$

By Theorem 7.7, this equation has solutions if and only if

$$\gcd(a, n) | 1.$$

That is, $\gcd(a, n)$ must equal 1. □

Note: Since all of $1, 2, 3, \ldots, n-1$ are relatively prime to n if and only if n is a prime, it follows that all non-zero elements of $\mathbb{Z}_n$ have an inverse if and only if n is a prime. That is, $\mathbb{Z}_p$ has exactly $(p-1)$ invertible elements. □

The proof of the theorem also indicates how to find the inverse of $\overline{a}$ in $\mathbb{Z}_n$.

Finding inverses in $\mathbb{Z}_n$

1. Use the Euclidean algorithm to calculate $\gcd(a, n)$.
2. **If** $\gcd(a, n) = 1$ use back-substitution to find integers x, y such that

   $$n\,x + a\,y = 1;$$

   otherwise $\overline{a}$ is not invertible modulo n.
3. Then $\overline{a}^{-1} = \overline{\text{rem}(y, n)}$.

Example 4. Find $\overline{17}^{-1}$ in $\mathbb{Z}_{43}$. First of all, 43 is prime so all inverses exist. We want to solve $17\,x \equiv 1 \pmod{43}$, that is, find integers x, y such that

$$43\,x + 17\,y \;=\; 1.$$

Applying the Euclidean algorithm gives $x = 2$ and $y = -5$ as a solution pair. So $y = -5$ solves the Diophantine equation. Now we have to find a remainder modulo 43 congruent to this: $\text{rem}(-5, 43)$. But $-5 = 43 \times (-1) + 38$, so $\text{rem}(-5, 43) = 38$. So $\overline{17}^{-1} = \overline{38}$. □

Exercise. Solve $17\,x \equiv 5 \pmod{43}$.

(7.6) Computing remainders and solving congruences

It turns out that certain kinds of congruences are really important for cryptography. Solving them rests on some really nice mathematics.

Fermat's little theorem and Euler's extension

Theorem 7.14 (Fermat's little theorem) *Let p be prime. Then $a^{p-1} \equiv 1$ (mod p) if p does not divide a.*

Proof: Since p is not a divisor of a, the number $b = \text{rem}(a, p) \neq 0$ and hence $\bar{a} = \bar{b}$ is invertible in $\mathbb{Z}_p$. Next, consider the elements

$$\{\overline{b}, \overline{2\,b}, \ldots, \overline{(p-1)\,b}\}$$

of $\mathbb{Z}_p$. Since p is prime, each of $1, 2, \ldots, (p-1)$ is relatively prime to p, so each is invertible. In particular, each $\overline{r\,b} \neq \overline{0}$ (since otherwise, we would have $\overline{b} = \overline{r}^{-1}\overline{0} = \overline{0}$—contradicting $b \neq 0$). Moreover, each

$\overline{r\,b}$ is invertible (with inverse $(\overline{b})^{-1}(\overline{r})^{-1}$), and these $(p-1)$ elements are all distinct (if $\overline{r\,b} = \overline{s\,b}$ then $\overline{r} = \overline{r\,bb}^{-1} = \overline{s\,bb}^{-1} = \overline{s}$). Taken together, these facts show that the collection

$$\{\overline{b}, \overline{2\,b}, \ldots, \overline{(p-1)\,b}\}$$

are exactly the $(p-1)$ invertible elements of $\mathbb{Z}_p$. Thus, by multiplying all of these elements together:

$$\overline{1} \times \overline{2} \times \cdots \times \overline{(p-1)} = \overline{b} \times \overline{2\,b} \times \cdots \times \overline{(p-1)\,b} = \overline{1} \times \overline{2} \times \cdots \times \overline{(p-1)} \times \overline{b}^{p-1}.$$

Now, simply multiply both sides by $\overline{(p-1)!}^{-1}$ to obtain,

$$\overline{1} = \overline{b}^{p-1} = \overline{a}^{p-1};$$

that is, $1 \equiv a^{p-1} \pmod{p}$. □

Example 1. We will compute the remainder when 2^{323} is divided by the prime 13. That is:

$$2^{323} \text{mod} 13.$$

To do this, we use Fermat's theorem to write: $2^{12} \equiv 1 \pmod{13}$. Then, we divide 323 by 12

$$323 \;=\; 12 \times 26 + 11.$$

so that

$$\begin{aligned} 2^{323} &= 2^{12\times 26+11} \\ &= (2^{12})^{26} \times 2^{11} \\ &\equiv 1^{26} \times 2^{11} \pmod{13} \\ &= 2^{11}. \end{aligned}$$

It remains to compute the remainder when this is divided by 13. There are many ways to do this. For example, $2^5 = 32 \equiv 6 \pmod{13}$, so

$$\begin{aligned} 2^{11} = (2^5)^2 \times 2 &\equiv 6^2 \times 2 \pmod{13} \\ &= 72 \\ &\equiv 7 \pmod{13}. \end{aligned}$$

In summary,

$$2^{323} \;\equiv\; 2^{11} \;\equiv\; 7 \pmod{13},$$

so the remainder is 7. □

In fact, Fermat's theorem is a special case of Euler's theorem (where the restriction on p being prime is relaxed). Both are useful for computing remainders. We will finish off by establishing enough notation to state Euler's theorem, and defer it's application to the next section.

Definition. The **Euler phi-function** $\phi(n)$ is defined for each $n \in \mathbb{N}$ as the number of natural numbers between 1 and n relatively prime to n. □

Note: From our work above, $\phi(n)$ is the number of invertible elements in $\mathbb{Z}_n$. □

Example 2. To evaluate $\phi(12)$, note that $1, 5, 7, 11$ are the only integers between 1 and 12 relatively prime to n. So $\phi(12) = 4$. □

Example 3. If p is prime then $\phi(p) = p - 1$. □

Theorem 7.15 (Euler's theorem) *Fix an integer n. If* $\gcd(a, n) = 1$ *then*

$$a^{\phi(n)} \equiv 1 \textit{ (mod } n\textit{)}.$$

Computing remainders

We can now apply some of our techniques to solving congruence problems.

Example 4. Let's work out the remainder of $7 \times 26 + 14$ on division by 5.

$$\begin{aligned} 7 \times 26 + 14 &\equiv 2 \times 1 + 4 \pmod 5 \\ &= 6 \\ &\equiv 1 \pmod 5. \end{aligned}$$

Here we have replaced 7, 26 and 14 by things they're congruent to modulo 5 (their remainders in this case), and by the earlier Theorem 7.12, we can be sure the result is congruent to what we started with. So we know $7 \times 26 + 14 \equiv 1 \pmod 5$. So $7 \times 26 + 14$ and 1 have the *same remainder* on division by 5. This is obviously 1! So the remainder of $7 \times 26 + 14$ on division by 5 is 1 and we don't have to calculate $7 \times 26 + 14$ to compute its remainder modulo 5. □

Example 5. Let's compute the remainder of 51^4 on division by 8. Again, we'll replace 51 by its remainder, and repeatedly replace things by their remainders until we get an answer. (Note that we haven't yet replaced the index in a power by its remainder!)

$$\begin{aligned} 51^4 &\equiv 3^4 \pmod 8 \\ &= 9^2 \\ &\equiv 1^2 \pmod 8 \\ &= 1 \end{aligned}$$

so the remainder is 1. Again, we can do all of this without having to compute 51^4 itself. □

Example 6. If the time is now 1PM, what will the time be in 625 hours?
Answer: 1PM is 13 hours after midnight. Adding 24 hours does not change the time, so we want the remainder (mod 24) of $13 + 625$.

$$\begin{aligned} 13 + 625 &= 13 + 25^2 \\ &\equiv 13 + 1^2 \pmod{24} \\ &= 14, \end{aligned}$$

so the time will be 14 hours after midnight, or 2PM. □

Fermat's little theorem and Euler's extension also provide useful tricks for evaluating remainders.

Example 7. Find the remainder when 7^{123} is divided by 10. To solve this problem we will use Euler's theorem. Note that $1, 3, 7, 9$ are relatively prime to 10, so $\phi(10) = 4$. Next, since $\gcd(7, 10) = 1$, Euler's theorem gives $7^4 = 7^{\phi(10)} \equiv 1 \pmod{10}$. So

$$\begin{aligned} 7^{123} &= 7^{4 \times 30 + 3} \\ &= (7^4)^{30} \times 7^3 \\ &\equiv 7^3 \pmod{10} \\ &\equiv (-3)^3 \pmod{10} \\ &= 9 \times (-3) \\ &\equiv (-1) \times (-3) \pmod{10} \\ &= 3, \end{aligned}$$

so the remainder is 3. □

Solving congruences

If n is not too large, we can fairly easily solve congruences modulo n ("equations" in which congruence modulo n is used instead of equality) by trying all possibilities. One special form is the following:

$$p(x) \equiv q(x) \pmod{n}.$$

Here $p(x)$ and $q(x)$ are polynomials.

Any such congruence can be rewritten (by subtracting $q(x)$ from both sides) as $p(x) - q(x) \equiv 0 \pmod{n}$, so we just consider things of the form $p(x) \equiv 0 \pmod{n}$.

The trick is that we only have to try out the remainders $0, 1, 2, \ldots n-1$, since other cases can be reduced to these, using Theorem 7.12. For instance, $58 \equiv 2 \pmod{7}$, so $p(58) \equiv p(2) \pmod{7}$.

Example 8. Find all *integer* solutions to $5x^2 + 5x + 2 \equiv 2x^2 + 3x \pmod{7}$.
Solution: Let $p(x) = (5x^2 + 5x + 2) - (2x^2 + 3x) = 3x^2 + 2x + 2$. Then

a	0	1	2	3	4	5	6
$p(a)$	2	7	18	35	58	87	122
rem$(p(a), 7)$	2	0	4	0	2	3	3

So $p(1) \equiv p(3) \equiv 0 \pmod{7}$. So $x = 1, 3$ are solutions.

In fact, any other integer k congruent to 1 or 3 is also a solution, since then $p(k) \equiv p(1) \pmod{7}$ or $p(k) \equiv p(3) \pmod{7}$. Moreover, any *other* k will be congruent to one of the *other* possible remainders and so *cannot* be a solution. So the general solution is

$$x = 1 + 7t \text{ or } 3 + 7t, \quad t \in \mathbb{Z}.$$

□

Via our work on Diophantine equations, we can even solve congruences of the general form

$$a\,x \equiv b \pmod{n}$$

This congruence has an integer solution x exactly when $n|(a\,x - b)$, *i.e.* when $a\,x - b = k\,n$ for some integer k. So we look for x, k such that $a\,x - k\,n = b$, or, letting $y = -k$,

$$a\,x + n\,y = b.$$

Since this is simply a linear Diophantine equation in two unknowns, we can solve it to find x—-provided that gcd$(a, n)|b$ (recall Theorem 7.7). (We generally discover whether there is a solution during the calculation.)

Example 9. Solve the congruence equation

$$30\,x \equiv 6 \pmod{108}.$$

We must find integers x, y for which

$$30\,x + 108\,y = 6.$$

Note that 6 divides both 30 and 108, so we can simplify this equation to

$$5\,x + 18\,y = 1.$$

(Obvious common factors can always be divided out of Diophantine equations in this way without changing the solution set.) Then

$$\begin{aligned} 18 &= 5 \times 3 + 3 \\ 5 &= 3 \times 1 + 2 \\ 3 &= 2 \times 1 + 1 \\ 2 &= 1 \times 2 + 0. \end{aligned}$$

(Obviously 1 is the gcd.) Back-substituting:

$$\begin{aligned} 1 &= 3 - 2 \times 1 \\ &= 3 - 1(5 - 3 \times 1) \\ &= 3 \times 2 - 5 \times 1 \\ &= 2(18 - 5 \times 3) - 5 \times 1 \\ &= 5 \times (-7) + 18 \times 2. \end{aligned}$$

So we have $5\,x + 18\,y = 1$, with $x = -7, y = 2$. By Theorem 7.8, the general solution is

$$x \;=\; -7 + t \times \frac{18}{1} \;=\; 18\,t - 7, t \in \mathbb{Z}.$$

□

Proof of Euler's theorem

The proof is essentially the same as the proof of Fermat's theorem; we just need to be a little more careful about which elements are invertible in $\mathbb{Z}_n$.

Proof: There are $\phi(n)$ remainders modulo n having gcd of 1 with n, say $n_1, n_2, \ldots, n_{\phi(n)}$. The invertible elements of $\mathbb{Z}_n$ are thus:

$$\{\bar{n}_1, \bar{n}_2, \ldots, \bar{n}_{\phi(n)}\}.$$

But $1 = \text{gcd}(a, n) = \text{gcd}(\text{rem}(a, n), n)$, so letting $r = \text{rem}(a, n)$, $\bar{r}$ is invertible in $\mathbb{Z}_n$. It follows that

$$\bar{n}_1\bar{r}, \bar{n}_2\bar{r}, \ldots, \bar{n}_{\phi(n)}\bar{r}$$

are all distinct (since if $\bar{n}_i\bar{r} = \bar{n}_j\bar{r}$, then multiplying by $\bar{r}^{-1}$ gives $\bar{n}_i = \bar{n}_j$). Also, each $\bar{n}_i\bar{r}$ is invertible in $\mathbb{Z}_n$, since it has inverse $\bar{n}_i^{-1}\bar{r}^{-1}$. So, since there are $\phi(n)$ of them, it must also be that

$$\{\bar{n}_1\bar{r}, \bar{n}_2\bar{r}, \ldots, \bar{n}_{\phi(n)}\bar{r}\}$$

are the invertible elements. Thus, the $\bar{n}_i$ and $\bar{n}_j\bar{r}$'s can be matched up and

$$\begin{aligned} \bar{n}_1\bar{n}_2\cdots\bar{n}_{\phi(n)} &= (\bar{n}_1\bar{r})(\bar{n}_2\bar{r})\cdots(\bar{n}_{\phi(n)}\bar{r}) \\ &= (\bar{n}_1\bar{n}_2\cdots\bar{n}_{\phi(n)})\bar{r}^{\phi(n)}. \end{aligned}$$

Multiplying through by $\bar{n}_1^{-1}\bar{n}_2^{-1}\cdots\bar{n}_{\phi(n)}^{-1}$ gives $\bar{r}^{\phi(n)} = \bar{1}$ in $\mathbb{Z}_n$. Hence $r^{\phi(n)} \equiv 1 \pmod{n}$, so $a^{\phi(n)} \equiv 1 \pmod{n}$ since $a \equiv r \pmod{n}$. □

VIII ∘ Cryptography

Cryptosystems guarantee (?) the secure transmission of information, on the internet for example. The ideas of modern cryptography go back to the Second World War (*e.g.* the German Enigma code). Here we look at three cryptosystems. The first two are too simple to be practical but give an idea of the methods used, while the third is in use today.

(8.1) The shift cipher

Let's keep things simple and consider letter-by-letter encryption methods. We'll ignore punctuation and spacing, so there are just 26 characters to encrypt.

The letters in words are first converted to elements of $\mathbb{Z}_{26}$:

$$A \leftrightarrow \overline{0},\ B \leftrightarrow \overline{1},\ C \leftrightarrow \overline{2}, \ldots, Z \leftrightarrow \overline{25}.$$

Then, an **encryption function** is applied to the letters of a message (which are by now in numerical form). Choose a fixed non-zero $b \in \mathbb{Z}_{26}$. Then for all $x \in \mathbb{Z}_{26}$, a typical encryption function is

$$E(x) \ = \ x + b.$$

(Once encrypted, numbers can be translated back into letters, giving an encrypted message, although the message can just as well be left as a stream of elements of $\mathbb{Z}_{26}$.)

The receiver must then decrypt the message. For this, a **decryption function** $D(y)$ must be known. The key feature of a decryption function is that it "undoes" the effect of $E(x)$, so that $D(E(x)) = x$ for all $x \in \mathbb{Z}_{26}$.

For the shift cipher, this is achieved by adding $-b$ to the received message (viewed as elements of $\mathbb{Z}_{26}$). Thus we define $D(y) = y - b$. Then

$$D(E(x)) \ = \ E(x) - b \ = \ x + b - b \ = \ x,$$

so it does work!

Example 1. Let us use the simple shift cipher with $b = \overline{8}$; the effect of encryption is to shift forward by eight letters, so

$$E(x) = x + \overline{8}.$$

Let's encrypt the word "password".

- first convert to elements of $\mathbb{Z}_{26}$:
$$\overline{15}, \overline{0}, \overline{18}, \overline{18}, \overline{22}, \overline{14}, \overline{17}, \overline{3};$$
- then apply the encryption formula to each of these: add 8 (mod 26):
$$\overline{23}, \overline{8}, \overline{0}, \overline{0}, \overline{4}, \overline{22}, \overline{25}, \overline{11}.$$

(For the record, this translates into “xiaaewzl”.)

The receiver will know b, and will therefore just add $-b$ to each element of $\mathbb{Z}_{26}$ in the incoming stream. In $\mathbb{Z}_{26}$, $\overline{8} + \overline{18} = \overline{0}$, so $-b = \overline{18}$. So the decryption function in our example is

$$D(y) = y + \overline{18}.$$

From here, the receiver will do the following:

- apply the decryption function to each incoming “letter” (so $D(\overline{23}) \to \overline{23} + \overline{18} = \overline{15}$, (since $41 \equiv 15 \pmod{26}$):
$$\overline{15}, \overline{0}, \overline{18}, \overline{18}, \overline{22}, \overline{14}, \overline{17}, \overline{3}$$
- then transform back to letters of the alphabet:
$$\text{“password”}.$$

□

The simple shift cipher is a very weak form of encryption: there are only 25 possible encryption functions (we exclude $b = 0$). Such ciphers are easily cracked by trial and error (try each possible $-b \in \mathbb{Z}_{26}$ until an intelligible message is produced). Also, if **just one** original and encrypted letter pair is known, then b can be determined easily by subtraction and the cipher broken.

Example 2. Suppose that a simple shift cipher is used, and “d” is encrypted to “a”. Decrypt the message “irzh”. To solve this problem, we must first find the decrypt function: $D(y) = y - b$. We know that “d” is identified with the number 3, and “a” is identified with 0, so we have

$$3 = D(0) = 0 - b$$

since D must convert the code for “a” back into the code for “d”. Thus, $b = -3$ and $D(y) = y + 3$. To decrypt, note that the message has code:

$$8\ 17\ 25\ 7$$

which becomes

$$11\ 20\ 2\ 10$$

upon adding 3 to each entry $\pmod{26}$. This is the message “luck”. □

(8.2) The affine cipher

This is one level more complex than the shift cipher: we multiply by a constant and then add the shift factor (all modulo 26 as before).

The encryption function maps $\mathbb{Z}_{26} \to \mathbb{Z}_{26}$:

$$E(x) \ = \ a\,x + b, \ \ a, b \in \mathbb{Z}_{26}.$$

(The shift cipher has $a = \overline{1}$.)

Example 1. Let $E(x) = \overline{3}\,x + \overline{11}$ and encrypt: “hello”.
Solution: First h $\leftrightarrow \overline{7}$. Then

$$E(\overline{7}) \ = \ \overline{3} \times \overline{7} + \overline{11} \ = \ \overline{6}.$$

Continuing, we obtain $\overline{6}, \overline{23}, \overline{18}, \overline{18}, \overline{1}$. (For the record, this translates as the encrypted message "gxssb".) □

We need to be careful with the affine cipher since there may be no decryption function if $E(x)$ is not $1:1$:

Example 2. Suppose we use $E(x) \equiv 4\,x + 7 \pmod{26}$. Then "a" is identified with $\overline{0}$, so encodes to $\overline{7}$, but also "n" identifies with $\overline{13}$ so encodes to

$$E(13) \equiv 4 \times 13 + 7 \pmod{26} = 7.$$

□

In general, if two letters encrypt to the same thing, there is uncertainty in the decryption. This situation should be avoided, and luckily the mathematics of inverses modulo n tells us how:

Theorem 8.1 *The encryption function $E(x) = a\,x + b$ has a decryption function $D(y)$ if and only if a^{-1} exists in $\mathbb{Z}_{26}$. If a^{-1} does exist, the decryption function is $D(y) = a^{-1}(y - b)$.*

Proof: If a has no inverse, then $a\,x$ is never $\overline{1}$ for any $x \in \mathbb{Z}_{26}$, so the set $\{a\,x \mid x \in \mathbb{Z}_{26}\}$ has fewer than 26 elements. But there are 26 possibilities for x, so there must be unequal $x, x' \in \mathbb{Z}_{26}$ for which $a\,x = a\,x'$. Then also $E(x) = a\,x + b = a\,x' + b = E(x')$, so no decryption function will be able to distinguish between $E(x)$ and $E(x')$.

On the other hand, if a has an inverse, then letting $D(y) = a^{-1}(y - b)$, we see that

$$\begin{aligned} D(E(x)) &= a^{-1}(E(x) - b) \\ &= a^{-1}(ax + b - b) \\ &= a^{-1}ax \\ &= x, \end{aligned}$$

as required. □

This theorem generalizes to cases where we have other than 26 characters, say m: just work in $\mathbb{Z}_m$ rather than $\mathbb{Z}_{26}$.

Definition. We call (a, b) the **key** of the encryption scheme based on $E(x) = ax + b$. □

Example 3. It is know that affine cipher has key $(5, 18)$. Find the message which encrypts to be "spvoe". To solve this problem we need to find the decrypt key. The encryption function is given as

$$E(x) = 5\,x + 18 \pmod{26}$$

so the decrypt function will be

$$D(y) = (\bar{5})^{-1}(y - 18) \pmod{26}.$$

So, we need to solve $z = (\bar{5})^{-1}$; this can be done by solving the Diophantine equation

$$5\,z + 26\,n = 1.$$

Fortunately, the first step of the Euclidean algorithm produces

$$26 = 5 \times 5 + 1$$

so $1 = 5\,(-5) + 26$ and $(\bar{5})^{-1} \equiv -5 \equiv 21 (\text{mod } 26)$. The most convenient representation of the D is:

$$D(y) = (-5)\,(y - 18) \equiv (-5)\,(y + 8) \equiv (-40) - 5\,y \equiv 12 - 5\,y \pmod{26}.$$

Now, "spvoe" has code $18\,15\,21\,14\,4$ and

$$D(18) = 12 - 5\,(18) = -78 \equiv 0 \pmod{26},$$

so "s" decodes to "a". Similarly,

$$D(15) = 15,\ D(21) = 11\ , D(14) = 20 \text{ and } D(4) = 18.$$

The decoded message is thus $0\,15\,11\,20\,18$ or "aplus". □

It is not hard to show that any choice of (a, b) with a invertible will give a different $E(x)$.

By Theorem 7.13, a is invertible in $\mathbb{Z}_m$ if and only if $\gcd(a, m) = 1$. In the $\mathbb{Z}_{26}$ case, the following are relatively prime to 26:

$$1, 3, 5, 7, 9, 11, 15, 17, 19, 21, 23, 25.$$

So $\phi(26) = 12$ and any of $a = \overline{1}, \overline{3}, \overline{5}, \ldots$ can be used in $E(x)$. Any value of b will do. But we exclude $a = 1, b = 0$. So altogether there are $12 \times 26 - 1 = 311$ different possible $E(x)$ functions. (For the shift cipher there were 25.)

Also, knowing only what one letter encrypts to is not enough to break the cryptosystem, although two *may* be enough. So overall, the affine cipher is a lot more secure than the shift cipher, but is still very insecure!

(8.3) The RSA cryptosystem

We finish with a discussion of a highly successful cryptosystem which is still much-used, the RSA cryptosystem. It is "public key", which means that the encryption function is known and usable by anyone but not the decryption method. One simple secret piece of information is needed to allow decryption. This is unlike the two previous cases, where $D(x)$ can be found if $E(x)$ is known.

First, note that Euler's theorem can be re-phrased as follows:

$$\text{if } a \in \mathbb{Z}_n \text{ is invertible, then } a^{\phi(n)} = \overline{1}.$$

This also tells us that $a^{-1} = a^{\phi(n)-1}$ for all invertible a in $\mathbb{Z}_n$.

The RSA cryptosystem is best described stage by stage.

Setup: Choose two large (*i.e.* ≈ 50 digits long!) prime numbers p and q.

Let $n = p\,q$. It can be shown that $\phi(n) = (p-1)(q-1)$.

Pick a natural number $a < \phi(n)$ and find $b < \phi(n)$ for which $ab \equiv 1 \pmod{\phi(n)}$. Since

$$a\,b - k\,\phi(n) = 1$$

for some k, b can be found (if it exists) via the Euclidean algorithm. If a suitable b doesn't exist, just pick a different a until you have one that works.

Now, remember the integers n, a, b. Make n, a public (for encryption) and keep b secret (it is needed for decryption—see below).

Encryption: represent the message as a sequence of elements of $\mathbb{Z}_n$, *e.g.* let each sequence of ten characters be represented by a different $x \in \mathbb{Z}_n$. Encrypt using the function

$$E(x) = x^a$$

over $\mathbb{Z}_n$ (will need to use repeated doubling or similar for quick computation).

Decryption: apply the decryption function $D(x) = x^b$ to a received message.

We need to show $D(x) = x^b$ really is a decryption function.

Remember that $a\,b \equiv 1 \pmod{\phi(n)}$, so $a\,b = k\phi(n) + 1$ for some integer $k > 0$. So

$$\begin{aligned} D(E(x)) &= E(x)^b \\ &= (x^a)^b \\ &= x^{ab} \\ &= x^{k\phi(n)+1} \\ &= (x^{\phi(n)})^k \cdot x \\ &= \overline{1}^k \times x \text{ by Euler} \\ &= x. \end{aligned}$$

Actually, there is a small hole in this line of reasoning: Euler's theorem may not apply to x—this will be the case if x is not invertible in $\mathbb{Z}_n$. Still, remember that of the $n = p\,q$ elements of $\mathbb{Z}_n$, $\phi(n) = (p-1)(q-1)$ are invertible. As a proportion, this is

$$\frac{(p-1)(q-1)}{pq} \;=\; (1 - 1/p)(1 - 1/q) \;\approx\; 1,$$

since p, q are **huge**. So an arbitrary $x \in \mathbb{Z}_n$ will be invertible **almost certainly**. Occasionally this won't be the case and $D(E(x))$ could be wrong, but almost never.

To break the cryptosystem, one needs to discover b from knowledge of a, n alone. This can be done if $\phi(n)$ can be found, since $a\,b \equiv 1 \pmod{\phi(n)}$. But to find $\phi(n) = (p-1)(q-1)$, one must discover p and q, where $pq = n$.

Thus the problem reduces to being able to factorize n into its large prime factors, which is **EXTREMELY** hard. So we can have a high degree of confidence that the cryptosystem cannot be broken using existing technology.

So strangely, modular arithmetic, prime numbers, and number theory, once thought to be completely useless, turn out to provide the security system for the whole information economy. Financial transactions worth trillions of dollars now depend on this "useless" stuff.

IX ∘ Extra topics

(9.1) Application: Least squares model fi tting

In science or statistics, experimentation or data collection can produce numerical data which it is desirable to model somehow. Generally speaking, one has a bunch of data points $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and would like to be able to find an approximate relationship between the components. For example, if each $\mathbf{v}_i = (x_i, y_i)$ we might like to "fit" a function of the form

$$y = m\,x + c$$

in such a way that the line is a good fit for the data points; that is, we'd like to choose m, c such that $y_i \approx m\,x_i + c$ for all of the data pairs (x_i, y_i). We can model this as a projection problem in $\mathbb{R}^n$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n), \mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^n$. Then each component of the vector equation $m\,\mathbf{x} + c\,\mathbf{1}$ has the form $m\,x_i + c$. If we put $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, then we can express our approximation problem as

$$\mathbf{y} \approx m\,\mathbf{x} + c\,\mathbf{1}.$$

We will use the method of projection to choose m and c so that the vector $\mathbf{y} - (m\,\mathbf{x} + c\,\mathbf{1})$ of *residual errors* has the shortest length possible. We need to solve the projection equations:

$$\begin{aligned} 0 &= [\mathbf{y} - (m\,\mathbf{x} + c\,\mathbf{1})] \cdot \mathbf{x} \\ 0 &= [\mathbf{y} - (m\,\mathbf{x} + c\,\mathbf{1})] \cdot \mathbf{1}. \end{aligned}$$

Example. Find the best fit line $y = m\,x + c$ to the pairs of (x, y) data: $(1, 1), (2, 2.8), (3, 4.2), (4, 5.2)$. The relevant vectors are:

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 1 \\ 2.8 \\ 4.2 \\ 5.2 \end{pmatrix}.$$

The equations to solve are:

$$\begin{aligned} 0 &= [(1, 2.8, 4.2, 5.2) - m\,(1, 2, 3, 4) - c\,(1, 1, 1, 1)] \cdot (1, 2, 3, 4) \\ 0 &= [(1, 2.8, 4.2, 5.2) - m\,(1, 2, 3, 4) - c\,(1, 1, 1, 1)] \cdot (1, 1, 1, 1), \end{aligned}$$

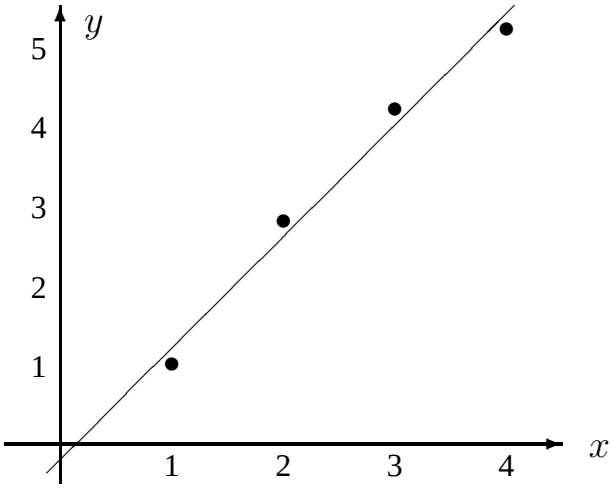
which are simply

$$\begin{aligned} m\,30 &+ c\,10 &&= 40 \\ m\,10 &+ c\,4 &&= 13.2. \end{aligned}$$

The solution is $m = 1.4, c = -0.2$, so the best fit line is

$$y = 1.4\,x - 0.2.$$

The data, and the best fit line, are depicted in the diagram below. □



(9.2) Matrices and linear transformations

Linear transformations are fundamental to the general theory of linear algebra, and are also extremely important in applications ranging from computer graphics to signal processing. They will be studied again in many mathematics and applications papers (especially MATH253).

Linear transformations T are functions which move vectors around with the special property that lines are mapped to lines. This means that if a line consists of points $\mathbf{x}$, then all of the image points $T(\mathbf{x})$ form a line. Basic examples of linear transformations are rotations and reflections, and we will see in this brief section that their actions can be easily computed via matrix algebra.

We will concentrate entirely on $\mathbb{R}^2$. Elements of $\mathbb{R}^2$ will be written as column vectors $\begin{pmatrix} x \\ y \end{pmatrix}$.

Matrix representation of linear transformations

Definition. With any 2×2 matrix A, there is an associated **linear transformation** T_A of $\mathbb{R}^2$, which takes a vector $\mathbf{x}$ in $\mathbb{R}^2$ to the vector $A\,\mathbf{x}$, also in $\mathbb{R}^2$. We call this the **linear transformation associated with** A.□

This definition makes perfect sense, since the laws of matrix algebra guarantee that lines are mapped to lines, and $T_A(\mathbf{x})$ is the product of a 2×2 matrix with a 2×1 vector, giving a 2×1 vector.

We can visualize the effect of linear transformations by examining their effect on the unit square, having corners at

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{p} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{r} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We will always have $\mathbf{0} \mapsto T(\mathbf{0}) = \mathbf{0}$, but also,

$$\mathbf{p} \mapsto \mathbf{p}' = T(\mathbf{p}), \mathbf{q} \mapsto \mathbf{q}' = T(\mathbf{q}), \mathbf{r} \mapsto \mathbf{r}' = T(\mathbf{r}).$$

The points represented by $\mathbf{0}, \mathbf{p}', \mathbf{q}', \mathbf{r}'$ can be drawn in another copy of the xy–plane. Since linear transformations map lines to lines, the line through $\mathbf{0}$ and $\mathbf{p}$ must be mapped to a line containing $\mathbf{0}$ and $\mathbf{p}'$. Similarly, the line through $\mathbf{p}$ and $\mathbf{q}$ gets mapped to a line through $\mathbf{p}'$ and $\mathbf{q}'$, and so on. Thus, the points $\mathbf{0}, \mathbf{p}', \mathbf{q}', \mathbf{r}'$ define the corners of geometrical object which is the image of the basic square under T.

Rotations

These rotate the points of $\mathbb{R}^2$ about the origin. The effect of a $45°$ rotation is depicted in Figure 9.1. To figure out the matrix for this transformation, we use a general formula. The matrix has the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

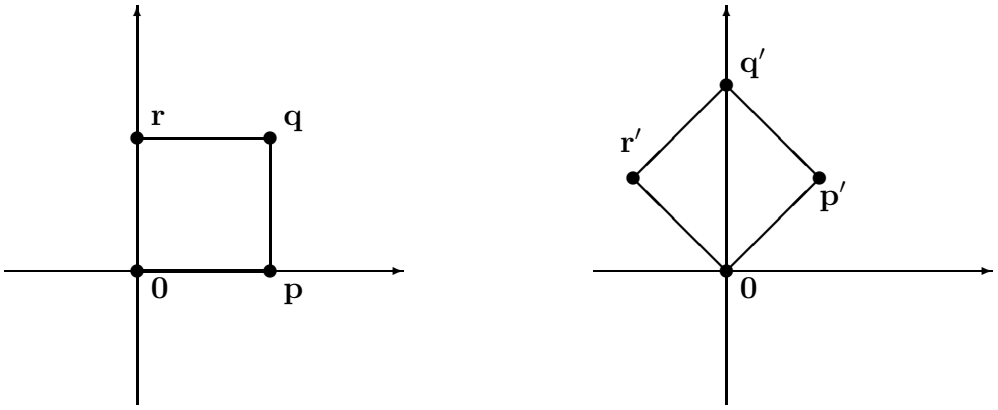


Figure 9.1: The effect on the unit square of a $45°$ anti-clockwise rotation.

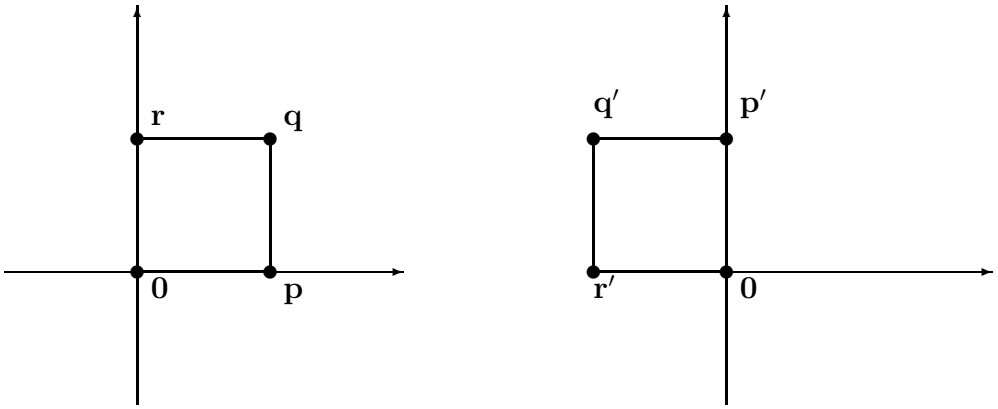


Figure 9.2: The effect on the unit square of a $90°$ clockwise rotation.

with θ the angle of rotation (measured anti-clockwise) about $\mathbf{0}$.

Example 1. So, a $45°$ anti-clockwise rotation has matrix:

$$\begin{pmatrix} \cos 45 & -\sin 45 \\ \sin 45 & \cos 45 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

□

Example 2. If $\theta = 90$ degrees, the matrix is

$$\begin{pmatrix} \cos 90 & -\sin 90 \\ \sin 90 & \cos 90 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

So if T is this rotation,

$$T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}$$

and

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

As always, $\mathbf{0} \mapsto \mathbf{0}$. This action is depicted in Figure 9.2. □

Reflections

These reflect the points of $\mathbb{R}^2$ about a particular line through $\mathbf{0}$, and change the orientation. The general matrix has the form

$$\begin{pmatrix} \cos\ 2\theta & \sin\ 2\theta \\ \sin\ 2\theta & -\cos\ 2\theta \end{pmatrix},$$

with θ the angle made with the positive x-axis, measured so that anti-clockwise is the positive direction.

Example 3. For example, if the line is $y = -x$, then $\theta = 135$ degrees, or $\frac{3\pi}{4}$ radians. So $2\theta = 270$ degrees, or -90 degrees. The matrix for the reflection is then

$$\begin{pmatrix} \cos\left(-90^\circ\right) & \sin\left(-90^\circ\right) \\ \sin\left(-90^\circ\right) & -\cos\left(-90^\circ\right) \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix},$$

since $\cos\left(-90^\circ\right) = \cos\ 90^\circ = 0$ and $\sin\left(-90^\circ\right) = -\sin\left(90^\circ\right) = -1$. So if T is the reflection, $T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ -x \end{pmatrix}$ and

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

As usual, $\mathbf{0} \mapsto \mathbf{0}$. □

Applying transformations in succession

Suppose that we first rotate by 45°, and then reflect in the y–axis. What kind of map do we obtain? It turns out that we work out the combined effect of several maps by matrix multiplication! Suppose that A and B are 2×2 matrices representing the maps T_A and T_B. Then,

$$T_B\left(T_A(\mathbf{x})\right) = B\ \left(T_A(x)\right) = B\ \left(A\,\mathbf{x}\right) = \left(B\,A\right)\mathbf{x},$$

so the combined transformation is represented by the matrix $B\,A$.

Example 4. Let us determine the overall effect of a reflection across the y-axis followed by a rotation by 45 degrees anti-clockwise. Since the y–axis is an anti-clockwise angle of 90° from the x–axis, the reflection matrix is

$$A = \begin{pmatrix} \cos 180^\circ & \sin 180^\circ \\ \sin 180^\circ & -\cos 180^\circ \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The rotation matrix is

$$B = \begin{pmatrix} \cos 45^\circ & \sin 45^\circ \\ \sin 45^\circ & \cos 45^\circ \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

So matrix for reflection followed by rotation is

$$B\,A = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

In general, a reflection followed by a rotation is a reflection through another axis, so we need to work out the angle θ. We must have:

$$\begin{pmatrix} \cos\ 2\theta & \sin\ 2\theta \\ \sin\ 2\theta & -\cos\ 2\theta \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

So $\cos\ 2\theta = -\frac{1}{\sqrt{2}}$ and $\sin\ 2\theta = -\frac{1}{\sqrt{2}}$. From the value of $\cos$, we find that $2\,\theta = \pm 135°$, and from the value of $\sin$ we have that $2\,\theta = -45$ or $-\ 135°$. Therefore, $2\,\theta = -135°$, and: *combined transformation is a reflection in the line* $\theta = 67.5\,°$ **clockwise** *from the* x*–axis*. (Note that a negative anti-clockwise angle is a positive clockwise angle.) □

Example 5. Let A and B be as in the previous example, but this time perform first the rotation T_B and then the reflection T_A. The matrix is

$$A\,B = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Setting this equal to a reflection matrix means

$$\begin{pmatrix} \cos\ 2\theta & \sin\ 2\theta \\ \sin\ 2\theta & -\cos\ 2\theta \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

So $\cos\ 2\theta = -\frac{1}{\sqrt{2}}$ and $\sin\ 2\theta = \frac{1}{\sqrt{2}}$. The only solution is $2\,\theta = 135°$, or $\theta = 67.5°$. □

(9.3) Eigenvectors

Many linear transformations allow some vectors to have a very special property: that their direction is unchanged by the application of the map.

Example 1. Consider the matrix A for a reflection T_A through the line $y = x$. This line makes an angle $\theta = 45°$ with the x–axis, so the matrix of rotation is

$$\begin{pmatrix} \cos 2\,\theta & \sin 2\,\theta \\ \sin 2\,\theta & -\cos 2\,\theta \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Geometrically, it is clear that any points lying on the line $y = x$ are not changed by the reflection. In terms of vectors, this line has representation $\{t\,\mathbf{v} | t \in \mathbb{R}\}$ where $\mathbf{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Indeed,

$$T_A(\mathbf{v}) = A\,\mathbf{v} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \mathbf{v}.$$

So the vector $\mathbf{v}$ is preserved by matrix multiplication with A. □

Example 2. Consider the composition of the shear with matrix $A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ and the dilation with matrix $B = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$. This map has matrix

$$B\,A = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ 0 & 2 \end{pmatrix}.$$

Notice that both transformations preserve the x–axis, so we would expect that the composed maps also preserve the x–axis. Now, the x–axis is generated by the vector $\mathbf{v} = \binom{1}{0}$, and we see that

$$B\,A\mathbf{v} = \begin{pmatrix} 3 & -1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 3\,\mathbf{v}.$$

Again, we see that the direction of $\mathbf{v}$ is preserved by the transformation. □

We might ask, given a linear transformation (or matrix), which directions (or vectors) are preserved by the action of the transformation. This question turns out to be a very important one in mathematics, and has a special name:

The eigenvalue problem

Given an $n \times n$ matrix A, can you find a number λ and an $n \times 1$ vector $\mathbf{v}$ such that

$$A\,\mathbf{v} = \lambda\,\mathbf{v}?$$

The *eigenvalue problem* is to find all such $\lambda, \mathbf{v}$ pairs.

Definition. A number λ is called an **eigenvalue** for A if there is a non-zero $n \times 1$ vector $\mathbf{v}$ such that

$$A\,\mathbf{v} = \lambda\,\mathbf{v}.$$

The vector $\mathbf{v}$ is called an **eigenvector** for λ. □

Note: Letting $\mathbf{v} = \mathbf{0}$, both sides of the equation are $\mathbf{0}$, regardless of the value of λ. Hence, we require $\mathbf{v} \neq \mathbf{0}$ in order to give the definition some content. □

Example 3. Let $A = \begin{pmatrix} 2 & 1 \\ -4 & 7 \end{pmatrix}$. Then

$$A \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

so $\lambda = 3$ is an eigenvalue of A with $(1, 1)$ a corresponding eigenvector. □

It is an amazing fact that every $n \times n$ matrix has at least one, and at most n, eigenvalues. The theory behind these facts is really nice, and we will introduce some of the ideas here; a more detailed study is deferred to MATH253. It is natural to ask: *how can we find all of the eigenvectors of a given square matrix?*

Suppose A is $n \times n$ and $\mathbf{v}$ is a non-zero $n \times 1$ column vector. If $A\,\mathbf{v} = \lambda\,\mathbf{v}$ for some number λ, then

$$\begin{aligned} \mathbf{0} &= A\,\mathbf{v} - \lambda\,\mathbf{v} \\ &= A\,\mathbf{v} - \lambda\,I_n\,\mathbf{v} \\ &= (A - \lambda\,I_n)\,\mathbf{v}. \end{aligned}$$

Thus, we are looking for non-zero solutions $\mathbf{v}$ to the matrix equation

$$(A - \lambda\,I_n)\,\mathbf{v} = 0.$$

By comparing this requirement with Theorem 2.3 ((1)≡(3)) and Theorem 3.2, we have:

Theorem 9.1 *Let A be an $n \times n$ matrix. Then λ is an eigenvalue of A if and only if $(A - \lambda\,I_n)$ is singular if and only if* $\det(A - \lambda\,I_n) = 0$.

So, we know precisely when λ is an eigenvalue (the matrix $(A - \lambda\,I_n)$ fails to be invertible), and we have an algebraic characterization too ($|A - \lambda\,I_n| = 0$). So if we evaluate $\det(A - \lambda\,I_n)$ with λ left as an unknown and set the result equal to zero, we should get an equation in λ which we can hopefully solve to find the eigenvalues.

Example 4. Find all eigenvalues of the 2×2 matrix $A = \begin{pmatrix} 5 & -6 \\ 2 & -2 \end{pmatrix}$. We will compute $\det(A - \lambda I_2)$, set it equal to zero, and solve for λ (if we can):

$$\begin{aligned} \det(A - \lambda I_2) &= \det\left(\begin{pmatrix} 5 & -6 \\ 2 & -2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ &= \det\left(\begin{pmatrix} 5 & -6 \\ 2 & -2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right) \\ &= \det\begin{pmatrix} 5-\lambda & -6 \\ 2 & -2-\lambda \end{pmatrix} \\ &= (5-\lambda)(-2-\lambda) - (-6)2 \\ &= -10 - 5\lambda + 2\lambda + \lambda^2 + 12 \\ &= \lambda^2 - 3\lambda + 2 \\ &= (\lambda - 1)(\lambda - 2). \end{aligned}$$

Now, this determinant is zero exactly when $\lambda = 1$ or 2, so these are the eigenvalues of A. □

Note: To find the corresponding eigenvectors in this example, one must solve the two matrix equations $A\,\mathbf{u} = 1\,\mathbf{u}$ and $A\,\mathbf{v} = 2\,\mathbf{v}$ for the unknown vectors $\mathbf{u}, \mathbf{v}$. This can be done by solving a system of linear equations. □

General method for the eigenvalue problem

By the theorem and example above, finding the eigenvalues of A relies on a particular polynomial.

Definition. Let A be an $n \times n$ matrix. Then the **characteristic polynomial** p for A is

$$p(\lambda) = |A - \lambda I_n|.$$

□

Corollary. *The eigenvalues of A are the roots of the characteristic polynomial p. That is, the numbers λ such that $p(\lambda) = 0$.*

Solving the eigenvalue problem

1. Use determinants to calculate $p(\lambda) = |A - \lambda I_n|$.
2. Find the values λ such that $p(\lambda) = 0$.
3. For each λ, use linear algebra to solve the equation $(A - \lambda I_n)\,\mathbf{v} = \mathbf{0}$.

Example 5. Find the eigenvectors in the previous example. *Solution:* We'll work first with $\lambda = 1$. Then

$$A - \lambda I_n = \begin{pmatrix} 5 & -6 \\ 2 & -2 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -6 \\ 2 & -3 \end{pmatrix}.$$

We need to solve

$$\begin{pmatrix} 4 & -6 \\ 2 & -3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The augmented matrix is

$$\left(\begin{array}{cc|c} 4 & -6 & 0 \\ 2 & -3 & 0 \end{array}\right)$$

and the row operations $R_2 \to R_2 - \frac{1}{2}R_1$, $R_1 \to \frac{1}{4}R_1$ give the RREF

$$\left(\begin{array}{cc|c} 1 & -\frac{3}{2} & 0 \\ 0 & 0 & 0 \end{array}\right).$$

The general solution to this system is $u_2 = t$, $u_1 = \frac{3}{2}\,t$. Taking $t = 2$ (any non-zero t would do), we obtain an eigenvector

$$\mathbf{u} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

It is easy to check that

$$A\,\mathbf{u} = \begin{pmatrix} 5 & -6 \\ 2 & -2 \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 1\,\mathbf{u},$$

so $\mathbf{u}$ is indeed an eigenvector for the eigenvalue 1. A similar calculation shows that $\mathbf{v} = \binom{2}{1}$ is an eigenvector for $\lambda = 2$. □

Final remarks

Most important eigenvalue problems are for matrices much bigger than 2×2. The determinant of a general $n \times n$ matrix A is used to compute the eigenvalues. But, $p(\lambda)$ turns out to be a polynomial of degree n, and finding zeros of degree n polynomials with $n > 2$ is much harder than the degree 2 case! In fact, the theory and practice of solving the eigenvalue problem is both broad and deep. Its applications penetrate many branches of pure and applied mathematics.

X ○ Exercises

1. Consider the line with slope 2 through the point $(x_0, y_0) = (2, 3)$. Write down formulas for this line: (a) using the point–slope formula; (b) as an algebraic equation; (c) in vector notation. Is the point $(-1, -2)$ on the line?

2. Consider the line described by the *algebraic equation*

$$2\,x + 3\,y = 7.$$

Write the line in the form $y = m\,x + c$ for suitable choices of m and c. Find a point (x_0, y_0) on the line.

3. Consider the line written in *vector notation* as:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + t \begin{pmatrix} 3 \\ -1 \end{pmatrix}, t \in \mathbb{R}$$

(a) Find an algebraic equation which describes the line; (b) write down a pair of parametric equations for x and y (as a function of t); (c) find a value of t which shows that the point $\binom{-4}{1}$ is on the line.

4. Determine whether each of the following systems is consistent. If a system is consistent, find the *general solution* and write it in vector notation.

(a)
$$\begin{aligned} x - 3\,y &= 4 \\ -4\,x + 2\,y &= 6 \end{aligned}$$

(b)
$$\begin{aligned} 2\,x - y &= -3 \\ 5\,x + 7\,y &= 4 \end{aligned}$$

(c)
$$\begin{aligned} 2\,x - 8\,y &= 5 \\ -3\,x + 12\,y &= 8 \end{aligned}.$$

5. The following systems of linear equations are in *echelon form*. Use back substitution to solve them, and write down the general solution in vector notation.

(a)
$$\begin{aligned} x + y + 2\,z &= 8 \\ y + 4\,z &= 3 \\ z &= -1 \end{aligned}$$

(b)
$$\begin{aligned} x + y + 2\,z &= 8 \\ y + 4\,z &= 3 \end{aligned}$$

(c)
$$\begin{aligned} x_1 + 4\,x_2 + x_4 &= 8 \\ x_3 - x_4 &= 2 \end{aligned}$$

(d)
$$x + y + 2\,z = 8.$$

6. A porcelain company manufactures ceramic cups and saucers. For each cup or saucer, a worker measures a fixed amount of material, and puts it in a forming machine where it is glazed and dried. On average, this takes three minutes per cup and two minutes per saucer. The materials for a cup cost 15cents, and a saucer 10cents. In exactly eight hours of work, a worker uses \$24 worth of materials. Can you determine how many cups and how many saucers are made?

7. Let $c \neq 0$. Find conditions on a and b such that the following system has infinitely many solutions:

$$\begin{array}{rcrcl} a\,x & + & b\,y & = & c \\ a\,x & - & b\,y & = & c \end{array}.$$

8. For which values of the constant k does the following system of equations have solutions?

$$\begin{array}{rcrcrcll} x & + & y & + & 2\,z & = & 1 & [1] \\ x & + & 2\,y & + & 3\,z & = & k & [2] \\ 2\,x & + & 3\,y & + & 5\,z & = & 3 & [3] \end{array}$$

[Hint: compare [1]+[2] with [3].]

9. Use any method to find **all** solutions of the following systems of **nonlinear** equations:

(a) $\begin{array}{rcl} x^2+y^2 & = & 5 \\ x\ +y & = & 2 \end{array}$

(b) $\begin{array}{rcl} x+y & = & 2 \\ x\times y & = & 1 \end{array}$

(c) $\begin{array}{rcl} x^2+2x+y^2-4y & = & 0 \\ x^2-4x+y^2+2y & = & 0 \end{array}$

(d) $\begin{array}{rcl} x^2-y^2 & = & 3 \\ x\ +y & = & 3 \end{array}$

[Hint: a graphical method may help.]

10. Find all solutions to the following pair of equations and interpret your result geometrically:

$$\begin{array}{rcrcl} 5\,x & + & 2\,y & = & 3 \\ 2\,x & + & 5\,y & = & 3 \end{array}.$$

11. Consider the setup of problem **6** (above), but now the materials costs are 25cents, 20cents and \$44 respectively. Find how many cups and how many saucers are made per worker.

12. Obtain the general solution to the following system, and write it in vector notation:

$$\begin{array}{rcrcrcrcl} 2\,x_1 & + & 3\,x_2 & - & x_3 & + & 4\,x_4 & = & 7 \\ & & & & x_3 & + & 2\,x_4 & = & 3 \end{array}.$$

13. Find conditions on a, b, c such that the system in problem **7** (above) has a unique solution.

14. Use an augmented matrix and Gaussian elimination to find all solutions (if there are any), to the given systems.

(a)
$$\begin{aligned} x_1 - 2\,x_2 + 3\,x_3 &= 11 \\ 4\,x_1 + x_2 - x_3 &= 4 \\ 2\,x_1 - x_2 + 3\,x_3 &= 10 \end{aligned}$$

(b)
$$\begin{aligned} 3\,x_1 + 6\,x_2 - 6\,x_3 &= 9 \\ 2\,x_1 - 5\,x_2 + 4\,x_3 &= 6 \\ -x_1 + 16\,x_2 - 14\,x_3 &= -3 \end{aligned}$$

(c)
$$\begin{aligned} x_1 + x_2 - x_3 &= 7 \\ 2\,x_1 - x_2 + 3\,x_3 &= 4 \\ 4\,x_1 + x_2 + x_3 &= 19 \end{aligned}$$

15. Solve the following systems

(a)
$$\begin{aligned} x + y + z &= 0 \\ 4\,x + 3\,y + 6\,z &= 0 \\ 3\,x - y + 11\,z &= 0 \end{aligned}$$

(b)
$$\begin{aligned} 3\,y - z + w &= -1 \\ x + y + z + 2\,w &= 8 \\ 2\,x + z &= 10 \end{aligned}$$

(c)
$$\begin{aligned} 5\,b - 8\,c + 3\,d &= 2 \\ a + 2\,b - 3\,c + d &= 4 \\ 2\,a - b + 2\,c - d &= 6 \end{aligned}$$

(d)
$$\begin{aligned} x_1 + x_3 - 2\,x_5 &= 1 \\ x_1 + x_2 + x_4 &= 3 \\ 2\,x_1 + 3\,x_2 + x_3 - 3\,x_4 - x_5 &= 3 \end{aligned}$$

16. Which of the follolwing augmented matrices is in RREF? For each system, either write down the general solution, or explain why it is inconsistent:

(a) $\left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 3 \end{array}\right)$ (c) $\left(\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$ (e) $\left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 1 & 5 \end{array}\right)$

(b) $\left(\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 2 \end{array}\right)$ (d) $\left(\begin{array}{ccc|c} 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$

17. Use elementary row operations to put the following matrix in RREF:

$$\begin{pmatrix} 1 & 0 & 2 & 1 & 4 \\ 1 & -3 & -1 & -1 & -4 \\ 1 & -1 & 1 & 2 & 3 \end{pmatrix}$$

18. An investor remarks to her stockbroker that all her stock holdings are in three companies, Delta Airlines, Hilton Hotels and McDonald's, and that two days ago the value of her stocks went down \$350 but yesterday the value increased by \$600. The broker recalls that two days ago, the price of Delta Airlines stock dropped by \$1 a share, Hilton Hotels dropped by \$1.50, but the price of McDonald's stock rose by \$0.50. The broker also remembers that yesterday the price of Delta Airlines rose \$1.50, there was a further drop of \$0.50 in Hilton Hotels stock, and McDonald's stock rose \$1. Show that the broker does not have enough information to calculate the number of shares the investor owns of each company's stock, but that when an investor says she owns 200 shares of McDonald's stock, the broker can calculate the number of shares of Delta Airlines and Hilton Hotels.

19. Let $\mathbf{a} = \begin{pmatrix} -3 \\ 1 \\ 4 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 5 \\ -4 \\ 7 \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}$, $A = \begin{pmatrix} 1 & 2 \\ -1 & 3 \\ 5 & 2 \end{pmatrix}$, $B = \begin{pmatrix} -2 & 1 \\ -7 & 0 \\ 4 & 5 \end{pmatrix}$, $C = \begin{pmatrix} -1 & 4 \\ -7 & 1 \\ 6 & 3 \end{pmatrix}$. Calculate:

(a) $\mathbf{a}+\mathbf{b}$ (b) $3\,\mathbf{b}$ (c) $-2\,\mathbf{c}$ (d) $2\,\mathbf{a}-5\,\mathbf{b}$ (e) $3\,\mathbf{b}-7\,\mathbf{c}+2\,\mathbf{a}$
(f) $3\,A$ (g) $A+B$ (h) $2\,C-5\,A$ (i) $0\,B$ (j) $2\,A-3\,B+4\,C$

20. Find the transpose of the given matrices:

(a) $\begin{pmatrix} -1 & 4 \\ 6 & 5 \end{pmatrix}$ (b) $\begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}$ (c) $\begin{pmatrix} 2 & -1 & 0 \\ 1 & 5 & 6 \end{pmatrix}$.

21. Suppose a, b, c are constants (i.e. some fixed real numbers). Show that the following system is consistent only if $c = 2\,a - 3\,b$:

$$\begin{aligned} 2\,x_1 - x_2 + 3\,x_3 &= a \\ 3\,x_1 + x_2 - 5\,x_3 &= b \\ -5\,x_1 - 5\,x_2 + 21\,x_3 &= c \end{aligned}$$

22. Prove the following facts about $n \times n$ matrices:

(a) if A and B are symmetric then $A + B$ is symmetric;
(b) $\frac{1}{2}(A + A^T)$ is symmetric [hint: you'll need $(a\,A)^T = a\,(A^T)$];
(c) if A is upper triangular then A^T is lower triangular.

23. The following are augmented matrices which describe systems of equations with variables x, y and z. In each case interpret the matrix (as equations in x, y, z) and solve the system (where possible).

a) $\left(\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 5 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right)$ b) $\left(\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 5 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right)$ c) $\left(\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 5 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$

24. For the following problems: (i) Write them in augmented matrix form; (ii) use Gaussian elimination to put them in an EF; (iii) proceed by Gauss-Jordan elimination to get RREF; (iv) write down the general solution (or show that the system is inconsistent).

(a) $\begin{aligned} 2\,x + y - 2\,z &= 10 \\ 3\,x + 2\,y + 2\,z &= 1 \\ 5\,x + 4\,y + 3\,z &= 4 \end{aligned}$ (b) $\begin{aligned} x + 2\,y - 3\,z &= 6 \\ 2\,x - y + 4\,z &= 2 \\ 4\,x + 3\,y - 2\,z &= 14 \end{aligned}$ (c) $\begin{aligned} x + 2\,y - 3\,z &= -1 \\ 3\,x - y + 2\,z &= 7 \\ 5\,x + 3\,y - 4\,z &= 2 \end{aligned}$

25. Let A, B and C be the matrices from problem **19** (above). Find a matrix D such that $A+2\,B-3\,C+D$ is the 3×2 zero matrix.

26. Write out the system of equations represented by the augmented matrix $\left(\begin{array}{ccc|c} 1 & 1 & -1 & 7 \\ 4 & -1 & 5 & 4 \\ 6 & 1 & 3 & 20 \end{array}\right)$.

27. Write the equations from problem **26** as a matrix equation $A\,\mathbf{x} = \mathbf{b}$ for suitable matrices $A, \mathbf{x}, \mathbf{b}$.

28. Calculate: $\begin{pmatrix} 3 & -2 & 1 \\ 4 & 0 & 6 \\ 5 & 1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

29. Calculate A^2, A^3, A^4, A^5 when $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$.

30. A company pays salaries to its management team, and gives shares as an annual bonus. Last year, the CEO received \$150000 and 5000 shares, each of three divisional directors received \$110000 and 3000 shares, and the chief operating officer received \$90000 and 2000 shares. (a) Express the payments to the senior managers (in both cash and shares) as a 2×3 matrix; then (b) put the number of managers of each rank in a column vector of suitable size; and (c) use matrix multiplication to calculate the cash and share cost of the company's executive remuneration scheme.

31. Consider the matrix equation $A^2 + 3A + 2I = 0$. We could factorise this as the matrix equation $(A+I)(A+2I) = 0$. Hence apparently the only solutions are $A = -I$ and $A = -2I$. This argument is false. Verify that $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -2 \end{smallmatrix}\right)$ is another solution. Can you find the defect in the reasoning above which led us to this false conclusion?

32. Below are listed a number of matrices.

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \\ 4 & 4 & 4 \end{pmatrix} \qquad B = \begin{pmatrix} 2 & 1 & 3 & 1 \\ 1 & 1 & 0 & 2 \end{pmatrix} \qquad C = \begin{pmatrix} -1 & 1 \\ -1 & 2 \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix} \qquad E = (1, 1, 1, 1) \qquad F = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

Determine the following products if they exist.

(a) AB **(b)** BA **(c)** EA **(d)** A^2 **(e)** DF **(f)** FD **(g)** BED **(h)** BAD

33. Let C be as in question **32**. Find all matrices X such that $XC = CX$.

34. For a real number x, $x^2 = x$ implies $x(x-1) = 0$, so $x = 0$ or 1.

(a) Find a square matrix A which is not the zero matrix and isn't I_2, but for which $A^2 = A$. (Hence the above rule doesn't work for matrices.)

(b) Show however that if $A^2 = A$, then A must be either I_2 or non–invertible (*singular*).

35. Find a non-zero solution to $\begin{pmatrix} 2 & -1 \\ -4 & 2 \end{pmatrix} \mathbf{x} = \mathbf{0}$. Is A invertible?

36. Find the inverses of the matrices $S = \begin{pmatrix} 3 & -1 \\ 2 & -1 \end{pmatrix}$ and $T = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 1 \\ 2 & 1 & 4 \end{pmatrix}$.

37. Verify that: $X = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 2 & 0 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}^{-1} = \frac{1}{2} \begin{pmatrix} -1 & 1 & -1 & 1 \\ -3 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 4 & -2 & 0 & 0 \end{pmatrix}$.

38. Let A, B, C, D, E, F be the matrices from problem **32**. Calculate the following expressions (if they exist):

(a) $(B + E)A$ **(b)** $A(D + F)$ **(c)** $C^2 - I$ **(d)** $(C - I)(C + I)$

39. Show that the matrix $\left(\begin{smallmatrix} 3 & 4 \\ -2 & -3 \end{smallmatrix}\right)$ is its own inverse.

40. Compute the inverse of the matrix $A = \begin{pmatrix} 1 & -2 & -1 \\ -2 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$. Hence solve $A\,\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ 7 \end{pmatrix}$.

41. Evaluate the following determinants in two ways:

(a) by expanding along rows and/or columns throughout, and

(b) by using row operations.

(i) $\begin{vmatrix} 2 & 1 & 4 \\ 4 & 3 & 6 \\ 0 & 1 & 3 \end{vmatrix}$ (ii) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 0 & 3 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 2 & 4 & 6 & 5 \end{vmatrix}$.

42. Let I_3 be the 3×3 identity matrix. Perform each of the following row operations on I_3 to obtain matrices A, B, C, D respectively. In each case, calculate the determinant of the new matrix by a suitable cofactor expansion, and comment on the result.

(a) $R_3 \to R_3 - 2\,R_1$

(b) $R_2 \to R_2 + R_3$

(c) $R_1 \to -2\,R_1$

(d) $R_2 \leftrightarrow R_3$

43. Let $A = \begin{pmatrix} -1 & 2 \\ -4 & 7 \end{pmatrix}$, $B = \begin{pmatrix} -3 & 2 \\ 1 & 1 \end{pmatrix}$, $C = \begin{pmatrix} 4 & -2 \\ 3 & 0 \end{pmatrix}$. Compute

$$\det(A), \det(B), \det(C), \det\left(A^2\right), \det(ABC), \det\left(A^{26}\right).$$

44. Show using determinants that if $AB = 0$ (where A, B are $n \times n$ matrices), then either A or B must be singular (*ie.* has no inverse).

45. A square matrix A is called *nilpotent* if $A^k = 0$ for some $k > 0$. Show that if A is nilpotent then $\det(A) = 0$.

46. A square matrix A is called *idempotent* if $A^2 = A$. What are the possible values of $\det(A)$ if A is idempotent?

47. Find the area of the triangle in 2-space with corners at $(1,-1), (2,3)$ and $(4,1)$.

48. Compute the volume of the parallelepiped in 3-space with one corner at $(1,-1,1)$ and the three adjacent corners at $(2,1,2), (-1,0,-1)$ and $(0,1,1)$. [Hint: translate the parallelepiped back to the origin, and find vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ which describe the adajacent corners when one of the corners is $\mathbf{0}$.]

49. Compute the cross product of the vectors $(1,4,-2)$ and $(-1,3,1)$.

50. Consider the triangle in the plane whose corners have coordinates: (x_1, y_1), (x_2, y_2), (x_3, y_3). Show that the area of the triangle is

$$\pm\frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}.$$

51. Find the determinants of each matrix below

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}; \quad B = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 3 \end{pmatrix}; \quad C = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 3 \\ 2 & 1 & 0 & 0 \end{pmatrix}.$$

52. Recall that $\det(A\,B) = \det(A)\ \det(B)$ for all $n \times n$ matrices A and B. Use this information to show that

(a) $\det(A^{-1}) = \frac{1}{\det(A)}$ for all invertible matrices A;

(b) $\det(A^{-1}\,B\,A) = \det(B)$ for all B and invertible A.

53. Prove the associative law of vector addition for vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in $\mathbb{R}^2$:

$$\mathbf{u} + (\mathbf{v} + \mathbf{w}) \ = \ (\mathbf{u} + \mathbf{v}) + \mathbf{w}.$$

54. Prove the following law for the dot product of vectors in $\mathbb{R}^2$:

$$\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}.$$

55. Given that $\mathbf{u} = (1,2,1)$, $\mathbf{v} = (-1,1,3)$, and $\mathbf{w} = (2,-1,3)$ calculate the following wherever possible. In cases where the expression cannot be calculated, explain why.

(a) $2\mathbf{u} - 3\mathbf{v}$	(e) $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$	(i) $\mathbf{v} \cdot (\mathbf{w} \times \mathbf{u})$
(b) $\mathbf{u} \cdot \mathbf{w}$	(f) $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$	(j) $(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{u}$
(c) $\mathbf{u} \times \mathbf{v}$	(g) $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$	(k) $(\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})$
(d) $(\mathbf{u} + \mathbf{v}) \times \mathbf{w}$	(h) $(\mathbf{u} \cdot \mathbf{v}) \times \mathbf{w}$	(l) $(\mathbf{u} + \mathbf{v}) \times (\mathbf{u} - \mathbf{v})$

56. Let $\mathbf{u} = 2\mathbf{i} - 3\mathbf{j} + 4\mathbf{k}$, $\mathbf{v} = \mathbf{i} - \mathbf{j} + 2\mathbf{k}$. Find (using the cross product):

(a) a vector perpendicular to both $\mathbf{u}$ and $\mathbf{v}$;

(b) the area of the parallelogram defined by $\mathbf{u}$ and $\mathbf{v}$; and

(c) the sine of the angle between $\mathbf{u}$ and $\mathbf{v}$.

57. Consider the triangle in 3-space which has as vertices the points P, Q, R with respective position vectors of $(1, 0, 1), (-1, 1, 0)$ and $(2, 1, 2)$. Find vectors $\mathbf{u}$ and $\mathbf{v}$ describing the sides of the triangle which meet at P. Use a vector product to work out the angle at P, and thus calculate the area of the triangle. Use a similar method to calculate the angles at Q and R. Do you get the same answer for the area? What do the angles add up to?

58. $P = (8, -1, 2)$ and $Q = (5, -4, 8)$. Find the vector parametric equation of the line through P and Q.

59. Let $\mathbf{u} = (1, 2, 1)$. Find two (non-zero) vectors $\mathbf{v}$ and $\mathbf{w}$ such that $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are mutually perpendicular to each other.

60. Let $\mathbf{u}$ and $\mathbf{v}$ be non-zero, non-parallel vectors and let:

$$\begin{aligned} \mathbf{p} &= -2\,\mathbf{u} + 5\,\mathbf{v} & \qquad \mathbf{r} &= 2\,\mathbf{u} - \mathbf{v} \\ \mathbf{q} &= \quad 2\,\mathbf{u} + 3\,\mathbf{v} & \qquad \mathcal{S} &= 8\,\mathbf{u}. \end{aligned}$$

(a) Determine whether or not the point with position vector $\mathbf{p}$ is on the line through the points with position vectors $\mathbf{q}$ and $\mathbf{r}$;

(b) Determine whether or not $\mathcal{S} - \mathbf{p}$ is parallel to $\mathbf{r}$.

61. Find a Cartesian equation for the plane containing the points $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

62. Suppose that $\mathbf{u}$ and $\mathbf{v}$ are non-parallel vectors of the same length. Show that $\mathbf{u} - \mathbf{v}$ is perpendicular to $\mathbf{u} + \mathbf{v}$. (*Hint:* an algebraic proof is preferred!)

63. Find the angle at the corner Q of the triangle with corners at points P, Q, R having position vectors $(1, 2, 1)$, $(0, -2, 1)$ and $(-1, 3, -2)$ respectively.

64. Find a (non-zero) vector perpendicular to both $(1, -2, 1)$ and $(1, 3, 2)$.

65. Do the lines $(1, 1, 3) + t\,(2, 1, 1)$ and $(5, 6, 4) + s\,(1, 2, 1)$ lie in a single plane?

66. Use the cross product to help you find the parametric vector equation of the line through the point $(1, 1, 1)$ and perpendicular to the plane given by

$$\mathbf{r} \;=\; \mathbf{r}_0 + t_1(2, 0, 1) + t_2(-1, -2, 3).$$

67. Let $A = (1, 1, 1)$, $B = (2, -4, -2)$ and $C = (10, 2, 7)$. Find a vector parametric equation for the plane through A, B and C.

68. In general a plane can be defined by a single Cartesian equation in x, y, z. Usually, two planes intersect in a line, so one way to specify a line is to give two simultaneous equations. Find a parametric vector equation for the line given by:

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

69. Find the point of intersection of the line with parametric vector equation $\mathbf{r} = (1, -1, 1) + t(-1, 2, -3)$ with the plane having Cartesian equation $3x - 4y + z = 2$.

70. Find a vector parallel to the line described by $x + 2y + 3z = 6$ and $x + 3y + 4z = 7$.

71. Show that the line with equation $(x, y, z) = (4, -1, 3) + t\,(-3, 3, 0)$ intersects the line with equation $(x, y, z) = (5, 4, 1) + s\,(2, 1, -1)$. Find the coordinates of the point of intersection, and determine the angle between the two lines.

72. Find the point of intersection of the three planes $x+y+z = 1$, $x+2y+3z = 2$ and $3x+2y+2z = 3$.

73. If $\mathbf{u} = (1, 3, -10)$ and $\mathbf{v} = (2, 3, 6)$ find the components $\mathbf{u}_{\parallel}$ and $\mathbf{u}_{\perp}$ of $\mathbf{u}$ which are parallel and perpendicular to $\mathbf{v}$.

74. Find the minimum distance between the point $\mathbf{w} = (3, 4, 5)$ and any point on the line $(x, y, z) = (1, 3, -2) + t(0, -2, 1)$, $t \in \mathbb{R}$.

75. Find the distance between the point with position vector $\mathbf{w} = (-1, 0, 3)$ and the plane with Cartesian equation $x + 2y - 2z = 11$.

76. Use the general method of projection onto a plane to prove that if the plane is given by $a\,x+b\,y+c\,z = k$, then the distance from (w_1, w_2, w_3) to the plane is given by:

$$\frac{|a\,w_1 + b\,w_2 + c\,w_3 - k|}{\sqrt{a^2 + b^2 + c^2}}.$$

77. Use the cross product to help you find the Cartesian equation of the plane through the points $(1, 1, 1)$, $(2, 1, 4)$ and $(5, 0, 1)$.

78. For what value of α does the plane $3x - 6y + 4z = \alpha$ contain the point with position vector $(1, 2, 5)$?

79. Let $\mathbf{w} = (1, 2, 3)$ and let L be the line $(x, y, z) = (6, 3, 6) + t\,(4, 3, 1)$, $t \in \mathbb{R}$. Find $\text{proj}_L\mathbf{w}$ and the minimum distance from $\mathbf{w}$ to L.

80. Using the method of projection (onto a plane), find the distance between the point $P = (1, 1, 11)$ and the plane $x - 2\,y + 2\,z = 3$.

81. Use induction to prove that for any real number $a \geq 0$: $(1 + a)^n \geq (1 + na)$ for all $n \geq 1$.

82. What is wrong with the following "proof" by induction that all balls have the same colour?

For $n > 0$, let $P(n)$ be the proposition that any set of n balls have the same colour.
Obviously $P(1)$ is true!
Now assume $P(k)$ is true for some $k > 0$. Let

$$S = \{b_1, b_2, \ldots, b_{k+1}\}$$

be a set of $k + 1$ balls. Then the two subsets $\{b_1, b_2, \ldots, b_k\}$ and $\{b_2, b_3, \ldots, b_{k+1}\}$ each have k elements so by the inductive assumption, all balls in each of these two smaller sets have the same colour. Clearly then, all balls in their union must have the same colour also. So by the Principle of Induction, the desired result follows!

83. Prove by induction that

$$1 \times 2 + 2 \times 3 + \cdots + n(n + 1) = \frac{1}{3}n(n + 1)(n + 2)$$

for all integers $n > 0$.

84. Prove: $1^3 + 2^3 + 3^3 + \ldots + n^3 = \dfrac{1}{4} n^2 (n+1)^2$.

85. Prove that $5^{2n+1} + 2^{2n+1}$ is divisible by 7 for all $n \geq 0$.

86. Prove by induction that $17\, n^3 + 103\, n$ is divisible by 6 for all $n \in \mathbb{N}$.

87. Prove that $a^2 - 1$ is divisible by 8 for all odd integers a.

88. Give a formal inductive proof that the sum of the interior angles of a convex polygon with n sides is $(n-2)\,\pi$ (radians). You may assume that the result is true for a triangle. (A *convex* polygon is one where all the interior angles are smaller than π radians.)

Hint: you can cut a convex n-gon into a convex $(n-1)$-gon and a triangle.

89. Use induction to prove that $1 + 2 + 3 + \ldots + n = \dfrac{1}{2} n(n+1)$.

90. Prove that $n\,(n^2 + 5)$ is divisible by 6 for all integers $n \geq 1$.

91. Suppose a sequence of numbers $a_1, a_2, a_3, \ldots$ is defined recursively by

$$a_1 = 3 \text{ and } a_{n+1} = \frac{a_n}{a_n + 1},$$

so the first few terms of the sequence are $3, \frac{3}{4}, \frac{3}{7}, \ldots$. Use induction to prove that

$$a_n = \frac{3}{3\,n - 2}.$$

92. Let the sequence $\{a_n\}$ be defined recursively by

$$a_n = 6\, a_{n-1} - 9\, a_{n-2}, \quad a_1 = 0, a_2 = 9.$$

Use strong induction to prove that $a_n = 3^n\,(n-1)$.

93. Recall the Fibonacci recurrence: $R_{n+1} = R_n + R_{n-1}$. Find a matrix A such that

$$\begin{pmatrix} R_n \\ R_{n+1} \end{pmatrix} = A \begin{pmatrix} R_{n-1} \\ R_n \end{pmatrix}.$$

Prove by induction that

$$\begin{pmatrix} R_n \\ R_{n+1} \end{pmatrix} = A^n \begin{pmatrix} R_0 \\ R_1 \end{pmatrix}.$$

94. Find a $k \times k$ matrix A such that kth order recurrence

$$a_n = b_1\, a_{n-1} + \cdots + b_k\, a_{n-k}$$

can be written as

$$\begin{pmatrix} a_{n+1-k} \\ \vdots \\ a_{n+1} \end{pmatrix} = A \begin{pmatrix} a_{n-k} \\ \vdots \\ a_n \end{pmatrix}.$$

Suppose that you knew a formula for powers of the matrix A. How would this help you to solve the recurrence?

95. Use strong induction to prove the least integer principle. [Hint: let $P(n)$ be the proposition that every nonempty subset of $\mathbb{N}$ containing n has a least element.]

96. Use induction to prove the principle of strong induction. [Hint: let S be the set of $n \in \mathbb{N}$ for which $P(n)$ is true and and let $Q(n)$ be the proposition that $\{1, 2, \ldots, n\} \subseteq S$. Use the hypotheses of strong induction to prove (using ordinary induction) that $Q(n)$ is true for every $n \in \mathbb{N}$.] Deduce that induction and strong induction are equivalent.

97. Simplify the following expressions involving complex numbers

(a) $(2 + 3i) + (-4 + i)$

(b) $(2 + 3i) \times (-4 + i)$

(c) $(2 + 3i) \div (-4 + i)$

(d) $|2 + 3i|$

(e) $\arg(-4 + i)$

(f) $6(1 + i)(1 - i)$

98. Let $z = 2 + 2i$. Convert z to polar form and hence evaluate z^5, expressing your final answer in rectangular form.

99. Prove that $\operatorname{cis}\theta_1 \cdot \operatorname{cis}\theta_2 = \operatorname{cis}(\theta_1 + \theta_2)$ by using the following trigonometric "addition theorems":

$$\cos(\theta_1 + \theta_2) = \cos\theta_1 \cdot \cos\theta_2 - \sin\theta_1 \cdot \sin\theta_2,$$

$$\sin(\theta_1 + \theta_2) = \sin\theta_1 \cdot \cos\theta_2 + \cos\theta_1 \cdot \sin\theta_2.$$

Deduce that for any pair $z_1, z_2 \in \mathbb{C}$, $\arg(z_1 z_2) = \arg(z_1) + \arg(z_2)$.

100. Let $z = -5 - 5i$.

(a) Write z in the form $r \operatorname{cis}\theta$ for suitable r and θ.

(b) Solve the equation $w^5 = -5 - 5i$. Give all solutions.

(c) Draw the solutions to part (b) on the complex plane.

101. Solve the following system of linear equations involving complex numbers. The method is identical to the method for real numbers, but the arithmetic gets a lot messier.

$$\begin{array}{rll} iw & +(1-i)z & = 2 + 3i \\ w & +(2+i)z & = 1 - i \end{array}$$

102. Write each of the following complex numbers in the form $r \operatorname{cis}\theta$.

(a) $3 - 6\,i$

(b) $5 - i$

(c) $\frac{3}{5} + \frac{4}{5}\,i$.

103. Solve in complex numbers the following equations. Give all solutions accurate to 4 significant figures.

(a) $z^7 = 3 - 6\,i$

(b) $z^3 = 5 - i$

(c) $z^8 = -1$.

104. Solve the following quadratic equation for the complex variable z: $\sqrt{3}\,z^2 + 2i\,z - i = 0$. Your answer should be in rectangular form.

105. Use the result of problem **99** above and induction to prove De Moivre's formula:

$$(r\,\text{cis}\,\theta)^n \;=\; r^n\,\text{cis}\,(n\,\theta) \text{ for all } n \in \mathbb{N}.$$

106. Find all solutions to: $z^3 = -2 + 2\,i$.

107. Show that $a^2 - 1$ is divisible by 8 for all odd integers a.

108. Express 960 and 468 as products of primes and hence calculate the number of distinct divisors each has. Hence also compute lcm$(960, 468)$ and gcd$(960, 468)$.

109. For $a = 8451$ and $b = 2277$, express each as a product of primes and compute the number of distinct divisors each has. Use the product of prime representations to compute lcm(a, b) and gcd(a, b).

110. If gcd$(a, b) = 21$, lcm$(a, b) = 630630$, and $a = 2310$, find b.

111. (a) Apply the Euclidean algorithm to find gcd$(966, 320)$, showing your setting out.

(b) Go on to find integers x, y such that $966\,x + 320\,y =$ gcd$(966, 320)$.

112. Use the Euclidean algorithm to help you find all solutions (if there are any) of the Diophantine equation

$$231\,x + 70\,y = 28.$$

113. Are there any integer solutions to $13242\,x - 123\,y = 5$?

114. Use the Euclidean Algorithm to solve the Diophantine equation $5326\,x - 10705\,y = 2$ where x and y are integers. Use the general solution to find the solution which has the smallest positive value of x.

115. (a) Use the Euclidean algorithm to find gcd$(576594, 256347)$.

(b) Use back substitution to find two integers a and b with

$$576594\,a + 256347\,b = \text{gcd}(576594, 256347).$$

116. Use the Euclidean Algorithm to find the general solution of the following Diophantine equations where x and y are integers

(a) $4523\,x + 3781\,y = 21$

(b) $836\,x - 17346\,y = 11$

117. Now we know how to solve one linear Diophantine equation. What would you need to do to solve a system of them? Illustrate your approach by finding the general solution to

$$\begin{array}{rcrcrcr} x & + & 5\,y & + & z & = & 11 \\ 2x & + & 11\,y & + & 5\,z & = & 2 \end{array}$$

where x, y and z are all integers!

118. Let $p, q, r, s \in \mathbb{N}$ and $x, y \in \mathbb{Q}$ be such that

$$q\,x = p \text{ and } s\,y = r.$$

Use standard properties of addition and multiplication (ie. no fractions) to prove that

$$q\,s\,(x+y) = p\,s + q\,r.$$

Deduce a formula for addition in $\mathbb{Q}$.

119. Let n be a natural number with a, b, c, d integers. Show that if $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$ then $ac \equiv bd \pmod{n}$. (Hint: model your proof on the one in the lectures for showing $a + c \equiv b + d$.)

120. If a, b are integers with remainders of 3 and 5 when divided by 7, find the remainder of $24\,a + 16\,b^3$ when divided by 7.

121. Find the remainder of 9^{999} on division by 7.

122. Show using congruences that $8^n - 3^n$ is divisible by 5 for all integers $n \geq 1$.

123. Find the remainder when 27^{475} is divided by 14.

124. What are the invertible elements of $\mathbb{Z}_{15}$? Find their inverses.

125. Since 131 is prime, all non-zero elements of $\mathbb{Z}_{131}$ have inverses. Using the Euclidean algorithm (or otherwise) find the inverse of 43.

126. Find all integer solutions to the congruence equations:

(a) $6x \equiv 2 \pmod{8}$

(b) $6x \equiv 2 \pmod{88}$.

127. Solve the polynomial $x^2 + x + 8 = 0 \pmod{10}$.

128. Which $n \in \{1, 2, 3, \ldots, 19\}$ are relatively prime to 20? Hence find $\phi(20)$.

129. Use Fermat's or Euler's theorems to find the remainders when:

(a) 3^{3962} is divided by 37;

(b) 53^{242} is divided by 143;

(c) $8^{123456789}$ is divided by 15.

130. Find all integer solutions of the congruence equation $2\,x^2 + 3\,x + 4 \equiv 0 \pmod{6}$.

131. This question concerns $\mathbb{Z}_{24}$.

(a) List all the invertible elements of $\mathbb{Z}_{24}$.

(b) Find $\bar{7}^{-1}$ in $\mathbb{Z}_{24}$ if it exists.

(c) Find the Euler number $\phi(24)$.

(d) Hence compute the remainder when $(11)^{13947}$ is divided by 24.

132. List the invertible elements of $\mathbb{Z}_{14}$. Hence calculate $\phi(14)$ and evaluate $\bar{5}^{1042}$ in $\mathbb{Z}_{14}$.

133. Solve the system of simultaneous congruences

$$\begin{aligned} x + 3y &\equiv 4 \pmod 5 \\ 3x + 2y &\equiv 3 \pmod 5 \end{aligned}$$

Hint: Just go ahead and do a Gaussian elimination. But use modular arithmetic, *i.e.* work in $\mathbb{Z}_5$. To divide, you just multiply by the inverse in $\mathbb{Z}_5$.

134. A shift cipher is used to encrypt a message which comes out as:

WEZI SYV WSVPW

It is known that the first letter of the unencrypted message was "S". Decrypt the rest of the message.

135. An affine cipher based on $\mathbb{Z}_{26}$ has encryption formula $E(x) = \overline{9}x + \overline{17}$.

(a) Show $\overline{9}^{-1}$ exists in $\mathbb{Z}_{26}$, and find it.

(b) Hence give the decryption formula $D(x)$ corresponding to $E(x)$.

(c) Use it to decrypt the four-letter message: "WRXX." (Remember that $\bar{0}$ corresponds to A, $\bar{1}$ to B, $\bar{2}$ to C, and so forth.)

136. If numbers are encrypted by raising to the power of 17 modulo 111, find an exponent b so that raising to the power of b modulo 111 decrypts them.

137. In an affine cipher, the letters "A",...,"Z" are encoded by the integers $0, \ldots, 25$ and then encrypted by a function

$$E(x) \equiv a\,x + b \pmod{26}.$$

For a certain choice of a and b, the letter "E" gets encrypted to be "R" and the letter "V" gets encrypted to be "Q". Crack this code by finding the decrypt function $D(y) = a^{-1}\,(y - b)$. Decrypt the message "FMMOVSR".

138. A toy RSA cryptosystem! Let $n = 5 \times 7 = 35$. Work through an RSA cryptosystem as follows.

- compute $\phi(n)$
- check that $a = 11$ is such that $\overline{a}$ has an inverse in $\mathbb{Z}_{\phi(n)}$
- compute the inverse, i.e. find b such that $ab \equiv 1 \pmod{\phi(n)}$
- encrypt the message "top secret!" by translating single letters into elements of $\mathbb{Z}_n$ in the usual way, with also a space represented by $\overline{26}$ and "!" by $\overline{27}$ (the other elements don't get used here but could correspond to other forms of punctuation), using

$$E(x) = x^a \text{ in } \mathbb{Z}_n$$

- decrypt using

$$D(x) = x^b \text{ in } \mathbb{Z}_n$$

Did decryption work in each case? Why not? Will this be a problem in practice if a *much* larger n is used, do you think?

139. An RSA encrypted message is sent with the published key numbers $a = 7$ and $n = 19519$. These are of course absurdly small for an RSA code - so we should be able to break it easily. Break the code and find the decryption key! Demonstrate that you have done so by first encrypting and then decrypting the number 2.

140. Consider the matrices

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},\ B = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix},\ \text{and } C = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

(a) Give geometrical descriptions of the maps T_A, T_B, T_C associated with A, B, C respectively.

(b) Calculate A^2 and B^2 and interpret your results in terms of linear maps.

(c) Calculate BC and CB and interpret your results in terms of linear maps.

(d) Calculate ABC and interpret your result in terms of linear maps.

141. Let T_1 be a reflection about the x-axis and T_2 a reflection about the line with Cartesian equation $y = x$.

(a) Compute the matrices A and B of T_1 and T_2 respectively.

(b) Show that both $T_1 \circ T_2$ and $T_2 \circ T_1$ are rotations and determine the angle of rotation in each case.

(c) The product of any two reflections is a rotation. Show this, and determine the angle of this rotation.

142. Write down the 2×2 matrices for the following linear maps of the plane, and draw the image of the unit square under each.

(a) Rotation through 60 degrees anticlockwise.

(b) Reflection across the line $y = \sqrt{3}x$.

143. Let

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \text{ and } B = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(a) Describe geometrically T_A and T_B.

(b) Determine using matrix algebra the effect of T_A followed by T_B.

144. Show that if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.

145. Find all eigenvalues for the matrix $A = \begin{pmatrix} -12 & 7 \\ -7 & 2 \end{pmatrix}$.

146. Consider $A = \begin{pmatrix} 0 & 0 & 5 \\ 3 & -4 & 0 \\ 4 & 3 & 0 \end{pmatrix}$. Write down the characteristic polynomial for A.

147. Find all eigenvalues for the matrix

$$A = \begin{pmatrix} 2 & 0 \\ 3 & -2 \end{pmatrix}.$$

Go on to find the eigenvectors corresponding to those eigenvalues.