Mathematics 1E2 (Maths for Engineers 2) 2006–07

Colm Ó Dúnlaing

April 30, 2007

Contents

1 Gauss-Jordan elimination

Gauss-Jordan elimination is a foolproof method for solving any number of simultaneous linear equations.

Example. Solve

$$x + 2y = 3 \quad \text{and} \quad 3x + 4y = 7.$$

(1.1) The first step is to write out the coefficients in a rectangular array. This array is called the *augmented matrix* for the system.

$$\begin{array}{ccc} 1 & 2 & 3 \\ 3 & 4 & 7 \end{array}$$

Each row of the augmented matrix represents an equation. We can alter the augmented matrix by the following operations. These are allowed, because they change the *equations*, not the *set of solutions*.

- **Scale** a row by any **nonzero** constant a.
- **Subtract** from one row a multiple of a different row.
- **Swap** two rows.

The above operations are called **elementary row operations** (EROs).

(1.2) The next step is to bring the matrix to a standard form, called *reduced row-echelon form*, using only EROs. This procedure is called *Gauss-Jordan elimination.*

```
Example.

1 2 3 ..... R1
3 4 7 subtract 3*R1

1  2  3 subtract 2*R2
0 -2 -2 scale by (-1/2), result is R2

1  0  1
0  1  1
```

Finished; it is now obvious that $x = 1$ and $y = 1$.

(1.3) Definition *An $m \times n$ matrix A is a rectangular array of numbers containing m rows and n columns. Usually one uses square brackets to mark its boundary, such as*

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 7 \end{bmatrix}.$$

Sometimes one writes $A_{m\times n}$ to emphasise its dimensions.
We use the notation — and this differs from the textbooks — A_j to mean the j-th column of A.
We use the notation I for the $n \times n$ identity matrix: e.g., if $n = 3$ then

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

With this notation, and $n = 3$,

$$I_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, I_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, I_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The k-th column A_k of A is a leading column *if for some i, (i,k) element of A is nonzero, but for all $j < k$, the (i,j)-th element of A is zero.*

$A_{m\times n}$ is in reduced row-echelon form *(RREF) if for some k, where $k \leq m$ and $k \leq n$, the leading columns of A are*

$$I_1, \ldots, I_k,$$

in that order.

(1.4) Lemma *In an RREF matrix the number k of leading columns equals the number of nonzero rows, and the bottom $m - k$ rows are entirely zero.*

Proof. Since the matrix contains $I_1, \ldots, I_k$, there is at least one nonzero entry in each of the first k rows.

If any other row were nonzero, then the leftmost nonzero entry in that row would be in another leading column, different from $I_1, \ldots, I_k$. ∎

For example,

```
0   0   1   0   2   0   0
0   0   0   1   1   0   0
0   0   0   0   0   1   1
0   0   0   0   0   0   0
0   0   0   0   0   0   0
```

We observe there are 3 nonzero rows, and the leading columns are columns numbered 3, 4, and 6.

(1.5) Gauss-Jordan elimination (GJE). Let A be an $m \times n$ matrix. The following procedure converts A to RREF.

- Let $k = 0$. This will be the number of leading columns produced, initially none.
- For $1 \leq j \leq n$, consider the j-th column A_j. If in this column the bottom $m - k$ entries are zero, pass on to the next column. Otherwise...
- Replace k by $k + 1$.
- Make sure the k-th entry of A_j is nozero by swapping the k-th row with a row below it, if necessary.
- Make sure the k-th entry of A_j is 1 by scaling, if necessary.
- Make sure all other entries of A_j are zero by subtracting multiples of the k-th row, if necessary.
- At this point, k has been incremented and $A_j = I_j$, as required.

```
Example of Gauss-Jordan Elimination:
inspect column no. 1
   1   3  -8  -4 -16 = R1
  -2  -8  20   8  34 subtract -2* R1
  -2  -5  14   7  28 subtract -2* R1
   2   4 -13  -9 -34 subtract 2* R1


inspect column no. 2
   1   3  -8  -4 -16 subtract 3*R2
   0  -2   4   0   2 *(-1/2) = R2
   0   1  -2  -1  -4 subtract R2
   0  -2   3  -1  -2 subtract -2*R2

inspect column no. 3
   1   0  -2  -4 -13
   0   1  -2   0  -1
   0   0   0  -1  -3 swap
   0   0  -1  -1  -4 swap

   1   0  -2  -4 -13 subtract (-2)*R3
   0   1  -2   0  -1 subtract (-2)*R3
   0   0  -1  -1  -4 *(-1) = R3
   0   0   0  -1  -3

inspect column no. 4
  1  0  0 -2 -5 subtract (-2)*R4
  0  1  0  2  7 subtract 2*R4
  0  0  1  1  4 subtract R4
  0  0  0 -1 -3 *(-1) = R4

Reduced Row-Echelon Form:
 1 0 0 0 1
 0 1 0 0 1
 0 0 1 0 1
 0 0 0 1 3
```

Interpreting the solutions. There were four equations in four unknowns. Suppose the unknowns were x_1, x_2, x_3, x_4. The RREF says:

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 1, \quad x_4 = 3.$$

There is a unique solution in this case.

```
Another example of Gauss-Jordan Elimination:
  -2  -4  -8 -14 -12 -42 -20
```

```
   3   6  12  21  19  66  31
   2   5   9  16  14  49  23
  -2  -1  -5  -8  -4 -15  -9
  -2  -3  -7 -12 -13 -44 -20

process column no. 1
scale row 1 by -1/2
subtract 3* row 1 from row 2
subtract 2* row 1 from row 3
subtract -2* row 1 from row 4
subtract -2* row 1 from row 5

  1  2  4  7  6 21 10
  0  0  0  0  1  3  1
  0  1  1  2  2  7  3
  0  3  3  6  8 27 11
  0  1  1  2 -1 -2  0

process column no. 2
swap rows 2, 3
subtract 2* row 2 from row 1
subtract 3* row 2 from row 4
subtract 1* row 2 from row 5

  1  0  2  3  2  7  4
  0  1  1  2  2  7  3
  0  0  0  0  1  3  1
  0  0  0  0  2  6  2
  0  0  0  0 -3 -9 -3

process column no. 5
subtract 2* row 3 from row 1
subtract 2* row 3 from row 2
subtract 2* row 3 from row 4
subtract -3* row 3 from row 5

Reduced Row-Echelon Form:
 1 0 2 3 0 1 2
 0 1 1 2 0 1 1
 0 0 0 0 1 3 1
 0 0 0 0 0 0 0
 0 0 0 0 0 0 0
```

(1.6) Interpreting the RREF. The above 5×7 matrix represented 5 equations in 6 unknowns. Let us assume the unknowns are $x_1, x_2, x_3, x_4, x_5, x_6$. The simple rule is:

The variables corresponding to *leading columns* may be expressed in terms of the *other* variables.

In the above example, the leading columns are numbered $1, 2,$ and 5. I like to relabel the 'nonleading' variables as $r, s, t \ldots$ and so on. The 'nonleading variables' are $x_3, x_4,$ and x_6. The equations can be rewritten

$$\begin{aligned} x_1 + 2x_3 + 3x_4 + x_6 = 2; \quad & x_1 = 2 - 2r - 3s - t \\ x_2 + x_3 + 2x_4 + x_6 = 1; \quad & x_2 = 1 - r - 2s - t \\ x_5 + 3x_6 = 1; \quad & x_5 = 1 - 3t. \end{aligned}$$

Of course, rows 4 and 5 of the RREF carry no information.

Here is another example.

```
 0   0   0   2   6
 0  -2  -6  -4 -14
 0   3   9   6  19

inspect column no. 2
swap rows 1, 2
scale row 1 by -1/2
subtract 3* row 1 from row 3

 0  1  3  2  7
 0  0  0  2  6
 0  0  0  0 -2

It is unnecessary to go any further, but if
you did you would get the following RREF:

0 1 3 0 0
0 0 0 1 0
0 0 0 0 1
```

It was unnecessary to go any further because the bottom row said

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

which has *no* solution. The equations have no solution. They are inconsistent.

2 Matrix algebra

There are rules for 'adding' and 'multiplying' matrices. Two matrices can be added if and only if they have the same dimensions, in which case the addition is entry-by-entry.

For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 3 & 5 \\ 7 & 9 & 11 \end{bmatrix} = \begin{bmatrix} 2 & 5 & 8 \\ 11 & 14 & 17 \end{bmatrix}.$$

By the way,

> The (i, j)-entry of a matrix A is the entry in its i-th row and j-th column.
> If we write $A = [a_{ij}]_{m\times n}$ it means that A is an $m \times n$ matrix and for $1 \le i \le m$, $1 \le j \le n$, a_{ij} is the (i, j)-entry of A.
> A **row vector** is a matrix of height 1.
> A **column vector** is a matrix of width 1.

Adding matrices is not of much interest in this course. Much more important is the strange rule for multiplying matrices.

(2.1) Definition *If A and B are matrices, then the matrix product AB is defined if and only if*

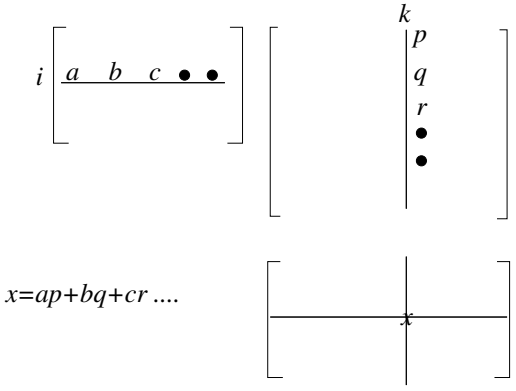
The width of A equals the height of B

and it is calculated as follows.

Suppose $A = [a_{ij}]_{\ell\times m}$ and $B = [b_{jk}]_{m\times n}$, then AB is an $\ell \times n$ matrix: $AB = [c_{ik}]_{\ell\times n}$, where for $1 \le i \le \ell, 1 \le k \le n$,

$$c_{ik} = \sum_j a_{ij}b_{jk}.$$

The (i, k)-entry in AB comes from the i-th row of A and the k-th column of B. The following picture may help:



For example

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}, \quad C = \begin{bmatrix} 3 & -1 \\ -6 & 2 \end{bmatrix}.$$

Then

$$AB = \begin{bmatrix} 7 & 14 \\ 15 & 30 \\ 23 & 46 \end{bmatrix}, \quad BC = {}^{1}\begin{bmatrix} -9 & 3 \\ -27 & 9 \end{bmatrix}, \quad \text{and} \quad CB = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

[1] This was wrong in earlier versions.

Notice that AC is defined but BA and CA are undefined.

(2.2) Remember the identity matrices I? They have the property that for an $m \times n$ matrix A, $IA = A$ (this I is an $m \times m$ matrix) and $AI = I$ (this time it is an $n \times n$ matrix).

Some matrices A have an *inverse* A^{-1}. Inverses only exist if A is square, and then not always.

If A is an $n \times n$ matrix and A^{-1} exists, then A^{-1} is an $n \times n$ matrix such that

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

There is a simple formula for A^{-1} when A is 2×2:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

There is a quite efficient method using GJE:

- Given an $n \times n$ matrix A, construct an $n \times 2n$ 'augmented matrix' by appending the n columns of the identity matrix to A.
- Bring this $n \times 2n$ matrix to RREF.
- If the first n columns of the RREF form the identity matrix, then the last n columns form A^{-1}. Otherwise the inverse does not exist.

Simple example.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} = \frac{1}{-2} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & \frac{-1}{2} \end{bmatrix}$$

```
1 2 1 0 = R1        1  2  1 0 - 2*R2          1  0  -2    1
3 4 0 1 - 3*R1      0 -2 -3 1 *(-1/2)=R2      0  1  3/2 -1/2   in RREF
```

3 Vectors in 2 dimensions

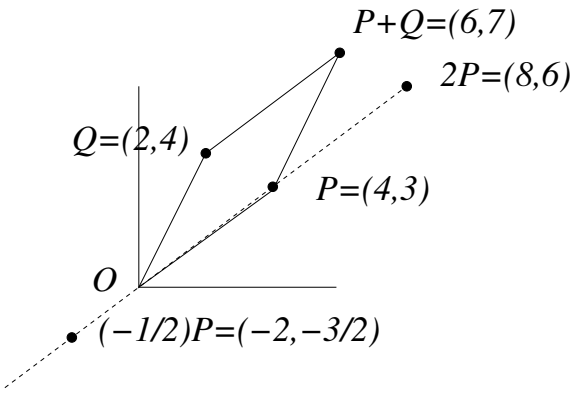
The cartesian coordinate system (in 2 dimensions) expresses points in the plane as pairs (x, y) of numbers.

Now, such pairs of numbers can be treated like row vectors and we can define the following

- **Scaling:** for example, $(0.5)(1, 2) = (0.5, 1)$.
- **Adding:** for example, $(1, 2) + (3, 4) = (4, 6)$.
- These two operations can be combined into one. Given numbers r, s and points $P = (a, b)$, $Q = (c, d)$,
$$rP + sQ = (ra + sc, rb + sd).$$
Such an operation is called *linear combination.*
- To emphasise that numbers are connected with scaling, they are called *scalars.*

- We write O for the origin of the coordinate system, i.e., $O = (0, 0)$.
- We write $|OP|$ for the distance between P and the origin. For example, $|(1, 3)| = \sqrt{10}$.

(3.1) How do we interpret these operations geometrically? See the diagram.



If $P = O$ or $r = 0$ then $rP = O$, the origin. Otherwise there is a unique line through OP and rP is on that line, its distance from the origin scaled by $|r|$ (the absolute value of r); if $r > 0$ then P and rP are on the same side of the origin and if $r < 0$ then they are on opposite sides.

The sum $P + Q$ is the fourth corner of the parallelogram whose other corners are O, P, and Q. This is called the *parallelogram law* of addition.

Scaling makes good sense, but the parallelogram law of addition is hard to absorb. There is another way of interpreting them.

(3.2) Definition *The* displacement *from P to Q is the directed line-segment from P to Q. It is written $\vec{PQ}$ and called a* vector.

There is a natural way to 'add' a point P and a vector $\vec{PQ}$:

$$P + \vec{PQ} = Q.$$

Two vectors are regarded as the same if they have the same length and direction.

If $P_1 = (1, 2)$, $Q_1 = (4, 3)$, $P_2 = (3, 4)$, and $Q_2 = (6, 5)$, then $P_1\vec{Q}_1 = P_2\vec{Q}_2$.

Then

$$P + Q = P + \vec{OQ}.$$

Explain this componentwise.

The equation

$$y = 2x + 3$$

describes a line in 2 dimensions. The line *is* the set of all points (x, y) such that $y = 2x + 3$.

In set notation, it is

$$\{(x, y) : \ y = 2x + 3\}.$$

A more general equation for a line is

$$ax + by = c.$$

If $b \neq 0$ this is equivalent to

$$y = \frac{c}{b} - \frac{a}{b}x.$$

Given two lines

$$ax + by = c, \quad dx + ey = f$$

their intersection is the set of all points (x, y) which satisfy both equations.

If the lines are parallel, then there are no solutions (if they are different lines) or infinitely many (if they are the same line).

If the lines are not parallel then there is a unique point of intersection and the equations have a unique solution.

The lines are *not* parallel if and only if the matrix

$$\begin{bmatrix} a & b \\ d & e \end{bmatrix}$$

is invertible, which is true if and only if

$$ae - bd \neq 0.$$

If $P \neq Q$ then there is a unique line through P and Q. It is

$$\{P + r\vec{PQ} : \ r \in \mathbb{R}\}.$$

($\mathbb{R}$ is the set of real numbers.) Generally if P is any point and $\vec{V}$ is any nonzero vector then

$$\{P + r\vec{V} : \ r \in \mathbb{R}\}$$

is the line through P parallel to the direction of $\vec{V}$.

(3.3) Parametrising the set of solutions to a system of linear equations. We consider a simple example:

$$x + 2y = 3; \quad 2x + 4y = 6.$$

These equations are consistent but not independent. Forming the augmented matrix in the usual way and applying GJE we get the RREF

```
1     2     3
0     0     0
```

which leads to the solution

$$x = 3 - 2r \quad \text{and} \quad y = r,$$

where r can be any real number. Put differently, the set of solutions is

$$\{(3, 0) + r(-2, 1) : \ r \in \mathbb{R}\}.$$

This is a *line in parametric form.*

Parametrisation can often be handy.[2] The equation $y = x^2$ describes a *parabola* which is the set of points

$$\{(x, y) : \ y = x^2\},$$

but the same curve set of points can be 'parametrised' as

$$\{(t^2, t) : \ t \in \mathbb{R}\}.$$

The equation $x^2 + y^2 = 1$ describes a circle

$$\{(x, y) : \ x^2 + y^2 = 1\}$$

which can be parametrised as

$$\{(\cos t, \sin t) : 0 \leq t < 2\pi\}.$$

The equation

$$4x^2 + y^2 = 4$$

describes an ellipse

$$\{(x, y) : \ 4x^2 + y^2 = 4\}$$

which can be parametrised as

$$\{(\cos t, 2\sin t) : \ 0 \leq t < 2\pi\}.$$

4 Vectors in 3 dimensions

We can have coordinate systems in 3 dimensions. Now we have (x, y, z) as coordinates. Think of x as distance east, y as distance north, and z as distance up.

In three dimensions as in two, a line can be parametrised as

$$\{P + r\vec{PQ} : \ r \in \mathbb{R}\}.$$

A plane, not parallel to the z-axis, can be written in the form

$$z = \ell x + my + c$$

or more generally as

$$ax + by + cz = d.$$

Again it can be parametrised, but this time with two unknowns, if we can find three points P, Q, R in the plane, so long as they are not collinear. Then the plane can be parametrised as

$$\{P + r\vec{PQ} + s\vec{PR} : \ r, s \in \mathbb{R}\}.$$

Given three linear equations in three unknowns, we can interpret the solution as the intersection of three planes. Usually they intersect in exactly one point, but the other possibilities are no points

[2] This is just background material, not part of the course.

(inconsistent), a line, a plane, or in one case, the whole of 3-dimensional space. The set of solutions can be calculated by GJE, and the result expressed in terms of $0, 1, 2$, or 3 parameters r, s, t.

For example, the single equation

$$x + 2y + 3z = 6$$

can be solved as follows.

```
1   2   3   6
```

This is already in RREF. The nonleading variables are y and z, and the solutions set are

$$x = 6 - 2r - 3s, \quad y = r, \quad z = s.$$

To express the result differently, write this as

$$(x, y, z) = (6 - 2r - 3s, r, s)$$

or rather, separating the constant part, the part involving r, and the part involving s,

$$(x, y, z) = (6, 0, 0) + r(-2, 1, 0) + s(-3, 0, 1).$$

The set of solutions is

$$\{(6, 0, 0) + r(-2, 1, 0) + s(-3, 0, 1) : \ r, s \in \mathbb{R}\}.$$

This is more-or-less what a parametrisation of the plane means.

Question: can you produce a P, a Q, and an R from the above?

Three equations

$$x + 2y + 3z = 6, \quad 2x + 4y + 6z = 12, \quad \text{and} \quad 3x + 6y + 9z = 18$$

have no more information than the first and their solution is the intersection of three planes, which are the same.

The equations

$$x + 2y + 3z = 6, \quad 4x + 5y + 6z = 15, \quad 7x + 8y + 9z = 24$$

describe three planes whose intersection is a line. The intersection of the first two is a line contained in the third. Applying GJE in the usual way we reduce the augmented matrix to RREF

```
1  0 -1  0
0  1  2  3
0  0  0  0
```

The solutions are

$$x = r, \quad y = 3 - 2r, \quad z = r.$$

This is a line

$$\{(r, 3 - 2r, r) : \ r \in \mathbb{R}\},$$

or equivalently

$$\{(0, 3, 0) + r(1, -2, 1) : \ r \in \mathbb{R}\}.$$

5 Coordinate systems

The following theorem is obviously important, and it doesn't belong here, but it should have been stated earlier. Therefore it is stated here.

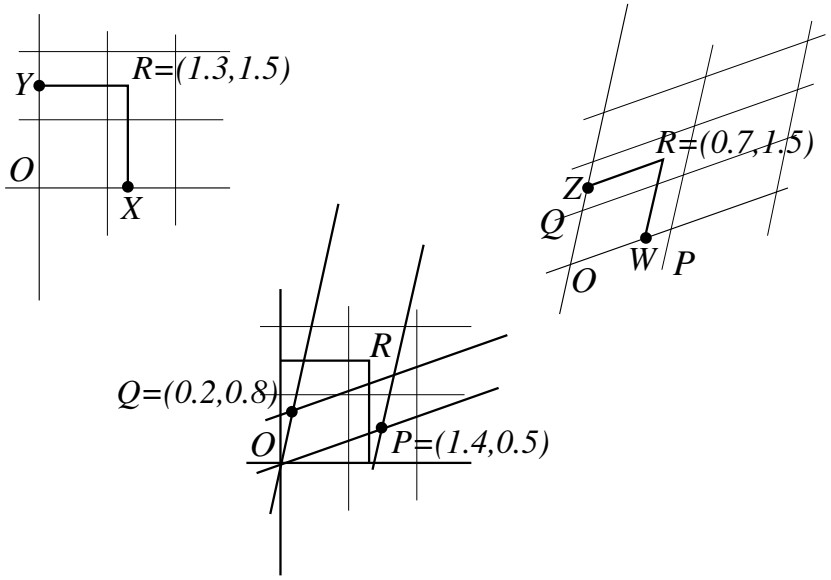
(5.1) Theorem *Let A be an $m \times n$ matrix. The equation*

$$AX = B$$

has a unique *solution X for* every *column vector B of height m if and only if A is an invertible square matrix.*

In this case the unique solution is $X = A^{-1}B$. (Proof omitted.)

Cartesian coordinates are set up by setting up two coordinate axes intersecting orthogonally at the origin O, and marking points at unit distance from O along these axes. The coordinate system suggests a division of the plane into squares.



The x- and y-coordinates of a point R are calculated by completing a *rectangle* $OXRY$ where X and Y are on the x- and y-axes, and measuring the (signed) distances of X and Y to O.

One can also have *oblique* coordinate systems where intersect at O but perhaps not orthogonally. The coordinates of R are calculated by completing a *parallelgram* $OWRZ$ and calculating the distances of W and Z from O.

Distances can be measured differently along these oblique axes. The oblique coordinate system suggests a division of the plane into parallelgrams.

Let P and Q be the points whose *new* coordinates are $(1, 0)$ and $(0, 1)$ respectively. By construction,

$R = \alpha P + \beta Q$ where (α, β) are the coordinates of R in the *new* system.

Of course the coordinates of a point in the new system, or the cartesian coordinate-system, are uniquely defined for each point R. This leads to the following important definition.

(5.2) Definition *A list $P_1, P_2, \ldots P_n$ of points in the plane forms an* ordered basis *for a coordinate system in the plane if for every point R there exist unique scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $R = \alpha_1 P_1 + \alpha_2 P_2 + \ldots \alpha_n P_n$.*

Fact. It can be shown that $n = 2$, and any two points P, Q will be a basis so long as O, P, Q are not collinear.

Algebraically, let $P = (a, b)$ and $Q = (c, d)$. Old and new coordinates are connected by

$$\alpha P + \beta Q = (x, y),$$

meaning

$$a\alpha + c\beta = x, \quad b\alpha + d\beta = y,$$

or

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

(5.3) Definition *Given a new basis $P_1, P_2, \ldots, P_n$ (if we are discussing the plane then $n = 2$, or if discussing 3-space then $n = 3$), let S be the matrix whose* columns *are the* standard cartesian *coordinates of $P_1, \ldots, P_n$. We call S the change-of-basis matrix.*[3]

The relation between *standard* coordinates $(x_1, \ldots, x_n)$ and *new* coordinates $(\alpha_1, \ldots, \alpha_n)$ of the same point R is

$$\alpha_1 P_1 + \ldots + \alpha_n P_n = (x_1, \ldots, x_n),$$

or in matrix terms

$$S \begin{bmatrix} \alpha_1 \\ \bullet \\ \bullet \\ \alpha_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \bullet \\ \bullet \\ x_n \end{bmatrix}.$$

Example. In the illustration, by eyeballing,

$$(x, y) = (1.3, 1.5), \quad P = (1.4, 0.5), \quad Q = (0.2, 0.8), \quad \text{and } (\alpha, \beta) = (0.7, 1.5).$$

According to our formula,

$$\begin{bmatrix} 1.4 & 0.2 \\ 0.5 & 0.8 \end{bmatrix} \begin{bmatrix} 0.7 \\ 1.5 \end{bmatrix} = \begin{bmatrix} 1.3 \\ 1.5 \end{bmatrix}.$$

The left-hand side is

$$\begin{bmatrix} 1.28 \\ 1.55 \end{bmatrix}.$$

Close enough.

Facts.
The same ideas work in 3 dimensions except of course that the cartesian coordinates form a triple (x, y, z).
It can be shown that every basis for 3-dimensional space contains exactly 3 points P, Q, R, and they form a basis if and only if O, P, Q, R are not coplanar.

[3] Any textbook on linear algebra will discuss this, but may disagree over which is S and which is S^{-1}.

6 Inner products

(6.1) In two dimensions one defines

$$\vec{OP} \cdot \vec{OQ} = |\vec{OP}||\vec{OQ}| \cos \widehat{POQ}.$$

Since $\cos\theta = 0$ if and only if θ is a right angle, assuming $P \neq O$ and $Q \neq O$,

The angle $\widehat{POQ}$ is a right angle if and only if $\vec{OP} \cdot \vec{OQ} = 0$.

(6.2) One can show, based on the **cosine rule** for triangles, that if $P = (a, b)$ and $Q = (c, d)$, then

$$\vec{OP} \cdot \vec{OQ} = ac + bd.$$

As a simple example, if $P = (2, 0)$ and $Q = (3, 3)$, then

$$\cos \widehat{POQ} = \frac{(2)(3) + (0)(3)}{2(3\sqrt{2}} = \frac{1}{\sqrt{2}},$$

so $\widehat{POQ} = \pi/4$.

(6.3) We can interpret the equation

$$ax + by = c$$

as describing the set of points

$$\{X : \ \vec{ON} \cdot X = \vec{ON} \cdot \vec{OP}$$

where $N = (a, b)$, a *normal* to the line, and P is any point on the line.

(6.4) If $P = (a, b)$, then the *positive normal* to P is obtained by rotating P $90°$ anticlockwise around O. Its coordinates are $N = (-b, a)$. Notice that $|\vec{ON}| = |\vec{OP}|$ and $\vec{ON} \cdot \vec{OP} = 0$.

(6.5) We can use this to produce an equation for the line through two points P and Q. Let N be the positive normal to $Q - P$. Then the equation is

$$\vec{ON} \cdot \vec{OX} = \vec{ON} \cdot \vec{OP}.$$

For example, let us give an equation for the line through $(1, 2)$ and $(3, 1)$.

$$P = (1, 2), \quad Q = (3, 1), \quad Q - P = (2, -1), \quad N = (1, 2).$$

The equation is

$$\vec{ON} \cdot \vec{OX} = \vec{ON} \cdot \vec{OP}.$$

Now if $X = (x, y)$, $\vec{ON} \cdot \vec{OX} = (1)x + (2)y = x + 2y$, and $\vec{ON} \cdot \vec{OP} = (1)(1) + (2)(2) = 5$:

$$x + 2y = 5.$$

(6.6) The same definition

$$\vec{OP} \cdot \vec{OQ} = |\vec{OP}||\vec{OQ}| \cos \widehat{POQ}$$

can be used in three dimensions. This time the cosine law yields the simple rule

If $P = (a, b, c)$ and $Q = (d, e, f)$ then $\vec{OP} \cdot \vec{OQ} = ad + be + cf$.

(6.7) The equation $ax + by + cz = d$, which describes a plane, can be intepreted in the form

$$\vec{ON} \cdot \vec{OX} = \vec{ON} \cdot \vec{OP},$$

where $N = (a, b, c)$ and P is some point in the plane.

(6.8) If P and Q are two points in 3-space, and O, P, Q are not collinear, then there is a unique plane containing O, P, and Q. There is a direction *normal* to the plane. Corresponding to the *positive normal* to a single point in two dimensions, we have the *cross product* $\vec{ON} = \vec{OP} \times \vec{OQ}$ in three dimensions, so that $\vec{ON}$ is perpendicular both to $\vec{OP}$ and to $\vec{OQ}$.

Suppose $P = (a, b, c)$ and $Q = (d, e, f)$. We want to calculate a point $N = (x, y, z)$ so that $\vec{OP} \cdot \vec{ON} = 0$ and $\vec{OQ} \cdot \vec{ON} = 0$.

$$ax + by = -cz \quad dx + ey = -fz.$$

This time we shall not apply GJE, but cross-multiply.

$$aex + bey = -cez \quad bdx + bey = -bfz.$$

Subtracting,

$$(ae - bd)x = (bf - ce)z.$$

If we choose $z = ae - bd$, then we get the cleanest solution.

$$x = bf - ce, \quad z = ae - bd.$$

Substituting for y,

$$a(bf - ce) + by = -c(ae - bd),$$
$$y = cd - af.$$

Thus the cross-product formula is, in a loose notation,

$$(a, b, c) \times (d, e, f) = (be - cf, cd - af, ae - bd).$$

There is a notation for 2×2 determinants, as follows:

$$\begin{vmatrix} w & x \\ y & z \end{vmatrix} = wz - xy.$$

(Despite the matrix-like notation the determinant is a single number.) Then

$$(a, b, c) \times (d, e, f) = \left(\begin{vmatrix} b & c \\ e & f \end{vmatrix}, - \begin{vmatrix} a & c \\ d & f \end{vmatrix}, \begin{vmatrix} a & b \\ d & e \end{vmatrix} \right).$$

Be careful of the sign in the middle.

Remark. If O, P, and Q are collinear then $\vec{OP} \times \vec{OQ} = \vec{O}$.

For example, give an equation for the plane passing through $P = (1, 2, 3), Q = (4, 3, 5), R = (7, 6, 5)$. The solution is

$$\vec{ON} \cdot \vec{OX} = \vec{ON} \cdot \vec{OP},$$

where $\vec{ON} = \vec{PQ} \times \vec{PR}$. In loose notation, $\vec{PQ} = (3,1,2)$ and $\vec{PR} = (6,4,2)$,

$$(3,1,2) \times (6,4,2) = ((1)(2) - (2)(4), (2)(6) - (3)(2), (3)(4) - (1)(6)) = (-6,6,6).$$

The equation is, in loose notation,

$$(-6,6,6) \cdot \vec{OX} = (-6,6,6) \cdot \vec{OP},$$

or

$$-6x + 6y + 6z = (-6)(1) + 6(2) + 6(3) = 24.$$

Gram-Schmidt orthogonalisation. It is often useful to have a basis U_1, U_2, U_3 in 3 dimensions such that $|\vec{OU_1}| = 1$, $|\vec{OU_2}| = 1$, and $|\vec{OU_3}| = 1$, and $\vec{OU_1}$, $\vec{OU_2}$, and $\vec{OU_3}$ are mutually perpendicular. Such a basis is called an *orthonormal basis.*

Gram-Schmidt orthogonalisation takes any basis P_1, P_2, and P_3, and produces an orthonormal basis U_1, U_2, U_3 such that U_1 is in the line OP_1 and U_2 is in the plane OP_1P_2.

The calculation is as follows. First *normalise* P_1,:

$$W_1 = P_1; U_1 = W_1/|\vec{OW_1}|.$$

$$|\vec{OU_1}| = \frac{|\vec{OW_1}|}{\vec{OW_1}|} = 1.$$

Next

$$W_2 = P_2 - \frac{\vec{OW_1} \cdot \vec{OP_2}}{\vec{OW_1} \cdot \vec{OW_1}} W_1.$$

$$\vec{OW_1} \cdot \vec{OW_2} = \vec{OW_1} \cdot \vec{OP_2} - \frac{\vec{OW_1} \cdot \vec{OP_2}}{\vec{OW_1} \cdot \vec{OW_1}} \vec{OW_1} \cdot \vec{OW_1} = 0.$$

Then normalise W_2: $U_2 = W_2/|\vec{OW_2}|$. As with U_1, $|\vec{OU_2}| = 1$. Finally,

$$W_3 = P_3 - \frac{\vec{OW_1} \cdot \vec{OP_3}}{\vec{OW_1} \cdot \vec{OW_1}} W_1 - \frac{\vec{OW_2} \cdot \vec{OP_3}}{\vec{OW_2} \cdot \vec{OW_2}} W_2.$$

Again it can be shown that $\vec{OW_1} \cdot \vec{OW_3} = \vec{OW_2} \cdot \vec{OW_3} = 0$. Normalise W_3: $U_3 = W_3/|\vec{OW_3}|$.

For example, apply Gram-Schmidt to $(1,2,3), (4,3,5), (7,6,5)$.

First $W_1 = (1,2,3)$. $U_1 = W_1/\sqrt{1+4+9} = (1,2,3)/\sqrt{14}$.

Next,

$$W_2 = (4,3,5) - \frac{(1,2,3) \cdot (4,3,5)}{(1,2,3) \cdot (1,2,3)}(1,2,3)$$
$$= (4,3,5) - \frac{4+6+15}{14}(1,2,3) = (4,3,5) - \frac{25}{14}(1,2,3)$$
$$= (\frac{31}{14}, \frac{-8}{14}, \frac{-5}{14}).$$

Normalise to get U_2.

Last,

$$W_3 = (7,6,5) - \frac{(1,2,3)\cdot(7,6,5)}{(1,2,3)\cdot(1,2,3)}(1,2,3) - \frac{(\frac{31}{14},\frac{-8}{14},\frac{-5}{14})\cdot(7,6,5)}{(\frac{31}{14},\frac{-8}{14},\frac{-5}{14})\cdot(\frac{31}{14},\frac{-8}{14},\frac{-5}{14})}(\frac{31}{14},\frac{-8}{14},\frac{-5}{14}) =$$
$$(7,6,5)\frac{34}{14}(1,2,3) - \frac{144}{1050}(31,-8,-5) =$$
$$\frac{1}{1050}(336,2352,-1680).$$

Then let U_3 be obtained by normalising W_3.

For this 3-dimensional example, there is a much easier way to compute U_3: simply let

$$\vec{OW_3} = \vec{OP_1} \times \vec{OP_2}.$$

In 3-dimensions, the choice of P_3 has little effect on U_3 (its only effect is on the sign of U_3). This gives

$$(1,7,-5).$$

$$U_1 = (1,2,3)/\sqrt{14}, \quad U_2 = (31,-8,-5)/\sqrt{1050}, \quad U_3 = (1,7,-5)/\sqrt{75}.$$

7 Linear regression

(7.1) Definition *If C is an $m \times n$ matrix $[c_{ij}]$, then its* transpose *C^T is the $n \times m$ matrix $[d_{ij}]$, where $d_{ij} = c_{ji}$. The transpose of a row-vector is a column-vector and vice-versa.*

This section is about fitting a straight-line graph to data. For example, we might want a straight-line graph which matches the following data closely:

$$(1,2)(3,5)(4,3)$$

These points are not collinear, so we need a compromise. Now the line has an equation

$$y = mx + c$$

and we want to find m and c to get the best fit, in some sense. The line is *parametrised* with two parameters m and c.

If we could get a straight-line solution then the following equations would be satisfied:

$$2 = m + c; \quad 5 = 3m + c; \quad 3 = 4m + c.$$

That is, we want a solution to the equations

$$\begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix}.$$

This has the form

$$AX = B$$

where $X = [m \;\; c]^T$.

If $U = [u_i]$ and $V = [v_i]$ are two column-vectors of height 2 or 3, then of course

$$U^T V = u_1 v_1 + u_2 v_2 (+ u_3 v_3).$$

In other words, $U^T V$ is the *dot product*.

Returning to $X = [m \;\; c]^T$. As m and c vary, the points

$$\left\{ A \begin{bmatrix} m \\ c \end{bmatrix} : m, c \in \mathbb{R} \right\}$$

parametrise a plane in three dimensions. If B is in the plane then the points $(1, 2), (3, 5), (4, 3)$ are collinear. They aren't, and B isn't. But we can calculate those values of m and c which produce a point *closest* to B. Again with $X = [m \;\; c]^T$, the *squared* distance from B to AX is

$$(AX - B)^T (AX - B).$$

We can calculate the closest point by a kind of calculus.

(7.2) As a kind of explanation, we suppose we want to minimise the value of

$$x^2 + x + 2.$$

For any given value of x we can test whether the value is minimal by considering values close to x. Let h be 'small.' Then the value becomes

$$(x + h)^2 + (x + h) + 2 = x^2 + x + 2 + 2hx + h + h^2.$$

The change in value is $h(2x + 1) + h^2$. If $2x + 1 \neq 0$ then one can choose a small h which makes the value even smaller. Hence $2x + 1 = 0$ is necessary for a minimum, and we conclude that $x = -1/2$ and the minimum value is 2.75.

Returning to our sophisticated problem, we imagine a small alteration H to X (recall again that $X = [m \;\; c]^T$).

$$(A(X + H) - B)^T (A(X + H) - B).$$

Properties of matrix multiplication allow us to expand this as follows.

$$\begin{aligned}(AX + AH - B)^T (AX + AH - B) = (AX - B + AH)^T (AX - B + AH) = \\ (AX - B)^T (AX - B) + (AH)^T (AX - B) + (AX - B)^T (AH) + H^T H.\end{aligned}$$

The squared distance has been changed by

$$(AH)^T (AX - B) + (AX - B)^T (AH) + H^T H.$$

Note at this point that if U and V are column vectors of the same height (3 in this case), then $U^T V = V^T U$. Take $U = AH$ and $V = AX - B$, so the change in squared distance is

$$2(AH)^T (AX - B) + H^T H.$$

We have not studied the properties of matrix multiplication and transposition, but the following steps are valid.

$$2(AH)^T(AX - B) + H^T H = H^T(2A^T(AX - B)) + H^T H.$$

As with the previous minimisation problem, this squared distance can be reduced (by choosing a suitable small H) unless

$$2A^T(AX - B) = O.$$

This gives us what we want:

$$A^T AX - A^T B = O \quad \text{so} \quad A^T AX = A^T B$$

and

$$X = (A^T A)^{-1} A^T B.$$

The same formula is accepted for *any number* of data points.

For this example

$$A^T A = \begin{bmatrix} 26 & 8 \\ 8 & 3 \end{bmatrix}, \quad \text{and} \quad A^T B = \begin{bmatrix} 29 \\ 10 \end{bmatrix}$$

and the solution is $m = 1/2, c = 2$. The estimated line is

$$y = (x/2) + 2.$$

This line passes through the points

$$(1, 2.5), (3, 3.5), (4, 4).$$

8 Linear maps and matrices

We begin by discussing *maps*. (The words 'mappings,' 'transformations,' and 'functions' are also used. One would expect 'function' to be more commonly used in connection with calculus, and so on.)

We are familiar with the graph of a function. So the graph of the function $y = x^2$ is the following set of points in the plane

$$\{(x, y) : \ y = x^2\},$$

a parabola.

A *map* $f : X \to Y$ is a rule or procedure which assigns to every element x of X a corresponding element $f(x)$ in Y. X and Y are just sets. We call X the *domain* of the map f and Y (which is somehow less important) the *codomain*.

The connection with the parabola is that we can interpret it as a *rule* which assigns to every real number x a unique real number y, namely x^2.

A useful shorthand is

$$f : \mathbb{R} \to \mathbb{R} : \ x \mapsto x^2.$$

The calculus is about differentiable maps, and linear algebra is much concerned with *linear* maps. The domain and codomain can be $\mathbb{R}$, or 2- or 3-dimensional space, or the set of column vectors of a certain height, or, basically, any set in which *linear combination* is defined and behaves in the familiar way. Remember a linear combination is simply an expression

$$\alpha P + \beta Q$$

or (more generally)

$$\alpha_1 P_1 + \ldots + \alpha_n P_n$$

where α, β, α_j are scalars.

(8.1) Definition *A map f is* linear *if for all points P, Q and scalars α, β,*

$$f(\alpha P + \beta Q) = \alpha f(U) + \beta f(V).$$

It is easy to deduce that for any linear combination $\alpha_1 P_1 + \ldots + \alpha_n P_n$,

$$f(\alpha_1 P_1 + \ldots + \alpha_n P_n) = \alpha_1 f(P_1) + \ldots + \alpha_n f(P_n).$$

Examples.

- The *identity maps* are always linear. They are defined by the rule $x \mapsto x$.
- Vertical projection onto the x-axis is linear: $(x, y) \mapsto x$.
- Orthogonal projection onto the line $x = y$ is linear: $(x, y) \mapsto ((x + y)/2, (x + y)/2)$.
- 90° anticlockwise rotation around O is linear, i.e., the 'positive normal' map $(x, y) \mapsto (-y, x)$.
- Anticlockwise rotation through *any fixed* angle ϕ around O is linear. This is best described in *polar coordinates*: $(r, \theta) \mapsto (r, \theta + \phi)$.
- Orthogonal reflection in the x-axis is linear: $(x, y) \mapsto (x, -y)$.
- Orthogonal projection onto the xy-plane in three dimensions is linear: $(x, y, z) \mapsto (x, y, 0)$.
- Orthogonal projection onto the x-axis is linear: $(x, y, z) \mapsto (x, 0, 0)$.
- Orthogonal projection in the xy-plane is linear: $(x, y, z) \mapsto (x, y, -z)$.
- Rotation through a fixed angle around any axis through O in 3 dimensions is linear.
- Projection onto any line *not* containing O is *not* linear.
- Projection onto any plane *not* containing O is *not* linear.
- Rotation around any axis *not* containing O is *not* linear.
- The map $(x, y, z) \mapsto (x^2, 0, 0)$ is *not* linear.
- Translation maps are *not* linear.

(8.2) Proposition *If f is a linear map, then f always carries straight lines into straight lines, and always carries O to O.*

(8.3) *Except in unusual cases, if f is not linear then either it does* not *take O to O or carries some straight line into a curve which is not a straight line.*

Matrices and linear maps. Suppose f is a linear map from the plane to the plane. Suppose $f((1,0)) = (a,b)$ and $f((0,1)) = (c,d)$. Then for any point (x,y),

$$f(x,y) = f(x(1,0) + y(0,1)) = x(f(1,0)) + yf((0,1)) = x(a,b) + y(c,d).$$

In terms of matrices, if we store coordinates in *column vectors*, the coordinates of $f(x,y)$ are

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

This reflects the fact that $(1,0),(0,1)$ is the ordered basis defining the cartesian coordinate system. More generally, for *any* ordered basis P, Q, if

$$f(P) = aP + bQ \quad \text{and} \quad f(Q) = cP + dQ,$$

then the following formula expresses the *new* coordinates of $f(X)$ in terms of the *new* coordinates (α,β) of X:

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

The same idea works, of course, in 3 dimensions.

Rotation through angle ϕ in two dimensions takes

$$(1,0) \mapsto (\cos\phi, \sin\phi) \quad \text{and} \quad (0,1) \mapsto (-\sin\phi, \cos\phi),$$

so the matrix for this map (in cartesian coordinates) is

$$\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Change of basis and matrices. Suppose S is the matrix associated with a new basis for 2- or 3-dimensional space, and A and A' are the matrices associated with the standard and new coordinate systems. Given a point X in *cartesian* coordinates,

$$S^{-1}X$$

are its *new* coordinates,

$$A'S^{-1}X$$

are the *new* coordinates of $f(X)$, and

$$SA'S^{-1}X$$

are the *cartesian* coordinates of $f(X)$. Therefore

$$A = SA'S^{-1}.$$

Often, given f, we can choose a coordinate system which makes A' easy or obvious, and use this formula to calculate A.

9 Determinants

(9.1) We have already seen a 2×2 determinant

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

Suppose that $P = (a, b)$, $Q = (c, d)$, and $N = (-b, a)$, the positive normal. Then the determinant equals

$$\vec{ON} \cdot \vec{OQ}.$$

This can be written as

$$|\vec{OP}||\vec{OQ}| \sin \widehat{POQ},$$

and can be interpreted as

$$\pm 2 * (\text{area of triangle } OPQ).$$

In particular, the determinant is zero iff O, P, and Q are collinear.

(9.2) We could have taken P and Q as the *columns* in the determinant — in general, the determinant of A and of A^T are the same.

(9.3) Determinants can be defined for square matrices of any size. For 3×3 matrices A we define $\det(A)$ as follows:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

If P, Q, R are the rows (or columns) of this matrix, then we can see that the determinant is

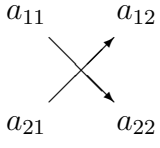
$$\vec{OP} \cdot (\vec{OQ} \times \vec{OR}).$$

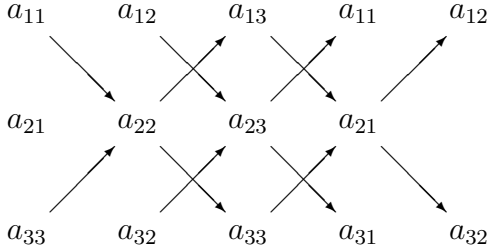
Thus the determinant is zero if and only if $\vec{OP}$ is perpendicular to the normal to the plane OQR — i.e., if and only if O, P, Q, R are coplanar.

It can be shown that

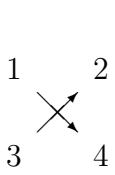
$$\vec{OP} \cdot (\vec{OQ} \times \vec{OR}) = \pm 6 * (\text{volume of tetrahedron } OPQR).$$

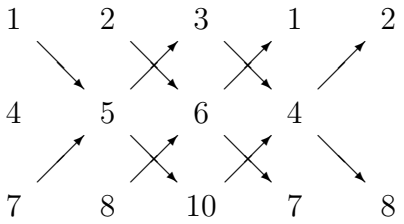
There is a simple layout which helps compute 2×2 and 3×3 determinants. 2×2 determinants are simple. The downward product is left unchanged in sign, the upward product is reversed in sign. In the 3×3 layout, the three downward products are left unchanged in sign, the three upward products are reversed in sign.





For example,





The 2×2 determinant is -2 as before. The 3×3 is

$$50 + 84 + 96 - 105 - 48 - 80 = -3.$$

Here is one way to define determinants of any order. If $A = [a_{ij}]_{n \times n}$ is an $n \times n$ matrix, let us temporarily use the notation $A \backslash i \backslash j$ to indicate the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and j-th column from A. Then

- If $n = 1$ then $\det(A) = a_{11}$.
- If $n > 1$ then
$$\det(A) = \sum_j (-1)^{1+j} \det(A \backslash i \backslash j).$$

For $n = 2, 3$ this leads to the same formulae as before. For $n = 4$ it expresses $\det(A)$ in terms of four 3×3 determinants, and so on.

(9.4) Effects of EROs on the determinant. **Scaling** scales the determinant. **Swapping** reverses its sign. **Subtracting** doesn't change the determinant. Also, if a square matrix $A = [a_{ij}]$ is *upper triangular* or *lower triangular*, meaning that $a_{ij} = 0$ if $i > j$ (respectively, $i < j$), then

$$\det(A) = a_{11} a_{22} \dots a_{nn},$$

the product of the main diagonal entries.

```
Example
   1   2   3  =R1            1   2   3   Upper triangular form.
   4   5   6  -4*R1          0  -3  -6   Determinant is (1)(-3)(1).
   7   8  10  -7*R1          0   0   1   No swapping; therefore
                                          original determinant is
   1   2   3                              -3
   0  -3  -6  =R2
   0  -6 -11  -2*R2

4x4 example:
  1  2 -1  0 =R1                1  2 -1  0
  0  0  1 -3                   0  1 -1  1 =R2
 -2 -3  1  1 +2*R1             0  0  1 -3
 -3 -7  6 -9 +3*R1             0 -1 -3 -9 +R2
```

```
1  2 -1  0
0  1 -1  1
0  0  1 -3 = R3
0  0  2 -8 -2*R3
```

```
1  2 -1  0
0  0  1 -3 swap
0  1 -1  1 swap
0 -1  3 -9
```

```
1  2 -1  0  UTF: det = -2
0  1 -1  1       1 swap, so
0  0  1 -3 determinant of
0  0  0 -2 original is 2.
```

For the 4×4 example, if we use cofactor expansion along the 1st row, we get

$$\begin{vmatrix} 0 & 1 & -3 \\ -3 & 1 & 1 \\ -7 & 6 & -9 \end{vmatrix} - 2 \begin{vmatrix} 0 & 1 & -3 \\ -2 & 1 & 1 \\ -3 & 6 & -9 \end{vmatrix} - \begin{vmatrix} 0 & 0 & -3 \\ -2 & -3 & 1 \\ -3 & -7 & -9 \end{vmatrix} = -1 - 12 + 15 = 2.$$

Of course, this involves computing the 3×3 determinants

$$\begin{vmatrix} 0 & 1 & -3 \\ -3 & 1 & 1 \\ -7 & 6 & -9 \end{vmatrix} = -1, \quad \begin{vmatrix} 0 & 1 & -3 \\ -2 & 1 & 1 \\ -3 & 6 & -9 \end{vmatrix} = 6, \quad \begin{vmatrix} 0 & 0 & -3 \\ -2 & -3 & 1 \\ -3 & -7 & -9 \end{vmatrix} = -15.$$

The examples have shown that the determinant can be calculated only on the basis that

- **Swapping** reverses sign,
- **subtracting** does not affect the determinant, and
- the determinant of an **upper triangular** matrix $[a_{ij}]_{n\times n}$ is the product

$$a_{11}a_{22}\cdots a_{nn}.$$

It can be shown that if $A = [a_{ij}]_{n\times n}$ and $1 \le i \le n$, the formula

$$\sum_j (-1)^{i+j} a_{ij} \det(A\backslash i\backslash j)$$

satisfies the above three properties, and therefore

$$\det(A) = \sum_j (-1)^{i+j} a_{ij} \det(A\backslash i\backslash j).$$

This is called **cofactor expansion** along the i-th row.

Because swapping rows changes sign, it follows that if two rows of A are equal then $\det(A) = 0$. If $i \neq k$ then

$$\sum_j a_{ij} (-1)^{i+j} \det(A\backslash k\backslash j)$$

can be regarded as the determinant of a matrix which is the same as A except that the i-th row of A has been duplicated in the k-th row, and hence is zero.

(9.5) Definition *If A is an $n \times n$ matrix, where $n \geq 2$, and $1 \leq i, j \leq n$, then the (i,j)-cofactor of A is*

$$(-1)^{i+j} \det(A \backslash i \backslash j),$$

and the **adjoint matrix** $\text{adj}(A)$ *is the* **transpose** *of the matrix of cofactors.*

By definition of $\text{adj}(A)$ and of matrix multiplication, the (i,k)-entry of the product $A\text{adj}(A)$ is

$$\sum (-1)^{k+j} a_{ij} \det(A \backslash k \backslash j) = \begin{cases} \det(A) & \text{if } i = k \\ 0 & \text{if } i \neq k. \end{cases}$$

In other words,

$$A\text{adj}(A) = \det(A) I,$$

and if $\det(A) \neq 0$ then

$$A \left(\frac{1}{\det(A)} \text{adj}(A) \right) = I.$$

Therefore if $\det(A) \neq 0$,

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A).$$

This is the **adjoint form of the inverse.**

In the 2×2 case it means the familiar formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

In the 3×3 case

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^{-1} = \begin{bmatrix} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \\ \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{bmatrix}.$$

The 3×3 formula is much easier to remember as follows:

- Call the three rows $\vec{OP}, \vec{OQ}, \vec{OR}$.
- The first **column** of $\text{adj}(A)$ is $\vec{OQ} \times \vec{OR}$.
- The second **column** of $\text{adj}(A)$ is $\vec{OR} \times \vec{OP}$.
- The third **column** of $\text{adj}(A)$ is $\vec{OP} \times \vec{OQ}$.
- Note that **columns** are formed, and note the **cyclic order** $\vec{OQ} \times \vec{OR}$, $\vec{OR} \times \vec{OP}$, and $\vec{OP} \times \vec{OQ}$.

For example, to calculate

$$\text{adj}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix},$$

the first **column** is

$$(4,5,6)\times(7,8,10) = (2,2,-3).$$

Using the $\vec{OP}\cdot(\vec{OP}\times\vec{OR})$ formula, $\det(A) = (1,2,3)\cdot(2,2,-3) = -3$. The second column is

$$(7,8,10)\times(1,2,3) = (4,-11,6),$$

and the third column is

$$(1,2,3)\times(4,5,6) = (-3,6,-3).$$

$$\text{adj}(A) = \begin{bmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{bmatrix}, \quad \text{so } A^{-1} = \begin{bmatrix} -\frac{2}{3} & -\frac{4}{3} & 1 \\ -\frac{2}{3} & \frac{11}{3} & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

10 Linear independence

Remember that $\mathbb{R}^m$ is the set of column vectors of height m.

Let $P_1, P_2, \ldots P_k$ be a list of points in 2- or 3-dimensional space, or a list of column vectors in $\mathbb{R}^m$. Another point (or column vector) P is said to **depend** on $P_1,\ldots,P_k$, if it can be expressed as a linear combination of $P_1,\ldots,P_k$.

- If $k=1$ then P depends on P_1 if and only if P is a scalar multiple of P_1. In 2- or 3-dimensional space this means that either $P_1 = P = O$ or $P_1 \neq O$ and P is on the line OP_1.
- In 2- or 3-dimensional space, if $k=2$ and O, P_1, P_2 are collinear, then P depends on P_1 and P_2 if and only if either $P_1 = P_2 = P = O$ or P is in the line containing O, P_1, and P_2.
- In 2- or 3-dimensional space, if $k=2$, and O, P_1, and P_2 are not collinear, then P depends on P_1 and P_2 if and only if P is in the plane OP_1P_2. This is always true in 2-dimensional space. (In this case P_1 and P_2 form a basis.)
- In 3-dimensional space, if $k=3$, and O, P_1, P_2, P_3 are not coplanar, then every point P depends on P_1, P_2, P_3 (which form a basis for 3-dimensional space).

(10.1) Definition *Given $k \geq 1$, a list $P_1,\ldots,P_k$ is* linearly independent *if either*
(i) $k=1$ *and $P_1 \neq O$, or*
(ii) $k>1$ *and for $1 \leq j \leq k$, P_j* **does not** *depend on the other 'points' $P_1,\ldots,P_{j-1},P_{j+1},\ldots,P_k$.*
Otherwise it is called linearly dependent.

It is easy to prove the following:

(10.2) Theorem *Given $k > 1$, a list $P_1, \ldots, P_k$ is linearly dependent if and only if there exist scalars $\alpha_1, \ldots, \alpha_k$,* **not all zero,** *such that*

$$\alpha_1 P_1 + \ldots + \alpha_k P_k = O.$$

Therefore the list is linearly independent if and only if the only solution $\alpha_1, \ldots, \alpha_k$ to the equation

$$\alpha_1 P_1 + \ldots + \alpha_k P_k = O$$

is $\alpha_1 = \alpha_2 = \ldots \alpha_k = 0$. (No proof.)

(10.3) Corollary *If the 'points' $P_1, \ldots, P_k$ are the columns of an $m \times k$ matrix A, then the columns are linearly independent if and only if the only solution to the equations*

$$AX = O$$

is $X = O$.

For example, test the points $(1, 2, 3), (4, 5, 6), (7, 8, 9)$ for linear independence (they aren't), by solving the equations whose augmented matrix is

```
1   4   7   0   =R1            1   4   7   0   -4*R2
2   5   8   0   -2*R1           0  -3  -6   0  *(-1/3)   =R2
3   6   9   0   -3*R1           0  -6 -12   0   +6*R2

1   0  -1   0
0   1   2   0
0   0   0   0  in rref
```

There is a nonzero solution because the third column is not a leading column: it depends on the leading columns. It is fairly clear that

$$C_3 = -C_1 + 2C_2,$$

where these are the columns of the RREF. The same relation holds among the columns of A, so

$$(7, 8, 9) = 2(4, 5, 6) - (1, 2, 3).$$

This shows that $(7, 8, 9)$ is in the plane containing O and the other two points. Equivalently,

$$-(1, 2, 3) + 2(4, 5, 6) - (7, 8, 9) = 0.$$

This shows that these three points are linearly dependent, and

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

On the other hand, the three points $(1, 2, 3), (4, 5, 6), (7, 8, 10)$ are linearly independent:

```
 1   4   7   0   =R1              1   0  -1   0   +1*R3
 2   5   8   0   -2*R1            0   1   2   0   -2*R3
 3   6  10   0   -3*R1            0   0   1   0   =R3

 1   4   7   0   -4*R2            1   0   0   0
 0  -3  -6   0  *(-1/3)=R2        0   1   0   0
 0  -6 -11   0   +6*R2            0   0   1   0  in rref
```

The calculation shows that the only solution to $AX = O$ is $X = 0$, so the columns of A, and the three points given, are linearly independent.

Actually, the fourth column is almost irrelevant to the calculation. In general,

(10.4) Theorem *Let A be an $m \times n$ matrix, A' its RREF. Then the leading columns of A' are linearly independent. Each non-leading column of A' depends on the leading columns to its left.*

If A' has no non-leading columns then the columns of A are linearly independent.

If A' has one or more leading columns, then there exists a nonzero solution to $A'X = O$. Then $AX = O$ also, so the columns of A are linearly dependent. (No proof.)

11 Theory

11.1 Matrix multiplication

(11.1) Lemma *Given $A_{\ell\times m}$ and $B_{m\times n}$, $1 \le i \le \ell$, and $1 \le j \le n$, let R_i be the i-th row of A and C_j be the j-th column of B. Then* (i) *the (i, j)-entry of AB equals the matrix product R_iC_j,* (ii) *the i-th row of AB equals R_iB, and* (iii) *the j-th column of AB equals AC_j.*

Proof omitted. Follows easily from the definition of matrix multiplication. ∎

(11.2) Lemma *Given three matrices A, B, C, the following are equivalent:*
(a) *$A(BC)$ is defined;* (b) *$(AB)C$ is defined;* (c) *AB and BC are both defined.*

Proof. Suppose $A_{k\times\ell}$, $B_{m\times n}$, and $C_{p\times q}$.
If (a) holds, then, because BC is defined, $n = p$, and because $A(BC)$ is defined, $\ell = m$. Thus, $A_{k\times\ell}$, and $B_{\ell\times n}$, so AB is defined, and (c) holds.
If (b) holds, then, because AB is defined, $\ell = m$, and because $(AB)C$ is defined, $n = p$. Thus $B_{\ell\times n}$ and $C_{n\times q}$: BC is also defined, and (c) holds.
If (c) holds then $\ell = m$ and $n = p$, so both $(AB)C$ and $A(BC)$ are defined, so (a) and (b) both hold. **Q.E.D.**

(11.3) Lemma (matrix multiplication is associative.) *If either $(AB)C$ or $A(BC)$ is defined, so both are defined, then $A(BC) = (AB)C$.*

Proof. Write

$$A = [a_{hi}]_{\ell\times m}, B = [b_{ij}]_{m\times n}, C = [c_{jk}]_{n\times p}, D = BC = [d_{ik}]_{m\times p},$$
$$E = AB = [e_{hj}]_{\ell\times n}, F = AD = [f_{hk}]_{\ell\times p}, \quad \text{and} \quad G = EC = [g_{hk}]_{\ell\times p}.$$

We need to show that $F = G$. For $1 \le h \le \ell$, $1 \le k \le p$,

$$f_{hk} = \sum_{i=1}^{m} a_{hi}d_{ik} = \sum_{i=1}^{m} a_{hi}\left(\sum_{j=1}^{n} b_{ij}c_{jk}\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{hi}b_{ij}c_{jk}\right),$$

from a generalised distributive law for real numbers: products distribute across sums. Again,

$$g_{hk} = \sum_{j=1}^{n} e_{hj}c_{jk} = \sum_{j=1}^{n}\left(\sum_{i=1}^{m} a_{hi}b_{ij}\right)c_{jk} = \sum_{j=1}^{n}\left(\sum_{i=1}^{m} a_{hi}b_{ij}c_{jk}\right).$$

The two sums at the end of these calculations represent the same list of numbers being added in different orders, so they have the same value, and $f_{hk} = g_{hk}$, as required. **Q.E.D.**

We have seen that matrix multiplication is not commutative: $AB \neq BA$ in general. But it is associative, that is, $A(BC) = (AB)C$. Also it is distributive:

(11.4) Lemma (distributive laws for matrices). *If B and C have the same dimensions and AB and AC are defined, then $A(B + C) = AB + AC$. Also if BD and CD are defined then $(B + C)D = BD + CD$ (Proof omitted).* ∎

11.2 Inverse matrices

(11.5) Definition *A square matrix A is* invertible *if there exists another matrix B such that $AB = I$ and $BA = I$.*

(11.6) Lemma (i) *Suppose A, B, and C are square matrices such that $AB = I$ and $CA = I$. Then $B = C$.*

(ii) *In particular, if A possesses an inverse, then that inverse is unique.*

Proof. (i) Since multiplication is associative,

$$(CA)B = C(AB),$$

that is,

$$IB = CI, \quad \text{i.e.,} \quad B = C,$$

as required.

(ii) If B and C are both inverses for A then $B = C$ by (i). **Q.E.D.**

Since A has a unique inverse if it exists, we use the notation A^{-1} in the usual way.

(11.7) Lemma *Let A and B be square matrices of the same size, and suppose that A is invertible.*
(i) *A^{-1} is invertible, and $(A^{-1})^{-1} = A$.*
(ii) *AB is invertible if and only if B is invertible. In this case, $(AB)^{-1} = B^{-1}A^{-1}$ and $(BA)^{-1} = A^{-1}B^{-1}$.*
(iii) *More generally, if $A_1 \cdots A_k$ is a product of invertible square matrices, then*

$$(A_1 \cdots A_k)^{-1} = A_k^{-1} \cdots A_1^{-1}.$$

Proof. (i): Trivial.
(iia) Suppose B is also invertible. Let $C = B^{-1}A^{-1}$. Then

$$ABC = ABB^{-1}A^{-1} = AIA^{-1} = AA^{-1} = I, \quad \text{and}$$
$$CAB = B^{-1}A^{-1}AB = B^{-1}IB = B^{-1}B = I,$$

so AB is invertible with inverse C, as required.
(iib) If AB is invertible, then so is A^{-1} by (i), and so is $A^{-1}AB = B$ by (iia), as required.
(iii) follows by induction on k. **Q.E.D.**

11.3 GJE and elementary matrices

(11.8) Lemma *The only (square) invertible matrices in RREF are the identity matrices.*

Proof. Let A' be an $m \times m$ matrix in RREF. If $A' = I$ then of course A' is invertible. If $A' \neq I$, then, because it is square and in RREF, not all columns are leading columns, and the bottom row of A', call it R_m, is entirely zero. For any other $m \times m$ matrix B, the bottom row of $A'B$ is R_mB (Lemma 11.1), hence entirely zero, so $A'B \neq I$. Therefore A' is not invertible. **Q.E.D.**

Suppose that $e()$ represents an elementary row operation on matrices of height m, i.e., scale, swap, or subtract: $e(A)$ is the matrix got by applying the operation to a matrix A (of height m).

(11.9) Lemma *Given $A_{m\times n}$ and $B_{n\times p}$, and an ERO $e()$ on matrices of height m,*

$$e(AB) = e(A)B.$$

Partial proof. Suppose $e(A)$ means: scale the k-th row by c.
For $1 \leq r \leq m$, Let R_i be the i-th row of A. Then R_iB is the i-th row of AB (Lemma 11.1).
The i-th row of $e(A)$ is

$$\begin{cases} R_i & \text{if } i \neq k \\ cR_i & \text{if } i = k, \end{cases}$$

the i-th row of $e(A)B$ is

$$\begin{cases} R_iB & \text{if } i \neq k \\ cR_iB & \text{if } i = k, \end{cases}$$

and the i-th row of $e(AB)$ is

$$\begin{cases} R_iB & \text{if } i \neq k \\ cR_iB & \text{if } i = k. \end{cases}$$

Thus $e(AB) = e(A)B$ in this case. The other cases are similar. ∎

In particular

$$e(IA) = e(I)A,$$

so the $m \times m$ matrix $e(I)$ is called the *elementary matrix* for the ERO $e()$.

(11.10) Lemma *Every elementary matrix, and every product of elementary matrices, is invertible.*

Proof. Let $e()$ be an ERO on matrices of height m. Every ERO is reversible, so there exists another ERO $f()$ such that $e(f(A)) = A$ and $f(e(A)) = A$ for all matrices A. Let $E = e(I)$ and $F = f(I)$.

$$EF = e(I)f(I) = e(If(I)) = e(f(I)) = I,$$

and similarly $FE = I$, so E is invertible and $F = E^{-1}$. Thus every elementary matrix is invertible. Given a product of elementary matrices, since each matrix is invertible, so is the product (Lemma 11.7). **Q.E.D.**

(11.11) Corollary *If A is an $m \times n$ matrix and A' the RREF (properly speaking, an RREF, but actually it is unique) of A, then there exists an invertible $m \times m$ matrix P such that*

$$A' = PA \quad \textit{and} \quad P^{-1}A' = A.$$

Proof. Suppose A' is produced by a sequence of k EROS $e_1, \ldots, e_k$, with elementary matrices $E_1, \ldots, E_k$. Let $P = E_k E_{k-1} \cdots E_1$. Then

$$A' = e_k(e_{k-1}(\ldots e_1(A)\ldots)) = e_k(e_{k-1}(\ldots (E_1A)\ldots)) = \ldots = E_k(E_{k-1}\ldots(E_1A)\ldots) = PA,$$

and $P^{-1}A' = P^{-1}PA = A$. **Q.E.D.**

(11.12) Theorem *Let A be a matrix and A' the[4] RREF of A. Then A is invertible if and only if A' is an identity matrix.*

Proof. If A is not a square matrix then A is not invertible and A' is not an identity matrix. So we assume A and A' are $m \times m$ matrices.

$A' = PA$ where P is invertible, so A' is invertible if and only if A is invertible (Lemma 11.7). But A' is invertible if and only if $A' = I$ (Lemma 11.8), so A is invertible if and only if $A' = I$. **Q.E.D.**

This gives a rationale for the GJE method of calculating inverses.

(11.13) Lemma *Let A be an $m \times n$ matrix, A' its RREF. For $1 \leq \ell \leq n$, the first ℓ columns of A' are the RREF for the first ℓ columns of A. (Proof omitted.)* ∎

[4] A' is unique, though that will not be proved.

Let A be an $m \times m$ matrix. Let us write

$$[A, I]$$

for the $m \times 2m$ matrix whose columns are those of A followed by those of the identity matrix I.

If we bring this matrix to RREF, there is an $m \times m$ matrix P such that the reduced matrix is

$$P[A, I] = [PA, PI] = [PA, P].$$

Since $[PA, PI]$ is in RREF, PA is the RREF of A (Lemma 11.13). If A is invertible, then $PA = I$, and, by Lemma 11.6, $P = A^{-1}$. If A is not invertible then PA is not the identity matrix. Summarising,

(11.14) Theorem *The GJE calculation of inverse matrices always works.* ∎

(11.15) Lemma (i) *If A and B are $m \times m$ matrices, and AB is invertible, then A and B are both invertible.*

In particular, if $AB = I$, then $A = B^{-1}$, $B = A^{-1}$, and $BA = I$.

Proof. (i) Let P be a matrix such that PA is the (an) RREF of A. Since P and AB are invertible, their product PAB has an inverse, call it C:

$$PABC = I,$$

so

$$(PA)(BC) = I.$$

Therefore the bottom row of PA is nonzero. Since PA is square and in RREF, this means that $PA = I$ and A is invertible.

Since A and AB are invertible, B is invertible by Lemma 11.7 (iii).

(ii): since I is invertible, A and B are invertible, and $AB = I$, which implies that $B = A^{-1}$ (Lemma 11.6). **Q.E.D.**

(11.16) Lemma *Let A be an $m \times n$ matrix which is not invertible, so it has an RREF A' which is not an identity matrix.*

We consider equations of the form $AX = Y$ where $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$.

Either (i) *not all columns of A' are leading columns, in which case the equation $AX = O$ has infinitely many solutions, or* (ii) *the bottom row of A' is entirely zero, in which case the equation $AX = Y$ sometimes has no solution.*[5]

Proof. Let A_j be the j-th column of A and A'_j the j-th column of A'. Let P be an invertible square matrix such that $PA = A'$. Therefore

$$PA_j = A'_j \quad \text{for } 1 \leq j \leq n.$$

[5] These cases are not mutually exclusive.

In case (i), we first show that the equation $A'X = O$ has infinitely many solutions. Let A'_k be the first non-leading column of A'. Possibly $k = 1$ and the first column of A' is entirely zero. In this case the first column of A is entirely zero, and

$$A \begin{bmatrix} t \\ 0 \\ \bullet \\ \bullet \\ \bullet \\ 0 \end{bmatrix} = \begin{bmatrix} t \\ 0 \\ \bullet \\ \bullet \\ \bullet \\ 0 \end{bmatrix},$$

for any $t \in \mathbb{R}$. Otherwise $k > 1$ and $A'_1, \ldots, A'_{k-1}$ are the first $k-1$ columns of the $m \times m$ identity matrix. Since A'_k is not a leading column, only its top $k-1$ entries can be nonzero, so

$$A'_k = \begin{bmatrix} \alpha_1 \\ \bullet \\ \bullet \\ \alpha_{k-1} \\ 0 \\ \bullet \\ \bullet \\ 0 \end{bmatrix},$$

so for any $t \in \mathbb{R}$

$$t(-\alpha_1 A'_1 + \ldots + -\alpha_{k-1} A'_{k-1} + A'_k) = O,$$

i.e., $A'X = O$ where

$$X = \begin{bmatrix} -\alpha_1 t \\ \bullet \\ \bullet \\ -\alpha_{k-1} t \\ t \\ 0 \\ \bullet \\ \bullet \\ 0 \end{bmatrix}.$$

Thus in case (i), the equation $A'X = O$ has infinitely many solutions. But whenever $A'X = O$, $P^{-1}A'X = O$, so $AX = O$. Therefore the equation $AX = O$ has infinitely many solutions.

In case (ii) the bottom row of A' is entirely zero. For any $X \in \mathbb{R}^n$, the bottom element of

$$A'X$$

is zero. Let Z be any column vector whose bottom element is nonzero. Therefore the equation

$$A'X = Z$$

has no solution. Let $Y = P^{-1}Z$. If $AX = Y$, then $PAX = Z$, i.e., $A'X = Z$, which is impossible. Therefore the equation $A'X = Y$ has no solution. **Q.E.D.**

(11.17) Corollary *Let A be an $m \times n$ matrix. Then A is invertible if and only if any (and all) of the following conditions hold: for all column vectors Y of height m, the equation $AX = Y$*
(i) *has a unique solution,*
(ii) *$m = n$ and the equation $AX = O$ has at most one solution, and*
(iii) *$m = n$ and the equation $AX = Y$ has at least one solution,*

Proof. Suppose A is invertible. Then $m = n$, and for any column vectors X and Y of height m,

$$AX = Y \iff A^{-1}AX = A^{-1}Y \iff X = A^{-1}Y,$$

so the equation $AX = Y$ has the unique solution $X = A^{-1}Y$. Thus (i), (ii), and (iii) hold.

Suppose A is not invertible. Then by Lemma 11.16, either the equation $AX = Y$ sometimes has no solutions or sometimes has many solutions. In either case, (i) is false.

Suppose A is not invertible, but $m = n$. The RREF A' of A is not an identity matrix, and is square. This implies two things: not all columns are leading columns, *and* the bottom row is entirely zero. Hence both (ii) and (iii) are false. **Q.E.D.**

11.4 Linear independence and bases

Remember $\mathbb{R}^n$ is the set of column vectors of height n.

The notion of a point *depending* on a list of points, or a list being *independent,* or a list forming an *ordered basis,* or the *coordinates* of a point with respect to an ordered basis, are important ideas in Linear Algebra. When the 'set of points' is one of these sets $\mathbb{R}^n$, these notions can be verified with simple calculations based on GJE.

Suppose $S_1, \ldots, S_k$ is a list drawn from $\mathbb{R}^n$. Let S be the matrix whose columns are $S_1, \ldots, S_k$ (in that order).

(11.18) Theorem *The columns $S_1, \ldots, S_k$ form an (ordered) basis for $\mathbb{R}^n$ if and only if S is an invertible matrix.*

Proof. These form an ordered basis if and only if for any $X \in \mathbb{R}^n$, there exist unique real numbers $\alpha_1, \ldots, \alpha_k$ such that

$$\alpha_1 S_1 + \ldots + \alpha_k S_k = X.$$

If $Y \in \mathbb{R}^k$ is the column-vector $[\alpha_1, \ldots, \alpha_k]^T$, then the equation

$$SY = X$$

has a unique solution, for all $X \in \mathbb{R}^n$. This is true if and only if S is invertible (Corollary 11.17 (i)). **Q.E.D.**

(11.19) Corollary *Every basis for $\mathbb{R}^n$ contains exactly n column vectors.* ∎

Remark. If you study the definitions, you will see that $S_1, \ldots, S_k$ is linearly independent if and only if for all X the equation $SY = X$ has *at most one* solution.

There is a notion of $S_1, \ldots, S_k$ *spanning* $\mathbb{R}^n$. This means that *every* $X \in \mathbb{R}^n$ depends on $S_1, \ldots, S_k$, or that for any $X \in \mathbb{R}^n$, there exist $\alpha_1, \ldots, \alpha_k$ such that

$$\alpha_1 S_1 + \ldots + \alpha_k S_k = X.$$

Equivalently, for any $X \in \mathbb{R}^n$, the equation

$$SY = X$$

has *at least* one solution, perhaps many. It is easy to deduce the following:

(11.20) Theorem *Let $S_1, \ldots, S_n$ be a list of n vectors in $\mathbb{R}^n$. Then the following are equivalent:* (i) *the vectors form a basis for $\mathbb{R}^n$,* (ii) *they are linearly independent, and* (iii) *they span $\mathbb{R}^n$.* ∎

11.5 Positive normal and cross products

Given $P = (a, b)$, it can be written as $(r\cos\theta, r\sin\theta)$ with $r = \sqrt{a^2 + b^2}$ and $0 \leq \theta < 2\pi$. Another point $Q = (c, d)$ can be written as $(s\cos\varphi, s\sin\varphi)$.

Let $N = (-b, a)$ be the positive normal to P. Then

$$\vec{ON} \cdot \vec{OQ} = ad - bc = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Using trigonometry,

$$\vec{ON} \cdot \vec{OQ} = rs\cos\widehat{NOQ} = rs\sin(\varphi - \theta).$$

This is the area of the parallelogram O, P, Q, with sign reversed if Q is to the right of OP. Thus,

(11.21) Lemma *Given $P = (a, b)$ and $Q = (c, d)$,*

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

equals ±(area of paralleogram $O, P, P + Q, Q$). ∎

Given $P = (a, b, c)$ and $Q = (d, e, f)$, the cross-product

$$\vec{OP} \times \vec{OQ} = \left(\begin{vmatrix} b & c \\ e & f \end{vmatrix}, -\begin{vmatrix} a & c \\ d & f \end{vmatrix}, \begin{vmatrix} a & b \\ d & e \end{vmatrix} \right).$$

When this was introduced, the following property was partially proved: let $\vec{ON} = \vec{OP} \times \vec{OQ}$. Then $\vec{ON}$ is perpendicular both to $\vec{OP}$ and $\vec{OQ}$.

It is fairly easy to show that $N \neq O \iff O, P, Q$ are not collinear (i.e., P and Q are linearly independent).

If O, P, Q are not collinear, then the cyclic order of OPQ is anticlockwise when viewed from one side of the plane and clockwise when viewed from the other. It is anticlockwise when viewed from N, though we shall not try to prove that.

Suppose $N \neq O$. It can be written as $(r\cos\alpha, r\cos\beta, r\cos\gamma)$ where α, β, γ are the angles ON makes with the x-, y-, and z-axes, respectively. Let A be the area of the parallelogram $O, P, P+Q, Q$. Again, $P = (a, b, c)$ and $Q = (d, e, f)$. The vertical projection of this parallelogram onto the xy-plane is a parallelogram with corners $O, (a, b, 0), (a, b, 0) + (d, e, 0), (d, e, 0)$, and the area of this projected parallelogram is

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

But this equals $A\cos\gamma$. Similarly for the xz- and yz- planes. Thus

$$N = A(\cos\alpha, \cos\beta, \cos\gamma).$$

That is, $|\vec{ON}|$ is the area of the parallelogram $O, P, P+Q, Q$. Summarising:

(11.22) Lemma *Given points P and Q in three dimensions, let N be the point such that $\vec{ON} = \vec{OP} \times \vec{OQ}$. Then $\vec{ON}$ is perpendicular both to $\vec{OP}$ and $\vec{OQ}$, is zero if and only if they are linearly dependent; if nonzero then N is on the 'positive' side of the plane OPQ. Also, $|\vec{ON}|$ is the area of the parallelogram $O, P, P+Q, Q$.* ∎

11.6 Determinants

The notation is changed from earlier in the notes.

(11.23) Definition *Let A be an $n \times n$ matrix and $1 \leq i, j \leq n$. Then*

$$A \setminus (\{i\} \times \{j\})$$

is the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and the j-th column. If $1 \leq i_1, i_2, j_1, j_2 \leq n$ and $i_1 \neq i_2$ and $j_1 \neq j_2$, then

$$A \setminus (\{i_1, i_2\} \times \{j_1, j_2\})$$

is the $(n-2) \times (n-2)$ matrix obtained by deleting rows i_1 and i_2 and columns j_1 and j_2.

(11.24) Definition *Given $A = [a_{ij}]_{n\times n}$*

$$\det(A) = \begin{cases} a_{11} & \text{if } n = 1, \text{ and (recursively)} \\ \sum_{j=1}^{n} a_{1j}(-1)^{1+j} \det(A \setminus (\{1\} \times \{j\})), & \text{if } n > 1. \end{cases}$$

(11.25) Lemma *Given $A_{n\times n}$ and $1 \leq j_1 \leq n$ and $1 \leq j_2 \leq n-1$,*

$$(A \setminus (\{1\} \times \{j_1\})) \setminus (\{1\} \times \{j_2\}) = \begin{cases} A \setminus (\{1,2\} \times \{j_1, j_2\}) & \text{if } j_1 > j_2 \\ A \setminus (\{1,2\} \times \{j_1, j_2+1\}) & \text{if } j_1 \leq j_2. \end{cases}$$

(Proof omitted).

(11.26) Lemma *Given $A = [a_{ij}]_{n\times n}$, where $n > 1$, let A' be obtained by swapping rows 1 and 2 in A. Then $\det(A') = -\det(A)$.*

Proof.

$$\det(A) = \sum_{j_1=1}^{n} a_{1j_1}(-1)^{1+j_1} \det(A \setminus (\{1\} \times \{j_1\})),$$

and

$$\det(A \setminus (\{1\} \times \{j_1\})) = \sum_{j_2=1}^{n-1} a_{2k_2}(-1)^{1+j_2} \det((A \setminus (\{1\} \times \{j_1\})) \setminus (\{2\} \times \{j_2\})),$$

where $k_2 = j_2$ if $j_2 < j_1$ and $k_2 = j_2 + 1$ if $j_2 \geq j_1$.

By Lemma 11.25,

$$(A \setminus (\{1\} \times \{j_1\})) \setminus (\{2\} \times \{j_2\}) = A \setminus (\{1,2\} \times \{j_1, k_2\}).$$

Separating the cases $j_2 < j_1$ and $j_2 \geq j_1$, and applying Lemma 11.25, $\det(A)$ is the sum of the following two quantities:

$$\sum_{j_1=1}^{n} a_{1j_1}(-1)^{1+j_1} \sum_{j2<j_1} a_{2j_2}(-1)^{1+j_2} \det(A \setminus (\{1,2\} \times \{j_1, j_2\})),$$

and

$$\sum_{j_1=1}^{n} a_{1j_1}(-1)^{1+j_1} \sum_{j2\geq j_1} a_{2k_2}(-1)^{1+j_2} \det(A \setminus (\{1,2\} \times \{j_1, k_2\})).$$

In the latter sum, $k_2 = j_2 + 1$. We actually change the labelling so k_2 is replaced by j_2, so j_2 must be replaced by $j_2 - 1$, and $j_2 > j_1$. The latter sum becomes

$$\sum_{j_1=1}^{n} a_{1j_1}(-1)^{1+j_1} \sum_{j2>j_1} a_{2j_2}(-1)^{j_2} \det(A \setminus (\{1,2\} \times \{j_1, j_2\})).$$

Note the difference in sign. The sum can now be written

$$\sum_{j_1>j_2} (-1)^{1+j_1+j_2} a_{1j_1} a_{2j_2} \det(A \setminus (\{1,2\} \times \{j_1, j_2\})) - \sum_{j_1>j_2} (-1)^{1+j_1+j_2} a_{1j_1} a_{2j_2} \det(A \setminus (\{1,2\} \times \{j_1, j_2\}))$$

Change the first expression by interchanging j_1 and j_2:

$$\sum_{j_1<j_2} (-1)^{1+j_1+j_2} a_{1j_2} a_{2j_1} \det(A \setminus (\{1,2\} \times \{j_1, j_2\})).$$

and combine the two expressions, noting the difference in sign:

$$\sum_{j_1<j_2} (-1)^{1+j_1+j_2} (a_{1j_2} a_{2j_1} - a_{1j_1} a_{2j_2}) \det(A \setminus (\{1,2\} \times \{j_1, j_2\})).$$

Now it is obvious that if the first and second rows are changed, this expression is reversed in sign.
Q.E.D.

(11.27) Lemma *If $A'_{m\times m}$ is obtained from A by swapping* any *two rows in A, then* $\det(A') = -\det(A)$.

Proof. By induction (base case is when $m = 2$). If the rows being swapped do not include the first, then induction applies directly. If the first and second, that has been covered. Otherwise, say the first row is swapped with the k-th where $k > 2$. Consider the following steps. Swap first with second: sign change. Swap second with k-th: sign change, by induction. Swap second with first: sign change. The end result is first and k-th swapped. **Q.E.D.**

(11.28) Corollary *If A contains two equal rows, then* $\det(A) = 0$.

Proof. Let A' be obtained from A by swapping two equal rows. Then $A' = A$, but $\det(A') = -\det(A)$, so $\det(A) = -\det(A)$ and $\det(A) = 0$. **Q.E.D.**

(11.29) Corollary *Given $A_{m\times m}$ and $1 \le i \le m$,*

$$\det(A) = \sum_j a_{ij}(-1)^{i+j}\det(A \setminus (\{i\} \times \{j\})).$$

Proof. Using the notation $R_s(B)$ for the s-th row of a matrix B, let B be the matrix obtained from A as follows:

$$\begin{cases} R_1(B) = R_i(A), \\ R_s(B) = R_{s-1}(A), \quad 2 \le s \le i, \\ R_s(B) = R_s(A) \quad s > i. \end{cases}$$

B is obtained from A by beginning at the i-th row and moving it up to the top row by $i-1$ swaps, so

$$\det(A) = (-1)^{i-1}\det(B).$$

Applying the definition, and writing $B = [b_{rs}]$,

$$\det(A) = (-1)^{i-1}\sum_j (-1)^{1+j} b_{1j}\det(B \setminus (\{1\} \times \{j\})).$$

But $b_{1j} = a_{ij}$ and $B \setminus (\{1\} \times \{j\}) = A \setminus (\{i\} \times \{j\})$, and we get

$$\sum_j a_{ij}(-1)^{i+j}\det(A \setminus (\{i\} \times \{j\}))$$

as required. **Q.E.D.**

(11.30) Corollary

$$A\mathrm{adj}(A) = \det(A)I.$$

(Proof immediate from cofactor expansion, Corollary 11.29.) ∎

(11.31) Lemma *If A' is obtained from A by scaling a row by a constant c, then* $\det(A') = c\det(A)$.

Proof. Suppose the i-th row is scaled by c. Use cofactor expansion along this row.

$$\det(A') = \sum_{ij} (-1)^{i+j} (ca_{ij}) \det(A \setminus (\{i\} \times \{j\})) = c \det(A).$$

Q.E.D.

(11.32) Corollary *If A contains a row of zeroes, then* $\det(A) = 0$.

Proof. Scale the zero row by zero; $A = A'$ in the above lemma, and $\det(A) = 0 \det(A)$. **Q.E.D.**

(11.33) Lemma *If $A'_{m\times m}$ is obtained from A by a SUBTRACT ero, then* $\det(A') = \det(A)$.

Proof. Suppose $c \times (k)$-th row is subtracted from the i-th row.

$$\det(A') = \sum_j (a_{ij} - ca_{kj})(-1)^{i+j} \det(A' \setminus (\{i\} \times \{j\}))$$

$$= \sum_j a_{ij}(-1)^{i+j} \det(A \setminus (\{i\} \times \{j\})) - c \sum_j a_{kj}(-1)^{i+j} \det(A \setminus (\{i\} \times \{j\})).$$

The latter sum is the determinant of a matrix in which rows i and k are equal, so it is zero. Therefore $\det(A') = \det(A)$.

(11.34) Lemma *If $A_{m\times m}$ is a matrix containing a column of zeros then* $\det(A) = 0$.

Proof. By induction on m. Trivial in the case $m = 1$. As usual, write $A = [a_{ij}]_{m\times m}$.
Induction: suppose the k-th column of A is entirely zero.

$$\det(A) = \sum_j a_{ij}(-1)^{1+j} \det(A \setminus (\{1\} \times \{j\})).$$

In this sum, if $j = k$ then $a_{1j} = 0$, and if $j \neq k$ then $A \setminus (\{1\} \times \{j\})$ contains a column of zeroes, and its determinant is zero by induction. Hence $\det(A) = 0$. **Q.E.D.**

(11.35) Corollary *If $A_{m\times m}$ is an upper triangular matrix then its determinant is the product of its main diagonal entries.*

Proof by induction. If $m = 1$ then the determinant is a_{11} which is also the main diagonal entry.
Induction:

$$\det(A) = \sum_j (-1)^{1+j} a_{1j} \det(A \setminus (\{1\} \times \{j\})).$$

In this sum, if $j \neq 1$, then the leftmost column of $A \setminus (\{1\} \times \{j\})$ is zero, hence its determinant is zero, and

$$\det(A) = a_{11} \det(A \setminus (\{1\} \times \{1\})),$$

the product of a_{11} by the determinant of a smaller upper-triangular matrix, and the inductive step follows easily. **Q.E.D.**

11.7 Determinants and invertibility

Let $e()$ be an ERO on matrices of height m, and $E = e(I)$ the corresponding elementary matrix. Since $\det(I) = 1$, Lemmas 11.27, 11.31, and 11.33 imply that

$$\det(E) = \begin{cases} c & \text{if } e() \text{ SCALEs a row by } c \\ -1 & \text{if } e() \text{ SWAPs two rows} \\ 1 & \text{if } e() \text{ is a SUBTRACT operation.} \end{cases}$$

Therefore the determinant of an elementary matrix is always nonzero, and by the same lemmas,

$$\det(EB) = \det(E)\det(B)$$

for every $m \times m$ matrix B. If $P = E_k \cdots E_1$ is a product of elementary matrices, then it follows by induction that

$$\det(PB) = \det(E_k) \cdots \det(E_1)\det(B).$$

Therefore

$$\det(P) = \det(PI) = \det(E_1) \cdots \det(E_1)(1)$$

and

$$\det(PB) = \det(P)\det(B).$$

So if $A' = PA$ is the RREF of A then

$$\det(A) = 0 \iff \det(A') = 0.$$

(11.36) Corollary *A is invertible iff* $\det(A) \neq 0$.

Proof. If A is invertible, then $A' = I$ with determinant 1. If A' is not invertible, then the bottom row of A' is zero, and $\det(A) = 0$. **Q.E.D.**

(11.37) Corollary *If A is invertible then*

$$A^{-1} = \frac{1}{\det(A)}\text{adj}(A).$$

Proof. If A is invertible, then $\det(A) \neq 0$. By Corollary 11.30,

$$A\frac{1}{\det(A)}\text{adj}(A) = I,$$

so by Lemma 11.15

$$A^{-1} = \frac{1}{\det(A)}\text{adj}(A)$$

(11.38) Theorem *If A and B are $m \times m$ matrices, then*

$$\det(AB) = \det(A)\det(B).$$

Proof. This time let $A = QA'$ where A' is the RREF of A and $Q = P^{-1}$. Q is also a product of elementary matrices so

$$\det(QA') = \det(Q)\det(A').$$

For any $m \times m$ matrix B,

$$\det(AB) = \det(QA'B) = \det(Q)\det(A'B).$$

Either A is invertible or it isn't. If A is invertible, then $A' = I$ and $A = Q$. In this case

$$\det(AB) = \det(A)\det(B).$$

Otherwise, $\det(A) = \det(A') = 0$. Also, the bottom row of $A'B$ is zero, so $\det(A'B) = 0$, so $\det(AB) = \det(Q)\det(A'B) = 0$, and

$$\det(AB) = 0 = \det(A)\det(B)$$

in this case also. **Q.E.D.**

(11.39) Lemma *A square matrix A is invertible if and only if so is A^T, and $(A^T)^{-1} = (A^{-1})^T$ (exercise).* ∎

(11.40) Corollary *Given $A_{m\times m}$,* $\det(A^T) = \det(A)$.

Proof. If $\det(A) = 0$, then A is not invertible, so neither is A^T and $\det(A^T) = 0$ (Corollary 11.36 and Lemma 11.39).

For the general case, we use induction on m. The base case is trivial. For the induction, since $\det(A) \neq 0$, the adjoint form of the inverse is correct, so

$$\text{cof}(A)A = \det(A)I.$$

Taking the $(1,1)$-entry of these matrices,

$$\det(A) = \sum_j (-1)^{1+j} \det(A \setminus (\{j\} \times \{1\}))a_{j1}.$$

If $B = A^T = [b_{ij}]$, then by induction, $\det(A \setminus (\{j\} \times \{1\} = \det(B \setminus (\{1\} \times \{j\}))$, and $b_{1j} = a_{j1}$. Therefore

$$\det(A) = \sum_j b_{1j} \det(B \setminus (\{1\} \times \{j\})) = \det(B).$$

Q.E.D.

11.8 Eigenvalues and eigenvectors

(11.41) Definition *Let A be a square matrix and λ a variable. The* characteristic polynomial *of A is*

$$\det(\lambda I - A)$$

An eigenvalue *of A is a particular value λ such that $\det(\lambda I - A) = 0$, and an* eigenvector *of A is a column vector X such that*

$$X \neq O \quad \textit{and} \quad AX = \lambda X$$

for some scale-factor λ.

(11.42) Theorem (i) *If X is an eigenvector then the scale-factor λ is an eigenvalue.* (ii) *If λ is an eigenvalue, then there exists a corresponding eigenvector.*

Proof. (i) $\lambda X = AX$ so $\lambda IX = AX$ and $(\lambda I - A)X = O$. The equation $(\lambda I - A)X = O$ has a nonzero solution, so $\lambda I - A$ is not invertible (Corollary 11.17), and $\det(\lambda I - A)$ is zero.

(ii) If $\det(\lambda I - A) = 0$, then $\lambda I - A$ is not invertible, so there exists an nonzero vector X such that $(\lambda I - A)X = O$ (Corollary 11.17). Then $\lambda IX = AX$, so $\lambda X = AX$, and X is an eigenvector with corresponding scale-factor λ. **Q.E.D.**

(11.43) Lemma *If S and A are $m \times m$ matrices and S is invertible, then*

$$\det(\lambda I - S^{-1}AS) = \det(\lambda I - A).$$

Hence A and $S^{-1}AS$ have the same eigenvalues. (Proof omitted).

12 Eigenvalues and eigenvectors

The idea is to choose a basis in which a matrix becomes simpler. Eigenvectors are on axes which are mapped to themselves when multiplied by A. Eigenvectors must be nonzero. This means that for some scalar λ and nonzero vector X,

$$\lambda X = AX.$$

In the 2×2 case this would mean

$$\lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

or

$$\begin{bmatrix} \lambda - a & -b \\ -c & \lambda - d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Geometrically, $[x_1\ x_2]^T$ is nonzero and perpendicular to both rows, which is only possible if the rows are proportional, or equivalently

$$(\lambda - a)(\lambda - b) - cd = 0.$$

Of course, the general formula is that

$$\det(\lambda I - A) = 0.$$

The number λ is called an *eigenvalue.*

For example, with

$$A = \begin{bmatrix} 1 & 2 \\ 18 & 1 \end{bmatrix}$$

the equation becomes

$$(\lambda - 1)^2 - 36 = 0,$$

so $\lambda = -5, 7$.

When $\lambda = 7$,

$$7I - A = \begin{bmatrix} 6 & -2 \\ -18 & 6 \end{bmatrix}$$

with corresponding eigenvector $[2 \;\; 6]^T$, and when $\lambda = -5$,

$$-5I - A = \begin{bmatrix} -6 & -2 \\ -18 & -6 \end{bmatrix}$$

with corresponding eigenvector $[2 \;\; -6]^T$.

The change-of-basis formula says that the the transformed matrix, with respect to the ordered basis of eigenvectors, is

$$S^{-1}AS$$

where S is the matrix whose columns are the eigenvectors

$$S = \begin{bmatrix} 2 & 2 \\ 6 & -6 \end{bmatrix}.$$

The transformed matrix is diagonal, very simple:

$$S^{-1}AS = \begin{bmatrix} \frac{1}{4} & \frac{1}{12} \\ \frac{1}{4} & -\frac{1}{12} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 18 & 1 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 2 & -6 \end{bmatrix} =$$
$$\begin{bmatrix} \frac{1}{4} & \frac{1}{12} \\ \frac{1}{4} & -\frac{1}{12} \end{bmatrix} \begin{bmatrix} 14 & -10 \\ 42 & 30 \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix}.$$

In general, we cannot expect a diagonal matrix. The following matrix cannot be diagonalised

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

That is, no choice of matrix S makes $S^{-1}AS$ into a diagonal matrix.

Similarity transformations are transformations of the form

$$A \mapsto S^{-1}AS.$$

(12.1) Proposition *If an $n \times n$ matrix A has n distinct eigenvalues then it can be diagonalised. That is, it can be transformed to a diagonal matrix by a similarity transformation.*

In this case, if λ is an eigenvalue, then the cofactors of any row of $\lambda I - A$ give an eigenvector.

Also, if A is symmetric, then it can be diagonalised, by an orthogonal similarity transformation. (No proof). ∎

One application is to evaluate high powers of A. This is because $(S^{-1}AS)^n = S^{-1}A^nS$. Or, if $S^{-1}AS = D$ is diagonal, then it is easy to evaluate D^n, and $A^n = SD^nS^{-1}$.

For example,

$$\begin{bmatrix} 1 & 2 \\ 18 & 1 \end{bmatrix}^6 = \begin{bmatrix} 2 & 6 \\ 2 & -6 \end{bmatrix} \begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix}^6 \begin{bmatrix} \frac{1}{4} & \frac{1}{12} \\ \frac{1}{4} & -\frac{1}{12} \end{bmatrix} =$$
$$\begin{bmatrix} 2 & 6 \\ 2 & -6 \end{bmatrix} \begin{bmatrix} 117649 & 0 \\ 0 & 15625 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{12} \\ \frac{1}{4} & -\frac{1}{12} \end{bmatrix} =$$
$$\begin{bmatrix} 66637 & 17004 \\ 153036 & 66637 \end{bmatrix}$$

Here is a 3×3 **example.** Calculate the eigenvalues of the matrix

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

$(2-\lambda)(\lambda^2 - 2) = 0$, so $\lambda = 2, \pm\sqrt{2}$.

Calculate corresponding eigenvectors.

Eigenvalue 2: array is

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

Cofactors of first row: $[2\ 0\ 0]^T$.

Eigenvalue $\sqrt{2}$:

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

Take cofactors of third row: $[0 \quad 2-2\sqrt{2} \quad \sqrt{2}-2]^T$.

Eigenvalue $-\sqrt{2}$:

$$\begin{bmatrix} 2+\sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 2 \\ 0 & 1 & \sqrt{2} \end{bmatrix}.$$

Take cofactors of second row: $[0 \quad -2-2\sqrt{2} \quad 2+2\sqrt{2}]^T$.

Example of a symmetric matrix.

Calculate the eigenvalues of the symmetric matrix

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{22}{9} & \frac{2\sqrt{5}}{9} \\ 0 & \frac{2\sqrt{5}}{9} & \frac{23}{9} \end{bmatrix}.$$

$$(2-\lambda))(\frac{22}{9}-\lambda)(\frac{23}{9}-\lambda) - \frac{20}{81}) = (2-\lambda)(\frac{506}{81} - 5\lambda + \lambda^2 - \frac{20}{81})$$
$$= (2-\lambda)(6 - 5\lambda + \lambda^2) = (2-\lambda)(3-\lambda)(2-\lambda).$$

Calculate eigenvectors for two different eigenvalues of A.

$$A - 2I = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{4}{9} & \frac{2\sqrt{5}}{9} \\ 0 & \frac{2\sqrt{5}}{9} & \frac{5}{9} \end{bmatrix}.$$

Clearly $[1 \ 0 \ 0]^T$ is an eigenvector with eigenvalue 2.

$$A - 3I = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{-5}{9} & \frac{2\sqrt{5}}{9} \\ 0 & \frac{2\sqrt{5}}{9} & \frac{-4}{9} \end{bmatrix}.$$

Take the cofactors of the second row: $[0 \ \ -4/9 \ \ -2\sqrt{5}/9]^T$, an eigenvector with eigenvalue 3.

Calculate an orthonormal basis of eigenvectors, relative to which the map $X \mapsto AX$ has a diagonal matrix.

Normalise the second vector (also, it is tidier with sign reversed): $[0 \ \ 2/3 \ \ \sqrt{5}/3]^T$. For a third vector, take the cross product of the first two:

$[0 \ \ -\sqrt{5}/3 \ \ 2/3]^T$. This is already normalised.

Summarising:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{\sqrt{5}}{3} \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -\frac{\sqrt{5}}{3} \\ \frac{2}{3} \end{bmatrix}.$$

(12.2) Proposition *In general, the eigenvalues of A are complex, and anyway one cannot always diagonalise, but one can bring A to upper triangular form by a similarity transformation.*

Example with repeated eigenvalues. Calculate the eigenvalues of the matrix

$$A = \begin{bmatrix} 10 & -5 & 2 \\ 13 & -6 & 3 \\ 6 & -4 & 4 \end{bmatrix}$$

(At least one of them is a small integer.)

Using cofactor expansion along the first row to evaluate $\det(A - \lambda I)$, we get

$$\begin{aligned} &(10-\lambda)((-6-\lambda)(4-\lambda)+12) + 5(13(4-\lambda)-8) + 2(-52-6(-6-\lambda)) \\ &= (10-\lambda)(\lambda+6)(\lambda-4)+12) + 5(52-13\lambda-18) + 2(-52+36+6\lambda) \\ &= (10-\lambda)(\lambda^2+2\lambda-12) + 5(34-13\lambda) + 2(6\lambda-16) \\ &= (\lambda^3+8\lambda^2+32\lambda-120) + 170 - 65\lambda + 12\lambda - 32 \\ &= -\lambda^3+8\lambda^2+32\lambda-120+170-65\lambda+12\lambda-32 \\ &= -\lambda^3+8\lambda^2-21\lambda+18. \end{aligned}$$

The remainder test shows that $2-\lambda$ is a factor. Divide:

$$(2-\lambda)(9-6\lambda+\lambda^2) = (2-\lambda)(3-\lambda)^2.$$

Thus the eigenvalues (with multiplicities) are $2, 3, 3$.

Calculate eigenvectors for two different eigenvalues from Question 1.

First, if $\lambda = 2$, $A - \lambda I$ is

$$\begin{bmatrix} 8 & -5 & 2 \\ 13 & -8 & 3 \\ 6 & -4 & 2 \end{bmatrix}.$$

Take cofactors of the first row: $[-4 \ \ -8 \ \ -4]^T$, which can be rescaled to $[1 \ \ 2 \ \ 1]^T$. Next, with $\lambda = 3$, $A = 3I$ is

$$\begin{bmatrix} 7 & -5 & 2 \\ 13 & -9 & 3 \\ 6 & -4 & 1 \end{bmatrix}.$$

Take cofactors of the first row: $[3 \ \ 5 \ \ 2]^T$.

Find a basis which brings the matrix in Question 1 into Upper Triangular Form.

We can take $v_1 = [1\ 2\ 1]^T$ and $v_2 = [3\ 5\ 2]^T$. For v_3 it is probably enough to choose any vector making v_1, v_2, v_3 a basis. Let us choose $v_3 = [0\ 0\ 1]^T$. Let us calculate Av_3 in terms of the *new* basis. Of course, $Av_3 = [2\ 3\ 4]^T$. What is $(A - 3I)v_3$? $[2\ 3\ 1]^T$. That is, $(A - 3I)v_3 = v_2 - v_1$ and $Av_3 = -v_1 + v_2 + 3v_3$. This gives the new matrix

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

13 Probability and statistics

Probability deals with uncertainty and repetition. If a coin is tossed many times, the number of heads should roughly equal the number of tails, though one cannot predict the outcome of a single toss.

An *event* is the result of an observation or experiment, or the results of several observations or experiments. For example, if a coin is tossed four times, there are sixteen possible results.

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT.

Where an event is as simple as possible, we call it *indecomposable.* Tossing a coin *once* produces an indecomposable event. There are two outcomes: H or T.

An indecomposable event is called a *point* and the collection of points is called *sample space.*

A *probability distribution* on a sample space $x_1, \ldots, x_n$ is an assignment of *a probability* p_i to each point x_i. Requirements are

$$p_i \geq 0, \qquad \text{and} \quad \sum_i p_i = 1.$$

The probability of a point is an estimate of how often it will occur in repeated observations. Tossing a coin, we assume H and T will both occur about 50% of the time when repeating many tosses. If they don't, you would just say that the coin is 'biased' or being tossed 'unfairly,' or that the experiment had an 'unusual' outcome.

Probability is usually based on counting.

Recall the 'n choose r' notation.

$$\binom{n}{r}$$

is the number of subsets of r elements in a set of n elements.

Equivalently, it is the number of *ascending* subsequences of length r out of the sequence $1, 2, \ldots, n$.

Also,

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}.$$

Consider poker, without jokers: 52 cards. Number the 52 cards in a pack in any order (e.g., 2 of diamonds to ace, clubs, hearts, spades). A poker hand is a subsequence of length 5. There are

$$\binom{52}{5} = 2598960$$

poker hands. Four of these are royal straight flushes. Your chance of being dealt a RSF is $1/649740$. $13 * 48$ are fours. Your chance of being dealt fours is $1/4065$. And so on. If we play a million games of poker, we get fours about 200 times and RSFs once or twice. That is, most times we play a million games we get fours about 200 times and RSFs once or twice.

Of course, probability is of enormous practical importance in connection with laboratory experiments, quality control, insurance, and so on.

Events are *independent* if the probability of one occurring does is not affected by the other occurring. For example, in four coin-tosses, the events 'First toss is heads' and 'Two heads come up' are independent, but 'first toss is heads' and 'one head comes up' are not independent.

This is because the first event is

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT,

probability 0.5, and the second is

HHTT, HTHT, HTTH, THHT, THTH, TTHH,

probability 0.375.

If the first event occurs, the probability of the second event also occurring is still 0.375. If it doesn't occur, the probability of the second is still 0.375.

The third event is

HTTT, THTT, TTHT, TTTH,

probability $1/4$. If the first event occurs, then the probability of the third event is $1/8$. The events are not independent.

Multiplicative principle. Where you have several independent events $E_1, E_2, \ldots$ the probability of $E_1, E_2, \ldots$ is the product of the probabilities of each event.

Consider a manufacturing process. Suppose that p is the probability of a certain component being defective. Out of n of these components, what is the probability that exactly r are defective?

If there are r defective, they form a subset of size r, and there are $\binom{n}{r}$ such subsets.

We need to assume that each component has probability p of being defective, independent of the other components. The probability of *not* being defective is $1 - p$. With that assumption, and the multiplicative principle, a given size-r subset has probability $p^r(1-p)^{n-r}$ of coinciding with the set of defectives, and the probability of r defectives is

$$\binom{n}{r} p^r (1-p)^{n-r}.$$

This distribution is called the *binomial* distribution. Note that it adds up to 1.

Example. Suppose $p = 0.1$. In a batch of 20, what is the probability that 3 are defective? The answer is simply

$$\binom{20}{3}(0.1)^3(0.9)^{17}.$$

This requires a calculator. The answer is about .19.

Here is a related distribution. Suppose that customers are arriving at a queue at random, but the probability of a customer arriving in an infinitesimal time-interval dt is λdt, where λ is a positive constant.

We note a property of the exponential function. It is well-known that

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

and that

$$(1 + x/n)^n = 1 + x + \frac{n-1}{n}\frac{1}{2!}x^2 + \frac{(n-1)(n-2)}{n^2}\frac{1}{3!}x^3 \dots$$

This explains the following limit

$$\lim_{n\to\infty}(1 + x/n)^n = e^x.$$

Imagine a time-interval of length t divided into a large number n of intervals each of length t/n. The chance of a customer arriving in the i-th interval is $\lambda t/n$. The chance of r customers arriving in the total interval follows a binomial distribution

$$\binom{n}{r}\left(\frac{\lambda t}{n}\right)^r\left(1 - \frac{\lambda t}{n}\right)^{n-r}.$$

This can be written as

$$\frac{n(n-1)\cdots(n-r+1)}{n\cdot n\cdots n}\frac{1}{(1-\lambda t/n)^r}\frac{1}{r!}(\lambda t)^r\left(1 - \frac{\lambda t}{n}\right)^n.$$

As $n \to \infty$, this converges to

$$e^{-\lambda t}\frac{(\lambda t)^r}{r!}.$$

Example. Cars arrive at a tollbooth at the rate of 2/minute. What is the probability that fewer than 4 will arrive in the next 3 minutes?

Answer. $\lambda = 2$ (when t is measured in minutes).

$$e^{-2*3}\sum_{r=0}^{3}\frac{(2*3)^r}{r!} = e^{-6}(\frac{1}{1} + \frac{6}{1} + \frac{36}{2} + \frac{216}{6}) = 61e^{-6} = .1512$$

The probability that exactly 6 will arrive is .1606, and the probability that fewer than 6 will arrive is .4456.

Mean and variance. Two important features of a probability distribution p_r are its *mean*, which is *defined as*

$$\sum rp_r$$

and its *variance* defined as

$$\sum (r-\mu)^2 p_r,$$

where μ is the mean. These depend on the probability distribution p_r, of course.

For example, in a binomial distribution $B(m,p)$, (meaning probability p, with m trials) the mean is

$$\sum_r r\binom{m}{r}p^r(1-p)^{m-r}.$$

Now

$$r\binom{m}{r} = r\frac{m(m-1)\cdots(m-r+1)}{r(r-1)\cdots 1} = m\frac{(m-1)\cdots(m-r+1)}{(r-1)\cdots 1} = m\binom{m-1}{r-1},$$

so long as $r \geq 1$. For $r = 0$ the value is zero. Therefore the mean is

$$m\sum_{r=1}^{m}\binom{m-1}{r-1}p^r(1-p)^{m-r}.$$

This is

$$mp\sum_{r=1}^{m}\binom{m-1}{r-1}p^{r-1}(1-p)^{(m-1)-(r-1)},$$

and the sum is just $(p+1-p)^{m-1} = 1$, so the mean is p.

Using similar tricks, the variance of $B(m,p)$ is

$$mp(1-p).$$

For the Poisson distribution with parameter λ, the mean is

$$\sum_{r\geq 0} r\frac{\lambda^r}{r!}e^{-\lambda} = \lambda\sum_{r\geq 1}\frac{\lambda^{r-1}}{(r-1)!}e^{-\lambda} = \lambda.$$

The variance is

$$\sum_r (r-\lambda)^2\frac{\lambda^r}{r!}e^{-\lambda} = \lambda,$$

also, using similar tricks.

In a general probability distribution, if the variance were zero, there would be no probability to discuss, since every trial would result in the same answer (equal to the mean). Generally, low variance means most of the probability is concentrated around the mean.

Mean and average. If we take n independent samples x_i from a distribution, their average

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

is also subject to a probability distribution. It has the same mean as the original distribution, but the variance is divided by n. This is reflected in the

Law of large numbers. For any $\epsilon > 0$, the probability that

$$\bar{x} - \mu > \epsilon$$

goes to 0 as $n \to \infty$.

In fact, there is a very curious theorem, the **Central Limit Theorem**, which says that the *distribution* of these averages converges to the same distribution (the Normal distribution) for large n, and that depends only on the mean and variance.

14 Event independence and Bayes' Theorem

The word 'event' means something like a 'set of possible outcomes.' Often, one compares different events.

We have already seen an example with four tosses of a coin. The events A: 'first toss is heads' and B: '2 heads come up' are independent but A and and C: '1 head comes up' are not.

Obviously the events 'first toss is heads' and 'first toss is tails' are not independent. They are examples of 'mutually exclusive' events. Their *joint probability* is zero.

Event A is

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT.

If we confined our attention only to coin-tossing trials yielding this event, the points in the restricted space would all have the same probability, only twice as big. This we call the *conditional probability* distribution. Generally we use the notation

$$\text{Prob}(B)$$

for the probability of an event B. The *conditional probability* of B given A is written

$$\text{Prob}(B|A)$$

and defined as

$$\frac{\text{Prob}(B \text{ and } A)}{\text{Prob}(A)}.$$

With the above examples, $\text{Prob}(B|A) = 3/8$ and $\text{Prob}(C|A) = 1/8$.

Events A and B are *independent* if

$$\text{Prob}(B|A) = \text{Prob}(B),$$

or equivalently

$$\text{Prob}(B \text{ and } A) = \text{Prob}(B)\text{Prob}(A).$$

Bayes's Theorem says

$$\text{Prob}(A)\text{Prob}(B|A) = \text{Prob}(B)\text{Prob}(A|B),$$

and is obvious from the definitions.

Suppose a sample space is partitioned into events $S_1, \ldots, S_k$. Assume they are disjoint in the sense that no two have any points in common, and exhaustive in the sense that every point belongs to some S_j. Since they are disjoint, if $i \neq j$ then $\text{Prob}(S_i \text{ and } S_j) = 0$. One says they are *mutually exclusive*.

(14.1) Lemma Distribution formula with disjoint events. *For any event B,*

$$\text{Prob}(B) = \sum_j \text{Prob}(B \text{ and } S_j)$$

and

$$\text{Prob}(B) = \sum_j \text{Prob}(S_j)\text{Prob}(B|S_j). \quad \blacksquare$$

Example. In a factory, assembly lines A and B make a certain product at the same rate. A has 10% defective,B 5%. What is the probability that a random defective product is from A?

Answer. Let A, B be the events that a product came from A,B respectively, and D be the event that a product is defective. By assumption, $\text{Prob}(A) = \text{Prob}(B) = 1/2$. By the above formulae,

$$\text{Prob}(D) = \text{Prob}(A)\text{Prob}(D|A) + \text{Prob}(B)\text{Prob}(D|B) = (0.5)(0.1) + (0.5)(0.05) = 0.075.$$

$$\text{Prob}(A|D) = \frac{\text{Prob}(D|A)\text{Prob}(A)}{\text{Prob}(D)} = \frac{0.05}{0.075} = \frac{2}{3}.$$

Example. Same general scenario, 3 assembly lines A,B,C. But A make products twice as fast as B, and B and C produce at the same rate. If A,B,C produce 5, 10, and 15% defectives, what is the probability of a defective product having come from B?

Answer. The information about the rates of production gives the first line of probabilities in the table below, the conditional probabilities give the second, and the general formula give the joint probabilities. Then the distribution formula gives the numbers in the last line, whence the probability of D is $.875$.

event	A	B	C
prob	$1/2$	$1/4$	$1/4$
conditional D	.05	.10	.15
D and	.025	.025	.0375

The answer we seek, $\text{Prob}(B|D)$, is $\text{Prob}(B \text{ and } D)/\text{Prob}(D)$ = $.025/.0875 = .2857$.

15 Gaussian (normal) distribution

The Gaussian distribution is continuous — the 'points' or irreducible events are real numbers. With mean μ and standard deviation σ, the distribution is

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

To begin with, total probability is

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-u^2} du = 1.$$

(The integral is well-known.)

The mean is

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2}dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x+\mu)e^{-x^2/2\sigma^2}dx = \mu,$$

since $\int_{-\infty}^{\infty} xe^{-x^2/2\sigma^2}dx = 0$ — it is an 'odd function.'

The variance is

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^2e^{-(x-\mu)^2/2\sigma^2}dx.$$

Replacing the variable x by $(x-\mu)/(\sigma\sqrt{2})$, we get

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(2\sigma^2)x^2e^{-x^2}\sigma\sqrt{2}dx.$$

The integral is

$$-\int xd(\frac{1}{2}e^{-x^2})dx = [x\frac{1}{2}e^{-x^2}] + \frac{1}{2}\int e^{-x^2}dx = \frac{\sqrt{\pi}}{2},$$

so the variance is

$$\sigma^2.$$

The Poisson distribution, for large λt, behaves like the normal distribution. Here is a partial explanation.

We make an approximation to

$$e^{-\lambda t}\frac{(\lambda t)^r}{r!}$$

on the assumption that λt is large and $(r-\lambda t)/(\lambda t)$ is relatively small.

Let $z = (r-\lambda t)/(\lambda t)$.

There is a well-known approximation to $r!$, **Stirling's Formula,** when r is large

$$r! \approx \sqrt{2r\pi}\left(\frac{r}{e}\right)^r.$$

Substitute this into the Poisson distribution and take logs.

$$\ln p_r \approx -\lambda t + r\ln(\lambda t) - r(\ln r - 1) + R,$$

where $R = \ln(1/(\sqrt{2\pi r}))$. Our assumption about λt and r allow us to approximate R by $\sqrt{2\pi\lambda t}$. We use the approximation $\ln(1+z) \approx z - z^2/2$.

$$-\lambda t + r\ln(\lambda t) - r\ln r + r + R = \lambda tz - r(\ln(r/\lambda t)) + R \approx$$
$$\lambda tz(1-(1+z)(1-z/2)) + R = \lambda tz(-z/2 + z^2/2) + R \approx -\lambda tz^2/2 + R.$$

Exponentiate and substitute μ for λt and σ^2 for the same. Then

$$p_r \approx \frac{1}{\sigma\sqrt{2\pi}}e^{-(r-\mu)^2/2\sigma^2}.$$

This is a Gaussian distribution.

This is related to a very strong result, the **Central Limit Theorem.**

(15.1) Theorem *Given a large number $x_1, \ldots, x_N$ of independent results from the same distribution, which has mean μ and variance V, the average*

$$\overline{x} = \frac{x_1 + \ldots + x_N}{N}$$

has a probability distribution converging to Gaussian with mean μ and variance V/N. ∎

16 Estimation

This is where we want to estimate properties of the distribution from samples $x_1, \ldots, x_N$.

(16.1) Definition *A* random variable *is simply a function f whose domain is a sample space.*
If the probability distribution over this space is written p_r, then the expectation of f, $E(f)$, *is*

$$E(f) = \sum_r p_r f(r),$$

and in the continuous case (probability distribution $p(x)$) it is

$$E(f) = \int p(x) f(x) dx.$$

Thus $E(x)$ is the mean, and $E(x - E(x))^2$ is the variance.
We also write $V(x)$ for the variance.

Mean. Generally speaking, there is pretty well only one sensible way to estimate the mean: use the average $\overline{x}$.

Variance. The formula

$$\frac{1}{N} \sum_i (x_i - \overline{x})^2$$

looks good, but is slightly inaccurate.

The correct estimate for $V(x)$ is

$$\frac{1}{N-1} \sum_i (x_i - \overline{x})^2$$

(details omitted).

Maximum likelihood estimators. This is where one is trying to discover the probability distribution by observing a sample. The Maximum Likelihood Principle (MLP) is: choose the distribution which makes the observation most likely.

Example. To estimate the number of fish in a lake, 1000 fish were caught, 'tagged,' and returned to the lake. Later, another 1000 fish were caught, inspected, and returned. 100 of them were found to be tagged. Using the MLP, estimate the number of fish in the lake.

Let N be the true number of fish in the lake. After tagging, the probability that a random fish will be tagged is $p = 1000/N$. Another sample of 1000 fish is taken. Strictly speaking, the number of

tagged fish in the second sample follows a 'hypergeometric distribution,' such as applies to hands in poker, but we treat it as a binomial distribution.

So the answer is that N such that

$$p_r(N) = \binom{1000}{r} p(N)^r (1 - p(N))^{1000-r}$$

is maximal, where $r = 100$ and $p(N) = 1000/N$.

In the general binomial distribution $B(n, p)$,

$$p_r = \binom{n}{r} p^r (1-p)^{n-r}.$$

Equate the derivative, with respect to p, to zero, and ignore the coefficient, getting

$$(1-p)^{n-r} r p^{r-1} - (n-r)(1-p)^{r-1} p^r = 0,$$

so

$$r(1-p) = (n-r)p, \quad r - rp = np - rp,$$

so $p = r/n$. With $p = 1000/N$, $r = 100$, and $n = 1000$,

$$\frac{1000}{N} = 0.1$$

and

$$N = 10,000.$$

17 Logic

We are concerned with **truth-functions**, functions whose values are the two **truth-values** $0, 1$ (for **false** and **true** respectively), and whose arguments are also truth-values.

Boolean variables are variables which are restricted to truth-values.

Certain truth-functions are well-known.

$$0 \mapsto 1, \quad 1 \mapsto 0$$

is simply *negation (not)*. If X is a Boolean variable then $\neg X$ is its negation. Negation can be represented in a **truth table** as follows

X	$\neg X$
0	1
1	0

$$(0,0) \mapsto 0, \quad (0,1) \mapsto 0, \quad (1,0) \mapsto 0, \quad (1,1) \mapsto 1$$

is *conjunction (and)*. If X and Y are Boolean variables, $X \wedge Y$ represents their conjunction. Here is the truth table for conjunction.

X	Y	$X \wedge Y$
0	0	0
0	1	0
1	0	0
1	1	1

It can also be displayed in a table as follows.

$X \wedge Y$	0	1
0	0	0
1	0	1

Disjunction (or) is represented $X \vee Y$ and has the following table.

$X \vee Y$	0	1
0	0	1
1	1	1

Implication (if... then) is represented $X \Rightarrow Y$ and has the following table.

$X \Rightarrow Y$	0	1
0	1	1
1	0	1

It is just a way of connecting Boolean variables, and in fact $X \implies Y$ is equivalent to $(\neg X) \vee Y$ — the two expressions have the same truth-table. I believe it is called the Philonian conditional. It is the weakest kind of 'implication' which guarantees the following:

If X is true and $X \implies Y$ is true then Y is true.

This is easily checked from the truth-tables. Hence the following kind of reasoning is valid:

$$\begin{array}{ll} & X \\ & X \implies Y \\ \hline \therefore & Y \end{array}$$

This is called the *rule of Modus Ponens.* Note that from X and $Y \implies X$ you *cannot* deduce Y. These logical connectives ($\wedge, \vee, \neg$, etcetera) have some simple algebraic properties:

- $\wedge$ and $\vee$ are commutative and associative.
- $\wedge$ distributes over $\vee$, that is, $X \wedge (Y \vee Z) = (X \wedge Y) \vee (X \wedge Z)$.
- $\vee$ distributes over $\wedge$.
- $X \implies Y$ is equivalent to $(\neg X) \vee Y$.
- $X \wedge \neg X$ is always false and $X \vee \neg X$ is always true.
- $\neg(X \wedge Y) = (\neg X) \vee (\neg Y)$ and $\neg(X \vee Y) = (\neg X) \wedge (\neg Y)$. **(De Morgan laws.)**

From a mathematical point of view, the business of logic is, given a Boolean expression, is it always true no matter what the values of its Boolean variables? Is it always false?

From an engineering point of view, the business of logic is to construct digital electronic circuits which implement certain truth-functions. A circuit is usually equivalent to a Boolean expression.

To begin with, *every* truth-function can be realised by a Boolean expression using only $\wedge, \vee, \neg$.

(17.1) Definition *Sometimes we write $\overline{X}$ for $\neg X$. We also define $\overline{\neg X} = X$.*

A literal *is either a Boolean variable X or the negation $\overline{X}$ of a Boolean variable.*

A disjunctive normal formula (DNF) *is a Boolean expression of the form*

$$(L_1 \wedge L_2 \wedge \ldots \wedge L_k) \vee (L_{k+1} \wedge L_{k+2} \wedge \ldots L_\ell) \vee \ldots \vee (L_{r+1} \wedge L_{r+2} \wedge \ldots L_s)$$

where $L_1, \ldots, L_s$ are literals, not necessarily distinct.

For example

$$(\neg X) \vee Y$$

is a very simple DNF.

(17.2) Lemma *Every truth-function can be realised by a DNF.*

Proof. Suppose the truth-function has n arguments so it can be written as

$$f(X_1, \ldots, X_n).$$

There are actually 2^n different inputs, that is, there are 2^n different ways of assigning truth-values to these n variables.

If $f(X_1, \ldots, X_n)$ is false for all inputs then it can be realised by the DNF $X_1 \wedge \overline{X_1}$.

We may therefore assume that $f(X_1, \ldots, X_n)$ is true for some inputs.

Let

$$T_1, \ldots, T_n$$

be one of these inputs. For example, we could have $n = 3$ and $T_1, T_2, T_3 = 0, 1, 0$. There is a unique expression E of the form

$$(\pm X_1) \wedge (\pm X_2) \wedge \ldots \wedge (\pm X_n)$$

which is true if and only if $X_1 = T_1, X_2 = T_2$, etcetera. Namely,

$$E = L_1 \wedge \ldots \wedge L_n, \quad \text{where}$$

$$L_i = \begin{cases} X_i & \text{if } T_i = 1 \\ \overline{X_i} & \text{if } T_i = 0. \end{cases}$$

Call this expression $E_{T_1,\ldots,T_n}$.

Let

$$D = E_1 \vee E_2 \vee \ldots \vee E_N,$$

where $E_1, \ldots, E_N$ are the formulae $E_{T_1,\ldots,T_n}$, for all the inputs $T_1, \ldots, T_n$ such that $f(T_1, \ldots, T_n) = 1$. Then

$$\begin{aligned} & f(T_1, \ldots, T_n) = 1 \\ \Leftrightarrow \quad & E_{T_1,\ldots,T_n} \text{ occurs in } D \\ \Leftrightarrow\ & D \text{ is true when } X_1 = T_1, \ldots, X_n = T_n, \end{aligned}$$

so D realises f. **Q.E.D.**

(17.3) Definition *A formula is in* conjunctive normal form (CNF) *if it is of the form*

$$(L_1 \vee L_2 \vee \ldots \vee L_k) \wedge (L_{k+1} \vee L_{k+2} \vee \ldots L_\ell) \wedge \ldots \wedge (L_{r+1} \vee L_{r+2} \vee \ldots L_s)$$

where $L_1, \ldots, L_s$ are literals, not necessarily distinct.

(17.4) Corollary *Every truth-function can be realised by a CNF.*

Proof. Let D be a DNF realising not-$f(T_1, \ldots, T_n)$. The formula $\neg D$ is easily converted into a CNF using De Morgan's laws, and it realises f. **Q.E.D.**

Digital logic is about constructing electronic circuits to realise truth-functions, or several truth-functions together. In this context the Boolean variables are called *inputs*, the function values are called *outputs*, and the electronic components are called *gates.* The inputs and outputs are usually just voltages (within certain ranges, which are taken to mean 0 and 1 respectively).

There are a small number of gates used, corresponding to the familiar connectives: NOT-gates, AND-gates, and OR-gates. The NAND is not-and, $\neg(X \wedge Y)$, and the connective is written $X|Y$. This is called the *Sheffer stroke.*

Also there is a NOR gate, not-or, $\neg(X \vee Y)$, and the connective is written $X \downarrow Y$. This is called the *Pierce arrow.*

We shall consider digital logic later. First we consider a method of theorem-proving.

The goal of theorem-proving in zero-order logic is to show that a given expression is always true (a *tautology*) or always false (a *contradiction* or *inconsistent.*)

The obvious way to prove one or the other is to try all possible inputs (combinations of truth-values). If there are n inputs there are 2^n inputs to check, and this could be a large number.

As a ridiculous example, consider the CNF

$$(X_1 \vee \overline{X_1}) \wedge \ldots \wedge (X_n \vee \overline{X_n}).$$

It is rather obviously a tautology, and one should not inspect all 2^n inputs separately to see this.

A method which cannot be less efficient and often is more so is the *Davis-Putnam procedure* published (for first-order logic) in 1960 by Martin Davis and Hilary Putnam. Hilary Putnam is a visiting professor at UCD this year.

Given an expression $E(X_1, \ldots, X_n)$, if $n = 1$ then try the two truth-values and hence determine the answer.

Otherwise choose one of the variables, X_j, (some choices may be better than others). Let

$$E_0(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) = E(X_1, \ldots, X_{j-1}, 0, X_{j+1}, \ldots, X_n)$$

and

$$E_1(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) = E(X_1, \ldots, X_{j-1}, 1, X_{j+1}, \ldots, X_n).$$

Recursively, test if E_0 is a tautology. If it isn't, stop: the original expression isn't. Otherwise test of E_1 is a tautology. If it is, then the original expression is; if not, then the original expression isn't.

In the example, E_0 and E_1 are always the same, so if one saves the results of intermediate computations the total work done is quite low.

Another method is called *Robinson's Resolution Principle.* It can be applied to a DNF to test for a tautology and to a CNF to test for inconsistency (it is easy to test a CNF for tautology or a DNF for inconsistency). The method is essentially the same for each.

We consider testing a CNF for inconsistency.

The subformulae $L_i \vee L_{i+1} \vee \ldots \vee L_j$ are called *clauses.* One views the CNF as a *set* of clauses, and repeatedly adds *resolvents* to the set of clauses.

Given two clauses C and C', a *resolvent* of C and C' is constructed as follows. It is necessary that C contains a literal L whose complement $\overline{L}$ occurs in C'. In this case suppose

$$C = L_1 \vee \ldots \vee L_k \vee L \quad \text{and} \quad C' = L'_1 \vee \ldots \vee L'_m \vee \overline{L}$$

then the clause obtained by *resolving L and $\overline{L}$* is

$$L_1 \vee \ldots \vee L_k \vee L'_1 \vee \ldots \vee L'_m.$$

It is possible that $k = m = 0$, in which case the resolvent is not a conventional formula but is called the *empty* clause and written $\Box$.

(17.5) Proposition *A CNF is inconsistent if and only if the empty clause can be generated by resolution.*

For example, the following is a kind of justification of Modus Ponens: we show that

$$X, \overline{X} \vee Y, \overline{Y}$$

are inconsistent.

$$\begin{aligned} X, \overline{X} \vee Y, \overline{Y} \\ X, \overline{X} \vee Y, \overline{Y}, Y \\ X, \overline{X} \vee Y, \overline{Y}, Y, \Box \end{aligned}$$

Or we may present the proof by listing the disjuncts as they are supplied or generated.

Given the CNF

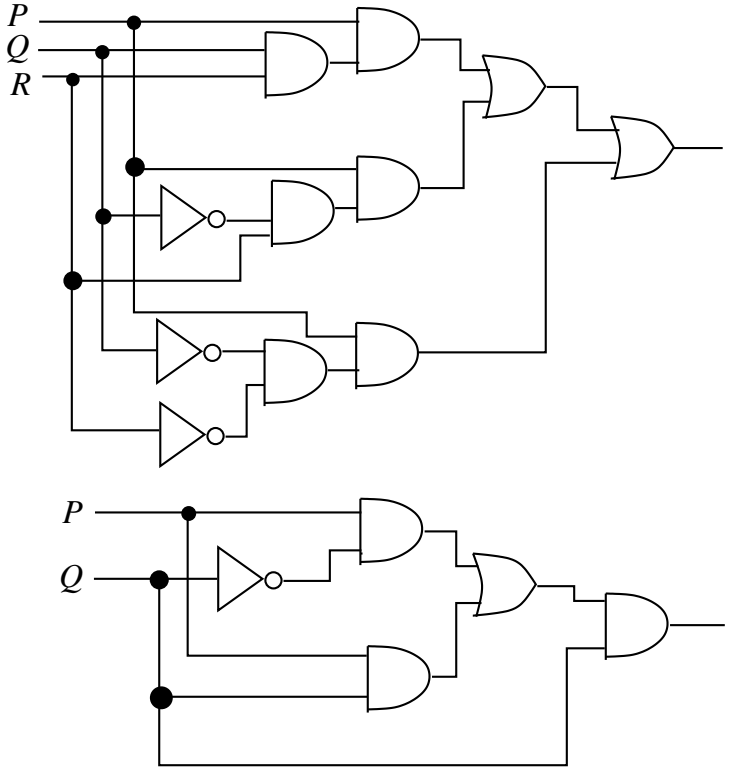
$$A \vee D, \quad \overline{A} \vee \overline{D}, \quad \overline{A} \vee B \vee C, \quad A \vee B \vee \overline{C}, \quad \overline{B}, \quad D \vee C, \quad \overline{D} \vee \overline{C}$$

here is a resolution refutation (proof of inconsistency).

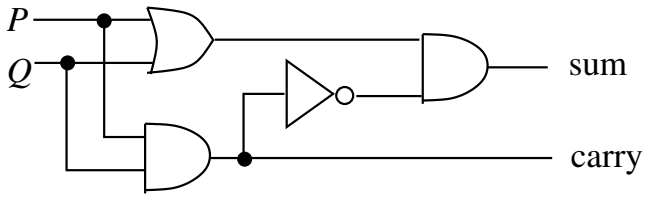
$$\begin{aligned}
A \vee B \vee C, \quad \overline{B} &\mapsto A \vee C \\
\overline{A} \vee B \vee \overline{C}, \quad \overline{B} &\mapsto \overline{A} \vee \overline{C} \\
A \vee C, \quad \overline{C} \vee \overline{D} &\mapsto A \vee \overline{D} \\
A \vee D, \quad A \vee \overline{D} &\mapsto A \\
A, \quad \overline{A} \vee \overline{D} &\mapsto \overline{D} \\
C \vee D, \quad \overline{D} &\mapsto C \\
\overline{A} \vee \overline{C}, \quad C &\mapsto \overline{A} \\
A, \quad \overline{A} &\mapsto \square
\end{aligned}$$

18 Digital circuits

Digital logic is about building circuits which realise Boolean functions. They have boolean inputs, generally measured by voltage, and one or more boolean outputs. Graphically, there are ways of depicting AND-gates, OR-gates, and NOT-gates. Multiple ANDs and ORs are also allowed. The input 'wires' are as shown: wires can be split as shown. (If they just cross they are separate.)



Binary arithmetic is very important. The first circuit is a *half-adder* which has two outputs, sum (modulo 2) and carry.



Computer arithmetic usually operates with binary numbers of fixed size. Typically, 32-bit numbers are used. A **bit** is a binary digit. We shall give small examples, in 4-bit arithmetic (equivalent to hexadecimal). The bit-patterns are

$$0000,\ 0001,\ 0010,\ \ldots\ 1111: \qquad 0,\ 1, 2, \ldots 15.$$

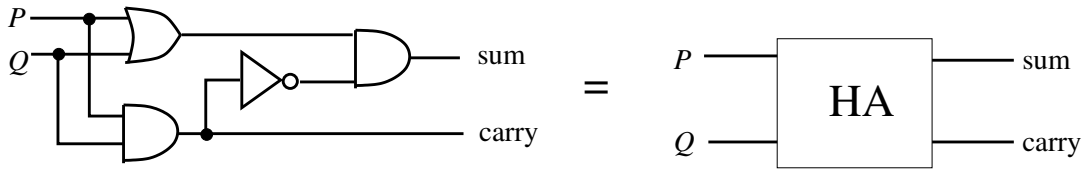
Addition is like decimal addition, only simpler.

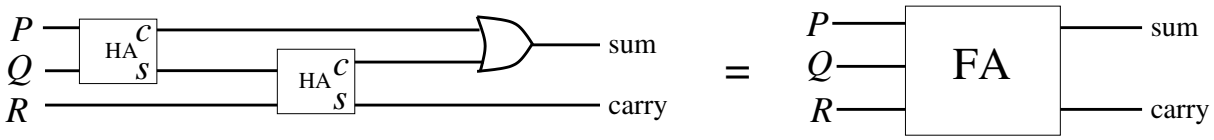
```
  1010
+ 0111
  ----
  1001
```

Or there may be a carry

```
  1010
+ 1101
  ----
  0111
 1     is carried.
```

In order to perform 4-bit addition, it helps to use a circuit for 3-bit addition. Such a circuit is illustrated below.





It produces the output in 2 bits, sum and carry (the order is for some reason reversed in this diagram). The input bits are labelled P, Q, R. The low-order output bit, labelled sum, is simply the mod-2 sum of P, Q, R, and is got by combining the sum output from a P, Q half-adder with R.

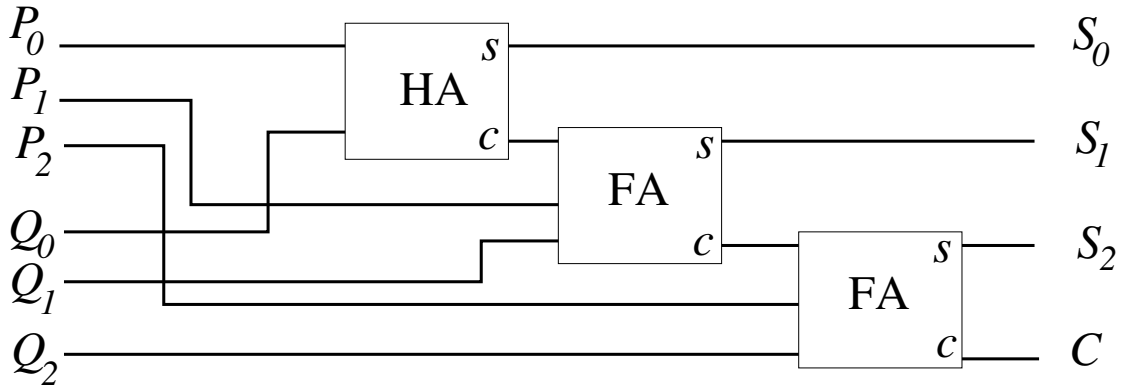
The high-order bit, labelled carry, is 1 if and only if at least 2 of P, Q, R are 1.

A formula for the carry is $(P \wedge Q) \vee ((P \vee Q) \wedge R)$.

$P \wedge Q$ is the carry from the first adder. If this is zero, then the sum bit in the first adder equals $P \vee Q$. This bit, together with R, is input to the second half-adder, and the output carry, in the case $P \wedge Q = 0$, is $(P \vee Q) \wedge R$.

Hence the full adder works.

These circuits can be combined to add binary numbers of fixed size, such as 16, 32, or 64 bits. The illustration shows a circuit to add two 3-bit numbers $P_2P_1P_0$ and $Q_2Q_1Q_0$. The result is $S_2S_1S_0$, plus a carry bit C indicating 'overflow.'



2s complement arithmetic. Suppose binary numbers are stored in 16 bits. It is customary to interpret numbers whose high-order bit is 1 as negative, and in fact, to a 16-bit number

$$n = b_{15}b_{14} \dots b_1b_0, \quad b_{15} = 1,$$

as $n - 2^16$. The range of 16-bit integers is

$$-2^{15} \dots 2^{15} - 1.$$

The same idea can be illustrated with 3-bit integers:

$$100 = -4, 101 = -3, 110 = -2, 111 = -1, 000 = 0, 001 = 1, 010 = 2, 011 = 3.$$

(18.1) Lemma *If x and y are 2s complement k-bit numbers in the correct range $-2^{k-1} \dots 2^{k-1} - 1$, and $x + y$ is in the correct range, then ordinary k-bit addition (ignoring the carry) will produce the correct result. (Proof needs a case-analysis, omitted).* ∎

The 2s complement of a k-bit number $n = b_{k-1}b_{k-2} \dots b_1b_0$ is $2^k - n$. For example, the 3-bit 2s complement of 2 is $8 - 2 = 6$. In binary

$$010 \mapsto 110.$$

The calculation can be taken in two steps. First, negate all bits in parallel. (This produces $111 \dots 111 - n = (2^k - 1) - n = (2^k - n) - 1$).

$$101$$

Then add 1, getting $2^k - n$.

$$101 + 001 = 110.$$

Let us reverse the process.

$$110 \mapsto 001 \mapsto 001 + 001 = 010.$$

19 ODEs and recurrences

We treat ODEs and recurrences together because, though they are very different, methods of solution are often very similar.

An ordinary differential equation is some relation between $x, y, y'(= dy/dx), y''(= d^2y/dx^2)$, etcetera. The solutions are differentiable functions y of x.

If an ODE involves x, y, y' and no higher derivatives, it is *first-order*. If it involves y'' but no higher derivatives, it is *second-order*.

A recurrence is some relation between n, y_n, y_{n+1}, y_{n+2}, etcetera. The solutions are functions y of $n = 0, 1, 2 \ldots$ — that is, sequences. First-order recurrences involve n, y_n, y_{n+1}. Second-order recurrences involve y_n and y_{n+2} but no higher-indexed member of the sequence y.

The simplest ODE is

$$\frac{dy}{dx} = f(x),$$

for which the solution is

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

The constant of integration is important. First-order ODEs (and recurrences) have not one solution, but infinitely many, parametrised by a 'constant of integration' C. The solutions to second-order ODEs and recurrences involve two 'constants of integration.'

One or two further pieces of information are enough to pin down the constants of integration.

For example,

$$\frac{dy}{dx} = \frac{1}{x}; \quad y(1) = 0$$

has the solution $y = \ln x$. The general solution is $y = \ln x + C$.

The recurrence

$$y_{n+1} = y_n + 1; \quad y(0) = 0$$

has the solution $y_n = n$. The general solution is $y_n = n + C$.

The recurrence analogue to an integration problem is a summation problem.

$$y_{n+1} - y_n = g_n,$$

where g_n is a known sequence. Write it in the form

$$y_{n+1} = y_n + g_n,$$

and you have a recipe for calculating y_n.

$$y_1 = y_0 + g_0,$$

$$y_2 = y_1 + g_1 = y_0 + g_0 + g_1,$$

$$y_3 = y_2 + g_2 = y_0 + g_0 + g_1 + g_2,$$

and so on. In general,

$$y_n = y_0 + \sum_{r=0}^{n-1} g_r.$$

This is what is meant by a summation problem. No information is given about y_0. It can have any value, so we write $y_0 = C$ and get

$$y_n = C + \sum_{r=0}^{n-1} g_r.$$

The recurrence has been reduced to a summation problem. The summation problem need not be easy.

The solution to

$$y_{n+1} - y_n = 0$$

is $y_n = C$. The solution to

$$y_{n+1} - y_n = 1$$

is

$$y_n = C + \sum_{r=0}^{n-1} 1 = C + n.$$

The solution to

$$y_{n+1} - y_n = n$$

is

$$y_n = C + \sum_{r=0}^{n-1} r = C + \frac{n(n-1)}{2}.$$

The solution to

$$y_{n+1} - y_n = n^2$$

is $y_n = C + \sum_{r=0}^{n-1} r^2 = C + \frac{(n-1)n(2n-1)}{6}$. This last is rather complicated to remember. There are formulae for

$$\sum_{r=0}^{n-1} r^d$$

for any nonnegative integer d, but they get increasingly complicated. On the other hand, a function of n similar to n^d is

$$n(n-1)\cdots(n-d+1) = d!\binom{n}{d} = P_d^n.$$

There is a simple formula for this sum, because

$$\begin{aligned} P_{d+1}^{n+1} - P_{d+1}^n = \\ (n+1)n(n-1)\cdots(n+1-(d+1)+1) - n(n-1)\cdots(n-d) = \\ (n+1-(n-d))(n(n-1)\cdots(n+1-d)) = (d+1)P_d^n. \end{aligned}$$

$$\begin{aligned} P_{d+1}^{n+1} - P_{d+1}^n + P_{d+1}^n - P_{d+1}^{n-1} + P_{d+1}^{n-1} - P_{d+1}^{n-2} + \ldots - P_{d+1}^0 \\ = P_{d+1}^{n+1} - 0 = \\ (d+1)P_d^n + (d+1)P_d^{n-1} \ldots = (d+1)\sum_{r=0}^{n} P_d^n. \end{aligned}$$

Therefore (changing the range of summation)

$$\sum_{r=0}^{n-1} P_d^r = (d+1)P_{d+1}^n.$$

For example,

$$\sum_{r=0}^{n-1} r(r-1) = \frac{n(n-1)(n-2)}{3}.$$

Therefore, since $n^2 = n(n-1) + n$,

$$\sum_{r=0}^{n-1} r^2 = \frac{n(n-1)(n-2)}{3} + \frac{n(n-1)}{2} = \frac{n(n-1)(2n-1)}{6}.$$

Another summation problem is a geometric series

$$\sum_{r=0}^{n-1} x^r.$$

If we call this sum s_{n-1}, then the well-known trick for calculating s_n is

$$\begin{aligned} s_{n-1} &= 1 + x + \ldots + x^{n-1} \\ xs_{n-1} &= x + x^2 + \ldots + x^{n-1} + x^n \\ (1-x)s_{n-1} &= 1 - x^n \end{aligned}$$

Therefore

$$s_{n-1} = \frac{1-x^n}{1-x}.$$

Here is a more complicated version of this trick.

$$\sum_{r=0}^{n-1} rx^{r-1}.$$

There are two ways to do this. One is to differentiate $\sum_r x^r$. The other is similar to the above: call the sum s_{n-1}.

$$\begin{aligned} s_{n-1} &= 1 + 2x + 3x^2 + \ldots + n - 1x^{n-2} \\ xs_{n-1} &= x + 2x^2 + 3x^3 + \ldots + (n-2)x^{n-2} + (n-1)x^{n-1} \\ (1-x)s_{n-1} = 1 + x + x^2 + \ldots x^{n-2} - (n-1)x^{n-1} &= \frac{1 - x^{n-1} - (1-x)(n-1)x^{n-1}}{1-x} \\ s_{n-1} &= \frac{1 - nx^{n-1} + x^n}{(1-x)^2}. \end{aligned}$$

20 Separable and linear ODEs and recurrences

A *separable* ODE is one which can be expressed in the form

$$g(y)\frac{dy}{dx} = h(x).$$

To solve it, follow these steps (they are easy to remember, but don't make much sense mathematically):

$$\begin{aligned} g(y)dy &= h(x)dx \\ \int g(y)dy &= \int h(x)dx + C \end{aligned}$$

For example

$$\begin{aligned} \frac{dy}{dx} &= \alpha y \\ \frac{dy}{y} &= \alpha dx \\ \int \frac{dy}{y} &= \int \alpha dx + c \\ \ln y &= \alpha x + c \\ y &= e^c e^{\alpha x} \\ y &= Ce^{\alpha x} \end{aligned}$$

Another example:

$$\begin{aligned} \frac{dy}{dx} &= -2xy \\ \frac{dy}{y} &= -2xdx \\ \ln y &= -x^2 + c \\ y &= e^c e^{-x^2} \\ y &= Ce^{-x^2} \end{aligned}$$

There seems to be no natural idea of 'separable recurrence.'

An important class of ODEs is *linear*. Linear first-order ODEs have the form

$$\frac{dy}{dx} + P(x)y = Q(x).$$

Here is a curious method which always works.

- If the right-hand side of a linear ODE is zero, it is called *homogeneous*.
- Let u be any nonzero solution to the homogeneous equation

$$\frac{du}{dx} + P(x)u = 0.$$

- Substitute $y = uv$ into the original equation. We get a separable differential equation for v.

For example,

$$\begin{aligned}
\frac{dy}{dx} - y &= x \\
\frac{du}{dx} - u &= 0 \\
\frac{du}{u} &= dx \\
\ln u = x + c \qquad &\text{take } c = 0 \\
\ln u &= x \\
u = e^x \qquad &\text{substitute } y = uv = e^x v \\
\frac{de^x v}{dx} - e^x v &= x \\
e^x \frac{dv}{dx} + e^x v - e^x v &= x \\
dv &= xe^{-x}dx \\
v = \int xe^{-x}dx = -\int x d(e^{-x}) = -xe^{-x} + \int e^{-x}dx &= (-x-1)e^{-x} \ldots + C \\
y = ve^x &= Ce^x - x - 1.
\end{aligned}$$

This method is called *variation of parameters* for the following reason. The general solution to the homogeneous equation is $y = uv$, where v is constant, the 'constant of integration,' a 'parameter.' So v becomes variable in the non-homogeneous equation.

Simple electrical circuits often produce linear ODEs. The simplest is the *RC-circuit* consisting of a resistor, a capacitor, and a voltage source, all connected in a circuit.

For the resistor, voltage is proportional to current

$$V = RI.$$

For a capacitor, voltage is proportional to charge, its derivative (d/dt) is proportional to current.

$$\frac{dV}{dt} = CI.$$

The current is the same at all points, so

$$\frac{dV}{dt} + \frac{C}{R}V = E$$

(E represents the voltage supplied). This is a typical linear ODE. For example, the supplied voltage is probably alternating,

$$E_0 \cos \omega t$$

$E_0 = 240$ volts or whatever, and $\omega = 100\pi$ inverse seconds, or whatever. Let $y = V/E_0$ and $\alpha = \frac{C}{R}$ so

$$\frac{dy}{dt} + \alpha y = \cos \omega t$$
$$\frac{du}{dt} + \alpha u = 0$$
$$\frac{du}{u} = -\alpha dt$$
$$\ln u = -\alpha t + c \qquad \text{take } c = 0$$
$$u = e^{-\alpha t}$$
$$\frac{de^{-\alpha t}v}{dt} + \alpha uv = \cos \omega t$$
$$e^{-\alpha t}\frac{dv}{dt} - \alpha e^{-\alpha t}v + \alpha e^{-\alpha t}v = \cos \omega t$$
$$\frac{dv}{dt} = e^{\alpha t} \cos \omega t$$
$$v = \int e^{\alpha t} \cos \omega t dt.$$

The last integral can be solved using integration by parts. This is laborious. Time is saved by using complex numbers, because

$$e^{\alpha t} \cos \omega t$$

is the real part of

$$e^{(\alpha + i\omega)t} = e^{\alpha t}(\cos \omega t + i \sin \omega t).$$

Now the ordinary formula for integrating e^{Ax} is valid even if A is complex. The integral is

$$\frac{e^{(\alpha + i\omega)t}}{\alpha + i\omega} + \text{const.}$$

Ignoring the constant, this is

$$\frac{e^{\alpha t}(\cos \omega t + i \sin \omega t)(\alpha - i\omega)}{\alpha^2 + \omega^2}$$

and we want the real part of this, which is

$$v = \frac{e^{\alpha t}(\alpha \cos \omega t + \omega \sin \omega t)}{\alpha^2 + \omega^2}.$$

Therefore

$$y = Ce^{-\alpha t} + \frac{\alpha \cos \omega t + \omega \sin \omega t}{\alpha^2 + \omega^2},$$

and V is just $E_0 y$.

Variation of parameters works well with recurrences. First, a linear homogeneous recurrence is one of the form

$$y_{n+1} - g_n y_n = 0,$$

$$y_1 = g_0 y_0$$
$$y_2 = g_1 y_1 = g_1 g_0 y_0$$
$$y_3 = g_2 y_2 = g_2 g_1 g_1 y_0$$

and the general solution is

$$y_n = C \prod_{r=0}^{r=n-1} g_r.$$

(This notation is similar to summation; it means the product of the g_r.) One example is

$$y_{n+1} - (n+1)y_n = 0.$$

The solution is

$$y_n = C \prod_{r=0}^{n-1} r = Cn!$$

This example is artificial, and the most useful recurrence is

$$y_{n+1} - \alpha y_n = 0$$

where α is constant. The solution is

$$y_n = C \prod_{r=0}^{n-1} \alpha = C\alpha^n.$$

The general nonhomogeneous linear first-order recurrence is

$$y_{n+1} - g_n y_n = h_n.$$

To solve it, solve the associated homogeneous recurrence for u_n (any nonzero solution will do) and substitute $y_n = u_n v_n$. This will lead to a summation problem.

For example, let us solve

$$y_{n+1} - 3y_n = n$$

(quite a hard problem).

$$\begin{aligned}
u_{n+1} - 3u_n = 0 \qquad u_n = c3^n; \quad \text{take } c = 1 \\
(3^{n+1}v_{n+1}) - 3(3^n v_n) = n \\
v_{n+1} - v_n = \frac{n}{3(3^n)} \\
v_n = \frac{1}{3}x \sum_{r=0}^{n-1} r x^{r-1} \qquad \text{where } x = 1/3 \\
= \frac{1}{9} \frac{1 - nx^{n-1} + (n-1)x^n}{(1-x)^2} \\
= \frac{1}{4}(1 - n(1/3)^{n-1} + (n-1)(1/3)^n) \ldots + C \\
y_n = c3^n + \frac{3^n - 3n + n - 1}{4} \\
= C3^n - \frac{2n+1}{4}.
\end{aligned}$$

All of these solutions can be checked.

21 Simultaneous linear ODEs

Here the differential equations describe functions $(x(t), y(t))$ of t ('time').

For example,

$$\frac{dx}{dt} = y; \frac{dy}{dt} = x.$$

In matrix terms, with $X = [x, y]^T$,

$$\frac{dX}{dt} = AX: \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Diagonalise the array. This is quite easy and we can produce an orthogonal system of eigenvectors

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

The matrix

$$A' = U^T A U = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

is diagonal, and we can use the eigenvectors to provide a new coordinate system

$$X' = U^T X.$$

Then

$$\frac{dX'}{dt} = \frac{dU^T X}{dt} = U^T \frac{dX}{dt} = U^T AX = U^T AUX' = A'X'.$$

Write $X' = [x', y']^T$. Then

$$\frac{dx'}{dt} = x'; \qquad \frac{dy'}{dt} = -y'.$$

We have 'decoupled' the two equations, and they are straightforward linear homogeneous ODEs:

$$x' = Ce^t, \quad y' = De^{-t}.$$

There are two 'constants of integration' which we can interpret as a 'starting point' (C, D) at $t = 0$ — in the new coordinate system.

For example, suppose $X = (1, 2)$ at $t = 0$.

$$\begin{bmatrix} C \\ D \end{bmatrix} = U^T \begin{bmatrix} C \\ D \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

In 'eigenvector' coordinates,

$$X' = \frac{1}{\sqrt{2}} \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \text{so } X = UX' = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{3e^t - e^{-t}}{2} \\ \frac{3e^t + e^{-t}}{2} \end{bmatrix}.$$

We can say something about the curves traced out from differing 'starting points.' They are easiest to describe in eigenvector coordinates.

$$(x', y') = (Ce^t, De^{-t})$$

so they satisfy the equation

$$x'y' = CD.$$

If $CD = 0$ this is a straight line, and if $CD \neq 0$ this is a hyperbola. (This is a hyperbola because the eigenvalues are ± 1. This is unusual; if the eigenvalues were 1 and -2, the solutions would not be hyperbolas.)

The general procedure to solve

$$\frac{dX}{dt} = AX$$

is to diagonalise A if possible. If this is possible, let S be a matrix with a basis of eigenvectors, so $A' = S^{-1}AS$ is diagonal, and change to eigenvector coordinates: $X' = S^{-1}X$. Then

$$\frac{dX'}{dt} = A'X'.$$

This is a system of homogeneous ODEs, with solutions

$$x'_j = C'_j e^{\lambda_j t},$$

where $X' = [x'_j]$ and C'_j are constants of integration.

Now this can be written as

$$X' = e^{A't}C'$$

so

$$X = Se^{A'}S^{-1}C,$$

where $C = SC'$. The exponential series is valid for diagonal matrices:

$$e^{A't} = I + A't + \frac{A'^2t^2}{2!} + \frac{A'^3t^3}{3!}\dots$$

so

$$\begin{aligned}\exp(S^{-1}AS) = I + (S^{-1}AtS) + \frac{(S^{-1}AtS)^2}{2!} + \frac{(S^{-1}AtS)^3}{3!}\dots \\ = S^{-1}(I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!}\dots)S \\ = S^{-1}e^{At}S.\end{aligned}$$

Conversely,

$$e^{At} = Se^{A't}S^{-1},$$

giving

(21.1) Theorem *The system of differential equations*

$$\frac{dX}{dt} = AX$$

has the (purely formal) family of solutions

$$X = e^{At}C.$$

This solution does not mention eigenvectors, but it's almost impossible to use without diagonalising A.

This is not acceptable as a solution because the only sensible way to calculate e^A is to diagonalise A.

Here is another example with complex eigenvalues.

$$\frac{dx}{dt} = x + y; \quad \frac{dy}{dt} = -x + y; \quad A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

We shall allow complex numbers in the solution.

The eigenvalues are $1 \pm i$ and here is a matrix S whose columns are eigenvectors

$$S = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix},$$

corresponding eigenvalues $1 \pm i$. We do not worry about orthonormal bases since complex numbers are involved (and the matrix A is not symmetric). By the adjoint matrix formula

$$S^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{i}{2} \\ \\ \frac{1}{2} & \frac{i}{2} \end{bmatrix}.$$

The general complex solution is

$$X = e^{At}C$$

where C is a complex vector. This is of no use in practice, but the general complex solution is easiest to describe in the eigenvector basis. It has the same form:

$$X' = e^{A't}C',$$

but

$$e^{A't} = \begin{bmatrix} e^{(1+i)t} & 0 \\ 0 & e^{(1-i)t} \end{bmatrix},$$

which allows the solution to be written simply as

$$x' = c'e^{(1+i)t}, \quad y' = d'e^{(1-i)t}.$$

This is a general *complex* solution. We can actually calculate e^{At}, which is real-valued.

$$e^{At} = Se^{A't}S^{-1} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \begin{bmatrix} e^{(1+i)t} & 0 \\ 0 & e^{(1-i)t} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{i}{2} \\ \\ \frac{1}{2} & \frac{i}{2} \end{bmatrix} =$$

$$e^t \begin{bmatrix} \frac{e^{it}+e^{-it}}{2} & \frac{e^{it}-e-it}{2i} \\ \\ -\frac{e^{it}-e^{-it}}{2i} & \frac{e^{it}+e^{-it}}{2} \end{bmatrix} = e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

The (real) solutions are

$$X = R(t) \begin{bmatrix} c \\ d \end{bmatrix}, \quad \text{where} \quad R(t) = e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

$R(t)$ is, or should be, e^{At}; c and d are (real) constants of integration. For each pair of (real) constants (c, d), the solution is a spiral passing through (c, d) at $t = 0$.

The derivation was complicated, so we should check the answer. We only need to check that the above formula $R(t)$ for e^{At} satisfies the system of ODEs:

$$\frac{dR(t)}{dt} = AR(t) \quad \text{and} \quad R(0) = I.$$

$$R(0) = e^0 \begin{bmatrix} \cos 0 & \sin 0 \\ -\sin 0 & \cos 0 \end{bmatrix} = I.$$

$$AR(t) = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} = e^t \begin{bmatrix} \cos t - \sin t & \cos t + \sin t \\ -\cos t - \sin t & \cos t - \sin t \end{bmatrix}.$$

To differentiate $R(t)$, we can use the formula

$$\frac{d}{dt}(uV) = u\frac{dV}{dt} + (\frac{du}{dt})V,$$

where u is a real-valued function and V is a matrix-valued function. The derivative of

$$e^t \begin{bmatrix} \cos t - \sin t & \cos t + \sin t \\ -\cos t - \sin t & \cos t - \sin t \end{bmatrix}$$

is therefore

$$e^t \begin{bmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{bmatrix} + e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

which is the same as $AR(t)$.

22 Higher order ODEs

(22.1) The D operator. Differentiation is a kind of linear map, and we can use the symbol D to denote this operation. This leads to some simple algebra. For example, if α and β are constants, then

$$(D - \alpha)(D - \beta) = D^2 - (\alpha + \beta)D + \alpha\beta,$$

in the sense that

$$(D-\alpha)(D-\beta)y = (D-\alpha)(\frac{dy}{dx} - \beta y) = \frac{d}{dx}\frac{dy}{dx} - \alpha\frac{dy}{dx} - \beta\frac{dy}{dx} + \alpha\beta y = D^2y - (\alpha+\beta)Dy + \alpha\beta y.$$

Obviously, the solution to

$$Du = v$$

is

$$u = \int v dx + \text{const.}$$

The following lemma is extremely useful.

(22.2) Lemma $D - \alpha = e^{\alpha x} D e^{-\alpha x}$.

Proof. For any function w of x,

$$\begin{aligned} e^{\alpha x} D e^{-\alpha x} w = \\ e^{\alpha x}((De^{-\alpha x})w + e^{-\alpha x} Dw) = \\ e^{\alpha x}((-\alpha e^{-\alpha x} w + e^{-\alpha x} Dw) = \\ -\alpha w + Dw = (D - \alpha)w. \end{aligned}$$

Q.E.D.

Example (which is familiar to us already). Solve $dy/dx - \alpha y = 0$.

$$(D - \alpha)y = 0; \quad D(e^{-\alpha x} y) = 0; \quad e^{-\alpha x} y = \text{constant},$$

since of course $Dw = 0$ if and only if w is constant. Hence $y = Ce^{\alpha x}$.

An equation of the form

$$\frac{d^2 y}{dx^2} + a\frac{dy}{dx} + by = g(x),$$

where a and b are constants and $g(x)$ is some function of x, is called a *linear second-order constant-coefficient ODE*.

If $g(x) \equiv 0$ then the ODE is *homogeneous*.

It can also be written in the form

$$(D^2 + aD + b)y = g(x).$$

To solve it, the *characteristic polynomial* of the above equation is $\lambda^2 + a\lambda + b$. Factorise it, putting it in the form $(\lambda - \alpha)(\lambda - \beta)$.

Then

$$\begin{aligned} (D^2 + aD + b)y = g(x) \iff \\ (D - \alpha)(D - \beta)y = g(x) \iff \\ e^{\alpha x} D e^{-\alpha x} e^{\beta x} D e^{-\beta x} = g(x). \end{aligned}$$

The general solution to linear constant-coefficient homogeneous second-order ODEs is as follows.

$$\begin{aligned} e^{\alpha x} D e^{(\beta - \alpha)x} D e^{-\beta x} y = 0 \\ e^{(\beta - \alpha)x} D e^{-\beta x} y = c \\ D e^{-\beta x} y = c e^{(\alpha - \beta)x} \end{aligned}$$

$$e^{-\beta x} y = \begin{cases} \frac{c}{\alpha - \beta} e^{(\alpha - \beta)x} + d & \text{if } \alpha \neq \beta \\ cx + d & \text{if } \alpha = \beta. \end{cases}$$

Hence the general solution is

$$y = \begin{cases} Ae^{\alpha x} + Be^{\beta x} & \text{if } \alpha \neq \beta \\ Ae^{\alpha x} + Bxe^{\alpha x} & \text{if } \alpha = \beta \end{cases}$$

Complex roots. This gives the general *complex* solution if α and β are conjugate complex (we would have to allow A and B to be complex numbers). In this case, take the real part of the general complex solution for the general real solution.

Examples.

$$(D^2 - D)y = 0$$

has the solutions

$$y = Ae^x + B$$

$$(D^2 + 2D + 10)y = 0$$

has the general complex solution

$$y = Ae^{-x}e^{3ix} + Be^{-x}e^{-3ix}.$$

Equivalently, the general complex solution is

$$y = Ae^{-x}\cos 3x + Be^{-x}\sin 3x$$

(A, B complex) and this gives the general real solution (A, B real).

$$(D^2 + 4D + 4)y = 0$$

has the general solution

$$y = Ae^{-2x} + Bxe^{-2x}.$$

Which α, which β? You get the correct answer, in whichever order you take the roots. In a non-homogeneous equation, if $e^{\alpha x}$ or $e^{\beta x}$ occurs on the right-hand-side, it helps to choose the order so it is $e^{\beta x}$ which occurs:

$$e^{\alpha x}De^{-\alpha x}e^{\beta x}De^{-\beta x}y = \ldots e^{\beta x}\ldots$$

This makes the integration problems as easy as possible.

Example: $D^2y - Dy = 1$.

$$\begin{aligned} D(D-1)y &= 1 \\ De^xDe^{-x}y &= 1 \\ e^xDe^{-x}y &= x \\ De^{-x}y &= xe^{-x} \\ &\text{use integration by parts}\ldots \\ e^{-x}y &= (-x-1)e^{-x} \\ y &= -x - 1 \text{ particular} \\ y &= A + Be^x - x - 1 \text{ general} \end{aligned}$$

$$
\begin{aligned}
(D-1)Dy &= 1 \\
e^x De^{-x} Dy &= 1 \\
De^{-x} Dy &= e^{-x} \\
e^{-x} Dy &= -e^{-x} \\
Dy &= -1 \\
y &= -x \text{ particular} \\
y &= A + Be^x - x \text{ general}
\end{aligned}
$$

The answers are the same (constants are different) but the second time there was no integration by parts.

Trigonometric functions on the RHS. If $\cos nx$ occurs in the right-hand side, replace it by e^{inx} and take the real part of the solution.

If $\sin nx$ occurs in the right-hand side, replace it by e^{inx} and take the imaginary part of the solution.

Example: $D^2y - Dy = e^{2x} \sin x$.

$$
\begin{aligned}
D(D-1)y &= e^{2x+ix} \\
De^x De^{-x} y &= e^{2x+ix} \\
e^x De^{-x} y &= \frac{e^{2x+ix}}{2+i} \\
De^{-x} y &= \frac{e^{ix+x}}{2+i} \\
e^{-x} y &= \frac{e^{ix+x}}{(2+i)(1+i)} \\
y &= \frac{e^{ix+2x}}{(2+i)(1+i)} \text{ particular complex solution}
\end{aligned}
$$

To get the imaginary part,

$$
\frac{e^{ix+2x}}{(2+i)(1+i)} = \frac{e^{ix+2x}}{1+3i} = e^{2x} \frac{(\cos x + i \sin x)(1-3i)}{1+9}.
$$

The imaginary part is

$$
\frac{e^{2x}(-3\cos x + \sin x)}{10}
$$

and to get the general real solution, add $A + Be^x$.

Example $(D^2 + 4D + 4)y = e^{-2x} \cos 2x$.

Take the real part of the solution to

$$
(D^2 + 4D + 4)y = e^{-2x+2ix}.
$$

Here the right-hand side involves one of the roots, which are $-2 \pm 2i$. Taking the preferred order, let $\alpha = -2 - 2i$ and $\beta = -2 + 2i$.

$$\begin{aligned} e^{\alpha x} D e^{-\alpha x} e^{\beta x} D e^{-\beta x} y &= e^{\beta x} \\ D e^{-\alpha x} e^{\beta x} D e^{-\beta x} y &= e^{\beta x - \alpha x} \\ e^{-\alpha x} e^{\beta x} D e^{-\beta x} y &= \frac{e^{\beta x - \alpha x}}{\beta - \alpha} \\ D e^{-\beta x} y &= \frac{1}{\beta - \alpha} \\ e^{-\beta x} y &= \frac{x}{\beta - \alpha} \\ y &= \frac{x e^{\beta x}}{\beta - \alpha} \end{aligned}$$

$$y = \frac{x e^{-2x}(\cos 2x + i \sin 2x)}{4i} \text{ particular complex}$$

This time the real part is easy to see:

$$\frac{x e^{-2x} \sin 2x}{4}$$

and the general real solution is

$$A e^{-2x} \cos 2x + B e^{-2x} \sin 2x + \frac{x e^{-2x} \sin 2x}{4}.$$

23 Higher order recurrences

A linear constant-coefficient second-order recurrence is an equation

$$y_{n+2} + a y_{n+1} + b y_n = g_n$$

where g_n is a known sequence. The same methods can be applied as with ODEs.

The (left) *shift operator* E shifts a sequence

$$y_0, y_1, y_2, \ldots$$

one place left:

$$y_1, y_2, \ldots$$

This is expressed by

$$(Ey)_n = y_{n+1}.$$

Then the recurrence can be given as

$$(E^2 + aE + b) y_n = g_n.$$

$E-1$ is the 'forward difference operator' similar in some ways to D. It is written as Δ, but we write everything in terms of the operator E. Remember that

$$\Delta y_n = (E-1)y_n = y_{n+1} - y_n = g_n$$

is a summation problem with solution

$$y_n = c + \sum_{r=0}^{n-1} g_r.$$

The recurrence

$$Ey_n = g_n$$

is easily solved and of no interest. We are interested in

$$(E-\alpha)y_n = g_n$$

where $\alpha \neq 0$.

(23.1) Lemma *If $\alpha \neq 0$ then*

$$\alpha^{n+1}(E-1)\alpha^{-n} = E - \alpha.$$

Proof.

$$\begin{aligned} \alpha^{n+1}(E-1)\alpha^{-n}y_n = \\ \alpha^{n+1}(\alpha^{-n-1}y_{n+1} - \alpha^{-n}y_n) = \\ y_{n+1} - \alpha y_n = (E-\alpha)y_n. \end{aligned}$$

Q.E.D.

This gives a method of solving (linear constant-coefficient) second-order recurrences. To solve

$$(E^2 + aE + b)y_n = g_n$$

- Factorise $\lambda^2 + a\lambda + b$ as $(\lambda - \alpha)(\lambda - \beta)$.
- The case where $b = 0$ so $\alpha = 0$ or $\beta = 0$ is a special case, easily solved and not of much interest.
- Assuming $b \neq 0$ so $\alpha, \beta \neq 0$, solve

$$\alpha^{n+1}(E-1)\alpha^{-n}\beta^{n+1}(E-1)\beta^{-n}y_n = g_n.$$

Homogeneous recurrences.

$$\begin{aligned} \alpha^{n+1}(E-1)\alpha^{-n}\beta^{n+1}(E-1)\beta^{-n}y_n &= 0 \\ (E-1)\alpha^{-n}\beta^{n+1}(E-1)\beta^{-n}y_n &= 0 \\ \alpha^{-n}\beta^{n+1}(E-1)\beta^{-n}y_n &= \text{const} \\ (E-1)\beta^{-n}y_n &= \text{const}\frac{(\alpha/\beta)^n}{\beta} \end{aligned}$$

$$\beta^{-n}y_n = \frac{\text{const}}{\beta}\sum_{r=0}^{n-1}(\alpha/\beta)^n = \begin{cases} A(\alpha/\beta)^n + B & \text{if } \alpha \neq \beta \\ Bn + A & \text{if } \alpha = \beta \end{cases}$$

so

$$y_n = \begin{cases} A\alpha^n + B\beta^n & \text{if } \alpha \neq \beta \\ A\alpha^n + Bn\alpha^n & \text{if } \alpha = \beta \end{cases}$$

Alternative forms can be given when α and β are conjugate complex.

For non-homogeneous recurrences, it is easiest to calculate a particular solution and add the general homogenous solution.

The most famous example is the Fibonacci numbers.

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$$

The recurrence is

$$(E^2 - E - 1)y_n = 0; y_0 = 0, y_1 = 1.$$

$$\alpha, \beta = \frac{1 \pm \sqrt{5}}{2}$$

$$y_n = A\left(\frac{1+\sqrt{5}}{2}\right)^n + B\left(\frac{1-\sqrt{5}}{2}\right)^n$$

$$y_0 = 0 : A + B = 0; B = -A$$

$$y_1 = 1 : A\left(\frac{1+\sqrt{5}}{2}\right) + B\left(\frac{1-\sqrt{5}}{2}\right) = 1$$

$$A(2\frac{\sqrt{5}}{2}) = 1$$

$$A = -B = \frac{1}{\sqrt{5}}$$

Actually, α is the well-known golden section, written $\phi = 1.62\ldots$ and $\beta = -1/\phi$, so

$$y_n = \frac{1}{\sqrt{5}}(\phi^n - (\frac{-1}{\phi})^n).$$

As n increases, the contribution from β^n decays to zero and y_{n+1}/y_n converges to the golden ratio.

Non-homogeneous recurrences. As with ODEs, if α^n occurs in the right-hand side, swap α and β — it reduces the work.

Example. $(E^2 - 3E + 2)y_n = 1$. The roots are 1 and 2, so we should take $\alpha = 2$ and $\beta = 1$.

$$(E-2)(E-1)y_n = 1$$

$$2^{n+1}(E-1)2^{-n}(E-1)y_n = 1$$

$$(E-1)2^{-n}(E-1)y_n = 2^{-n-1}$$

$$2^{-n}(E-1)y_n = \frac{1}{2}\sum_{r=0}^{n-1} 2^{-r} = 1 - \left(\frac{1}{2}\right)^n$$

(the 1 can be ignored)

$$(E-1)y_n = -1$$

$$y_n = -n \text{ particular}$$

$$y_n = -n + A2^n + B \text{ general}$$

If we take it in the other order

$$(E-1)(E-2)y_n = 1$$
$$(E-1)2^{n+1}(E-1)2^{-n}y_n = 1$$
$$2^{n+1}(E-1)2^{-n}y_n = n$$
$$(E-1)2^{-n}y_n = \frac{1}{4}n\left(\frac{1}{2}\right)^{n-1}$$
$$2^{-n}y_n = \frac{1}{4}\sum_{r=0}^{n-1} r\left(\frac{1}{2}\right)^{r-1}$$
$$2^{-n}y_n = \frac{1}{4}\frac{1-n(1/2)^{n-1}+(n-1)(1/2)^n}{(1-1/2)^2} = 1 - n\left(\frac{1}{2}\right)^{n-1} + (n-1)\left(\frac{1}{2}\right)^n$$

(the 1 can be ignored)

$$y_n = -2n + n - 1 = -n - 1 \quad \text{particular}$$
$$y_n = -n - 1 + A2^n + B$$

same as before except B becomes $B-1$.

Example where α^n and β^n don't occur in the RHS. At the same time, if $(E-1)$ is processed first then we get n^2 on the right-hand side which makes the second part harder.

$$(E^2 - 3E + 2)y_n = n$$
$$2^{n+1}(E-1)2^{-n}(E-1)y_n = n$$
$$(E-1)2^{-n}(E-1)y_n = \frac{1}{4}n\left(\frac{1}{2}\right)^{n-1}$$
$$2^{-n}(E-1)y_n = \frac{1}{4}\sum_{r=0}^{n-1} r\left(\frac{1}{2}\right)^{r-1} = 1 - n(1/2)^{n-1} + (n-1)(1/2)^n$$

ignore the 1

$$(E-1)y_n = (-n-1)$$
$$y_n = \sum_{r=0}^{n}(-r-1) = -\frac{n(n+1)}{2} \text{ particular}$$
$$y_n = A2^n + B - \frac{n(n+1)}{2} \text{ general}$$

Taken in the other order

$$(E-1)2^{n+1}(E-1)2^{-n}y_n = n$$
$$2^{n+1}(E-1)2^{-n}y_n = \frac{n(n-1)}{2}$$
$$(E-1)2^{-n}y_n = \frac{n(n-1)}{2}2^{-n-1}$$

We have no formula to help sum the right-hand-side, though of course one could be found.

Complex roots. If α and β are conjugate complex, then the answers can be expressed in terms of the real and complex parts of α^n and β^n. Sine and cosine functions can be used.

The right-hand side chosen makes the next example very laborious, far longer than would ever be asked in an exam.

$$(E^2 - 6E + 25)y_n = n.$$

The roots are $\alpha = 3 + 4i$ and $\beta = \overline{\alpha} = 3 - 4i$. If we get a particular complex solution, then its real part gives a particular real solution.

Our solution method amounts to solving the recurrence in two steps. In this example, the solution to the first step helps with the second.

$$\alpha^{n+1}(E-1)\alpha^{-n}\beta^{n+1}(E-1)\beta^{-n}y_n = n$$

So we first consider

$$\begin{aligned}
\alpha^{n+1}(E-1)\alpha^{-n}w_n &= n \\
(E-1)\alpha^{-n}w_n &= \frac{1}{\alpha^2}\frac{n}{\alpha^{n-1}} \\
\alpha^{-n}w_n &= \frac{1}{\alpha^2}\frac{1 - n(1/\alpha)^{n-1} + (n-1)(1/\alpha)^n}{(1-(1/\alpha))^2} \\
&\text{ignore the 1 in the numerator} \\
w_n = \frac{n-1-\alpha n}{(\alpha-1)^2} &= \frac{n(1-\alpha)-1}{(\alpha-1)^2} = \frac{n}{1-\alpha} - \frac{1}{(1-\alpha)^2}
\end{aligned}$$

Separately we consider ($\alpha \neq 0$)

$$\begin{aligned}
(E-\alpha)v_n &= 1 \\
\alpha^{n+1}(E-1)\alpha^{-n}v_n &= 1 \\
(E-1)\alpha^{-n}v_n &= \frac{1}{\alpha}\frac{1}{\alpha^n} \\
\alpha^{-n}v_n = \frac{1}{\alpha}\sum_{r=0}^{n-1}\frac{1}{\alpha^r} &= \frac{1}{\alpha}\frac{(1/\alpha)^n - 1}{(1/\alpha) - 1} \\
&\text{ignore the } -1 \text{ in the numerator} \\
v_n &= \frac{1}{1-\alpha}
\end{aligned}$$

v_n and w_n are particular solutions. Having these solutions available, we shall leave $(E-\alpha)(E-\beta)$ in that form.

$$(E-\alpha)(E-\beta)y_n = n$$

$$(E-\beta)y_n \frac{n}{1-\alpha} - \frac{1}{1-\alpha} \text{ (from above; using both formulae,)}$$

$$y_n = \frac{n}{(1-\alpha)(1-\beta)} - \frac{1}{(1-\alpha)(1-\beta)^2} - (\frac{1}{(1-\beta)(1-\alpha)^2} =$$

$$\frac{n}{(-2-4i)(-2+4i)} + \frac{1}{20}\left(-\frac{1}{-2-4i} - \frac{1}{-2+4i}\right) =$$

$$\frac{n}{20} + \frac{1}{100}.$$

This is a particular solution, and it is real-valued. The general solution is got by adding $A\alpha^n + B\beta^n$. But $\alpha = 5e^{i\gamma}$ and $\beta = 5e^{-i\gamma}$ where $\gamma = \tan^{-1}(4/3)$. The general real solution is

$$\frac{n}{20} + \frac{1}{100} + A5^n \cos(n\gamma) + B5^n \sin(n\gamma),$$

where A and B are real numbers.

24 Syllabus for theory question

This syllabus only refers to the theory question, question 4, on the paper. You should study the following definitions, lemmas, etcetera. You may need to state results given in other lemmas, but not repeat the lemmas.
Definition 11.5
Lemma 11.6
Lemma 11.7
Lemma 11.8
Corollary 11.11
Theorem 11.12
Lemma 11.13
Theorem 11.14
Corollary 11.17
Definition 11.23
Definition 11.24
Lemma 11.27
Corollary 11.28
Lemma 11.33
Corollary 11.35
Corollary 11.36
Theorem 11.38
Definition 11.41
Theorem 11.42