University of Bath
Department of Mechanical Engineering

ME10305 Mathematics 2

Dr D Andrew S Rees

Contents

1 ORDINARY DIFFERENTIAL EQUATIONS

Ordinary differential equations are equations which contain differentials, i.e. derivatives! I cannot imagine tht anyone would find that fact surprising. Therefore this section of the unit is devoted to solving equations which contain differentials. These equations are ordinary differential equations (ODEs), rather than partial differential equations (PDEs) which contain partial derivatives (see ME20021, Modelling Techniques 2).

Ordinary differential equations (ODEs) use dependent and independent variables. If y is a function of t, then y is called the **dependent** variable (since its value *depends* on the value of t) and t is the **independent** variable. The function is written in the form $y(t)$, or often just as y if the context implies that it is a function of t. Likewise, should y be a function of x, distance, then we could write it as $y(x)$, or often just as y if the context is clear.

1.1 Motivation

But before we begin all of this we need to answer the question:

What is the point of solving ODEs?

The briefest answer that I can think of is that virtually everything in science (i.e. anything within the known universe with the possible exception of the interior of a black hole) satisfies an equation or a set of equations of some sort. The topic of differential equations is an extremely vast and diverse part of science and engineering. The design of buildings, bridges, aircraft and power supplies, and the modelling and description of natural phenomena such as tides, cloud patterns, the buckling of rock layers, planetary orbits, our wretched UK weather, the aerodynamics of a cricket ball, and blood flow in the human body may all be modelled by differential equations.

One of the most straightforward examples is the equation describing the motion of a mass on a spring. If it is assumed that the force which is required to maintain an extension of length x of the spring from the equilibrium position is kx, then a mass at the end of such a spring which has that extension suffers a force of $-kx$. As Newton's law states that the rate of change of momentum is equal to the applied force, we immediately obtain the differential equation,

$$\underbrace{\frac{d}{dt}\left(m\frac{dx}{dt}\right)}_{\text{rate of change of momentum}} = \underbrace{-kx.}_{\text{restoring force}} \tag{1.2}$$

As the mass is usually constant for a mass/spring system, we obtain

$$m\frac{d^2x}{dt^2} = -kx. \tag{1.3}$$

In this (ODEs) part of the unit we will find out how to solve equations like this in order to determine exactly how this mass moves.

Note: that we will often abbreviate the second derivative in (1.3) as x'' for convenience.

1.2 Classification

Just as different types of algebraic equation (e.g. linear, quadratic, cubic, transcendental) have their different methods of solution, so do ordinary differential equations. Therefore it is necessary at the outset to classify the equation at hand in order to select which method is suitable for its solution. So our first chunk of technical stuff won't cover solutions at all.

We will classify ODEs in three different ways:

- IVP or BVP (i.e. Initial value problem or Boundary Value Problem),
- Linear or Nonlinear,
- The order of the equation.

Whilst we might have at best a diaphanous or wraith-like sense of what these mean, there remains the question,

What is the point of classifying ODEs?

The answer is that the methods which are needed to solve ODEs will depend very much on what the classification is. This is true not only for the analytical methods we shall be dealing with, but also for numerical/computational methods which will taught in ME20014 (Modelling Techniques 1) and ME20020 (Modelling Techniques 2) next year.

1.2.1 IVP or BVP?

An equation containing only a first derivative, such as

$$\frac{dy}{dt} + 3y = 0, \tag{1.4}$$

is incomplete because it requires an extra condition for it to be solved. Therefore a **boundary condition** such as,

$$y(0) = 1, \tag{1.5}$$

could be provided and then it is possible to provide a full solution, as we shall see later.

For first order equations like this, ones which have a first derivative, such a boundary condition is also called an **initial condition** and that the equation together with the initial condition is called an **Initial Value Problem** (IVP). These can also be called evolution equations because the solution then evolves in time from the initial condition.

That this is so may be made clearer (hopefully) by the following thought experiment. Suppose that we are at the given initial time (say $t = 0$) and y is equal to the given initial value (say $y = 1$). The substitution of these values of t and y into the ODE then gives us the value of y', which is the initial slope of y. If we now use that slope to predict what value y takes a very small time interval later, then that will give us a new y-value at the new time. So we have travelled a very small way along the tangent to the curve. Having the values of y and t at this slightly later time means that we can compute the new (but slightly different) value of y' and progress a little further. Thus we can repeat this procedure for as long as we wish, and it will trace out at least an approximation to the exact solution.

Actually, I have just described what is called Euler's method, a well-known numerical scheme for solving ODEs. It isn't particularly accurate — there are much much better ones available than this — but its accuracy

improves as the time steps are reduced in size. Nevertheless, this shows that the solution evolves from the **initial condition** and that the problem being solved (ODE plus initial condition) forms an IVP.

If instead we have the following ODE with a second derivative (which represents an undamped mass/spring system) together with two initial conditions,

$$\frac{d^2y}{dt^2} + 3y = 0 \qquad \text{subject to} \qquad y(0) = 1 \text{ and } y'(0) = 0, \tag{1.6}$$

then this too forms an IVP because both of the boundary conditions are given at the same point in time. An alternative thought experiment, a physical one in this instance, may be used to justify this classification as an IVP. The initial conditions correspond to a unit displacement and a zero velocity. These are precisely equivalent to what we might do in practice, namely to extend the spring to a given extension and to release it from rest at $t = 0$. Then what happens next is that the mass begins to move in the negative y-direction and the displacement from the initial state begins to decrease. So these two conditions are sufficient to determine the future evolution of the system. The system is an IVP.

In general, then, when all the Boundary Conditions are Initial Conditions (i.e. the values of y and a sufficient number of derivatives are given at the same point in time), then this will always be an IVP.

As an example of a BVP we may use the equation given in Eq. (1.6) but with the alternative boundary conditions,

$$y(0) = 1 \text{ and } y(1) = 0. \tag{1.7}$$

The most important feature of these conditions is that they are given at **two different points in time.** Formally this is called a **two-point boundary value problem** although it is exceptionally rare for BVPs to involve boundary conditions which involve three or more points. So one can get away with calling it a **boundary value problem**. Physically, we now have a mass/spring system which is subject to a unit initial displacement but with an unknown initial velocity. However, that velocity must be the correct one to yield a zero displacement when $t = 1$. Mathematically, this doesn't provide too much of a difficulty when compared with an equivalent IVP, but numerically it requires quite different techniques to find the solution: the initial velocity must be iterated upon in order to find the correct one which satisfies the condition at $t = 1$. IVPs do not need such iterations. So the choice of the method depends on the classification.

As a final example consider the following coupled pair of equations:

$$\frac{d^2x}{dt^2} + 2x - y = 0, \qquad \frac{d^2y}{dt^2} + 2y - x = 0. \tag{1.8}$$

These represent the motion of two masses in an undamped mass/spring system. If we had to solve this subject to the conditions:

$$x(0) = 1,\ x'(0) = 0,\ y(0) = 0,\ y'(0) = 1, \tag{1.9}$$

then the ODE and its boundary conditions form an IVP. In this case one mass (x) has a unit displacement with a zero velocity at $t = 0$, while the other has a zero displacement and a unit velocity which are also at $t = 0$.

On the other hand if we had to solve the same equation subject to the boundary conditions,

$$x(0) = 1,\ x'(0) = 0,\ y(1) = 0,\ y'(0) = 1, \tag{1.10}$$

or

$$x(0) = 1,\ x'(5) = 0,\ y(5) = 0,\ y'(0) = 5, \tag{1.11}$$

then both of these cases would form examples of BVPs.

1.2.2 Linearity and Nonlinearity

I offer the following definition of a linear equation which is precise, although it may need to be read quite a few times before it makes sense!

An equation or system of coupled equations is linear when all the dependent variables and their derivatives are multiplied either by constants or by functions of the independent variable, otherwise the equation or system is nonlinear.

If we confine ourselves to a single ODE for $y(t)$ with as many derivatives appearing as we wish, then the most complicated **linear equation** will take the form,

$$f_n(t)\frac{d^n y}{dt^n} + f_{n-1}(t)\frac{d^{n-1}y}{dt^{n-1}} + \cdots + f_2(t)\frac{d^2y}{dt^2} + f_1(t)\frac{dy}{dt} + f_0(t)y = F(t), \tag{1.12}$$

where $f_i(t)$ $(i = 0, 1, 2, \cdots, n)$ and $F(t)$ are given functions of t at worst. At the opposite extreme, the coefficients of y and its derivatives could be constants:

$$a_n\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_2\frac{d^2y}{dt^2} + a_1\frac{dy}{dt} + a_0 y = F(t), \tag{1.13}$$

and we will be solving equations like these later. The extension of this definition of linearity to systems of equations is straightforward: all of the ODEs making up the system must be linear. An example is Eq. (1.8).

Example 1.1: $\dfrac{dy}{dt} + ay = 0$ is linear. This is so because both y and y' are multiplied by constants.

Example 1.2: $\dfrac{d^2y}{dt^2} + ty = e^{-t}$ is is also linear. This is so because both y is multiplied by a function of t and y' is multiplied by a constant. The function of t on the right hand is irrelevant.

Example 1.3: $\dfrac{d^2A}{dx^2} + A - A^3 = 0$ is nonlinear because of the A^3, a power of the dependent variable.

Example 1.4: $\dfrac{d^3y}{dt^3} + \underbrace{y\dfrac{dy}{dt}}_{\text{nonlinear}} + \underbrace{t^5 y}_{\text{linear}} = 0$ is nonlinear. We have the product of two dependent variables.

Example 1.5: $\dfrac{d^2\theta}{dt^2} + \dfrac{g}{L}\sin\theta = 0$ is nonlinear because $\sin\theta \simeq \theta - \frac{1}{3!}\theta^3 + \cdots$, using Taylor's series.

Example 1.6: $y'' + ty - z = 0, \quad z''' + 2z - yz = 0$. The first equation is linear but the second is nonlinear due to the presence of yz, a product of dependent varables. Overall, this is a nonlinear system because the nonlinearity in the second equation effectively contaminates the full system.

Note: There is one big difference between linear and nonlinear systems. If there are two different solutions of a linear equation or system of equations, then the sum (or even a weighted sum) will also be a solution. However, this does not happen for nonlinear equations/systems. As an example, let both $y_1(t)$ and $y_2(t)$ satisfy the equation $y'' + Ky = 0$, and we shall test if $y_1 + y_2$ also satisfies it. So if we let $y = y_1 + y_2$ in the ODE, then

$$\begin{aligned} y'' + Ky &= (y_1'' + y_2'') + K(y_1 + y_2) \\ &= \cancel{(y_1'' + Ky_1)} + \cancel{(y_2'' + Ky_2)} \\ &= 0. \end{aligned} \tag{1.14}$$

If, on the other hand, we were to test this idea out with the equation, $y'' + y^2 = 0$, then we would obtain,

$$\begin{aligned} y'' + y^2 &= (y_1'' + y_2'') + (y_1 + y_2)^2 \\ &= \cancel{(y_1'' + y_1^2)} + \cancel{(y_2'' + y_2^2)} + 2y_1y_2 \\ &= 2y_1y_2 \\ &\neq 0 \quad \text{in general.} \end{aligned} \tag{1.15}$$

Therefore a sum of two solutions of a nonlinear ODE doesn't necessarily satisfy the same ODE.

Note: The fact that we can add two solutions of a linear ODE to obtain another solution will be used later when we solve Linear Constant-Coefficient ODEs.

1.2.3 The order of ODEs and systems of ODEs

The **order** of an equation is the order of the highest derivative appearing in that equation. Thus

$$\frac{dy}{dt} = -ay \quad \text{is of 1st order,} \tag{1.16}$$

while

$$\frac{d^3y}{dt^3} + \left(\frac{dy}{dt}\right)^{10} + y = 0 \qquad \text{is of 3rd order.} \tag{1.17}$$

In Eq. (1.17) one must not be put off by the 10th power appearing on the 1st derivative term; the 10th power does not affect the fact that the highest derivative is a 3rd derivative.

In many physical systems such as complex circuits or coupled mass/spring configurations more thn just than one differential equation is the rule, rather than the exception. For such systems **the order of the system is equal to the sum of the orders of the individual equations**. For example, the system,

$$\begin{aligned} m\frac{d^2x}{dt^2} + k(2x - y) &= 0 \\ m\frac{d^2y}{dt^2} + k(2y - x) &= 0, \end{aligned} \tag{1.18}$$

is of 4th order because it is composed of two 2nd order equations: $2 + 2 = 4$. If we have a system which is composed of one 1st order equation, two 3rd order equations and one 5th order equation, then the whole system is of 12th order simply because $1 + 3 + 3 + 5 = 12$.

This seems straightforward enough, but very occasionally there might be slight confusion. In the following system,

$$\begin{aligned} \frac{dx}{dt} + x - \frac{d^2y}{dt^2} &= t \\ \frac{d^3y}{dt^3} + 2y - x &= 0, \end{aligned} \tag{1.19}$$

the first equation (for x) is of 1st order while the second equation (for y) is of 3rd order. Hence the system is of 4th order. The presence of y'' in the first equation doesn't make that equation to be of 2nd order because this is the equation for x. The troublesome y'' is of lower order (a 2nd derivative) than the y''' (a third derivative) in the second equation. This illustrates the fact that, for physical systems, each equation in a system will be associated/identified with a unique dependent variable. However, I can assure you that it is very rare that such a potential conundrum will arise in real life.

1.2.4 Reduction to First order Form

The idea of *the order of a system of equations* may also be seen when such a system is reduced to what is called **first order form**.

For example, if we take Eq. (1.17) above, then we may transform it into first order form by first defining three new independent variables, y_1, y_2 and y_3, each of which are functions of t, according to

$$y_1 = y, \qquad y_2 = \frac{dy}{dt}, \qquad y_3 = \frac{d^2y}{dt^2}. \tag{1.20}$$

We now form 1st order equations for each of these three variables, either by using their definitions in Eq. (1.20) or by using the original Eq. (1.17) suitably translated into the new notation. Thus we obtain,

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 \\ \frac{dy_2}{dt} &= y_3 \\ \frac{dy_3}{dt} &= -y_1 - y_2^{10}, \end{aligned} \tag{1.21}$$

and therefore we have three 1st order equations replacing the original one 3rd order equation: $\mathbf{3 \times 1 = 1 \times 3}$. Note carefully that the first two of these equations come directly from the definitions in Eq. (1.20) while the third equation is essentially a translation of Eq. (1.17) into the new notation. So we have achieved the primary aim, namely first order equations for *each* of the three new dependent variables.

We may also reduce the system given in Eq. (1.18) to first order form. As both of these equations are of 2nd order, we need to use two new variables for each of the original dependent variables. We define

$$z_1 = x, \qquad z_2 = \frac{dx}{dt}, \qquad z_3 = y, \qquad z_4 = \frac{dy}{dt}, \tag{1.22}$$

and therefore the system (1.18) becomes,

$$\begin{aligned} \frac{dz_1}{dt} &= z_2 \\ m\frac{dz_2}{dt} &= -k(2z_1 - z_3) \\ \frac{dz_3}{dt} &= z_4 \\ m\frac{dz_4}{dt} &= -k(2z_3 - z_1). \end{aligned} \tag{1.23}$$

Thus a 4th order system which was composed of two 2nd order equations has been reduced to four 1st order equations: $\mathbf{2 \times 2 = 4 \times 1}$.

Note: This technique of reduction to first order form is especially useful when solving ODEs using numerical methods; this topic will be covered in ME20014 Modelling Techniques 1 next semester. The subscripts on z in Eq. (1.22), for example, represent array indices in Matlab or Fortran.

1.2.5 Full classifications and reduction to first order form

In this subsubsection we'll undertake a full classification on an ODE and on a system of ODEs and reduce each together with their boundary conditions to first order form.

Example 1.7: Consider the fifth order equation,

$$y'''' + t^2 y'''(1 - y') + yy'' = t, \tag{1.24}$$

subject to the boundary conditions,

$$y(0) = y'(0) = 0, \quad y''(0) = 1, \quad y'(1) = 1, \; y'''(1) = 0. \tag{1.25}$$

This may be seen to be a **5th order nonlinear BVP**. Given that it is of 5th order, we need to define five new variables to replace the original y. I like to tackle this simply by creating a Table of data, as follows.

Variable	Equation	BC($t = 0$)	BC($t = 1$)
$y_1 = y$	$y_1' = y_2$	$y_1(0) = 0$	•
$y_2 = y'$	$y_2' = y_3$	$y_2(0) = 0$	$y_2(1) = 1$
$y_3 = y''$	$y_3' = y_4$	$y_3(0) = 1$	•
$y_4 = y'''$	$y_4' = y_5$	•	$y_4(1) = 0$
$y_5 = y''''$	$y_5' = t - y_1 y_3 - t^2 y_4(1 - y_2)$	•	•

(1.26)

The bullet symbols have been used solely to indicate those initial and final conditions which haven't been specified. Thus we can see that there are no initial conditions for y_4 and y_5, and therefore these would need to be found so that the conditions for y_2 and y_4 at $t = 1$ are satisfied.

Example 1.8: Consider the following system of ODEs,

$$\begin{aligned} &\psi''' + 3\psi\psi'' - 2(\psi')^2 + \theta + N\phi = 0, \\ &\theta'' + 3a\,\psi\theta' = 0, \\ &\phi'' + 3b\,\psi\phi' = 0, \end{aligned} \tag{1.27}$$

where a, b and N are constants. These equations arise in the free convective boundary layer flow due to heat and solute being supplied into a fluid. The primes denote derivatives with respect to distance, x. The boundary conditions are that ψ and ψ' are zero at $x = 0$ while both θ and ϕ are equal to 1. As $x \to \infty$, ψ', θ and ϕ all tend to zero.

This is a seventh order system $(3 + 2 + 2)$ and so we need seven variables to replace the present dependent

variables, ψ, θ and ϕ. Three will be needed for ψ and two each for θ and ϕ. Therefore we get the following:

Variable	Equation	BC($x = 0$)	BC($x \to \infty$)
$v_1 = \psi$	$v_1' = v_2$	$v_1 = 0$	•
$v_2 = \psi'$	$v_2' = v_3$	$v_2 = 0$	$v_2 \to 0$
$v_3 = \psi''$	$v_3' = -3v_1v_3 + 2v_2^2 - v_4 - Nv_6$	•	•
$v_4 = \theta$	$v_4' = v_5$	$v_4 = 1$	$v_4 \to 0$
$v_5 = \theta'$	$v_5' = -3av_1v_5$	•	•
$v_6 = \phi$	$v_6' = v_7$	$v_6 = 1$	$v_6 \to 0$
$v_7 = \phi'$	$v_7' = -3bv_1v_7$	•	•

(1.28)

So this is a **7th order nonlinear boundary value problem**. The initial conditions for v_3, v_5 and v_7 are unknown and will need to be iterated upon in a numerical scheme to enable the the three boundary conditions as $x \to \infty$ to be satisfied — again, this numerical aspect is something that you won't need to worry about until next semester in ME20014, Modelling Techniques 1.

Finally we return the apparently awkward case given in (1.19) which is repeated here:

$$\frac{dx}{dt} + x - \frac{d^2y}{dt^2} = t$$
$$\frac{d^3y}{dt^3} + 2y - x = 0, \tag{1.29}$$

The reduction to first order form (without boundary conditions, for none were given) is as follows:

Variable	Equation
$v_1 = x$	$v_1' = t - v_1 + v_4$
$v_2 = y$	$v_2' = v_3$
$v_3 = y'$	$v_3' = v_4$
$v_4 = y''$	$v_4' = v_1 - 2v_2.$

(1.30)

While one might have thought that a serious problem will ensue by having a second y-derivative in the equations for x, this reduction to first order form has derivatives only on the left hand sides of the transformed equations.

1.3 A very simple 1st order ODE

The most straightforward case of an ODE is when

$$\frac{dy}{dt} = f(t) \tag{1.31}$$

and the solution is obtained by integrating both sides of the equation with respect to t to get

$$y = \int f(t)\, dt + c \tag{1.32}$$

where c is the constant of integration and the integral is an indefinite integral. The constant of integration is obtained by having an initial condition whereby the value of y at a given value of t is known. Therefore if we were solving

$$\frac{dy}{dt} = f(t) \qquad \text{subject to} \quad y(1) = A, \tag{1.33}$$

then the solution may be written in the form,

$$y = \int_1^t f(\alpha)\, d\alpha + A, \tag{1.34}$$

where α is a dummy variable of integration. The choice of the lower limit means that the value of the integral is zero when $t = 1$, and therefore the initial condition is seen to be satisfied.

Example 1.9: If we have the equation

$$\frac{dy}{dt} = t^2 \qquad \text{subject to } y(1) = 3, \tag{1.35}$$

then the solution is

$$y = \int_1^t \alpha^2\, d\alpha + 3 \quad = \quad \left[\tfrac{1}{3}\alpha^3\right]_1^t + 3 \quad = \quad \left[\tfrac{1}{3}t^3 - \tfrac{1}{3}(1)^3\right] + 3 \quad = \quad \tfrac{1}{3}t^3 + \tfrac{8}{3}. \tag{1.36}$$

An alternative (but more familiar) method is to use the indefinite integral, as in Eq. (1.32), and then to substitute the initial condition to obtain c. Thus

$$y = \tfrac{1}{3}t^3 + c \qquad \text{but} \qquad y(1) = 3 \implies 3 = \tfrac{1}{3}(1)^3 + c \qquad \implies \qquad c = \tfrac{8}{3} \qquad \Rightarrow \qquad y = \tfrac{1}{3}t^3 + \tfrac{8}{3}. \tag{1.37}$$

Generally people tend to use the latter method as it is less complicated.

Note: Really it's a bit of a cheat to call this an ODE when it is really an integration exercise in disguise.

1.4 Separation of variables

This technique applies to 1st order ODEs which take the form,

$$\frac{dy}{dt} = f(y)g(t), \tag{1.38}$$

where f and g are given functions. The general solution is obtained by **separating the variables,** i.e. by taking all y-dependent terms to one side and all t-dependent terms to the other, and then integrating. Thus the general solution follows in this manner:

$$\frac{dy}{dt} = f(y)g(t)$$

$$\implies \quad \frac{dy}{f(y)} = g(t)\, dt \tag{1.39}$$

$$\implies \quad \int \frac{dy}{f(y)} = \int g(t)\, dt.$$

This splitting of the dy/dt into a separate dy and dt is meant to be thought of in a "$\lim_{\delta t \to 0}$" sense, such as we saw many times in Maths 1.

So this method of solving a **variables-separable** equation reduces down to the finding of two integrals. The final solution is them obtained when an initial condition is applied. We shall apply this to a few examples.

Example 1.10: Solve the ODE, $\dfrac{dy}{dt} = 2\sqrt{y}\,\cos t$ subject to $y(0) = 0$.

Clearly this is a variables-separable ODE and therefore we may proceed as above:

$$\begin{aligned}
& \frac{dy}{dt} &&= 2\sqrt{y}\,\cos t && \\
\Longrightarrow \quad & \frac{dy}{\sqrt{y}} &&= 2\cos t\,dt && \\
\Longrightarrow \quad & \int \frac{dy}{\sqrt{y}} &&= \int 2\cos t\,dt && \\
\Longrightarrow \quad & 2\sqrt{y} &&= 2\sin t + c && c \text{ is arbitrary} \\
\Longrightarrow \quad & \sqrt{y} &&= \sin t + \tfrac{1}{2}c && \\
\Longrightarrow \quad & y &&= (\sin t + \tfrac{1}{2}c)^2 && \text{Now apply the Initial Condition} \ldots \\
\Longrightarrow \quad & y &&= \sin^2 t && y(0) = 0 \;\Rightarrow\; c = 0
\end{aligned} \tag{1.40}$$

Note: This is the typical way that a separation-of-variables problem proceeds. The final solution is as difficult to obtain as the integrals are which make it up. In the next two examples we will consider slightly different cases where the respective presence of a square root and of a logarithm will require some careful treatment.

Example 1.11: Solve the ODE, $y' = 3t^2/y$ subject to $y(1) = 2$.

We have

$$\begin{aligned}
\frac{dy}{dt} = \frac{3t^2}{y} \quad & \Longrightarrow \quad \int y\,dy = \int 3t^2\,dt \\
& \Longrightarrow \quad \tfrac{1}{2}y^2 = t^3 + c \\
& \Longrightarrow \quad y = \pm\sqrt{2t^3 + 2c}.
\end{aligned} \tag{1.41}$$

Application of the initial condition shows that $2 = \pm\sqrt{2 + 2c}$. The only way that this can be solved is for $c = 1$ to be chosen *and* the positive sign to be taken. The final solution is

$$y = \sqrt{2t^3 + 2}. \tag{1.42}$$

In our analysis we have evaluated the arbitrary constant right at the end. It is possible to do so at an earlier point. So if we have got as far as $\frac{1}{2}y^2 = t^3 + c$ and applied $y(1) = 2$, then we will obtain, $c = 1$, perhaps not surprisingly. Hence $y^2 = 2t^3 + 2$. Now we need to take the square root: $y = \pm\sqrt{2t^3 + 2}$, and then appeal a second time to the initial condition in order to confirm that we need the positive square root. This way of doing things does feel a little strange but it is entirely consistent with the first way.

Example 1.12: Solve the ODE, $y' = 2ty$, subject to $y(0) = -2$.

On separating the variables we get

$$\int \frac{dy}{y} = \int 2t\, dt, \tag{1.43}$$

and hence

$$\ln|y| = t^2 + c. \tag{1.44}$$

If we apply the Initial Condition at this point, then we obtain $c = \ln 2$. Setting $e^{\text{LHS}} = e^{\text{RHS}}$ gives

$$|y| = e^{t^2+\ln 2} \qquad \Longrightarrow \qquad |y| = 2e^{t^2} \qquad \Longrightarrow \qquad y = \pm 2e^{t^2}, \tag{1.45}$$

and again we need to invoke the Initial Condition a second time to confirm that we need the negative option. Hence the solution is $y = -2e^{t^2}$.

That was fine and in many ways the outcome (viz. the choosing of the correct sign) was determined in the same way as in Ex. 1.10. However, many people prefer to find y explicitly in terms of both t and the arbitrary constant before applying the initial condition, so lets do that. Equation (1.44) yields,

$$|y| = e^{t^2+c} = e^c e^{t^2}. \tag{1.46}$$

Now the application of $y(0) = -2$ gives $e^c = 2$, and hence $|y| = 2e^{t^2}$. Strictly, we can now remove the modulus signs and account for that with the introduction of a $\pm$ on the right hand side, and then we can choose the correct sign for the final answer, as before. But I have often seen the modulus signs just disappear when in the heat of the battle in the exam and then the initial condition gives $e^c = -2$, and thereby disaster follows!

My preferred safe route through a problem like this would be to use the following as the next line after Eq. (1.46):

$$y = e^{t^2+c} = Ae^{t^2} \tag{1.47}$$

where A is still arbitrary for now, but it does allow us to obtain the correct sign for y without any choice being made.

OK, this analysis has been far too wordy and disjointed, and therefore the best thing to do is for me to run this all again from scratch as a full analysis.

$$\begin{aligned}
& \frac{dy}{dt} && = 2ty \\
\Longrightarrow \quad & \int \frac{dy}{y} && = \int 2t\, dt \\
\Longrightarrow \quad & \ln|y| && = t^2 + c \\
\Longrightarrow \quad & |y| && = e^{t^2+c} \;=\; e^c\, e^{t^2} \\
\Longrightarrow \quad & y && = Ae^{t^2} && \text{replacement of the arbitrary constant} \\
\Longrightarrow \quad & y && = -2e^{t^2} && \text{using } A = -2 \text{ from the initial conditions.}
\end{aligned} \tag{1.48}$$

I hope this is also a safe route for you.

1.5 1st order linear equations

These equations fall into the general form,

$$\frac{dy}{dt} + P(t)y = Q(t), \tag{1.49}$$

where $P(t)$ and $Q(t)$ are given functions of t. There is a general method for solving these equations, but before it is presented let us consider the following example.

Example 1.13: Solve the equation,

$$\frac{dy}{dt} + \frac{2}{t}y = 5t^2. \tag{1.50}$$

Believe it or not, this equation is simplified slightly by multiplying both sides by t^2. On doing this we get

$$t^2\frac{dy}{dt} + 2ty = 5t^4. \tag{1.51}$$

Although this latest equation still doesn't look simple, the left hand side is the exact derivative of t^2y since $d(t^2y)/dt = t^2y' + 2ty$. Hence this equation may be rewritten in the form,

$$\frac{d}{dt}\left(t^2y\right) = 5t^4. \tag{1.52}$$

We may now integrate both sides to obtain

$$t^2y = t^5 + c \qquad \text{and hence} \qquad y = t^3 + ct^{-2} \tag{1.53}$$

is the solution, where c is an arbitrary constant.

Note: Once more, the ODE has been solved using an integration, but only after the ODE had been modified by multiplication by a function that I appeared to pluck out of thin air! So the above method is straightforward enough except for the reason why we chose to multiply throughout by t^2. Our progress was facilitated by choosing the correct function to make the left hand side equal to an exact differential. The question is: **How do we find that function?** The following subsection is a derivation of the formula for finding that function — this is for interest only, but it does explain why the weird formula works.

1.5.1 The Integrating Factor

If we return to the general equation given in Eq. (1.49), reproduced here:

$$\frac{dy}{dt} + P(t)y = Q(t),$$

then let the function we require be $F(t)$. After multiplication by F, we get,

$$Fy' + FPy = FQ. \tag{1.54}$$

Now we insist that the left hand side is an exact derivative. If it is, then it must be the derivative of Fy, given the presence of Fy' in Eq. (1.54). As the differential of Fy is $Fy' + F'y$, this means that that Eq. (1.54) must also be written precisely as,

$$Fy' + F'y = FQ. \tag{1.55}$$

So the terms in red in Eqs. (1.54) and (1.55) must be identical and hence we must have,

$$FP = F'. \tag{1.56}$$

Equation (1.56) may be rearranged to get

$$\frac{1}{F}\frac{dF}{dt} = P(t), \tag{1.57}$$

which is of variables-separable type although this may not be too obvious. Therefore

$$\int \frac{dF}{F} = \int P(t)\,dt \quad \Longrightarrow \quad \ln|F| = c + \int P(t)\,dt \quad \Longrightarrow \quad F = Ae^{\int P(t)\,dt}. \tag{1.58}$$

We have introduced the constant of integration, c, as usual, and then set $A = e^c$ as we did in Eq. (1.47).

Note: we always neglect to use A. This is because we shall be multiplying the original equation by F, and therefore A will multiply both sides of the resulting equation and may be cancelled. So in practice we always use the following formula for F, which is referred to as the **Integrating Factor**,

$$\boxed{F = e^{\int P(t)\,dt}.} \tag{1.59}$$

For the example ODE given above in Eq. (1.50), the Integrating Factor is

$$F = e^{\int (2/t)\,dt} = e^{2\ln t} = e^{\ln t^2} = t^2, \tag{1.60}$$

which is indeed the function by which we multiplied.

This method also yields a formula for the general solution,

$$y = \left[c + \int FQ\,dt\right]/F \qquad \text{where} \qquad F = e^{\int P(t)\,dt}, \tag{1.61}$$

but I would very definitely recommend remembering only the formula for F, and then proceeding as in the following examples.

Example 1.14: Solve the equation, $y' + \dfrac{3y}{t} = \dfrac{2}{t^2}$ subject to $y(1) = 2$.

The coefficient of y is $3/t$, and therefore the Integrating Factor is

$$e^{\int (3/t)\,dt} \;=\; e^{3\ln|t|} \;=\; e^{\ln|t^3|} = t^3. \tag{1.62}$$

Note: that this is one of the rare occasions when one doesn't need to worry about the modulus signs in a logarithm! So let us multiply the original ODE by t^3. We get

$$t^3y' + 3t^2y = 2t. \tag{1.63}$$

Our theory guarantees that the left hand side is an exact derivative of something. In this case it is the derivative of t^3y — check this using the product rule. So Eq. (1.63) becomes,

$$\begin{aligned} (t^3y)' &= 2t \\ \Longrightarrow \qquad t^3y &= t^2 + c \qquad \text{upon integrating} \\ \Longrightarrow \qquad y &= \frac{1}{t} + \frac{c}{t^3}. \end{aligned} \tag{1.64}$$

The application of the initial condition, $y(1) = 2$, yields $c = 1$ and therefore the final solution is,

$$y = \frac{1}{t} + \frac{1}{t^3}. \tag{1.65}$$

Example 1.15: Solve the equation, $y' + 2xy = 4xe^{-x^2}$, subject to $y(0) = 1$.

OK, the right hand side looks horrible, but let us just get on with the task of finding the Integrating Factor and leave any potential trouble until later. Also, note that this equation has x as the independent variable, not t, but this makes no difference at all to what we do. Given that the coefficient of y is $2x$ the Integrating Factor is,

$$e^{\int 2x\,dx} = e^{x^2}. \tag{1.66}$$

On multiplying the given ODE by the Integrating Factor we get,

$$e^{x^2}y' + 2xe^{x^2}y = 4x. \tag{1.67}$$

So the right hand side has simplified nicely but the left hand side has now become a mess. However, we are guaranteed that the left hand side is an exact derivative. So the rest of the analysis proceeds as follows.

$$\begin{aligned} e^{x^2}y' + 2xe^{x^2}y &= 4x \\ \implies \left(e^{x^2}y\right)' &= 4x \\ \implies e^{x^2}y &= 2x^2 + c \qquad \text{on integrating} \\ \implies y &= \left[2x^2 + c\right]e^{-x^2}. \end{aligned} \tag{1.68}$$

Application of the Initial Conditon, $y(0) = 1$, yields $c = 1$. Hence the required solution is,

$$y = \left[2x^2 + 1\right]e^{-x^2}. \tag{1.69}$$

Example 1.16: Solve the equation, $(\cot t)\,y' + y = \cot t \cos t$, subject to $y(0) = 1$.

The very very first thing to be done is to note that the coefficient of y' must be 1 so that we may apply the theory we derived earlier. Clearly we need to divide the full equation by $\cot t$. The equation becomes,

$$y' + (\tan t)\,y = \cos t. \tag{1.70}$$

The integrating factor is,

$$F = e^{\int \tan t\,dt} = e^{-\int(-\sin t/\cos t)\,dt} = e^{-\ln\cos t} = 1/\cos t. \tag{1.71}$$

Note how the integrand was manipulated to get it into an f'/f form, which necessitated the use of two minus signs. Multiplication by the Integrating Factor yields,

$$(\sec t)\,y' + (\tan t \sec t)\,y = 1. \tag{1.72}$$

As fearsome as this looks, again we can reduce our collective blood pressures by noting that the left hand side is, by design, an exact derivative. Therefore our analysis takes the following route:

$$\begin{aligned} (\sec t)\, y' + (\tan t \sec t)\, y &= 1 \\ \Longrightarrow \qquad \Big[(\sec t)\, y\Big]' &= 1 \\ \Longrightarrow \qquad (\sec t)\, y &= t + c \qquad\qquad \text{on integration} \\ \Longrightarrow \qquad y &= (t + c)\cos t. \end{aligned} \tag{1.73}$$

Application of the initial condition, $y(0) = 1$, gives $c = 1$ and hence the final solution is,

$$y = (t + 1)\cos t. \tag{1.74}$$

Example 1.17: Solve the equation, $(\cot t)\, y' - y = \cot t \cos t$, subject to $y(0) = 1$.

This is the same as the previous example except for the replacement of a plus by a minus. Clearly we need to attain a unit coefficient for the y' once more and therefore division by $\cot t$ yields,

$$y' - (\tan t)\, y = \cos t. \tag{1.75}$$

The integrating factor is,

$$F = e^{\int -\tan t\, dt} = e^{\int (-\sin t/\cos t)\, dt} = e^{\ln \cos t} = \cos t. \tag{1.76}$$

Multiplication by the Integrating Factor yields,

$$(\cos t)\, y' - (\sin t)\, y \;=\; \cos^2 t \;=\; \tfrac{1}{2}(1 + \cos 2t), \tag{1.77}$$

using a multiple angle formula. The rest of the analysis now follows:

$$\begin{aligned} (\cos t)\, y' - (\sin t)\, y &= \tfrac{1}{2}(1 + \cos 2t) \\ \Longrightarrow \quad \Big[(\cos t)\; y\Big]' &= \tfrac{1}{2}(1 + \cos 2t) \\ \Longrightarrow \quad (\cos t)\, y &= \tfrac{1}{2}t + \tfrac{1}{4}\sin 2t + c \qquad\qquad \text{upon integrating} \\ &= \tfrac{1}{2}(t + \sin t \cos t) + c \qquad\qquad \text{multiple angles again} \\ \Longrightarrow \quad y &= \tfrac{1}{2}(t \sec t + \sin t) + c \sec t. \end{aligned} \tag{1.78}$$

Now we apply the initial condition, $y(0) = 1$, to yield $c = 1$. Hence the final solution is,

$$y = \tfrac{1}{2}\Big[(t + 2)\sec t + \sin t\Big]. \tag{1.79}$$

So the change of one sign between Examples 1.16 and 1.17 results in two very different solutions.

1.6 Solution of homogeneous linear, constant coefficient ODEs

The most general n^{th} order ODE of this type may be written in the form

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots a_1 \frac{dy}{dt} + a_0 y = F(t), \tag{1.80}$$

where

- $F(t)$ is a given real function,
- all the a_i coefficients are real and
- a_n is nonzero (since otherwise it would not be an n^{th} order ODE!).

Given that $a_n \neq 0$, we could divide Eq. (1.80) by a_n to yield an n^{th} order ODE where $d^n y/dt^n$ has a unit coefficient.

Such equations (and indeed coupled systems of these equations) form the core of ODE theory and the modelling of physical systems. Mass/spring systems, electrical circuits, hydraulic circuits, vibrating structures of many different kinds including bridges, buildings and aircraft, may all be modelled by linear constant-coefficient ODEs. Therefore they assume a huge importance in Engineering. We will be revisiting ODE theory occasionally in later topics in this unit: Laplace Transforms, Fourier Series and Matrices. By the end of the unit you should have acquired a great facility in the various methods of their solution.

At the outset it should be noted that these ODEs may be split into two closely-related families depending on the function, $F(t)$, which is on the right hand side of Eq. (1.80). Here follows some terminology.

- When $F(t) = 0$ then Eq. (1.80) is called **homogeneous**.
- When the equation is homogeneous, nonzero solutions can only arise if at least one initial condition is nonzero.
- When $F(t) \neq 0$ then Eq. (1.80) is called **inhomogeneous**.
- The term, $F(t)$, is called the **forcing term** or the **inhomogeneous term**.
- When the equation is inhomogeneous, then the solutions will be nonzero even if all of the initial conditions are zero.

Some of these ideas will be unpacked later in the unit

My intention is to develop a unified approach to solving these equations through a sequence (not series, hah!) of examples with suitable comments to show how the exposition is evolving. We will begin with homogeneous ODEs and then extend our expertise to inhomogeneous ODEs afterwards.

Example 1.1: Solve the ODE $y' + ay = 0$.

Obviously $y = 0$ is a solution, but that's boring and there is a nonzero solution to be found.

Now this is the simplest possible 1st order ODE. It is linear and has no forcing term. Indeed, the theory from §1.5 could be used because it is of first order linear form. Actually, it is also of variables-separable form, and so the theory of §1.4 could also be applied. However, I wish to develop an approach which will work with all linear constant-coefficient equations whatever their order.

First, I'll rearrange the equation:

$$y' = -ay. \tag{1.81}$$

An equation is always the expression of a balance of some kind, but this equation also says that, whatever shape y has, then y' must have precisely the same shape. In other words, y and y' must be identical functions apart, perhaps, from their amplitudes. The only function which, when differentiated, yields exactly the same function is an exponential. Sines and cosines don't do this after one differentiation, and polynomials don't do it either. The question then is, which exponential? We may work it out by substituting $y = Ae^{\lambda t}$ into Eq. (1.81), where λ is to be found and where A is arbitrary. We get,

$$A\lambda e^{\lambda t} = -aAe^{\lambda t} \qquad \Longrightarrow \qquad A(\lambda + a)e^{\lambda t} = 0. \tag{1.82}$$

Given that the exponential cannot be zero and that $A = 0$ leaves us with only the boring zero solution, this means that,

$$\lambda + a = 0. \tag{1.83}$$

So $\lambda = -a$ and therefore the solution is,

$$y = Ae^{-at}, \tag{1.84}$$

where A is an arbitrary constant. If I had provided an initial condition then A could be found.

For example, if we had $y(0) = 2$ then (1.84) yields $A = 2$ and therefore the final solution is $y = 2e^{-at}$. Alternatively, if we were to have $y(1) = 2$ as the initial condition, then (1.84) yields $A = 2e^{a}$ and the final solution could be written in the form, $y = 2e^{-a(t-1)}$ — do check that out!

More generally, if $y(b) = c$ then $y = ce^{-a(t-b)}$.

As an exercise solve this example equation again, but using the separation-of-variables and first-order-linear methods in turn. This is worth doing just to see how these methods cope with the equation.

Example 1.2: Solve $y'' + 3y' + 2y = 0$.

Given our experience with Example 1.1 we shall use the same substitution. This is likely to work because *all* derivatives of $e^{\lambda t}$ are proportional to $e^{\lambda t}$. Therefore we obtain,

$$\begin{aligned}
& \lambda^2 e^{\lambda t} + 3\lambda e^{\lambda t} + 2e^{\lambda t} = 0 \\
\implies \quad & (\lambda^2 + 3\lambda + 2)e^{\lambda t} = 0 \\
\implies \quad & \underbrace{\lambda^2 + 3\lambda + 2 = 0}_{\text{Auxiliary Equation}} \qquad \text{since } e^{\lambda t} \text{ cannot be zero} \\
\implies \quad & (\lambda + 1)(\lambda + 2) = 0 \qquad \text{by factorisation} \\
\implies \quad & \lambda = -1, -2.
\end{aligned} \tag{1.85}$$

Note that the equation labelled, **Auxiliary Equation**, is also known as the **Indicial Equation** or even as the **Characteristic Equation**. All three terms are used by many people, so it's worth knowing their names (the terms, that is, not the people).

So we have two possible choices for λ, but which one should we choose? The answer is both, and we do so simultaneously. Therefore the general solution to the ODE is

$$y = Ae^{-t} + Be^{-2t}, \tag{1.86}$$

where A and B are arbitrary constants. If we wished to do so then we may show that Eq. (1.86) satisfies the original ODE by substitution into the ODE.

In practice we would be provided with two boundary/initial conditions to satisfy in order to find A and B. An example of an Initial Value Problem might be:

$$\begin{cases} y(0) = 0 \\ y'(0) = 1 \end{cases} \implies \begin{cases} A + B = 0 \\ -A - 2B = 1 \end{cases} \implies \begin{cases} A = 1 \\ B = -1 \end{cases} \implies y = e^{-t} - e^{-2t}. \tag{1.87}$$

An example of a Boundary Value Problem is

$$\begin{cases} y(0) = 1 \\ y(1) = 0 \end{cases} \implies \begin{cases} A + B = 1 \\ Ae^{-1} + Be^{-2} = 0 \end{cases} \implies \begin{cases} A = 1/(1-e) \\ B = -e/(1-e) \end{cases} \implies y = \frac{e^{-t} - e^{1-2t}}{1-e}. \tag{1.88}$$

The final solution did need a few lines of manipulation to get it into such a compact form.

Note that it is quite easy to write down the auxiliary equation because the coefficients of powers of λ correspond to the coefficients of the respective derivatives of y:

$$y'' + 3y' + 2y = 0 \quad \implies \quad \lambda^2 + 3\lambda + 2 = 0.$$

This also works the other way:

$$3\lambda^4 - 2\lambda^2 + \lambda + 6 = 0 \quad \implies \quad 3y'''' - 2y'' + y' + 6y = 0.$$

Therefore an homogeneous ODE has a unique auxiliary equation and vice versa; knowledge of one means that we have knowledge of the other.

Example 1.3: Solve the equation, $y''' - 2y'' - y' + 2y = 0$.

Proceeding a little more quickly, the substitution of $y = e^{\lambda t}$ yields the auxiliary equation,

$$\begin{aligned} \lambda^3 - 2\lambda^2 - \lambda + 2 &= (\lambda - 2)(\lambda^2 - 1) \\ &= (\lambda - 2)(\lambda + 1)(\lambda - 1) = 0, \end{aligned} \tag{1.89}$$

for which the roots are $\lambda = 2, 1, -1$. Hence the solution is

$$y = Ae^{2t} + Be^{t} + Ce^{-t},$$

where we have the three arbitrary constants, A, B and C.

Note: we could introduce further examples of this type which are of 4th order, 5th order and so on, and where the auxiliary equation has different roots, all real and all nonzero, but nothing new arises. So the above three examples illustrate the general/standard case. The rest of this section is devoted to exceptions to this general case. Clearly it will become increasingly difficult to find all the λ-values as the degree of the polynomial for λ increases. Given that an n^{th} order ODE yields an n^{th} order polynomial for λ and hence there will be n aribtrary constants in the geneal solution, the application of boundary conditions also becomes more difficult as the degree of the polynomial increases. For example, in the case of a 5th order ODE there will be five boundary conditions to satisfy and hence five algebraic equations to solve for the five unknown constants. Nasty.

Example 1.4: Solve the equation, $y'' + 2y' = 0$.

The auxiliary equation is $\lambda^2 + 2\lambda = 0$. So $\lambda(\lambda + 2) = 0$ which means that $\lambda = 0, -2$. Although one of these λ-values is zero, we may proceed as usual:

$$y \;=\; Ae^{0t} + Be^{-2t} \;=\; A + Be^{-2t}. \tag{1.90}$$

So the presence of the root, $\lambda = 0$, means that the corresponding solution is that y is equal to a constant.

Example 1.5: Solve the ODE, $y'' + 9y = 0$.

The auxiliary equation is $\lambda^2 + 9 = 0$ from which we obtain,

$$\lambda^2 = -9 \qquad \Longrightarrow \qquad \lambda = \pm 3j. \tag{1.91}$$

The auxiliary equation has purely imaginary roots. Surely this is a problem? No, it isn't, for we may again proceed as usual:

$$y = Ae^{3jt} + Be^{-3jt}. \tag{1.92}$$

Given our theory to date, this is the obvious way to write down the solution. However, we have real equations and we expect to have real solutions rather than complex ones. However, we have already met complex exponentials and so we may play around a little with these:

$$\begin{aligned} y &= Ae^{3jt} + Be^{-3jt} \\ &= A(\cos 3t + j \sin 3t) + B(\cos 3t - j \sin 3t) \\ &= (A + B)\cos 3t + (Aj - Bj)\sin 3t \\ &= C\cos 3t + D \sin 3t, \end{aligned} \tag{1.93}$$

where $C = A + B$ and $D = (A - B)j$. When first seen, the last line in Eq. (1.93) looks like sleight-of-hand. Indeed, it will be automatically assumed by almost everyone that A and B are real, possibly because they have

been so in all of the Examples before this one. It will also be assumed that C and D are real, but this is clearly inconsistent with the definitions of C and D in terms of A and B.

If A and B are both real, then C is real but D is purely imaginary. On the other hand, if A and B are complex conjugates of one another (which isn't crazy because the complex exponentials, e^{3jt} and e^{-3jt}, are complex conjugates of one another), then C and D are real. Specifically, we may take $A = C - Dj$ and $B = C + Dj$ and everything then the final solution in Eq. (1.93) is real.

Note 1: There are two main ways of writing down the solution for the present equation. They are

$$y = Ae^{3jt} + Be^{-3jt} \qquad \text{and} \qquad y = C\cos 3t + D\sin 3t.$$

The one which is chosen will almost always be the one involving sinusoids, although there are some circumstances when the one involving complex exponentials is better (see ME20021 Modelling Techniques 2).

Note 2: If we were to solve this equation with the *real* initial conditions, $y(0) = 1$ and $y'(0) = 0$, then the final solution will be real. This happens even if the complex exponential form of the general solution is used. Let us check this out. On taking the complex exponential form, $y = Ae^{3jt} + Be^{-3jt}$, we have

$$\begin{cases} y(0) = 1 \\ y'(0) = 0 \end{cases} \implies \begin{cases} A + B = 1 \\ 3j(A - B) = 0 \end{cases} \implies \begin{cases} A = \frac{1}{2} \\ B = \frac{1}{2} \end{cases} \implies y = \tfrac{1}{2}\left[e^{3jt} + e^{-3jt}\right] = \cos 3t. \tag{1.94}$$

On taking the sinusoidal form, $y = C\cos 3t + D\sin 3t$, we have

$$\begin{cases} y(0) = 1 \\ y'(0) = 0 \end{cases} \implies \begin{cases} C = 1 \\ 3D = 0 \end{cases} \implies y = \cos 3t. \tag{1.95}$$

So both forms of solution will yield the correct real answer, but the one involving sinusoids is a little quicker.

Example 1.6: Solve the ODE, $y'' + 4y' + 13y = 0$.

The auxiliary equation is,

$$\begin{array}{lll} & \lambda^2 + 4\lambda + 13 = 0 & \\ \implies & (\lambda + 2)^2 + 9 = 0 & \text{completing the square} \\ \implies & (\lambda + 2)^2 = -9 & \\ \implies & \lambda + 2 = \pm 3j & \\ \implies & \lambda = -2 \pm 3j. & \end{array} \tag{1.96}$$

We have a pair of values for λ which are a complex conjugate pair. These are fully complex, unlike those in Example 1.5 which were purely imaginary. Given that we have set $y = e^{\lambda t}$, the solution may be written as follows:

$$\begin{array}{lll} & y = Ae^{(-2+3j)t} + Be^{(-2-3j)t} & \\ \implies & y = e^{-2t}\left[Ae^{3jt} + Be^{-3jt}\right] & \\ \text{or} & y = e^{-2t}\left[C\cos 3t + D\sin 3t\right] & \text{c.f. Ex. 1.5.} \end{array} \tag{1.97}$$

Again, we have two forms of the solution, one involving complex exponentials and the other involving sinusoids. In general, should $\lambda = a \pm bj$ then the general solution would be,

$$y = e^{at}\left[Ae^{bjt} + Be^{-bjt}\right] \quad \text{or} \quad y = e^{at}\left[C\cos bt + D\sin bt\right]. \tag{1.98}$$

Solutions of this form are always associated with damped vibrating systems such as structures and stringed musical instruments. Hence a will always be negative for these applications.

Example 1.7: Solve the equation, $y'' + 4y' + 4y = 0$.

This is the first example of a really special case, one where the auxiliary equation has a repeated root. The auxiliary equation is

$$\begin{array}{ll} & \lambda^2 + 4\lambda + 4 = 0 \\ \Longrightarrow & (\lambda + 2)^2 = 0 \\ \Longrightarrow & \lambda = -2, -2. \end{array} \tag{1.99}$$

We have a repeated value for λ because the equation for λ has a repeated root. But how do we treat such cases?

Clearly we cannot use just $y = Ae^{-2t}$ because a second order ODE requires two boundary/initial conditions, but we have only one constant. So that guess is no good. However, it is clear that e^{-2t} must play some sort of role because $\lambda = -2, -2$. Perhaps the best thing would be to factor out the e^{-2t} dependence by means of the following substitution.

$$\begin{array}{lll} \text{Let} & y = e^{-2t}z(t) & \\ \Longrightarrow & y' = e^{-2t}[z' - 2z] & \text{(product rule and tidying up)} \\ \Longrightarrow & y'' = e^{-2t}[z'' - 4z' + 4z] & \text{(product rule and more tidying up).} \end{array} \tag{1.100}$$

Substitution of this into the ODE gives,

$$\underbrace{e^{-2t}[z'' - 4z' + 4z]}_{y''} + \underbrace{4e^{-2t}[z' - 2z]}_{4y'} + \underbrace{4e^{-2t}[z]}_{4y} = 0. \tag{1.101}$$

All but one of these terms then cancel leaving just,

$$e^{-2t}z'' = 0 \qquad \Longrightarrow \qquad z'' = 0. \tag{1.102}$$

Two successive integrations using the appropriate constants of integration yield $z = A + Bt$, from which we obtain the final solution,

$$y = (A + Bt)e^{-2t}. \tag{1.103}$$

Note 1: For future reference we are going to interpret the form of this solution in the following way. The constant term, A, corresponds to the first of the two repeated λ-values, while the Bt corresponds to the second λ-value. The extreme usefulness of this interpretation will be seen later in further examples.

Note 2: Although we have demonstrated what happens when $\lambda = -2, -2$, the same form of solution applies for other pairs of repeated λ-values. As a further example, the ODE $y'' + 10y' + 25y = 0$ has the auxiliary equation, $\lambda^2 + 10\lambda + 25 = 0$, for which $\lambda = -5, -5$ is the solution. Hence $y = (A + Bt)e^{-5t}$ is the solution of the ODE.

Note 3: In this example I used the substitution, $y = e^{-2t}z(t)$, solely to determine what solution corresponds to a repeated value of λ. I won't expect this to be done in an exam unless it is asked for specifically. Thus I would expect one to go immediately from the values of λ in Eq.(1.99) to the solution in Eq. (1.103).

Example 1.8: Solve the ODE, $y''' + 3y'' + 3y' + y = 0$.

The auxiliary equation for this ODE is,

$$\lambda^3 + 3\lambda^2 + 3\lambda + 1 = 0 \qquad \Longrightarrow \qquad (\lambda + 1)^3 = 0 \qquad \Longrightarrow \qquad \lambda = -1, -1, -1. \tag{1.104}$$

Thus the auxiliary equation has a triple root and we have three instances of $\lambda = -1$. The solution for y in this case is,

$$y = (A + Bt + Ct^2)e^{-t}, \tag{1.105}$$

where the A-term corresponds to the first instance of $\lambda = -1$, the Bt-term to the second instance and the Ct^2-term to the third instance.

Note 1: The correctness of Eq.(1.105) may be confirmed using the substitution, $y = e^{-t}z(t)$, which then yields the ODE, $z''' = 0$, for which its solution is $z = A + Bt + Ct^2$. Do check this.

Note 2: It may now be conjectured that an n-times repeated value of λ, will correspond to a polynomial of order $n - 1$, i.e. it has n terms. This is indeed correct. For example, should we have $(\lambda - 4)^6 = 0$ as the auxiliary equation, then the solution of equivalent ODE would be

$$(A + Bt + Ct^2 + Dt^3 + Et^4 + Ft^5)e^{4t}. \tag{1.106}$$

Example 1.9: Solve the ODE, $y'''' + 4y''' + 4y'' = 0$.

The auxiliary equation in this case is,

$$\lambda^4 + 4\lambda^3 + 4\lambda^2 = 0 \qquad \Longrightarrow \qquad \lambda^2(\lambda + 2)^2 = 0 \qquad \Longrightarrow \qquad \lambda = 0, 0, -2, -2. \tag{1.107}$$

This auxiliary equation has two pairs of repeated roots. The solution of the ODE is,

$$y = \underbrace{(A + Bt)e^{0t}}_{\lambda = 0, 0} + \underbrace{(C + Dt)e^{-2t}}_{\lambda = -2, -2}, \tag{1.108}$$

where the distinct values of λ, (i.e. 0 and -2) have been treated separately but each has been in the same way as before for twice-repeated roots. In addition, the $\lambda = 0$ components of the solution have been written in the form, e^{0t}, merely to show that $\lambda = 0$ has been used in the solution. Of course, it looks better to write Eq. (1.108) in the form,

$$y = A + Bt + (C + Dt)e^{-2t}. \tag{1.109}$$

Note 1: The main over-riding message from this section so far is that different values of λ (namely, zero, nonzero real and complex) may be treated in exactly the same way via exponentials. In the case of $\lambda = 0$ the corresponding function of t is a constant, and in the case of complex conjugate pairs, the complex exponentials may be replaced by a real exponential multiplied by the appropriate cosine and sine.

Note 2: A secondary but nevertheless important message is that any multiplicities in the values of λ transform into polynomial coefficients of the exponential, and that different values of λ act independently in this regard. Given that I am aware that this Note is quite tricky to understand, its message may be gleaned from the following crazy example.

Example 1.10: If the auxiliary equation for an ODE is $(\lambda+3)^4(\lambda+1)(\lambda+1+2j)^2(\lambda+1-2j)^2\lambda^3=0$, then what is the solution of the equivalent ODE?

The roots of the auxiliary equation are,

$$\lambda=-3,-3,-3,-3,\qquad -1,\qquad -1\pm 2j,-1\pm 2j,\qquad 0,0,0, \tag{1.110}$$

and therefore the solution of the equivalent 12^{th} ODE is,

$$y=\underbrace{(A+Bt+Ct^2+Dt^3)e^{-3t}}_{\lambda=-3,-3,-3,-3}+\underbrace{Ee^{-t}}_{\lambda=-1}+\underbrace{\big[(F+Gt)\cos 2t+(H+It)\sin 2t\big]e^{-t}}_{\lambda=-1\pm 2j,-1\pm 2j}+\underbrace{(J+Kt+Lt^2)}_{\lambda=0,0,0}. \tag{1.111}$$

For each different value of λ, namely -3, -1, $-1\pm 2j$ and 0, note how their multiplicities are reflected in the solution, especially for the complex conjugate pair.

Should you be interested the ODE is

$$y^{(12)}+17y^{(11)}+132y^{(10)}+628y^{(9)}+2026y^{(8)}+4750y^{(7)}+7860y^{(6)}+8964y^{(5)}+6345y''''+2025y'''=0, \tag{1.112}$$

and you may be relieved to know that I constructed the ODE from the factorised auxiliary equation, rather than the other way around!

1.6.1 A checklist of examples.

Some examples of how the solution of an ODE is related to the roots of the auxiliary equation and their multiplicity:

Roots	Solution
$2,\ -2$	$Ae^{2t}+Be^{-2t}$
$2,\ 3$	$Ae^{2t}+Be^{3t}$
$2,\ 3,\ 4,\ 5,\ 10$	$Ae^{2t}+Be^{3t}+Ce^{4t}+De^{5t}+Ee^{10t}$
$2,\ 2$	$(A+Bt)e^{2t}$
$2,\ 2,\ 4$	$(A+Bt)e^{2t}+Ce^{4t}$
$2,\ 2,\ 4,\ 4$	$(A+Bt)e^{2t}+(C+Dt)e^{4t}$
$2,\ 2,\ 2,\ 2,\ 4,\ 4,\ 5$	$(A+Bt+Ct^2+Dt^3)e^{2t}+(E+Ft)e^{4t}+Ge^{5t}$
$0,\ 2$	$A+Be^{2t}$
$0,\ 0,\ 0,\ 2$	$(A+Bt+Ct^2)+De^{2t}$
$\pm 2j$	$A\cos 2t+B\sin 2t$
$\pm 2j,\ \pm 5j,\ 2$	$A\cos 2t+B\sin 2t+C\cos 5t+D\sin 5t+Ee^{2t}$
$\pm 2j,\ \pm 2j$	$(A+Bt)\cos 2t+(C+Dt)\sin 2t$
$2\pm 3j$	$e^{2t}(A\cos 3t+B\sin 3t)$
$2\pm 3j,\ 2\pm 4j$	$e^{2t}(A\cos 3t+B\sin 3t)+e^{4t}(C\cos 5t+D\sin 5t)$
$2\pm 3j,\ 4\pm 5j$	$e^{2t}\big[(A+Bt)\cos 3t+(C+Dt)\sin 3t\big]$
$2\pm 3j,\ 2\pm 3j$	$e^{2t}\big[(A+Bt+Ct^2)\cos 3t+(D+Et+Ft^2)\sin 3t)\big]$
$2\pm 3j,\ 2\pm 3j,\ 2\pm 3j$	$e^{2t}\big[(A+Bt+Ct^2+Dt^3)\cos 3t+(E+Ft+Gt^2+Ht^3)\sin 3t)\big]$

1.6.2 A general note on the effect of the substitution $y = e^{ct}z(t)$

For background information only.

In Example 1.7, above, we considered the ODE, $y'' + 4y' + 4y = 0$, for which $\lambda = -2, -2$. Of course we now know how to write down the solution immediately; it is $y = (A + Bt)e^{-2t}$. This solution was found by using the sustitution, $y = e^{-2t}z(t)$, which yielded $z'' = 0$. The solution for z was found by integrating twice, but if we try to solve it using $z = e^{\sigma t}$ (note that I have used σ here, not λ), then the auxiliary equation for the z-ODE is now $\sigma^2 = 0$. The roots of this auxiliary equation are $\sigma = 0, 0$, and therefore the repeated $\sigma = 0$ yields the solution, $z = A + Bt$.

So we have $\lambda = -2, -2$ as the roots of the auxiliary equation for the y-ODE, and this transforms into $\sigma = 0, 0$ as the roots of the auxiliary equation for the z-ODE when using the substitution, $y = e^{-2t}z(t)$. So the act of stripping out a e^{-2t} factor changes the values of the roots of the auxiliary equation by 2. **This is no accident**, and the purpose of this subsection is to generalise this observation. First, a specific case and then we generalise.

Example 1.11: Solve the ODE, $y'' - 7y' + 12y = 0$. Then use the substitution, $y = e^t z(t)$, and solve the resulting ODE for z.

Letting $y = e^{\lambda t}$ in the ODE for y gives $\lambda^2 - 7\lambda + 12 = 0$. Factorisation gives $(\lambda - 3)(\lambda - 4) = 0$, and hence $\lambda = 3, 4$. The solution for y is, therefore,

$$y = Ae^{3t} + Be^{4t}. \tag{1.113}$$

The substitution, $y = e^t z(t)$, gives $z'' - 5z' + 6z = 0$ as the ODE for z. The auxiliary equation which follows the substitution $z = e^{\sigma t}$ is $\sigma^2 - 5\sigma + 6 = 0$. Hence $(\sigma - 2)(\sigma - 3) = 0$ which gives $\sigma = 2, 3$, and the solution for z is,

$$z = Ae^{2t} + Be^{3t}. \tag{1.114}$$

So we had $\lambda = 3, 4$ from the equation for y, while the substitution, $y = e^t z(t)$, means that $\sigma = 2, 3$ from the equation for z. So the act of factoring out e^t from y reduces the value of each of the roots of the auxiliary equation by 1.

Example 1.12: Solve the ODE, $y'' - (a + b)y' + aby = 0$. Then use the substitution, $y = e^{ct}z(t)$, and solve the resulting ODE for z.

I will merely summarize the results of the same process — it is worth checking this if you have time.

The auxiliary equation for y gives $\lambda = a, b$, and hence the solution,

$$y = Ae^{at} + Be^{bt}. \tag{1.115}$$

After substituting $y = e^{ct}z(t)$ into the ODE for y we obtain the ODE,

$$z'' - (a + b - 2c)z' + (a - c)(b - c)z = 0. \tag{1.116}$$

The substitution, $z = e^{\sigma t}$, yields the auxiliary equation the roots for which are, $\sigma = a - c, b - c$. Hence the solution for z is,

$$z = Ae^{(a-c)t} + Be^{(b-c)t}. \tag{1.117}$$

To summarise:

$$y = Ae^{at} + Be^{bt} \qquad \xrightarrow{y=e^{ct}z(t)} \qquad z = Ae^{(a-c)t} + Be^{(b-c)t}. \tag{1.118}$$

So the factoring out of e^{ct} decreases the roots of the auxiliary equation by c. This will also be true for a linear constant-coefficient ODE of any order.

1.7 Solution of inhomogeneous linear, constant coefficient ODEs

Now we turn to the solution of inhomogeneous equations such as Eq. (1.80) for which $F(t)$, the inhomogeneous or forcing term, is nonzero. This ODE is repeated here for convenience:

$$a_n\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots a_1\frac{dy}{dt} + a_0 y = F(t). \tag{1.119}$$

Analytical progress is assured for those cases where $F(t)$ takes exponential, sinusoidal or polynomial form. Other cases typically require numerical solution.

When solving inhomogeneous ODEs the resulting solution is composed of two parts, the **Complementary Function** and the **Particular Integral**. The Complementary Function (CF) is the full solution of the corresponding homogeneous equation while the Particular Integral (PI) is any solution of the full equation (but it is generally only that part which is intimately associated with the presence of $F(t)$). It is probably best to illustrate the roles of these two components of the full solution with a simple example.

Example 1.13: Solve the ODE, $y' + y = e^{2t}$.
This is solved in two parts:

- (i) Solve $y' + y = 0$. This yields the Complementary Function. This needs to be undertaken first — reasons to follow later.
- (ii) Solve the full equation, $y' + y = e^{2t}$, but focussing solely on the consequence of having a nonzero right hand side. This yields the Particular Integral.

The Complementary Function is the solution of $y' + y = 0$. This follows the ideas of §1.6, so the auxiliary equation is $\lambda + 1 = 0$. Therefore $\lambda = -1$ and hence,

$$y_{\mathrm{cf}} = Ae^{-t}. \tag{1.120}$$

For the Particular Integral we need to solve,

$$y' + y = e^{2t}, \tag{1.121}$$

but we have to concentrate on the contribution of e^{2t} to the solution. First we need to be reminded that, since differentials of e^{at} are proportional to e^{at}, then it makes some sense to let $y_{\mathrm{pi}} = Be^{2t}$ in order to find the value of B. Substitution into Eq. (1.121) yields,

$$\underbrace{2Be^{2t}}_{y'} + \underbrace{Be^{2t}}_{y} = \underbrace{e^{2t}}_{e^{2t}} \qquad \Longrightarrow \qquad 3B = 1 \qquad \Longrightarrow \qquad B = \tfrac{1}{3}. \tag{1.122}$$

So the Particular Integral is,

$$y_{\mathrm{pi}} = \tfrac{1}{3}e^{2t}. \tag{1.123}$$

Given that the type of ODEs which we are solving are linear, we may add together both the Complementary Function and the Particular Integral to obtain the General Solution:

$$y = y_{\mathrm{cf}} + y_{\mathrm{pi}} = Ae^{-t} + \tfrac{1}{3}e^{2t}. \tag{1.124}$$

Note: At this stage of the analysis the Complementary Function always has arbitrary constants as coefficients whereas the Particular Integral doesn't.

If, in addition, we had been given the initial condition, $y(0) = 1$, then it is straightforward to show that $A = \frac{2}{3}$. Hence the final solution would then be,

$$y = \tfrac{2}{3}e^{-t} + \tfrac{1}{3}e^{2t}. \tag{1.125}$$

Example 1.14: Solve the equation, $y'' + 3y' + 2y = e^{at}$. The value, a is an unspecified constant.

The Complementary Function is found by solving $y''+3y'+2y = 0$. The auxiliary equation is $\lambda^2+3\lambda+2 = 0$, and the roots of this are $\lambda = -1, -2$. Hence the CF is,

$$y_{\rm cf} = Ae^{-t} + Be^{-2t}. \tag{1.126}$$

The Particular Integral is found by solving the full ODE using the substitution, $y_{\rm pi} = Ce^{at}$. Hence,

$$Ce^{at}\Big[a^2 + 3a + 2\Big] = e^{at} \qquad \Longrightarrow \qquad C = \frac{1}{a^2+3a+2} = \frac{1}{(a+1)(a+2)}, \tag{1.127}$$

and therefore the PI is,

$$y_{\rm pi} = \frac{e^{at}}{(a+1)(a+2)} \tag{1.128}$$

The general solution is

$$y = y_{\rm cf} + y_{\rm pi} = Ae^{-t} + Be^{-2t} + \frac{e^{at}}{(a+1)(a+2)}. \tag{1.129}$$

Let us consider a few different values of a.

$$\begin{aligned}
a = -3 \quad &\Longrightarrow \quad y'' + 3y' + 2y = e^{-3t} \quad &&\Longrightarrow \quad y = Ae^{-t} + Be^{-2t} + \tfrac{1}{2}e^{-3t} \\
a = 10 \quad &\Longrightarrow \quad y'' + 3y' + 2y = e^{10t} \quad &&\Longrightarrow \quad y = Ae^{-t} + Be^{-2t} + \tfrac{1}{132}e^{10t} \\
a = 0 \quad &\Longrightarrow \quad y'' + 3y' + 2y = 1 \quad &&\Longrightarrow \quad y = Ae^{-t} + Be^{-2t} + \tfrac{1}{2} \\
a = -1.01 \quad &\Longrightarrow \quad y'' + 3y' + 2y = e^{-1.01\,t} \quad &&\Longrightarrow \quad y = Ae^{-t} + Be^{-2t} - \tfrac{10000}{99}e^{-1.01t} \\
a = -0.99 \quad &\Longrightarrow \quad y'' + 3y' + 2y = e^{-0.99t} \quad &&\Longrightarrow \quad y = Ae^{-t} + Be^{-2t} + \tfrac{10000}{101}e^{-0.99t}.
\end{aligned} \tag{1.130}$$

In these cases, and in almost every other case, the solution given in Eq. (1.129) works and is correct. However, there are difficulties when $a = -1$ or $a = -2$ for then the denominator of $y_{\rm pi}$ is zero and the coefficient of e^{at} becomes infinite. We may see the approach to the difficulties which arise when $a = -1$ by observing the numerical coefficient of the PIs in Eq. (1.130) when $a = -1.01$ and $a = -0.99$. These exceptional values of a are the roots of the auxiliary equation which we used to find $y_{\rm cf}$ — this is not an accident.

So when $a = -2$ then the forcing term is e^{-2t}, but this function is identical to one of the components of $y_{\rm cf}$ in (1.126). If we were to choose to think about all of this in terms of λ-values (as in $e^{\lambda t}$), then $y_{\rm cf}$ corresponds to $\lambda = -1, -2$ while the forcing term may be regarded as being *equivalent* to $\lambda = -2$, a second instance of this λ-value. When we encountered this type of situation in Example 1.7, a second instance of $\lambda = -2$

there was shown to lead to a solution of the form, te^{-2t}. The same is true here, but we'll consider a different example first to demonstrate this more convincingly before returning to the present example.

Example 1.15: Solve the equation, $y' + 2y = e^{-2t}$.

For this ODE the auxiliary equation for the Complementary Function is $\lambda + 2 = 0$ and hence $\lambda = -2$. The inhomogeneous term is e^{-2t} which may be said to be equivalent to a second instance of $\lambda = -2$. Therefore this ODE is a prototype of Example 1.14. We will solve this using two different methods.

Method 1: Being a first order linear equation we may solve this ODE using an Integrating Factor. This factor is $e^{\int 2\,dt}$ which is e^{2t}. So we shall multiply the ODE by e^{2t} and eventually obtain the final solution:

$$\begin{aligned}
& y' + 2y = e^{-2t} && \text{the original ODE} \\
\Longrightarrow \quad & e^{2t}\left[y' + 2y\right] = 1 && \text{multiplied by } e^{2t} \\
\Longrightarrow \quad & \left[e^{2t}\,y\right]' = 1 && \text{LHS is an exact derivative} \\
\Longrightarrow \quad & e^{2t}\,y = t + A && \text{integrating} \\
\Longrightarrow \quad & y = \underbrace{Ae^{-2t}}_{\text{CF}} + \underbrace{te^{-2t}}_{\text{PI}}
\end{aligned} \tag{1.131}$$

Although this method doesn't use the terms, Complementary Function and Particular Integral, I have indicated which terms are which in terms of the language of the present section. We see that the PI does turn out to be proportional to te^{-2t}, as we guessed it might be.

Given that all the available λ-values are equal to one another, we could also use the substitution, $y = e^{-2t}z(t)$, to obtain, $z' = 1$. This yields $z = t + A$, and hence we obtain the above solution for y.

Method 2: Now let us rerun this Example using a CF/PI approach. So we are solving,

$$\underbrace{y' + 2y}_{\lambda = -2} = \underbrace{e^{-2t}}_{\lambda = -2}\ , \tag{1.132}$$

where I have indicated the λ-values associated with the left hand side (i.e. the roots of the auxiliary equation) and the right hand side (i.e. the coefficient of t in the exponent).

Given that the CF has $\lambda = -2$ as the root of its auxiliary equation, we may state immediately that $y_{\mathbf{cf}} = Ae^{-2t}$. Given that the equivalent value of λ from the forcing term is $\lambda = -2$, a second instance, then we need to let $y_{\mathbf{pi}} = Bte^{-2t}$ and then find the value of B by substituting it into the full ODE:

$$\begin{aligned}
& \underbrace{Be^{-2t}(1 - 2t)}_{y_{\mathbf{pi}}'} + \underbrace{2Bte^{-2t}}_{2y_{\mathbf{pi}}} = e^{-2t} \\
\Longrightarrow \quad & Be^{-2t} = e^{-2t} && \text{all the terms involving } te^{-2t} \text{ cancel} \\
\Longrightarrow \quad & B = 1.
\end{aligned} \tag{1.133}$$

Hence $y_{\mathbf{pi}} = te^{-2t}$ and therefore we recover the solution given in Eq. (1.131).

Note: This latest example suggests that the repetition of a λ-value is treated in exactly the same way as in the last section, namely that increasing powers of t appear depending on how many repetitions there are. Clearly we haven't got a general proof of this, but the following few examples will show this in action.

Example 1.16: Solve the equation, $y'' + 3y' + 2y = e^{-2t}$. This is the $a = -2$ instance of Example 1.14.

Guided by Example 1.15, we'll write out the ODE and classify it according its λ-values. We have,

$$\underbrace{y'' + 3y' + 2y}_{\lambda = -1, -2} = \underbrace{e^{-2t}}_{\lambda = -2}, \tag{1.134}$$

The Complementary Function is $y_{\text{cf}} = Ae^{-t} + Be^{-2t}$, as found in Example 1.14. With regard to the Particular Integral, the substitution we need is now determined by the fact that the λ-value corresponding to the forcing term is the second instance of $\lambda = -2$, and therefore we have to set $y_{\text{pi}} = Cte^{-2t}$. We get,

$$\begin{aligned} & y'' + 3y' + 2y = e^{-2t} \\ \implies \quad & Ce^{-2t}\Big[\underbrace{-4 + 4t}_{y''} + \underbrace{3(1 - 2t)}_{3y'} + \underbrace{2t}_{2y} \Big] = e^{-2t} \\ \implies \quad & -Ce^{-2t} = e^{-2t} \quad \implies \quad C = -1. \end{aligned} \tag{1.135}$$

Hence $y_{\text{pi}} = -te^{-2t}$. The general solution is,

$$y = y_{\text{cf}} + y_{\text{pi}} = Ae^{-t} + Be^{-2t} - te^{-2t}. \tag{1.136}$$

If we had chosen to solve $y'' + 3y' + 2y = e^{-t}$, then the forcing term represents a second instance of $\lambda = -1$ and therefore we would need to use $y_{\text{pi}} = Cte^{-t}$. In this case it is worth checking that $C = 1$ is the correct value. The general solution in this case is,

$$y = y_{\text{cf}} + y_{\text{pi}} = Ae^{-t} + Be^{-2t} + te^{-t}. \tag{1.137}$$

As a further twist on this problem, if we had wished to solve $y'' + 3y' + 2y = ae^{-t} + be^{-2t}$ then each of the forcing terms should be considered separately, and then the general solution may be found to be,

$$y = y_{\text{cf}} + y_{\text{pi}} = Ae^{-t} + Be^{-2t} + ate^{-t} - bte^{-2t}. \tag{1.138}$$

Example 1.17: Solve the ODE, $y'' + 3y' + 2y = te^{-2t}$.

The left hand side of this ODE is the same as for Examples 1.14 and 1.16, and therefore the Complementary Function is the same. We note too that the λ-values forming the roots of the auxiliary equation are $\lambda = -1, -2$. But what should we make of the present forcing term?

In Example 1.7 we interpreted two instances of $\lambda = -2$ as being equivalent to e^{-2t} (for the first $\lambda = -2$) and te^{-2t} (for the second $\lambda = -2$). In the present Example we shall do the same, but the equivalence will be taken in the opposite direction. Therefore we shall adopt the point of view that the presence of te^{-2t} as the forcing term is equivalent to having $\lambda = -2, -2$. To illustrate this we may write the ODE with suitable labelling:

$$\underbrace{y'' + 3y' + 2y}_{\lambda = -1, -2} = \underbrace{te^{-2t}}_{\lambda = -2, -2} \quad . \tag{1.139}$$

As before the Complementary Function is given by $y_{\rm cf} = Ae^{-t} + Be^{-2t}$. Given that the forcing term is now to be regarded as the second and third instances of $\lambda = -2$, we need to set $y_{\rm pi} = (Ct + Dt^2)e^{-2t}$. My belief is that, if this is understood, then nothing else in this topic holds any fears apart from the length of the algebra.

I will omit much of the algebra for this example, but eventually we get to,

$$\begin{aligned} & y'' + 3y' + 2y \\ &= \left[\Big(3C + 2D - 4C\Big) + t\Big(2C + 6D - 6C - 4D + 4C - 4D\Big) + t^2\Big(2D - 6D + 4D\Big)\right] e^{-2t} \\ &= (2D - C - 2Dt)e^{-2t} \quad = \quad te^{-2t}. \end{aligned} \tag{1.140}$$

It is much to be recommended that the analysis leading to Eq. (1.140) is checked. So we see that all the t^2 terms cancel, and those involving C for the t-coefficents have also cancelled. This serves as a check that one's algebra is correct! This final right hand side should now be equal to the original forcing term, te^{-2t}. Hence $2D = -1$ (matching the coefficients of t) and $2D - C = 0$ (matching the constants). Hence $D = -\frac{1}{2}$ and $C = -1$. The Particular Integral is $y_{\rm pi} = (-t + \frac{1}{2}t^2)e^{-2t}$, and therefore the general solution is,

$$y = y_{\rm cf} + y_{\rm cf} = Ae^{-t} + Be^{-2t} + (-t - \tfrac{1}{2}t^2)e^{-2t}. \tag{1.141}$$

Example 1.18: Solve the ODE, $y''' + 5y'' + 8y' + 4y = te^{-2t}$.

I have contrived this example so that the auxiliary equation takes the form, $(\lambda + 1)(\lambda + 2)^2 = 0$, so that we have $\lambda = -1, -2, -2$. From this I reconstructed the left hand side of the ODE that is seen above, and then included the te^{-2t} on the right hand side. With labels, the ODE is,

$$\underbrace{y''' + 5y'' + 8y' + 4y}_{\lambda = -1, -2, -2} = \underbrace{te^{-2t}.}_{\lambda = -2, -2} \tag{1.142}$$

So we have two instances of $\lambda = -2$ in the auxiliary equation and the equivalent of another two from the forcing term. Hopefully it is now not too surprising that we shall take,

$$y_{\rm cf} = Ae^{-t} + (B + Ct)e^{-2t} \qquad \text{and} \qquad y_{\rm pi} = (Dt^2 + Et^3)e^{-2t}. \tag{1.143}$$

Here the values, A, B and C are arbitrary, while D and E may be found by substitution into the full ODE. These values turn out to be, $D = -\frac{1}{2}$ and $E = -\frac{1}{6}$, and therefore the full solution is,

$$y = y_{\text{cf}} + y_{\text{pi}} = Ae^{-t} + (B + Ct)e^{-2t} + (-\tfrac{1}{2}t^2 - \tfrac{1}{6}t^3)e^{-2t}. \tag{1.144}$$

The amount of algebra that is required to find the Particular Integral is quite large, but there is an easier route for this ODE. Given that there are so many instances of $\lambda = -2$, we may factor an e^{-2t} out using $y = e^{-2t}z(t)$ and this yields,

$$z''' - z'' = t. \tag{1.145}$$

Later we will find out how to deal with powers of t on the right hand side. The solution for z is quicker to obtain than the one for y. We will return to this ODE later as Example 1.26.

Example 1.19: Find the solutions of $y''' + 3y'' + 3y' + y = t^5e^{-t}$.

Yes, this is a seriously extreme example, but the CF is $y_{\text{cf}} = (A+Bt+Ct^2)e^{-t}$, given that $\lambda = -1, -1, -1$ from the auxiliary equation (see Example 1.8). The constants, A, B and C are arbitrary.

Given the t^5 multiplying the e^{-t} on the right hand side of the ODE, we have the equivalent of six further repetitions of $\lambda = -1$. So we may label the ODE as follows,

$$\underbrace{y''' + 3y'' + 3y' + y}_{\lambda = -1, -1, -1} \quad = \quad \underbrace{t^5e^{-t}}_{\lambda = -1 \text{ six times.}} \tag{1.146}$$

Note that e^{-t} is equivalent to one instance of $\lambda = -1$, te^{-t} to two, t^2e^{-t} to three and so on. Therefore we need to substitute

$$y_{\text{pi}} = (Dt^3 + Et^4 + Ft^5 + Gt^6 + Ht^7 + Jt^8)e^{-t} \tag{1.147}$$

in order to find the Particular Integral.

For this very extreme case, and upon noting that the only value that λ takes is -1, then we may solve the whole problem in one go by substituting $y = z(t)e^{-t}$. This shifts all the λ values from -1 to 0; see Example 1.12. The ODE transforms to

$$\begin{aligned} & z''' = t^5 \\ \Longrightarrow \quad & z = A + Bt + Ct^2 + \tfrac{1}{336}t^8 \qquad \text{using three integrations} \\ \Longrightarrow \quad & y = (A + Bt + Ct^2 + \tfrac{1}{336}t^8)e^{-t}. \end{aligned} \tag{1.148}$$

Therefore the constants introduced in Eq. (1.147) are

$$D = E = F = G = H = 0, \quad J = \tfrac{1}{336}. \tag{1.149}$$

A summary. OK, we need to pause briefly here to take stock of what has been achieved with all of these Examples. In §1.6 and in the present section so far I have attempted to introduce a unified way of determining from the ODE what forms are taken by the Complementary Function and by the Particular Integral. These are intimately associated with the λ-values. Hopefully, it has become clear that the form the Particular Integral takes depends on what has happened with the Complementary Function. This is why I have been focussed so strongly on the values of λ, as we have defined them here, and on their multiplicity. This is also why the Complementary Function *must* be found first.

As an attempt to describe this *unified way* one could say that repeated values of λ involve the use of increasing powers of t multiplying the associated function, $e^{\lambda t}$, the power increasing by 1 for every subsequent repetition. This happens for both the Complementary Function and the Particular Integral, but the counting must start with the auxiliary equation for the Complementary Function.

Here is a Table of examples of "what to do when" when faced with an ODE where all the λ-values are real quantities. I will take the value, $\lambda = 2$, to be the one which tends to be repeated, although the very last instance in the Table is slightly different. As always with such ODEs, the constants in the Complementary Function are arbitrary, whereas those of the Particular Integral need to be found.

ODE	λ(CF)	λ(PI)	CF	PI
$y' - 3y = e^{2t}$	3	2	Ae^{3t}	Be^{2t}
$y' - 2y = e^{3t}$	2	3	Ae^{2t}	Be^{3t}
$y' - 2y = e^{2t}$	2	2	Ae^{2t}	Bte^{2t}
$y' - 2y = te^{2t}$	2	$2, 2$	Ae^{2t}	$(Bt + Ct^2)e^{2t}$
$y' - 2y = t^2e^{2t}$	2	$2, 2, 2$	Ae^{2t}	$(Bt + Ct^2 + Dt^3)e^{2t}$
$y'' - 4y' + 3y = e^{2t}$	$1, 3$	2	$Ae^{t} + Be^{3t}$	Ce^{2t}
$y'' - 3y' + 2y = e^{3t}$	$1, 2$	3	$Ae^{t} + Be^{2t}$	Ce^{3t}
$y'' - 3y' + 2y = e^{2t}$	$1, 2$	2	$Ae^{t} + Be^{2t}$	Cte^{2t}
$y'' - 3y' + 2y = te^{2t}$	$1, 2$	$2, 2$	$Ae^{t} + Be^{2t}$	$(Ct + Dt^2)e^{2t}$
$y'' - 3y' + 2y = t^2e^{2t}$	$1, 2$	$2, 2$	$Ae^{t} + Be^{2t}$	$(Ct + Dt^2 + Et^3)e^{2t}$
$y'' - 4y' + 4y = e^{3t}$	$2, 2$	3	$(A + Bt)e^{2t}$	Ce^{3t}
$y'' - 4y' + 4y = e^{2t}$	$2, 2$	2	$(A + Bt)e^{2t}$	Ct^2e^{2t}
$y'' - 4y' + 4y = te^{2t}$	$2, 2$	$2, 2$	$(A + Bt)e^{2t}$	$(Ct^2 + Dt^3)e^{2t}$
$y'' - 4y' + 4y = t^2e^{2t}$	$2, 2$	$2, 2, 2$	$(A + Bt)e^{2t}$	$(Ct^2 + Dt^3 + Et^4)e^{2t}$
$y''' - 6y'' + 12y' - 8y = e^{3t}$	$2, 2, 2$	3	$(A + Bt + Ct^2)e^{2t}$	De^{3t}
$y''' - 6y'' + 12y' - 8y = e^{2t}$	$2, 2, 2$	2	$(A + Bt + Ct^2)e^{2t}$	Dt^3e^{2t}
$y''' - 6y'' + 12y' - 8y = te^{2t}$	$2, 2, 2$	$2, 2$	$(A + Bt + Ct^2)e^{2t}$	$(Dt^3 + Et^4)e^{2t}$
$y''' - 6y'' + 12y' - 8y = t^2e^{2t}$	$2, 2, 2$	$2, 2, 2$	$(A + Bt + Ct^2)e^{2t}$	$(Dt^3 + Et^4 + Ft^5)e^{2t}$
$y''' - 3y' - 2y = te^{2t}$	$-1, -1, 2$	$2, 2$	$(A + Bt)e^{-t} + Ce^{2t}$	$(Dt + Et^2)e^{2t}$

The next few Examples will consider how to deal with sinusoidal forcing terms.

Example 1.20: Solve the ODE, $y'' + 3y' + 2y = \cos bt$.

With regard to the Complementary Function things are straightforward. The auxiliary equation is $\lambda^2+3\lambda+2 = 0$ and hence $\lambda = -1, -2$. So the Complementary Function is $y_{\rm cf} = Ae^{-t} + Be^{-2t}$.

For the Particular Integral, one *could* argue as follows. If we were to set $y_{\rm pi}$ to be proportional to $\cos bt$, then the y' term in the ODE yields a sine and it means that our initial substitution is incorrect. So we cannot use just a cosine as the substitution. Therefore we need to use $y_{\rm pi} = C\cos bt + D\sin bt$ as the substitution. While this form for $y_{\rm pi}$ is indeed correct for this Example, in general we need to make sure that there are no repeated λ-values. We know from ME10304, Maths 1, that

$$\cos bt = \tfrac{1}{2}(e^{bjt} + e^{-bjt}). \tag{1.150}$$

Hence the right hand side of the ODE is equivalent to $\lambda = \pm bj$. Restating the ODE with labelling we have,

$$\underbrace{y'' + 3y' + 2y}_{\lambda = -1, -2} = \underbrace{\cos bt}_{\lambda = \pm bj}, \tag{1.151}$$

and hence there are no repeated factors.

We shall solve this ODE in two different ways. The first is to use the substitution, $y_{\rm pi} = C\cos bt + D\sin bt$, while the second will replace $\cos bt$ with e^{bjt} (which means that we have added an imaginary component to the forcing term) and then we take the real part of the resulting complex Particular Integral.

Method 1: The substitution of $y_{\rm pi} = C\cos bt + D\sin bt$ into Eq. (1.151) yields,

$$\underbrace{\Big[-b^2C\cos bt - b^2D\sin bt\Big]}_{y''} + 3\underbrace{\Big[-bC\sin bt + bD\cos bt\Big]}_{3y'} + 2\underbrace{\Big[C\cos bt + D\sin bt\Big]}_{2y} = \cos bt. \tag{1.152}$$

Now we collect like terms:

$$\begin{aligned} \cos bt \text{ terms:} \quad & (2-b^2)C + 3bD = 1, \\ \sin bt \text{ terms:} \quad & -3bC + (2-b^2)D = 0. \end{aligned} \tag{1.153}$$

This pair of simultaneous equations may be solved in the usual way to obtain,

$$C = \frac{2-b^2}{b^4+5b^2+4}, \qquad D = \frac{3b}{b^4+5b^2+4}, \tag{1.154}$$

and hence the PI is

$$y_{\rm pi} = \frac{(2-b^2)\cos bt + 3b\sin bt}{b^4+5b^2+4}. \tag{1.155}$$

An alternative way of solving Eqs. (1.153) is to recast them in matrix/vector form:

$$\begin{pmatrix} (2-b^2) & 3b \\ -3b & (2-b^2) \end{pmatrix} \begin{pmatrix} C \\ D \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{1.156}$$

Using the formula for the inverse matrix (see later in these notes) we have,

$$\begin{pmatrix} C \\ D \end{pmatrix} = \frac{1}{b^4+5b^2+4} \begin{pmatrix} (2-b^2) & -3b \\ 3b & (2-b^2) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{b^4+5b^2+4} \begin{pmatrix} (2-b^2) \\ 3b \end{pmatrix}, \tag{1.157}$$

which is identical to Eq. (1.154).

Note: For those who haven't studied matrices yet, this will all make sense by the end of the semester!

Method 2. This proceeds by replacing the forcing term, $\cos bt$, by e^{bjt}. So we have added the imaginary term, $j \sin bt$, to the right hand side. This subterfuge makes it much easier to find the Particular Integral, at least in terms of e^{bjt}, although we will then need to find the real part of this version of the Particular Integral to obtain the one that is needed. So we shall solve,

$$y'' + 3y' + 2y = e^{bjt} \tag{1.158}$$

using the substitution, $y_{\text{pi}} = Ce^{bjt}$. This yields,

$$Ce^{bjt}\left[-b^2 + 3bj + 2\right] = e^{bjt} \qquad \Longrightarrow \qquad C = \frac{1}{2 - b^2 + 3bj}. \tag{1.159}$$

Hence the Particular Integral is,

$$y_{\text{pi}} = \frac{e^{bjt}}{2 - b^2 + 3bj}. \tag{1.160}$$

The real part of this expression is the solution of $y'' + 3y' + 2y = \cos bt$, while the imaginary part is the solution of $y'' + 3y' + 2y = \sin bt$. We need the real part here, so

$$\begin{aligned}
\frac{e^{bjt}}{2 - b^2 + 3bj} &= \frac{\cos bt + j \sin bt}{2 - b^2 + 3bj} && \text{expanding the complex exponential} \\
&= \frac{(\cos bt + j \sin bt)(2 - b^2 - 3bj)}{(2 - b^2 + 3bj)(2 - b^2 - 3bj)} && \text{using complex conjugates} \\
&= \frac{(2 - b^2)\cos bt + 3b \sin bt + j(-3b \cos bt + (2 - b^2)\sin bt)}{(2 - b^2)^2 + 9b^2} && \text{multiplying out} \\
&= \left[\frac{(2 - b^2)\cos bt + 3b \sin bt}{b^4 + 5b^2 + 4}\right] + j\left[\frac{-3b \cos bt + (2 - b^2)\sin bt}{b^4 + 5b^2 + 4}\right].
\end{aligned} \tag{1.161}$$

The real part of this final expression is what we were aiming for, although the imaginary part is an added bonus, namely the solution corresponding to having $\sin bt$ as the forcing term.

Note: This rather lengthy Example shows that you have a choice of methods for solving equations with sinusoidal forcing terms. One requires the solution of simultaneous equations (a task that is made a little quicker by using the matrix/vector variant), while other dives off into complex numbers. It is worth practicing both ways a few times to see which you find to be quicker and more reliable.

Note: There are some interesting comments that may be made about the physical meaning of this example. First, this is an over-damped system because the Complementary Function consists solely of decaying exponentials. This behaviour is what happens with many types of door restraints, such as those in 4 East! So the amplitude would naturally rather than to decay in an oscillatory manner.

This damped system is being perturbed by $\cos bt$ and the PI is of most interest because it is what remains when the transient, the CF, has decayed. When the perturbations have a very low frequency then $b \ll 1$. We may therefore write the PI in the form,

$$y_{\text{pi}} = \frac{(2 - b^2)\cos bt + 3b \sin bt}{b^4 + 5b^2 + 4} \simeq \tfrac{1}{2}\cos bt,$$

where the greyed-out terms are almost negligible compared with the terms that have remained black. Physically, the velocity and acceleration are negligible and we obtain an in-phase response.

Very fast perturbations correspond to $b \gg 1$. Hence,

$$y_{\mathrm{pi}} = \frac{(2-b^2)\cos bt + 3b\sin bt}{b^4+5b^2+4} \simeq -\frac{\cos bt}{b^2}.$$

This response corresponds to y'' dominating the left hand side of the ODE (i.e. y and y' are negligible). We have obtained a very small amplitude and an out-of-phase response.

Both of these responses may be confirmed using a mass attached to the end of a long elastic string/band. When the string is moved up and down with a very low frequency then the mass follows passively, but when it is jiggled up and down at a high frequency then the mass hardly moves at all but the movement that it does make is out of phase with your hand.

The next three examples will bring us slowly to a point where we will be able to deal with repeated complex values of λ in the context of inhomogeneous ODEs.

Example 1.21: Solve the ODE, $y'' + 9y = \cos bt$, where b is an unspecified constant.

This equation represents the effect of a periodic forcing on an undamped mass/spring system. The auxiliary equation for the Complementary Function is $\lambda^2 + 9 = 0$ and hence $\lambda = \pm 3j$. The ODE, with labelling, is,

$$\underbrace{y'' + 9y}_{\lambda = \pm 3j} = \underbrace{\cos bt}_{\lambda = \pm bj}, \tag{1.162}$$

and therefore we do not have any repeated values of λ unless $b = 3$. In this example we shall assume, therefore, that $b \neq 3$, and then we'll treat the special case, $b = 3$, in the next example.

Given that $\lambda = \pm 3j$, the Complementary Function is

$$y_{\mathrm{cf}} = A\cos 3t + B\sin 3t. \tag{1.163}$$

The arguments used in Example 1.20 about how to choose the form of the Particular Integral also apply here, and therefore we could let $y_{\mathrm{pi}} = C\cos bt + D\sin bt$. In the present case, the ODE doesn't have a first derivative, and therefore we could let $y_{\mathrm{pi}} = C\cos bt$ without any problem. This is because the equations for C and D decouple. Nevertheless, we shall use the full substitution in order to illustrate these comments. Equation (1.162) yields,

$$(-b^2+9)C\cos bt + (-b^2+9)D\sin bt = \cos bt, \tag{1.164}$$

and therefore we obtain,

$$\begin{array}{ll} \cos bt \text{ terms:} & (-b^2+9)C = 1 \\ \sin bt \text{ terms:} & (-b^2+9)D = 0. \end{array} \tag{1.165}$$

So the equations for C and D have decoupled, and therefore $C = 1/(9-b^2)$ and $D = 0$. The Particular Integral is,

$$y_{\mathrm{pi}} = \frac{\cos bt}{9-b^2}, \tag{1.166}$$

and the general solution is,

$$y = y_{\mathrm{cf}} + y_{\mathrm{pi}} = A\cos 3t + B\sin 3t + \frac{\cos bt}{9 - b^2}. \tag{1.167}$$

This solution is valid when $b \neq 3$ but the amplitude of the Particular Integral becomes infinite as $b \to 3$. An alternative solution is required when $b = 3$, and this is considered in the next Example.

Example 1.22: Solve the ODE, $y'' + 9y = \cos 3t$.

The auxiliary equation has roots, $\lambda = \pm 3j$, and the forcing term may be regarded as being equivalent to $\lambda = \pm 3j$. Therefore both $\lambda = 3j$ and $\lambda = -3j$ are repeated. Knowing how repeated values of λ are dealt with when λ is real, this means that we could write,

$$y_{\mathrm{cf}} = Ae^{3jt} + Be^{-3jt}, \qquad y_{\mathrm{pi}} = t\Big[Ce^{3jt} + De^{-3jt}\Big]. \tag{1.168}$$

Using and extending the results of Example 1.5 means that we may rewrite this solution in the form,

$$y_{\mathrm{cf}} = A\cos 3t + B\sin 3t, \qquad y_{\mathrm{pi}} = t\Big[C\cos 3t + D\sin 3t\Big], \tag{1.169}$$

where the values of A, B, C and D in Eq. (1.168) are not the same as the ones in Eq. (1.169). The general solution is

$$y = y_{\mathrm{cf}} + y_{\mathrm{pi}} = A\cos 3t + B\sin 3t + t\Big[C\cos 3t + D\sin 3t\Big], \tag{1.170}$$

and substitution of this into the ODE yields $C = 0$ and $D = \frac{1}{6}$. Hence the solution is,

$$y = y_{\mathrm{cf}} + y_{\mathrm{pi}} = A\cos 3t + B\sin 3t + \tfrac{1}{6}t\sin 3t. \tag{1.171}$$

Note: that the Particular Integral here has an amplitude which grows linearly in time. This is quite typical of undamped systems which are perturbed at one of its resonant frequencies. Generally, structures tend to be at least lightly damped and therefore this growth eventually attentuates leaving behind what could still be a rather large response but at least it doesn't continue to grow unboundedly.

Example 1.23: Solve the ODE, $y'' + 6y' + 25y = te^{-3t}\cos 4t$.

This is a rather strange example due to the form of the forcing term, but first we need to consider the roots of the auxiliary equation before making sense of that forcing term. The auxiliary equation is,

$$\lambda^2 + 6\lambda + 25 = 0 \quad \Longrightarrow \quad (\lambda + 3)^2 + 16 = 0 \quad \Longrightarrow \quad \lambda = -3 \pm 4j, \tag{1.172}$$

and hence

$$y_{\mathrm{cf}} = e^{-3t}\Big[A\cos 4t + B\sin 4t\Big] \tag{1.173}$$

is the Complementary Function. Now we see that the forcing term in the ODE is equivalent to a second and a third instance of $\lambda = -3 \pm 4j$. Repeating the ODE with labels, we have,

$$\underbrace{y'' + 6y' + 25y}_{\lambda = -3 \pm 4j} \quad = \quad \underbrace{te^{-3t}\cos 4t.}_{\lambda = -3 \pm 4j, -3 \pm 4j} \tag{1.174}$$

So the Particular Integral takes the form,

$$y_{\mathrm{pi}} = e^{-3t}\Big[(Ct + Dt^2)\cos 4t + (Et + Ft^2)\sin 4t\Big]. \tag{1.175}$$

I have to admit that I haven't determined what values C, D, E and F take, but it is possible to find them if you have a day or two free.

Finally, we turn to having polynomials as forcing functions. Although we haven't yet considered these explicitly, the process of determining the Particular Integral is no different from when we have nonzero real values of λ. All one has to do is to keep in mind that there is a ghostly e^{0t} present even if it is not written down. I will offer three examples of this.

Example 1.24: Solve $y' + 3y = 6$.

If we label this ODE with the appropriate values of λ, then we have

$$\underbrace{y' + 3y}_{\lambda = -3} \quad = \quad \underbrace{6}_{\lambda = 0} \quad . \tag{1.176}$$

These values of λ are different and hence we may write $y_{\rm cf} = Ae^{-3t}$ and, to find the Particular Integral, we let $y_{\rm pi} = Be^{0t} = B$. Substitution into the ODE yields $B = 2$, and hence the general solution is,

$$y = y_{\rm cf} + y_{\rm pi} = Ae^{-3t} + 2. \tag{1.177}$$

A simple initial condition such as, $y(0) = 1$, yields $A = -1$, and hence the final solution is $y = 2 - e^{-3t}$.

Note that we have already seen the solution of a second order ODE with a constant forcing term in the $a = 0$ cases in Eq. (1.130).

Example 1.25: Solve $y' + 3y = t^2$.

In this example the forcing term is equivalent to three instances of $\lambda = 0$, i.e. we may label the equation as follows:

$$\underbrace{y' + 3y}_{\lambda = -3} \quad = \quad \underbrace{t^2}_{\lambda = 0, 0, 0} \quad . \tag{1.178}$$

The Complementary Function is the same as for Example 1.24, but in view of the triple instance of $\lambda = 0$, we need to let $y_{\rm pi} = B + Ct + Dt^2$. After substitution, we obtain, $D = \frac{1}{3}$, $C = -\frac{2}{9}$ and $B = \frac{2}{27}$, in turn. Hence the general solution is,

$$y = y_{\rm cf} + y_{\rm pi} = Ae^{-3t} + \tfrac{2}{27} - \tfrac{2}{9}t + \tfrac{1}{3}t^2. \tag{1.179}$$

Example 1.26: Solve $z''' - z'' = t$.

This was the equation we obtained at the end of Example 1.18 and now we have the tools to solve it. Let us rewrite the ODE with labelling:

$$\underbrace{z''' - z''}_{\lambda = 1, 0, 0} \quad = \quad \underbrace{t}_{\lambda = 0, 0} \quad , \tag{1.180}$$

and hence we will write,

$$z = z_{\rm cf} + z_{\rm pi} = \underbrace{Ae^t + B + Ct}_{\text{CF}} + \underbrace{Dt^2 + Et^3}_{\text{PI}}. \tag{1.181}$$

Substitution into the ODE yields, $D = -\frac{1}{2}$ and $E = -\frac{1}{6}$, and hence the general solution is,

$$z = z_{\rm cf} + z_{\rm pi} = Ae^t + B + Ct - \tfrac{1}{2}t^2 - \tfrac{1}{6}t^3. \tag{1.182}$$

This final solution is consistent with the one obtained in Example 1.18.

Final remarks

Throughout the whole of these two sections on homogeneous and inhomogeneous ODEs I have striven to emphasize a common approach to finding the forms for both the Complementary Function and the Particular Integral. This approach involves the identification of all of the $\boldsymbol{\lambda}$-values which are associated with the CF and the PI and also their multiplicity. Imaginary or complex values are usually interpreted in terms of the appropriate sines and cosines, while zero values involve powers of $\boldsymbol{t}$ beginning with the zeroth power, i.e. 1, but otherwise everything works within this uniform approach.

Therefore I hope that it is possible in all situations to identify both $\boldsymbol{y_{\mathbf{cf}}}$ and $\boldsymbol{y_{\mathbf{pi}}}$, with the latter being decided upon *after* the former. Thereafter it is a matter of determining the unknown coefficients which are associated with the PI, and this algebraic tedium usually takes more than half of the time needed for the full solution. Should initial and/or boundary conditions be given, then yet more tedium is involved in computing the formerly arbitrary constants in the Complementary Function.

Post Script

Those who have studied ODEs before will notice the omission of Bernoulli's equation and of equidimensional equations. These appear in the problem sheets.

A large omission here is the solution of systems of linear constant-coefficient ODEs. Well, the ODE notes so far cover five lectures' worth of content and I think we need to give it a rest for a moment. Although we will touch briefly on the solution of ODEs and systems of ODEs in the Laplace Transform section, we'll do a much more general analysis in the Matrices section in a few weeks' time using eigenvalues and eigenvectors.

2 LAPLACE TRANSFORMS

2.1 Motivation

This is the first topic this academic year where I am pretty sure that only a small handful at most will have met it before and so some motivation is needed. So what are Laplace Transforms used for? Here's a short list:

(i) to solve linear constant-coefficient ODEs;

(ii) to "translate" electrical/hydraulic circuits into a form which is the *equivalent* of an ODE or system of ODEs;

(iii) to solve feedback sytems;

(iv) to solve some of the Partial Differential Equations which arise in fluid machanics and heat transfer.

In this unit we will definitely be covering (i) because of the present emphasis on ODEs, and we will also be developing the background theory behind (ii).

Items (ii) and (iii) form part of ME20013 Systems and Control next year.

Item (iv) isn't covered using Laplace Transforms, but the solution of PDEs will be via Fourier Transforms in ME20021 Modelling Techniques 2.

The chief objectives in these four lectures are to acquire some facility in the use the Laplace Transform itself, to acquire some of the terminology that is used in Control Theory (not that we will be doing Control Theory) and to use Laplace Transforms to solve some ODEs. You'll also meet some weird animals in the Laplace Transform zoo, namely the unit impulse and the unit step function.

2.2 What is the Laplace Transform?

The Laplace Transform is an integral which changes a function of time, $y(t)$, say, into $Y(s)$, a function of s. The variable, s, is known as the **Laplace Transform variable.** It may be described informally as being equivalent to a time derivative with a hint of initial condition! Sounds mad, but you'll see why later.

If one has a constant-coefficient ODE for $y(t)$, then the first step in its solution via Laplace Transforms is to **take the Laplace Transform** of that equation. This process results in an algebraic system for $Y(s)$, the transform of the dependent variable, which is then solved. Finally the process of taking the Inverse Laplace Transform takes place; this results in a function of time which, somewhat magically, is the desired solution of the original equation. Diagrammatically, we could summarise this process as follows:

$$\text{ODE for } y(t) \longrightarrow \text{algebraic equation for } Y(s) \longrightarrow \text{solve for } Y(s) \longrightarrow \text{solution for } y(t).$$

The final step, namely the recovery of $y(t)$ from $Y(s)$, is known as **taking the Inverse Laplace Transform**.

If you were studying this topic in a Mathematics Department, then taking of the inverse Laplace Transform would require the use of a contour integral in the complex plane called the Bromwich Contour Integral. This certainly does need to be done for quite complicated types of $Y(s)$, but these are ones which we won't meet, fortunately. Instead, we will be creating a toolchest of results that may be used to invert the sorts of $Y(s)$ which arise when solving linear constant-coefficient ODEs. So no worries then.

The Laplace Transform of the function $y(t)$ is defined in the following way,

$$\boxed{\mathcal{L}[y(t)] \quad = \quad \int_0^\infty y(t)e^{-st}\,dt \quad = \quad Y(s).} \tag{2.2}$$

Note that the symbol, $\mathcal{L}[\ ]$, merely means **'Laplace Transform of'** whatever happens to be in the square bracket, and this is how it is stated when talking about it. Thus the Laplace Transform process changes the function $y(t)$ into a new function $Y(s)$.

Once one has $Y(s)$, the corresponding $y(t)$ is found using,

$$\mathcal{L}^{-1}[Y(s)] \quad = \quad y(t). \tag{2.3}$$

Here the symbol, $\mathcal{L}^{-1}[\ ]$, means the **'Inverse Laplace Transform of'** whatever happens to be in the square bracket. At this stage I will say no more, but this concept is much easier than you might fear it will be.

2.3 Some examples of Laplace Transforms

Example 2.1: Find $\mathcal{L}[1]$.

This may be done simply by writing down the definition of the Laplace Transform and performing the integration:

$$\mathcal{L}[1] = \int_0^\infty 1\,e^{-st}\,dt = \Big[-\frac{e^{-st}}{s}\Big]_0^\infty = \frac{1}{s}. \tag{2.4}$$

In this integration it is important to note that the answer is correct only when $s > 0$. If s had been negative then the integral would be infinite. Thus $s > 0$ is an existence condition for the transform. All Laplace Transforms must have a range of values of s for which the integral exists. That being said, it is not often that one needs to consider this aspect of Laplace Transforms for the functions which we will be transforming.

Example 2.2: Find $\mathcal{L}[e^{-at}]$.

Again, by definition, we have

$$\mathcal{L}[e^{-at}] = \int_0^\infty e^{-at}\,e^{-st}\,dt = \int_0^\infty e^{-(s+a)t}\,dt = \frac{1}{s+a}. \tag{2.5}$$

This Laplace Transform exists when $s + a > 0$, i.e. when $s > -a$. When $a = 0$ we recover the result of Example 2.1. Other examples are when $a = 1$ and $a = -2$ we get

$$\mathcal{L}[e^{-t}] = \frac{1}{s+1} \qquad \text{and} \qquad \mathcal{L}[e^{2t}] = \frac{1}{s-2}. \tag{2.6}$$

Hopefully it is really easy to see that $\mathcal{L}[Ae^{-at}] = A/(s+a)$ so that multiplication of a function by a constant is equivalent to multiplication of its Laplace Transform by the same constant. This will be used freely with no further comment.

Example 2.3: Find $\mathcal{L}\big[\cos at\big]$.

By definition we have,

$$\mathcal{L}\big[\cos at\big] = \int_0^\infty \cos at\, e^{-st}\, dt = \frac{s}{s^2+a^2}. \tag{2.7}$$

This integral could have been performed using either integration by parts in the usual way, or by changing $\cos at$ into e^{ajt} and then taking the real part of the final answer. In addition we have,

$$\mathcal{L}\big[\sin at\big] = \frac{a}{s^2+a^2}. \tag{2.8}$$

The easiest proof of the results in (2.7) and (2.8) follows:

$$\begin{aligned}
&\mathcal{L}\big[\cos at + j\sin at\big] \\
&\quad = \mathcal{L}\big[e^{ajt}\big] \\
&\quad = \int_0^\infty e^{ajt}\, e^{-st}\, dt \\
&\quad = \int_0^\infty e^{-(s-aj)t}\, dt && \text{combining the exponentials} \\
&\quad && \text{(careful with the } aj \text{ term)} \qquad (2.9)\\
&\quad = \frac{1}{s-aj} && \text{yes, this works for a complex exponent} \\
&\quad = \frac{s+aj}{s^2+a^2} && \text{using the complex conjugate} \\
&\quad = \left(\frac{s}{s^2+a^2}\right) + j\left(\frac{a}{s^2+a^2}\right)
\end{aligned}$$

As can be seen, the **real** part and the **imaginary** part have been colour-coded.

Example 2.4: Find $\mathcal{L}\big[t\big]$.

Using one integration by parts we obtain,

$$\mathcal{L}\big[t\big] = \int_0^\infty t\, e^{-st}\, dt = \frac{1}{s^2}. \tag{2.10}$$

This is worth checking. The Laplace Transforms for higher powers of t are given by,

$$\begin{aligned}
\mathcal{L}\big[t^2\big] &= \int_0^\infty t^2\, e^{-st}\, dt = \frac{2}{s^3}, \\
\mathcal{L}\big[t^3\big] &= \int_0^\infty t^3\, e^{-st}\, dt = \frac{6}{s^4} = \frac{3!}{s^4}, \qquad (2.11)\\
\mathcal{L}\big[t^n\big] &= \int_0^\infty t^n\, e^{-st}\, dt = \frac{n!}{s^{n+1}}.
\end{aligned}$$

All of these were covered in ME10304 Mathematics 1.

Example 2.5: Find $\mathcal{L}\left[te^{-at}\right]$.

Application of the definition of the Laplace Transform yields,

$$\mathcal{L}\left[te^{-at}\right] = \int_0^\infty t\, e^{-at}\, e^{-st}\, dt = \int_0^\infty t\, e^{-(s+a)t}\, dt = \frac{1}{(s+a)^2}. \tag{2.12}$$

Again, this integral may be found using integration by parts. However, it is interesting (and quite consequential) to note that the role which is played by $\boxed{s}$ in the integration in Eq. (2.10) is identical to the role played by $\boxed{s+a}$ in Eq. (2.12): both s and $s+a$ are constants.

This is a foretaste of what we be calling the s-shift theorem a little later.

2.4 Laplace Transforms of derivatives

Although we have derived many useful Laplace Transform results, we aren't yet in a position to solve some linear constant-coefficient ODEs. Therefore we need to find the Laplace Transforms of some derivatives in order to enable us to do this.

Example 2.6: Find $\mathcal{L}\left[y'(t)\right]$.

We start with the statement that $\mathcal{L}\left[y(t)\right] = Y(s)$. Using the definition of the Laplace Transform we have,

$$\begin{aligned}
\mathcal{L}\left[y'\right] &= \int_0^\infty \underbrace{y'}_{I}\ \underbrace{e^{-st}}_{D}\, dt \\
&= \underbrace{\left[y\right]}_{I_1}\ \underbrace{\left[e^{-st}\right]_0^\infty}_{D_0} - \int_0^\infty \underbrace{\left[y\right]}_{I_1}\ \underbrace{\left[-se^{-st}\right]}_{D_1} dt \qquad \text{one integration by parts} \\
&= -y(0) + s\int_0^\infty y\, e^{-st}\, dt \qquad \text{where } ye^{-st} \to 0 \text{ as } t \to \infty \\
&= sY - y(0).
\end{aligned} \tag{2.13}$$

In other words, $\mathcal{L}\left[y\right] = Y \implies \mathcal{L}\left[y'\right] = sY - y(0)$.

Note 1: We have assumed that $ye^{-st} \to 0$ as $t \to \infty$ in the above analysis. Even if y were a growing exponential, then there is always a value of s above which this limit is satisfied.

Note 2: If one were trying to work out if there is a physical meaning for s, then one might say that multiplication by s is essentially equivalent to a time derivative but there is an additional contribution from the initial condition for y at $t = 0$ — the hint of initial condition! This fits well with solving ODEs that are Initial Value Problems.

With two and three integrations by parts, one may also find that,

$$\mathcal{L}\left[y''\right] = s^2Y - y'(0) - sy(0), \tag{2.14}$$

and

$$\mathcal{L}\left[y'''\right] = s^3Y - y''(0) - sy'(0) - s^2y(0), \tag{2.15}$$

and so on. I will leave these as exercises for you to do, but these results now make it very clear that Laplace Transforms will be well-suited to solve Initial Value Problems where all the initial conditions are at $t = 0$.

An alternative way of deriving Eq. (2.14) is to apply the result of Eq. (2.13) recursively. If we define $v = y'$ then

$$\begin{aligned} \mathcal{L}[v'] &= s\mathcal{L}[v] - v(0) && \text{using Eq. (2.13) on } v \\ \Longrightarrow \quad \mathcal{L}[y''] &= s\mathcal{L}[y'] - y'(0) && \text{on translating back to } y \\ \Longrightarrow \quad \mathcal{L}[y''] &= s\Big(s\mathcal{L}[y] - y(0)\Big) - y'(0) && \text{using Eq. (2.13) on } y \\ &= s^2Y - y'(0) - sy(0). \end{aligned} \tag{2.16}$$

A similar trick works for $\mathcal{L}[y''']$ and higher derivatives.

In addition, given that integration is the inverse process to differentiation, it might not come as a surprise that,

$$\mathcal{L}\left[\int_0^t y(\tau)\,d\tau\right] = \frac{Y(s)}{s}. \tag{2.17}$$

In this result the variable τ is a dummy variable of integration. Proof of this will require integration by parts again, but the integral we see in Eq. (2.17) will need to be differentiated where use is made of the result,

$$\frac{d}{dt}\left[\int_0^t y(\tau)\,d\tau\right] = y(t). \tag{2.18}$$

This will form a question on a problem sheet.

2.5 Solutions of ODEs

We will now consider three examples of ODEs which will be solved using Laplace Transforms. Use will be made of some of the above results in order to illustrate the process of applying the inverse Laplace Transform.

Example 2.7: Solve the ODE, $y' + 2y = e^{-t}$, subject to the initial condition, $y(0) = 0$.

In the language I used in the ODEs section, the Complementary Function is characterised by $\lambda = -2$ while the forcing term is equivalent to $\lambda = -1$. From that point of view the writing down of $y_{\mathbf{cf}}$ and $y_{\mathbf{pi}}$ is straightforward, but let us see what happens with Laplace Transforms.

We will solve this equation by applying Laplace Transforms to each term in turn. We already know the following:

$$\begin{aligned} \mathcal{L}[y'] &= -y(0) + sY(s) && \text{(see Example 2.6)} \\ \mathcal{L}[2y] &= 2Y \\ \mathcal{L}[e^{-t}] &= \frac{1}{s+1}. && \text{(see Example 2.2)} \end{aligned} \tag{2.19}$$

Therefore we may write the following,

$$\begin{aligned}
& y' + 2y &&= && e^{-t} && \\
\Longrightarrow \quad & sY - y(0) + 2Y &&= && \frac{1}{s+1} && \\
\Longrightarrow \quad & (s+2)Y &&= && \frac{1}{s+1} && \text{since } y(0) = 0 \\
\Longrightarrow \quad & Y &&= && \frac{1}{(s+1)(s+2)} && \\
\Longrightarrow \quad & Y &&= && \frac{1}{(s+1)} - \frac{1}{(s+2)} && \text{using partial fractions} \\
\Longrightarrow \quad & y &&= && e^{-t} - e^{-2t} && \text{taking the inverse LT}
\end{aligned} \tag{2.20}$$

That final step used Eq. (2.5), namely $\mathcal{L}\left[e^{-at}\right] = 1/(s+a)$, with $a = 1$ and $a = 2$.

Example 2.8: Solve the equation, $y' + y = e^{-t}$, subject to the initial condition, $y(0) = c$, where c is a known, but unspecified, constant.

This equation is almost identical to that of Example 2.7, but the right hand side forcing term is now proportional to the Complementary Function of the ODE. In terms of λ-values, the auxiliary equation yields $\lambda = -1$ and the forcing term is also equivalent to $\lambda = -1$. So it will be of interest to see how Laplace Transforms cope with this special case with a repeated λ-value.

On taking Laplace Transforms of each term in the ODE we obtain,

$$sY - c + Y = \frac{1}{s+1}, \tag{2.21}$$

and therefore

$$Y = \frac{1}{(s+1)^2} + \frac{c}{s+1}. \tag{2.22}$$

The first term may be inverted using Example 2.5 with $a = 1$, and hence

$$y = te^{-t} + ce^{-t}. \tag{2.23}$$

In this solution the first term is the Particular Integral, while the second is the Complementary Function. Clearly the Laplace Transform takes this situation in its stride with no additional difficulties, but we need previously-derived results such as the one given by Example 2.5 as part of the toolchest of results to draw upon. But let us now consider a second order ODE.

Example 2.9: Solve the equation $y'' + 4y = 5e^{-t}$ subject to $y(0) = 0$ and $y'(0) = -1$.

On using the formula for $\mathcal{L}[y'']$ which is given in Eq. (2.16), the ODE transforms into

$$\underbrace{s^2Y - y'(0) - sy(0)}_{y''} + \underbrace{4Y}_{4y} = \underbrace{\frac{5}{s+1}}_{5e^{-t}}, \tag{2.24}$$

which may be rearranged into the form,

$$(s^2 + 4)Y + 1 = \frac{5}{s+1}. \tag{2.25}$$

Therefore Y is given by,

$$Y = \frac{5}{(s+1)(s^2+4)} - \frac{1}{s^2+4}. \tag{2.26}$$

We may now proceed either by expanding the first fraction using the method of partial fractions, or else combining the two fractions together to get $(4 - s)/[(s + 1)(s^2 + 4)]$, and then using the method of partial fractions. Either way, we obtain,

$$Y = \frac{1}{s+1} - \frac{s}{s^2+4}. \tag{2.27}$$

Results already derived (Example 2.2 with $a = 1$ and Example 2.3 with $a = 2$) are now sufficient to invert this expression; we get

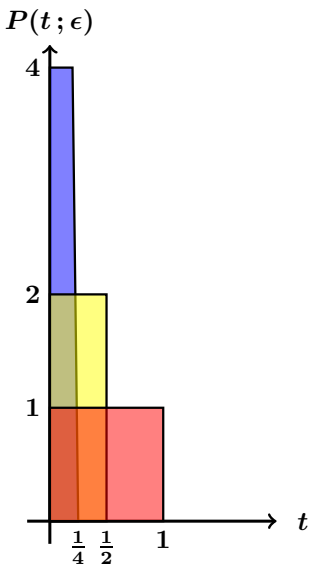
$$y = e^{-t} - \cos 2t. \tag{2.28}$$

2.6 The unit impulse

2.6.1 The definition

This is one of two unusual functions which are of great utility in Laplace Transforms. The other is the unit step function which is considered a little later.

We define $P(t\,;\epsilon)$ to be the **unit pulse of duration, ϵ**, beginning at $t = 0$. It has a **unit area**, and therefore Fig 2.1 shows three specific examples, while Eq. (2.29) gives the mathematical definition.



$$P(t\,;\epsilon) = \begin{cases} 1/\epsilon & (0 < t < \epsilon) \\ \\ 0 & (\epsilon < t). \end{cases} \tag{2.29}$$

Figure 2.1. Showing $P(t\,;\epsilon)$, for $\epsilon = 1$, $\frac{1}{2}$, and $\frac{1}{4}$.

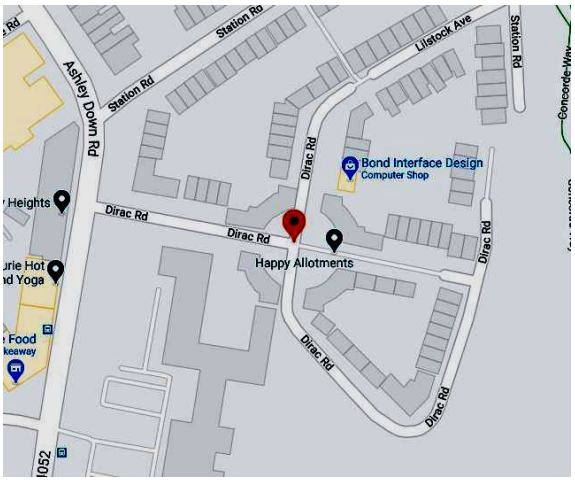
The **unit impulse** is what is obtained when ϵ becomes infinitesimally small. In this limit, $\epsilon \to 0$, the unit pulse now has an infinite strength over an interval of length zero, but the total area remains equal to 1 by definition.

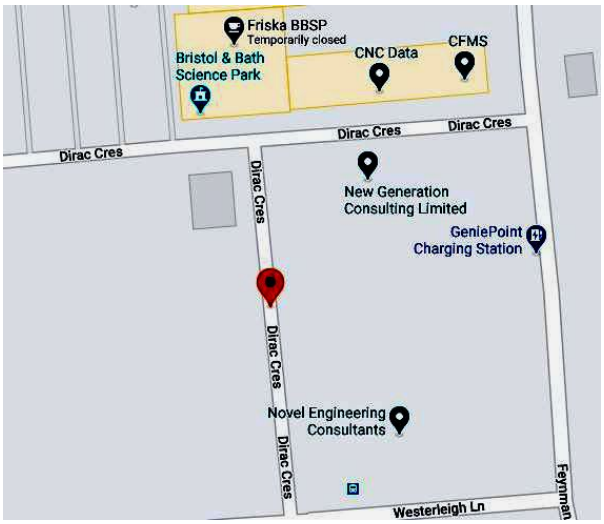
The unit impulse is also known as the **delta function** or, in physics contexts, as the **Dirac delta function**. It is denoted by $\delta(t)$ and may be defined formally as,

$$\delta(t) = \lim_{\epsilon \to 0} P(t\,;\epsilon). \tag{2.30}$$

The unit impulse is used to as an idealised model of an impact, and has an important role in many areas of physics, mathematics and different branches of engineering,

Given the extremely tall and very narrow form of the Dirac delta function, it is surprising how poorly the Bristol-born Dirac has been commemorated by the Bristol and South Gloucestershire Councils:





Courtesy of Google Maps

2.6.2 The Laplace Transform of the unit impulse

We now need to find the Laplace Transform of the delta function, and we shall do this by first taking the Laplace Transform of the unit pulse, and then taking the $\epsilon \to 0$ limit. The Laplace Transform of the unit pulse of duration, ϵ, is

$$\begin{aligned}\mathcal{L}\big[P(t\,;\epsilon)\big] &= \int_0^\infty P(t\,;\epsilon)\, e^{-st}\, dt \\ &= \int_0^\epsilon \frac{1}{\epsilon} e^{-st}\, dt + \int_\epsilon^\infty 0\, e^{-st}\, dt \qquad P(t\,;\epsilon) = 0 \text{ when } t > \epsilon \\ &= \frac{1}{\epsilon}\left[\frac{1 - e^{-\epsilon s}}{s}\right].\end{aligned} \tag{2.31}$$

Now we may let $\epsilon \to 0$. This may be done by letting $\epsilon \to 0$ in Eq. (2.31), above. To do this we need to use the Taylor's series expansion of $e^{-\epsilon s}$ in terms of ϵ:

$$e^{-\epsilon s} = 1 - \epsilon s + \frac{(\epsilon s)^2}{2!} - \frac{(\epsilon s)^3}{3!} + \frac{(\epsilon s)^4}{4!} \cdots . \tag{2.32}$$

When Eq. (2.32) is substituted into Eq. (2.31) we get,

$$\begin{aligned}\mathcal{L}\big[P(t\,;\epsilon)\big] &= \frac{1}{\epsilon}\Big[\frac{1 - (1 - \epsilon s + (\epsilon s)^2/2! - (\epsilon s)^3/3! \cdots)}{s}\Big] \\ &= \frac{\epsilon s - (\epsilon s)^2/2! + (\epsilon s)^3/3! \cdots}{\epsilon s} \\ &= 1 - \epsilon s/2! + (\epsilon s)^2/3! \cdots , \\ &\to 1 \qquad \text{as } \epsilon \to 0.\end{aligned} \tag{2.33}$$

Therefore,

$$\boxed{\mathcal{L}\big[\delta(t)\big] = 1.} \tag{2.34}$$

This value could also have been obtained using l'Hôpital's rule:

$$\mathcal{L}\big[\delta(t)\big] = \lim_{\epsilon\to 0} \frac{1 - e^{-\epsilon s}}{\epsilon s} \overset{l'H}{=\!=} \lim_{\epsilon\to 0} \frac{s e^{-\epsilon s}}{s} = 1, \tag{2.35}$$

on taking derivatives of both the numerator and denominator with respect to ϵ.

Note: While there has been a lot of derivation here, it is only Eq. (2.34) that needs to be remembered.

2.6.3 The unit impulse and integration

The unit impulse also has the property that,

$$\begin{aligned}\int_{-\infty}^{\infty} g(t)\,\delta(t)\,dt &= \int_{-\infty}^{\infty} g(0)\,\delta(t)\,dt && \text{since } g(t)=g(0) \text{ where } \delta(t) \text{ is nonzero}\\ &= g(0)\int_{-\infty}^{\infty}\delta(t)\,dt && \text{elementary property of integrals}\\ &= g(0)\times 1 \quad = \quad g(0).\end{aligned} \tag{2.36}$$

In other words the integral of a function multiplied by the unit impulse is equivalent to picking out the value of $g(t)$ when $t=0$. Bizarrely this is the easiest possible integral!

The apparent sleight of hand with the first equals sign above is motivated by the fact that the function, $g(t)$ and the value $g(0)$ are equal when $t=0$ (see the red disk in Fig. 2.2a, below), and this implies that $g(t)\delta(t)=g(0)\delta(t)$ because $\delta(t)=0$ when $t\neq 0$.

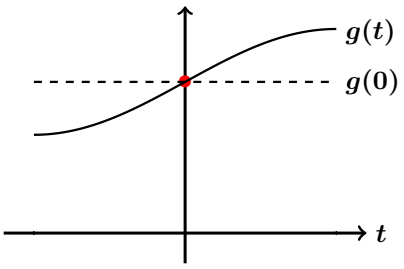


Figure 2.2a. Comparing $g(t)$ and $g(0)$.

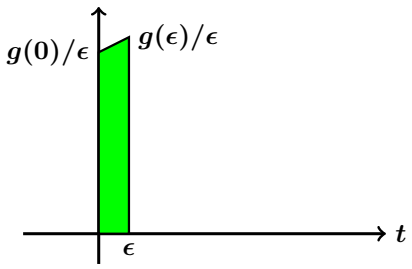


Figure 2.2b. Illustration of the trapezium rule.

An alternative proof may be provided by returning to the unit pulse, $P(t\,;\epsilon)$. When ϵ is very small, then we may approximate the integral of $g(t)\delta(t)$ using the trapezium rule (see Fig 2.2b, above):

$$\int_{-\infty}^{\infty} g(t)P(t\,;\epsilon)\,dt \quad\simeq\quad \underbrace{\left[\frac{1}{2}\left(\frac{g(\epsilon)}{\epsilon}+\frac{g(0)}{\epsilon}\right)\right]}_{\text{mean height}}\times\underbrace{\epsilon}_{\text{width}} \quad=\quad \tfrac{1}{2}[g(\epsilon)+g(0)]. \tag{2.37}$$

As $\epsilon\to 0$ this quantity tends towards $g(0)$ as well.

More generally, the unit impulse does not have to be located at $t=0$. When it is centred at $t=a$ it is denoted by $\delta(t-a)$ and it is therefore infinite when $(t-a)=0$. We now have the result,

$$\int_{-\infty}^{\infty} g(t)\,\delta(t-a)\,dt \quad=\quad \int_{-\infty}^{\infty} g(a)\,\delta(t-a)\,dt \quad=\quad g(a)\int_{-\infty}^{\infty}\delta(t-a)\,dt \quad=\quad g(a). \tag{2.38}$$

So the conclusion is that the integral of a function of t multiplied by a unit impulse is precisely equal to the value of the function at the point where the impulse is located.

Following on from this, we may find the Laplace Transform of $\delta(t-a)$:

$$\mathcal{L}\big[\delta(t-a)\big] = \int_0^\infty \delta(t-a)\, e^{-st}\, dt = e^{-as}, \tag{2.39}$$

and we may recover Eq. (2.36) when $a = 0$. Strictly speaking, this result applies only when $a \geq 0$. Should a be negative, then the impulse happens outside of the range of integration, and in that case the Laplace Transform is zero. We may state this mathematically as,

$$\mathcal{L}\big[\delta(t-a)\big] = \begin{cases} e^{-as} & (a \geq 0) \\ 0 & (a < 0) \end{cases} \tag{2.40}$$

Note: Again this has been a heavy subsubsection with multiple derivations. These derivations have been necessary in order that we may be convinced by the following results:

$$\boxed{\int_{-\infty}^{\infty} g(t)\, \delta(t-a)\, dt = g(a)} \qquad \text{and} \qquad \boxed{\mathcal{L}\big[\delta(t-a)\big] = e^{-as} \ \text{ when } a \geq 0.}$$

2.7 The solution of ODEs where the forcing term is a unit impulse

We shall have three examples of this type of ODE problem. The first is a 1st order ODE, while the others are of 2nd order.

Example 2.10: Solve $y' + ay = \delta(t)$ subject to $y(0) = 0$.

We do not yet have the technique to be able to find the Particular Integral that corresponds to a unit impulse, but we can use Laplace Transforms to solve this example. Again, using known results, we obtain the transformed version of the equation:

$$\begin{aligned} y' + ay = \delta(t) \qquad &\Longrightarrow \qquad \Big(sY - y(0)\Big) + aY = 1 \\ &\Longrightarrow \qquad (s+a)Y = 1, \\ &\Longrightarrow \qquad Y = \frac{1}{s+a}, \\ &\Longrightarrow \qquad y = e^{-at} \qquad \text{using Eq. (2.5).} \end{aligned} \tag{2.41}$$

Note 1: The final solution, $y = e^{-at}$, is an example of what is known as the **impulse response** (or the **unit impulse response function**) for the system represented by the left hand side of the ODE. Crudely, it is how a system at rest then reacts to a unit impulse.

Note 2: The expression for Y is an example of what is known as the **Transfer Function** of the system. In a very real sense the Transfer Function encapsulates the full properties of the system. It is interesting to see that the reciprocal of the Transfer Function (i.e. $s+a$) is identical (account being taken for notation) with the Auxiliary equation, $\lambda + a = 0$, for the ODE. This is not a coincidence. There's a little more later.

Note 3: The exciting of a system by means of a unit impulse means that the given initial condition appears to have been violated. The original ODE has the initial condition, $y(0) = 0$, but the solution that we have derived satisfies $y(0) = 1$. Is this a contradiction?

Well, is it a contradiction? We'll need to analyse this in a little bit of detail, again just so that we can trust the solution that we have found. **Note:** that the rest of this page is for information only and serves solely to justify the Laplace Transform solution of Example 2.10.

Let us return to the unit pulse of duration, ϵ, and solve $y' + ay = P(t\,;\epsilon)$ instead. The objective here is to solve for the system's response to this pulse and then to let $\epsilon \to 0$ once more. Perhaps we need to rewrite the ODE as follows,

$$y' + ay = \begin{cases} \dfrac{1}{\epsilon} & (0 \le t \le \epsilon) \\ \\ 0 & (\epsilon < t) \end{cases} \tag{2.42}$$

We'll minimise the detail of this analysis, but essentially (i) we solve for y in the range, $0 \le t \le \epsilon$, (ii) then determine the value of $y(\epsilon)$ which will then be used as an initial condition for (iii) the solution for y in the range, $\epsilon < t$. This solution is given by,

$$y = \begin{cases} \dfrac{1}{a\epsilon}(1 - e^{-at}) & (0 \le t \le \epsilon) \\ \\ \dfrac{1}{a\epsilon}(1 - e^{-a\epsilon})e^{-a(t-\epsilon)} & (\epsilon < t) \end{cases} \tag{2.43}$$

where the solutions have been written in a way where it is easy to check that the two formulae for y agree at $t = \epsilon$. For the sake of illustration we have set $a = 1$ in Figure 2.3 where the solutions for three different pulse durations are shown.

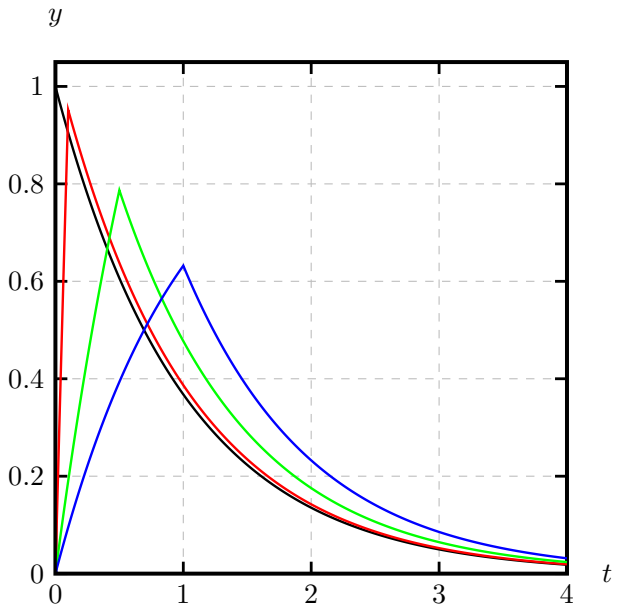


Figure 2.3. Solutions given by Eq. (2.43) for $\epsilon = 0$ (unit impulse) $\epsilon = 0.1$, $\epsilon = 0.5$ and $\epsilon = 1$.

In this Figure we see that the solutions corresponding to the various unit pulses converge towards that for the unit impulse as $\epsilon \to 0$. More specifically we see that the initial rise from $y = 0$ at $t = 0$ becomes increasingly steep in this limit. Equation (2.43) confirms this since $y'(0) = 1/\epsilon$. Although this is an unusual limit, the Laplace Transform solution given in Eq. (2.41) is indeed correct even though it appears to violate the given initial conditions; it is merely the case that there is an extremely rapid variation in y over the infinitesimally short duration of the impulse.

Example 2.11: Solve the equation $y'' + (a+b)y' + aby = \delta(t)$ subject to $y(0) = y'(0) = 0$.

On taking the Laplace Transform of the equation we get,

$$
\begin{aligned}
& s^2Y + (a+b)sY + abY = 1 && \text{noting that } y = y' = 0 \text{ at } t = 0 \\
\Longrightarrow \quad & \Big[s^2 + (a+b)s + ab\Big]Y = 1 \\
\Longrightarrow \quad & Y = \frac{1}{s^2 + (a+b)s + ab} \\
\Longrightarrow \quad & Y = \frac{1}{(s+a)(s+b)} && (2.44) \\
\Longrightarrow \quad & Y = \frac{1}{b-a}\left[\frac{1}{s+a} - \frac{1}{s+b}\right] && \text{using partial fractions} \\
\Longrightarrow \quad & y = \left[\frac{e^{-at} - e^{-bt}}{b-a}\right] && \text{taking the inverse Laplace Transform.}
\end{aligned}
$$

In the light of the initial condition violation which we encountered in Example 2.10, let us check whether the same happens here. We may easily determine mathematically that $y(0) = 0$, and $y'(0) = 1$. Thus the **impulse response** for a second order ODE with zero initial conditions has a nonzero first derivative at $t = 0$, so it is the initial condition for the first derivatve that has been violated this time.

Although I shall not prove it, the general case is that the initial condition for the $(n-1)^{\text{st}}$ derivative is violated when solving an n^{th} order ODE.

In the following Figure we show what the solution given by Eq. (2.44) looks like for the case, $a = 1$ and $b = 2$. The maximum value of y is 0.25 and this is attained when $t = \ln 2 = 0.69315$. The initial slope may also be found analytically and it is $y'(0) = 1$.

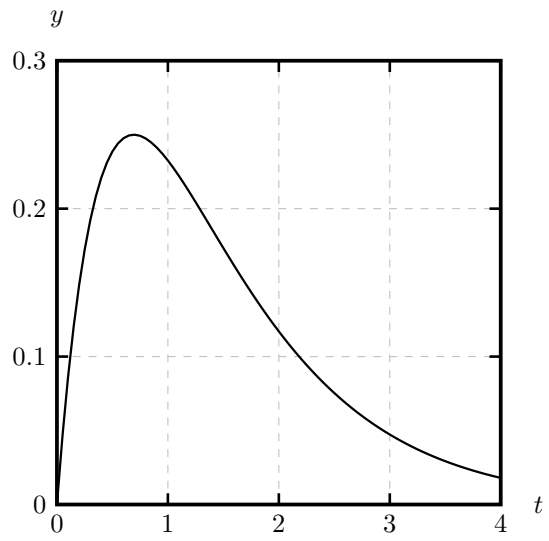


Figure 2.4. Solutions given by Eq. (2.44) when $a = 1$ and $b = 2$.

A final example involves an undamped mass/spring system.

Example 2.12: Solve the ODE, $my'' + ky = \delta(t)$, subject to $y(0) = y'(0) = 0$. Here m is mass and k, the spring stiffness.

We rearrange the equation for the ODE slightly to the form,

$$y'' + (k/m)y = (1/m)\delta(t), \tag{2.45}$$

and the result of taking the Laplace Transform is,

$$\left(s^2 + \frac{k}{m}\right)Y = \frac{1}{m}, \tag{2.46}$$

where the zero initial conditions have been accounted for. We now proceed as follows:

$$\begin{aligned}
& \left(s^2 + \frac{k}{m}\right)Y = \frac{1}{m} \\
\implies \quad & Y = \frac{1/m}{s^2 + k/m} && \text{which reminds us of the LT of a sine} \\
\implies \quad & Y = \frac{1}{\sqrt{km}} \times \frac{\sqrt{k/m}}{s^2 + k/m} && \text{now use Ex. 2.2 with } a = \sqrt{k/m} \\
\implies \quad & y = \frac{1}{\sqrt{km}} \sin\sqrt{\frac{k}{m}}\, t.
\end{aligned} \tag{2.47}$$

This final solution certainly satisfies $y(0) = 0$, as did the solution in Example 2.11. However, we have

$$y' = \frac{1}{m}\cos\sqrt{\frac{k}{m}}\, t, \tag{2.48}$$

and therefore $y'(0) = 1/m$. So we have another violation of the $y'(0) = 0$ initial condition. In this case the physical meaning of this violation tells us that the initial velocity which is induced by the unit impulse decreases as the mass, m, increases. Clearly this is physically correct — think of the different responses of a ping pong ball and a cricket ball.

We may also say that the solution given in Eq. (2.47) satisfies $my'(0) = 1$ and, given that y' is the velocity, we can say that **the unit impulse imparts a unit momentum to the system.**

2.7.1 Some observations

Given that the presence of the unit impulse as a forcing term causes a violation of an initial condition, there is the scope to use our experience of solving these equations to write down an alternative versions of the ODEs and initial condition without the presence of the unit impulse. Thus for Example 2.10,

$$\begin{aligned}
& y' + ay = \delta(t), \qquad y(0) = 0, \\
\text{and} \qquad & \\
& y' + ay = 0, \qquad\quad y(0) = 1,
\end{aligned} \tag{2.49}$$

have the same solutions. Likewise for Example 2.11:

$$\begin{aligned}
& y'' + (a+b)y' + ab = \delta(t), \qquad y(0) = 0, \quad y'(0) = 0, \\
\text{and} \qquad & \\
& y'' + (a+b)y' + ab = 0, \qquad\quad y(0) = 0, \quad y'(0) = 1,
\end{aligned} \tag{2.50}$$

have the same solution. The more general case given in Example 2.12 (where the y'' term doesn't have a unit coefficient):

$$\begin{aligned} &my'' + ky = \delta(t), \qquad y(0) = 0, \quad y'(0) = 0, \\ \text{and} \qquad & \\ &my'' + ky = 0, \qquad\quad y(0) = 0, \quad my'(0) = 1, \end{aligned} \tag{2.51}$$

also have the same solution. As a final example which hasn't been covered above, the following two ODEs have identical solutions:

$$\begin{aligned} &my'' + ky = c\delta(t), \qquad y(0) = a, \quad y'(0) = b, \\ \text{and} \qquad & \\ &my'' + ky = 0, \qquad\quad y(0) = a, \quad y'(0) = b + c/m. \end{aligned} \tag{2.52}$$

Check carefully whether this last case makes sense, given the preceding analyses.

2.8 The Unit Step Function

2.8.1 Definition

This is the second of the two special functions that we'll consider and this one is sketched in Figure 2.5, below. It is denoted either by $u(t)$ or by $H(t)$; we will use $H(t)$. The former notation is because the step is a unit step, from 0 to 1. The latter notation comes from its alternative name, the Heaviside step function, named for the British physicist/engineer/mathematician, Oliver Heaviside, whose entry on Wikipedia is worth reading for many reasons!

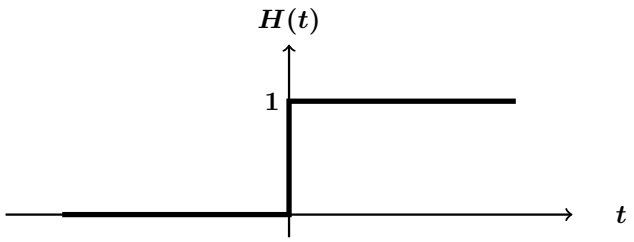


Figure 2.5. The unit step function.

Mathematically, the unit step function is defined as

$$H(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t > 0). \end{cases} \tag{2.53}$$

The step rise does not need to occur at $t = 0$. Its equivalent at $t = a$ is denoted by $H(t - a)$ and this is shown in Figure 2.6.

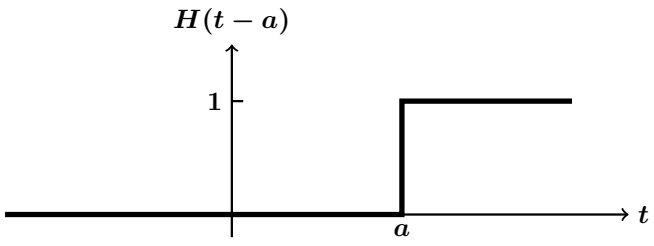


Figure 2.6. The unit step function at $t = a$.

This is defined by,

$$H(t - a) = \begin{cases} 0 & (t - a < 0) \\ 1 & (t - a > 0). \end{cases} \tag{2.54}$$

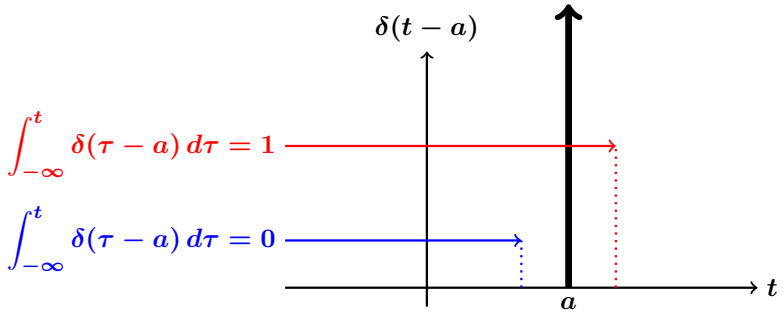
There is a very strong link between the unit impulse and the unit step function. The two main relationships may be expressed as,

$$H(t) = \int_{-\infty}^{t} \delta(\tau)\, d\tau \qquad \text{and} \qquad \delta(t) = \frac{dH(t)}{dt}, \tag{2.55}$$

or, more generally, as,

$$H(t-a) = \int_{-\infty}^{t} \delta(\tau - a)\, d\tau \qquad \text{and} \qquad \delta(t-a) = \frac{dH(t-a)}{dt.} \tag{2.56}$$

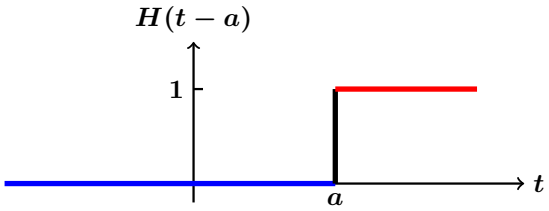
Figure 2.7 shows how the unit step function may be obtained by integrating the unit impulse. When $t < a$, which is represented by the blue line, the range of integration doesn't include the location of the unit impulse at $t = a$, and therefore the integral is zero. On the other hand, when $t > a$, which is represented by the red line, the range of integration includes the location of the unit impulse, and therefore the integral is 1.



$$\boxed{H(t-a) = \int_{-\infty}^{t} \delta(\tau - a)\, d\tau}$$

Figure 2.7. The unit impulse at $t = a$.

Figure 2.8 shows the unit step function and, in particular, we see that the slope is zero when $t \neq a$ (the red and the blue sections) and is infinite at $t = a$ (the black section). Pure mathematicians might have more to say about this but we will run with idea as being self-evident.



$$\boxed{\delta(t-a) = \frac{dH(t-a)}{dt}.}$$

Figure 2.8. The unit step function at $t = a$.

2.8.2 The Laplace Transform of the Unit Step Function

We will now consider the Laplace Transform of the unit step function because this has some use in the solution of ODEs. The Laplace Transform of $\boldsymbol{H(t-a)}$ is given by

$$\begin{aligned}
\mathcal{L}\big[H(t-a)\big] &= \int_0^\infty H(t-a)\,e^{-st}\,dt \\
&= \int_0^a 0\times e^{-st}\,dt + \int_a^\infty 1\times e^{-st}\,dt \qquad \text{splitting the integral} \\
&= 0 + \int_a^\infty 1\times e^{-st}\,dt \\
&= \left[-\frac{e^{-st}}{s}\right]_a^\infty = \frac{e^{-as}}{s}.
\end{aligned} \tag{2.57}$$

Note: that the use of $a=0$ in the above shows that $\mathcal{L}\big[H(t)\big] = 1/s$. If this seems familiar then it is because we also showed that $\mathcal{L}\big[1\big] = 1/s$ in Example. 2.1. Although it appears that we have two functions with the same transform, the functions are identical within the range of integration of the Laplace Transform: $H(t)=1$.

We will revisit both the unit impulse and the unit step function later.

2.9 The shift theorem in s

This is almost trivial! Here is a combined statement and proof of the theorem. Given that,

$$\mathcal{L}\big[f(t)\big] = \int_0^\infty f(t)\,e^{-st}\,dt = F(s), \tag{2.58}$$

then

$$\begin{aligned}
\mathcal{L}\big[f(t)e^{-at}\big] &= \int_0^\infty f(t)e^{-at}e^{-st}\,dt \qquad \text{by definition} \\
&= \int_0^\infty f(t)e^{-(s+a)t}\,dt \\
&= F(s+a).
\end{aligned} \tag{2.59}$$

If this proof seems to be too quick, then compare the role that $\boxed{s}$ plays in Eq. (2.58) with the role played by $\boxed{(s+a)}$ in Eq. (2.59). A simple statement of the theorem is,

$$\boxed{\mathcal{L}\big[f(t)\big] = F(s) \qquad \Longrightarrow \qquad \mathcal{L}\big[f(t)\,e^{-at}\big] = F(s+a).} \tag{2.60}$$

The simplicity of this theorem is demonstrated in the next two Examples.

Example 2.13: Find the Laplace Transform of te^{-at}.

We start by noting that we already know that $\mathcal{L}[t] = 1/s^2$ from Example 2.4. Therefore the s-shift theorem tells us that,

$$\mathcal{L}[te^{-at}] = 1/(s+a)^2. \tag{2.61}$$

This is the same answer as we found in Example 2.5, but there we undertook the integration explicitly rather than by using a previously-known result and the s-shift theorem.

Example 2.14: Find the Laplace Transform of $\cos bt\ e^{-at}$.

Example 2.3 shows that $\mathcal{L}[\cos bt] = s/(s^2+b^2)$. So the s-shift theorem tells us that,

$$\mathcal{L}[\cos bt] = \frac{s}{s^2+b^2} \qquad \Longrightarrow \qquad \mathcal{L}[\cos bt\ e^{-at}] = \frac{s+a}{(s+a)^2+b^2}. \tag{2.62}$$

So every instance of s in the left hand equation is replaced by $(s+a)$ in the right hand equation.

Similarly, we may state that,

$$\mathcal{L}[\sin bt] = \frac{b}{s^2+b^2} \qquad \Longrightarrow \qquad \mathcal{L}[\sin bt\ e^{-at}] = \frac{b}{(s+a)^2+b^2}. \tag{2.63}$$

Both the left hand sides for both Eqs. (2.62) and (2.63) have been taken from Example 2.3.

We may also use this shift theorem to find inverse transforms and, indeed, this is its main use. The following is a typical example which arises from the solution of an ODE.

Example 2.15: Solve the ODE, $y'' + 4y' + 13y = 0$, subject to $y(0) = 1$ and $y'(0) = -4$.

On applying the Laplace Transform to the ODE we obtain,

$$\Big[s^2Y - y'(0) - sy(0)\Big] + 4\Big[sY - y(0)\Big] + 13Y = 0. \tag{2.64}$$

Using the given initial conditions this simplifies to,

$$(s^2 + 4s + 13)Y - s = 0, \tag{2.65}$$

and therefore,

$$Y = \frac{s}{s^2+4s+13}. \tag{2.66}$$

The determination of y, i.e. the inverse Laplace Transform of Y, will require us to coerce Y into a form which may then be inverted immediately. This will involve the completion of the square for the denominator, the s-shift theorem and the following Laplace Transforms for $\cos bt$ and $\sin bt$ from Example 2.3:

$$\mathcal{L}[\cos bt] = \frac{s}{s^2+b^2}, \qquad \mathcal{L}[\sin bt] = \frac{b}{s^2+b^2}.$$

First, we note that $s^2+4s+13=(s+2)^2+9=(s+2)^2+3^2$. The presence of the $(s+2)$ tells us that the s-shift theorem is soon to be used. So we have,

$$\begin{aligned} Y &= \frac{s}{s^2+4s+13} \\ &= \frac{s}{(s+2)^2+3^2} && \text{on completing the square} \\ &= \frac{s+2}{(s+2)^2+3^2} - \frac{2}{(s+2)^2+3^2} && \text{to get } (s+2) \text{ everywhere} \\ &= \frac{s+2}{(s+2)^2+3^2} - \tfrac{2}{3}\left[\frac{3}{(s+2)^2+3^2}\right] && \text{now for inversion.} \end{aligned} \tag{2.67}$$

The last step involved placing a 3 in the numerator of the fraction comprising the second quotient. Now every term is just an s-shift away from the Laplace Transforms of a cosine and a sine, respectuvely. Therefore the s-shift theorem tells us that

$$\mathcal{L}^{-1}\left[\frac{s+2}{(s+2)^2+3^2}\right] = e^{-2t}\cos 3t, \tag{2.68}$$

and

$$\mathcal{L}^{-1}\left[\tfrac{2}{3}\,\frac{3}{(s+2)^2+3^2}\right] = \tfrac{2}{3}e^{-2t}\sin 3t. \tag{2.69}$$

Therefore we may write the final solution of the ODE as,

$$y = e^{-2t}\cos 3t - \tfrac{2}{3}e^{-2t}\sin 3t = e^{-2t}\left[\cos 3t - \tfrac{2}{3}\sin 3t\right]. \tag{2.70}$$

2.10 The shift theorem in t

We shall begin with the following sketch of a function, $f(t)$ (black curve), and how it may be shifted sideways to the right but not retain any information about $f(t)$ when $t<0$ (red curve). This uses $H(t-a)$ as a multiplier to filter out the unwanted information (the dotted line). So $f(t)$ has undergone a t-shift, and $f(t-a)H(t-a)$ contains exactly the same information as $f(t)$ does in the range, $0\le t<\infty$.

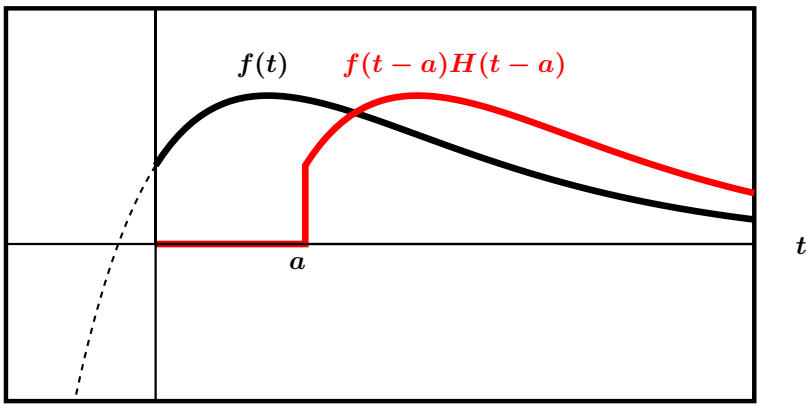


Figure 2.9. Showing a typical $f(t)$ and $f(t-a)H(t-a)$.

Now let us find the Laplace Transform of $f(t-a)H(t-a)$:

$$\begin{aligned}
\mathcal{L}\big[H(t-a)f(t-a)\big] &= \int_0^\infty f(t-a)\,H(t-a)\,e^{-st}\,dt \\
&= \int_a^\infty f(t-a)\,H(t-a)\,e^{-st}\,dt && \text{because the integrand is zero in } 0 \le t < a \\
&= \int_a^\infty f(t-a)\,e^{-st}\,dt && \text{because } H(t-a)=1 \text{ in } a \le t < \infty \\
&= \int_0^\infty f(\tau)e^{-s(\tau+a)}\,d\tau && \text{using } \tau = t-a\text{; also note the lower limit} \\
&= e^{-sa}\underbrace{\int_0^\infty f(\tau)\,e^{-s\tau}\,d\tau}_{F(s)} && e^{-sa} \text{ is a constant with respect to } \tau \\
&= e^{-sa}F(s).
\end{aligned} \tag{2.71}$$

Therefore we have shown that

$$\boxed{\mathcal{L}\big[H(t-a)f(t-a)\big] = e^{-sa}F(s),} \tag{2.72}$$

which is the Shift Theorem in t.

Example 2.16: Find $\mathcal{L}^{-1}\left[\dfrac{e^{-as}}{s^2}\right]$.

The presence of the exponential tells us that the t-shift theorem needs to be used. The presence of the $1/s^2$ tells us that the basic time-dependent function which lies beneath all of this is t, because $\mathcal{L}\big[t\big] = 1/s^2$. So therefore we apply the t-shift theorem and obtain,

$$\mathcal{L}\big[t\big] = \frac{1}{s^2} \qquad \Longrightarrow \qquad \mathcal{L}\big[(t-a)H(t-a)\big] = \frac{e^{-as}}{s^2}. \tag{2.73}$$

Therefore the answer to the original question is,

$$\mathcal{L}^{-1}\left[\frac{e^{-as}}{s^2}\right] = (t-a)H(t-a). \tag{2.74}$$

Example 2.17 Find the inverse Laplace Transform of $e^{-as}/(s+b)^2$.

This is an example of the use of both shift theorems. The exponential tells us that we need the t-shift theorem. The presence of the $(s+b)$ tells us that the s-shift is needed. From all this mess we can see that the underlying function is again, $1/s^2$. The only question now is, which is the better way to go? Should we use the s-shift theorem first or the t-shift theorem first? We'll do it both ways and everyone can then make up their own minds. But in both cases we lead off with the Laplace Transform of t.

If we start with the s-shift theorem, then we get,

$$\mathcal{L}\big[\underbrace{\ t\ }_{f(t)}\big] = \underbrace{\frac{1}{s^2}}_{F(s)}$$

$$\implies \quad \mathcal{L}\big[\underbrace{te^{-bt}}_{f(t)e^{-bt}}\big] = \underbrace{\frac{1}{(s+b)^2}}_{F(s+b)} \qquad \text{applying the } s\text{-shift theorem}$$

$$\implies \quad \mathcal{L}\big[\underbrace{te^{-bt}}_{f(t)}\big] = \underbrace{\frac{1}{(s+b)^2}}_{F(s)} \qquad \text{redefining the labelling}$$

$$\implies \quad \mathcal{L}\big[\underbrace{H(t-a)\times(t-a)e^{-b(t-a)}}_{H(t-a)f(t-a)}\big] = \underbrace{\frac{e^{-as}}{(s+b)^2}}_{e^{-as}F(s)} \qquad \text{applying the } t\text{-shift theorem} \tag{2.75}$$

where we have meticulously changed every t in te^{-bt} to $(t-a)$ in the last line.

On the other hand, if we start with the t-shift theorem, then we get,

$$\mathcal{L}\big[\underbrace{\ t\ }_{f(t)}\big] = \underbrace{\frac{1}{s^2}}_{F(s)}$$

$$\implies \quad \mathcal{L}\big[\underbrace{(t-a)H(t-a)}_{f(t-a)H(t-a)}\big] = \underbrace{\frac{e^{-as}}{s^2}}_{e^{-as}F(s)} \qquad \text{applying the } t\text{-shift theorem}$$

$$\implies \quad \mathcal{L}\big[\underbrace{(t-a)H(t-a)}_{f(t)}\big] = \underbrace{\frac{e^{-as}}{s^2}}_{F(s)} \qquad \text{relabelling} \tag{2.76}$$

$$\implies \quad \mathcal{L}\big[\underbrace{(t-a)e^{-bt}H(t-a)}_{f(t)e^{-bt}}\big] = \underbrace{\frac{e^{-a(s+b)}}{(s+b)^2}}_{F(s+b)} \qquad \text{applying the } s\text{-shift theorem}$$

$$\implies \quad \mathcal{L}\big[(t-a)e^{-b(t-a)}H(t-a)\big] = \frac{e^{-as}}{(s+b)^2} \qquad \text{multiplying both sides by } e^{ab}$$

and we obtain the same final line as before.

Hence,

$$\mathcal{L}^{-1}\left[\frac{e^{-as}}{(s+b)^2}\right] = H(t-a)\times(t-a)e^{-b(t-a)}. \tag{2.77}$$

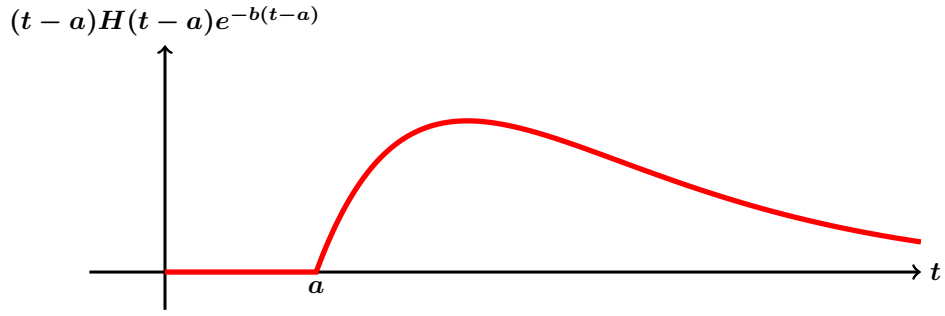


Figure 2.10. Sketch of $\mathcal{L}^{-1}[e^{-as}/(s+b)^2] = (t-a)H(t-a)e^{-b(t-a)}$

2.11 Convolution Theorem

Given the two functions $f(t)$ and $g(t)$ and their respective transforms, $F(s)$ and $G(s)$, then the convolution of f and g is defined by

$$f * g = \int_0^t f(\tau)g(t-\tau)\,d\tau = \int_0^t f(t-\tau)g(\tau)\,d\tau, \tag{2.78}$$

where it should be noted that each of these two integrals results in the same function of t. We may say that $f * g$ is **the convolution of f and g.**

It is interesting to note what happens to the arguments of f and g in the above integrals. In the first integral we see that, while the argument of $f(\tau)$ increases from 0 to t, the argument of $g(t-\tau)$ decreases from t to 0. This seems unusual but this behaviour arises elsewhere and for more obvious reasons. Perhaps the simplest example is the calculation of probabilities associated with rolling two standard dice. For example, the probability of throwing, say, a 9 in total is given by

$$\begin{aligned} P(9 \text{ with 2 dice}) &= P(3)\times P(6) + P(4)\times P(5) + P(5)\times P(4) + P(6)\times P(3) \\ &= \sum_{n=3}^{6} P(n) \times P(9-n). \end{aligned} \tag{2.79}$$

where I am assuming that my notation is self-explanatory. So we end up with a convolution sum where n increases while $9-n$ decreases.

However, it is more important to explain why it is even necessary to have the concept of convolution here at all. The reason is the following:

$$\mathcal{L}\big[f * g\big] = F(s)G(s), \tag{2.80}$$

i.e. that **the transform of the convolution is the product of the transforms.** I will not give a proof of this here, but it is relegated to an Appendix at the end of the Laplace Transform section should you be curious. This result may also be written in the form,

$$\mathcal{L}^{-1}[F(s)G(s)] = f * g, \tag{2.81}$$

which is the form that is generally needed in practice.

We will illustrate this by following the fate and fortune of the following two ODEs:

$$\boxed{\begin{array}{c} y' + 2y = f(t) \\ \\ y(0) = 0 \end{array}} \quad \text{and} \quad \boxed{\begin{array}{c} z'' + 4z = f(t) \\ \\ z(0) = z'(0) = 0 \end{array}} \tag{2.82}$$

So we have two physical systems, one for $y(t)$ the other for $z(t)$, both of which are at rest for $t < 0$, and then each of which is set into motion solely by the forcing function, $f(t)$. These may be solved by first applying the Laplace Transform:

$$\boxed{(s+2)Y = F} \quad \text{and} \quad \boxed{(s^2+4)Z = F,} \tag{2.83}$$

These are solved easily for Y and Z:

$$\boxed{Y = F \times \frac{1}{s+2}} \quad \text{and} \quad \boxed{Z = F \times \frac{1}{s^2+4},} \tag{2.84}$$

both of which are products of functions of s and so we need the Convolution Theorem.

2.11.1 Some examples of the use of the Convolution Theorem

Before returning to Eqs. (2.83) and (2.84) we shall consider two Examples of the use of the Convolution Theorem.

Example 2.18: Given that $\mathcal{L}\big[t\big] = 1/s^2$, use the Convolution Theorem to find $\mathcal{L}^{-1}[1/s^4]$.

Clearly $1/s^4 = (1/s^2) \times (1/s^2)$ and therefore we may use the Convolution Theorem as given by Eq. (2.81). Formally, we shall let $F = G = 1/s^2$, which means that $f = g = t$. Equation (2.81) gives,

$$\begin{aligned} \mathcal{L}^{-1}\left[\frac{1}{s^4}\right] = \mathcal{L}^{-1}[FG] \;&=\; f * g = \int_0^t f(\tau)\, g(t-\tau)\, d\tau \\ &= \; t * t \\ &= \; \int_0^t \tau(t-\tau)\, d\tau \\ &= \; \Big[\frac{\tau^2}{2}t - \frac{\tau^3}{3}\Big]_0^t \\ &= \; t^3/6. \end{aligned} \tag{2.85}$$

Example 2.19 Use the Convolution Theorem to find the inverse Laplace Transform of

$$\frac{1}{s+1} \times \frac{1}{s+2}.$$

This particular function of s was obtained back in Example 2.7 when we were solving the ODE, $y' + 2y = e^{-t}$ subject to $y(0) = 0$. At that point we employed Partial Fractions to simplify this function of s before applying the Inverse Laplace Transform. Now we shall use the convolution theorem instead, and we really ought to obtain the same function of t.

We already know that

$$\mathcal{L}\left[e^{-t}\right] = \frac{1}{(s+1)} = F(s) \qquad \text{and} \qquad \mathcal{L}\left[e^{-2t}\right] = \frac{1}{(s+2)} = G(s),$$

and therefore we have $f(t) = e^{-t}$ and $g(t) = e^{-2t}$ to use in the Convolution Theorem. Then the required $\mathcal{L}^{-1}[FG] = f * g$. We'll use both versions of the convolution integral:

$$\begin{aligned}
f * g &= e^{-t} * e^{-2t} & f * g &= e^{-t} * e^{-2t} \\
&= \int_0^t \underbrace{e^{-\tau}}_{f(\tau)} \times \underbrace{e^{-2(t-\tau)}}_{g(t-\tau)}\, d\tau & &= \int_0^t \underbrace{e^{-(t-\tau)}}_{f(t-\tau)} \times \underbrace{e^{-2\tau}}_{g(\tau)}\, d\tau \\
&= \int_0^t e^{-2t}e^{\tau}\, d\tau & &= \int_0^t e^{-t}e^{-\tau}\, d\tau \\
&= e^{-2t}\int_0^t e^{\tau}\, d\tau & &= e^{-t}\int_0^t e^{-\tau}\, d\tau \\
&= e^{-2t}\Big[e^{\tau}\Big]_0^t & &= e^{-t}\Big[-e^{-\tau}\Big]_0^t \\
&= e^{-2t}\Big[e^{t}-1\Big] & &= e^{-t}\Big[-e^{-t}+1\Big] \\
&= e^{-t} - e^{-2t}. & &= e^{-t} - e^{-2t}.
\end{aligned} \tag{2.86}$$

This is the same expression as was given in Eq. (2.20).

Note: In this Example we had a choice of which integral to use in the definition of the convolution integral in Eq. (2.78). Both have been applied and they are of equal difficulty, but there are some cases where one of the choices is either more useful or quicker than the other.

Example 2.20: Solve the ODE, $y' + 2y = f(t)$ subject to $y(0) = 0$ using the Convolution Theorem.

We have already started this problem back in Eq. (2.82), but we'll run it fully here. On taking the Laplace Transform (noting that $y(0) = 0$) we obtain,

$$sY + 2Y = F \qquad \Longrightarrow \qquad Y = F \times \frac{1}{s+2}, \tag{2.87}$$

which is a product and therefore a perfect target for the Convolution Theorem.

We may let $G(s) = 1/(s+2)$ and therefore $g(t) = e^{-2t}$. Now we face the choice of which integral to use. Here's the 'wrong' one:

$$y(t) = f(t) * e^{-2t} = \int_0^{\tau} f(t-\tau)e^{-2\tau}\, d\tau. \tag{2.88}$$

Whilst this a perfectly good expression for $y(t)$, it isn't as good as the other one. This second one is,

$$y(t) = f(t) * e^{-2t} = \int_0^t f(\tau)e^{-2(t-\tau)}\, d\tau = e^{-2t}\int_0^t f(\tau)e^{2\tau}\, d\tau. \tag{2.89}$$

On the face of it, this doesn't necessarily look better, but this is the form of solution which we would obtain by treating the original ODE as a first-order-linear ODE and using the Integrating Factor. For this ODE the Integrating Factor is e^{2t}, and when the ODE is multiplied by it, we obtain,

$$\begin{aligned} e^{2t}(y' + 2y) &= e^{2t} f(t) \\ \implies \quad (e^{2t} y)' &= e^{2t} f(t) && \text{exact derivative} \\ \implies \quad e^{2t} y &= \int_0^t e^{2\tau} f(\tau)\, d\tau && \text{using } y(0) = 0 \\ \implies \quad y &= e^{-2t} \int_0^t e^{2\tau} f(\tau)\, d\tau. \end{aligned} \tag{2.90}$$

Note: If $f(t) = \delta(t)$, the unit impulse, then $F(s) = 1$. Therefore Eq. (2.87) becomes

$$Y = \frac{1}{s+2}, \tag{2.91}$$

which is the **Transfer Function** for the system (i.e. for $y' + 2y$), and its inverse Laplace Transform is

$$y = e^{-2t}, \tag{2.92}$$

is the **unit impulse response function**. This expression for y is obtained by substituting for $f(t)$ into Eq. (2.89).

Example 2.21: Solve the ODE, $z'' + 4z = f(t)$ subject to $z(0) = z'(0) = 0$ using the Convolution Theorem.

Again we have already started this problem back in Eq. (2.82), and again we'll run it fully here.

On taking the Laplace Transform (noting that $z(0) = z'(0) = 0$) we obtain,

$$(s^2 + 4)Z = F \qquad \implies \qquad Z = F \times \frac{1}{s^2 + 4}. \tag{2.93}$$

If we let $G(s) = 1/(s^2 + 4)$ then Eq. (2.8) gives its inverse as $g(t) = \frac{1}{2} \sin 2t$; check that one carefully. Hence the inverse Laplace Transform of Z is,

$$z = \tfrac{1}{2} \sin 2t * f(t) = \tfrac{1}{2} \int_0^t f(\tau) \sin 2(t - \tau)\, d\tau) = \tfrac{1}{2} \int_0^t f(t - \tau) \sin 2\tau\, d\tau. \tag{2.94}$$

I am not sure which is the better form of the convolution integral to use here, but both are fine. On the other hand, if we have $f(t) = \delta(t)$, then the first integral is better because it is slightly easier to use the result of integrating with the delta function. Once more, $f(t) = \delta(t) \Rightarrow F(s) = 1$, and hence the **Transfer Function** is $Z = 1/(s^2 + 4)$ and the **unit impulse response function** is $z = \frac{1}{2} \sin 2t$.

2.12 Solving systems of ODEs

We'll consider just one of these, and it will also involve the unit impulse. These become quite rapidly more difficult as the order of the system increases, and it will eventually be necessary to use numerical methods to tackle these. The example given here involves a pair of first order ODEs. There turns out to be more than one way of organising one's workings for this, and I will attempt to make each route as clear as possible.

Example 2.22: Solve the system of ODEs,

$$y' + y - z = 0, \qquad z' + 2y + 4z = \delta(t), \qquad \text{subject to } y(0) = z(0) = 0. \tag{2.95}$$

Method 1. This involves the elimination of one of the dependent variables to leave a second order ODE. So the elimination of z between the two yields, $y'' + 5y' + 6y = \delta(t)$, which has to be solved subect to $y(0) = y'(0) = 0$. Taking Laplace Transforms of this ODE, accounting for the initial conditions, yields,

$$(s^2 + 5s + 6)Y = 1 \qquad \Longrightarrow \qquad Y = \frac{1}{s^2 + 5s + 6} = \frac{1}{(s+2)(s+3)} = \frac{1}{s+2} - \frac{1}{s+3}. \tag{2.96}$$

Here we have used partial fractions to simplify the denominator, although the Convolution Theorem could also be used. Hence a simple Inverse Laplace Transform of these two components yields,

$$y = e^{-2t} - e^{-3t}. \tag{2.97}$$

One then needs to find z as well. Although one could also find a second order ODE for z which would then be transformed, this is a lot of work and very much a waste of time. Given that we have y, it is very much quicker to substititute this into the first 1st order ODE, namely $y' + y - z = 0$. to find z. Hence we get,

$$z = -e^{-2t} + 2e^{-3t}. \tag{2.98}$$

Method 2. This begins with taking the Laplace Transform of both ODEs simultaneously. This yields,

$$(s+1)Y - Z = 0, \qquad (s+4)Z + 2Y = 1. \tag{2.99}$$

At this point we may eliminate Z to obtain, not surprisingly,

$$Y = \frac{1}{(s+2)(s+3)}, \tag{2.100}$$

and from here on we follow the analysis of Method 1.

Method 3. This third method uses the language of matrix/vector equations. First we take Laplace Transforms to obtain Eq. (2.99), and then this is written in matrix/vector form.

$$\begin{pmatrix} s+1 & -1 \\ 2 & s+4 \end{pmatrix} \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.101}$$

Missing out a line of working, we may premultiply both sides by the inverse matrix, and this yields,

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \frac{1}{s^2 + 5s + 6} \begin{pmatrix} s+4 & 1 \\ -2 & s+1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad = \qquad \frac{1}{s^2 + 5s + 6} \begin{pmatrix} 1 \\ s+1 \end{pmatrix}, \tag{2.102}$$

where $s^2 + 5s + 6$ is the determinant of the matrix in Eq. (2.101). Now we need to use partial fractions:

$$Y = \frac{1}{s^2 + 5s + 6} = \frac{1}{s+2} - \frac{1}{s+3} \quad \Longrightarrow \quad y = e^{-2t} - e^{-3t} \tag{2.103}$$

and

$$Z = \frac{s+1}{s^2 + 5s + 6} = -\frac{1}{s+2} + \frac{2}{s+3} \quad \Longrightarrow \quad z = -e^{-2t} + 2e^{-3t}. \tag{2.104}$$

In the solutions, Eqs. (2.97) and (2.98), we notice that, while $y(0) = 0$, as required, we also have $z(0) = 1$ from this solution. This violation of the initial condition is to be expected because it is the z-equation that has the unit impulse as the forcing term.

2.13 Appendix: Proof of $\mathcal{L}[f * g] = F(s)G(s)$

This is for information and background only.

We start by writing down the definition of the Laplace Transform of $f * g$:

$$\mathcal{L}[f * g] = \int_0^\infty \Big[\int_0^t f(\tau)g(t-\tau)\,d\tau\Big]e^{-st}\,dt. \tag{2.105}$$

The proof proceeds by interchanging the order of integration, but this is not a rectangular region in (t, τ)–space and therefore it is not straightforward. In the inner integral, and for a given given value of t, τ varies between 0 and t. This is illustrated by the horizontal arrow in Fig. 2.11, and therefore the whole region of integration is given by the yellow shading.

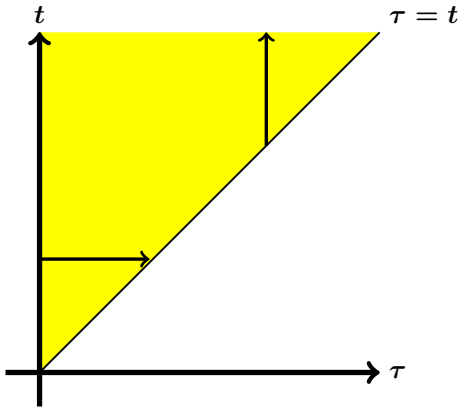


Figure 2.11. The yellow shading denotes the region of integration in Eq. (2.85).

When we interchange the order of integration, then we see from Fig. 2.11 that, for a fixed value of τ, t varies from τ to infinity (see the vertical arrow in Fig. 2.11). Hence Eq. (2.105) becomes,

$$\mathcal{L}[f * g] = \int_0^\infty \Big[\int_\tau^\infty f(\tau)g(t-\tau)e^{-st}\,dt\Big]\,d\tau. \tag{2.106}$$

Now the aim is to change the range of integration in the inner intergral so that the lower limit is zero. So we shall change variable from t to $\hat{t}$ where $\hat{t} = t - \tau$. Thus $d\hat{t} = dt$, and the lower limit in the inner integral will change from $t = \tau$ to $\hat{t} = 0$. So we get

$$\begin{aligned}
\mathcal{L}[f * g] &= \int_0^\infty \int_0^\infty f(\tau)g(\hat{t})e^{-s(\hat{t}+\tau)}\,d\hat{t}\,d\tau \\
&= \int_0^\infty \int_0^\infty \underbrace{f(\tau)e^{-s\tau}}_{\text{function of } \tau}\ \underbrace{g(\hat{t})e^{-s\hat{t}}}_{\text{function of } \hat{t}}\ d\hat{t}\,d\tau \\
&= \Big[\int_0^\infty f(\tau)e^{-s\tau}\,d\tau\Big] \times \Big[\int_0^\infty g(\hat{t})e^{-s\hat{t}}\,d\hat{t}\Big] \\
&= F(s)\,G(s).
\end{aligned} \tag{2.107}$$

Poetry in motion.

2.14 Standard results and theorems involving Laplace Transforms

I have listed below a set of results and theorems which are useful. Those theorems and transforms whose derivation could be in the exam are typeset in **red**. Not all of these results form part of the unit, so the items in the black font will either be quoted on the exam paper itself, or else will not be part of the exam.

Definition: $\boldsymbol{F(s) = \mathcal{L}[f(t)] = \int_0^\infty f(t)e^{-st}\,dt}$

$\boldsymbol{\mathcal{L}\left[\frac{df}{dt}\right] = sF(s) - f(0)}$

$\boldsymbol{\mathcal{L}\left[\frac{d^2f}{dt^2}\right] = s^2F(s) - f'(0) - sf(0)}$

$\boldsymbol{\mathcal{L}\left[\frac{d^3f}{dt^3}\right] = s^3F(s) - f''(0) - sf'(0) - s^2f(0)}$

$\boldsymbol{\mathcal{L}[tf(t)] = -\frac{dF}{ds}}$

$\boldsymbol{\mathcal{L}[t^2f(t)] = \frac{d^2F}{ds^2}}$

$\boldsymbol{\mathcal{L}[t^nf(t)] = (-1)^n\frac{d^nF}{dx^n}.}$

$\boldsymbol{\mathcal{L}\left[\int_0^t f(x)dx\right] = \frac{1}{s}F(s)}$

Scaling theorem: $\boldsymbol{\mathcal{L}[f(at)] = \frac{1}{a}F\left(\frac{s}{a}\right)}$

s-shift theorem: $\boldsymbol{\mathcal{L}[e^{-at}f(t)] = F(s+a)}$

t-shift theorem: $\boldsymbol{\mathcal{L}[H(t-a)f(t-a)] = e^{-sa}F(s)}$

Convolution theorem: $\boldsymbol{\mathcal{L}[f*g] = F(s)G(s)}$ where $\boldsymbol{f*g = \int_0^t f(\tau)g(t-\tau)d\tau = \int_0^t f(t-\tau)g(\tau)d\tau}$

Initial value theorem: $\boldsymbol{\lim_{s\to\infty} sF(s) = f(0)}$

Final value theorem: $\boldsymbol{\lim_{s\to 0} sF(s) = \lim_{t\to\infty} f(t)}$

2.15 Table of standard Laplace Transforms of functions

Each of the following could appear on the exam paper — all are derivable directly by applying the Laplace Transform definition, but some may also be derived using one of the shift theorems or the convolution theorem.

$f(t)$	$F(s)$	$f(t)$	$F(s)$
1	$\frac{1}{s}$	$e^{-at}\cos bt$	$\frac{s+a}{(s+a)^2+b^2}$
t	$\frac{1}{s^2}$	$\sinh bt$	$\frac{b}{s^2-b^2}$
t^2	$\frac{2}{s^3}$	$\cosh bt$	$\frac{s}{s^2-b^2}$
t^n	$\frac{n!}{s^{n+1}}$	$e^{-at}\sinh bt$	$\frac{b}{(s+a)^2-b^2}$
e^{at}	$\frac{1}{s-a}$	$e^{-at}\cosh bt$	$\frac{s+a}{(s+a)^2-b^2}$
e^{-at}	$\frac{1}{s+a}$	$t\sin bt$	$\frac{2bs}{(s^2+b^2)^2}$
$t^n e^{-at}$	$\frac{n!}{(s+a)^{n+1}}$	$t\cos bt$	$\frac{s^2-b^2}{(s^2+b^2)^2}$
$\sin bt$	$\frac{b}{s^2+b^2}$	$H(t)$	$\frac{1}{s}$
$\cos bt$	$\frac{s}{s^2+b^2}$	$H(t-a)$	$\frac{e^{-sa}}{s}$
$e^{-at}\sin bt$	$\frac{b}{(s+a)^2+b^2}$	$\delta(t-a)$	$e^{-as}\ (a>0)$

3 MATRICES

This chapter is designed to be an introduction to matrices and their manipulation, the evaluation of determinants, various methods of solving simultaneous linear equations with matrix methods, and finally, the determination of eigenvalues and eigenvectors which is set into the context of the solution of systems of ODEs.

3.1 Terminology

There is nothing which is better designed to undermine one's hard-earned reputation as a skillful engineer than the use of dodgy terminology or incorrect words. Here, specifically, I am talking about singulars and plurals related to the word, **matrix**. Here's a Table of the good and the bad:

	Correct	Incorrect
Singular	MATRIX	MATRICEE
Plural	MATRICES	MATRIXES

Given that many are studying Aerospace Engineering, the same may be said of **vortex**:

	Correct	Incorrect
Singular	VORTEX	VORTICEE
Plural	VORTICES	VORTEXES

and

	Correct	Incorrect
Singular	INDEX	INDICEE
Plural	INDICES	INDEXES

No doubt the same could be said for codex, suffix, prefix and appendix.

3.2 Matrix representation of simultaneous linear equations

Let us take the following pair of simultaneous equations for x and y:

$$\begin{aligned} 4x + 2y &= 3 \\ 3x - y &= 4 \end{aligned}. \tag{3.1}$$

These equations may be written in what is called matrix/vector form:

$$\underbrace{\begin{pmatrix} 4 & 2 \\ 3 & -1 \end{pmatrix}}_{\text{matrix}} \underbrace{\begin{pmatrix} x \\ y \end{pmatrix}}_{\text{vector}} = \underbrace{\begin{pmatrix} 3 \\ 4 \end{pmatrix}}_{\text{vector}}. \tag{3.2}$$

The square array of numbers is a **matrix**, while the other two entities are **vectors**. The numbers in the matrix are precisely the same as the various coefficients of x and y in Eq. (3.1), and that is the whole point. This vector/matrix notation is a shorthand notation and this will make life much easier for certain calculations.

The left hand side of Eq. (3.2) looks like a multiplication, but how does it work? Compare the following:

$$4x + 2y = 3 \qquad \text{and} \qquad \begin{pmatrix} \mathbf{4} & \mathbf{2} \\ 3 & -1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{3} \\ 4 \end{pmatrix} \tag{3.3}$$

and also

$$3x - y = 4 \qquad \text{and} \qquad \begin{pmatrix} 4 & 2 \\ \mathbf{3} & \mathbf{-1} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{3} \\ 4 \end{pmatrix}. \tag{3.4}$$

So the first entry in the right hand side vector is given by the scalar product of row 1 of the matrix (treated like a vector) and the vector, (x, y). Likewise, the second entry in the solution vector is given by the scalar product of row 2 of the matrix and (x, y). Each of these are illustrated by the characters in red in Eqs (3.3) and (3.4), respectively.

This shorthand notation applies for any number of equations in any number of unknowns. For example the following set of 3 equations in 4 unknowns and the matrix/vector equivalent:

$$\begin{aligned} x + 2y + 3z + 4w &= 0 \\ 3x + 2y + z &= 1 \\ 5x - 5y + z + 2w &= 2 \end{aligned} \qquad \begin{pmatrix} 1 & 2 & 3 & 4 \\ \mathbf{3} & \mathbf{2} & \mathbf{1} & \mathbf{0} \\ 5 & -5 & 1 & 2 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \end{pmatrix}. \tag{3.5}$$

Note that we are not interested at this stage in whether such systems of equations have solutions, although Eq. (3.2) has one solution and Eq. (3.5) has infinitely many, given that we have 3 equations in 4 unknowns. Once again, the multiplication process has been illustrated using a red font.

3.3 Classification of matrices

The shape of a matrix is important for various reasons. In Eq. (3.2) the matrix is square, has 2 rows and 2 columns, and is said to be a $\mathbf{2} \times \mathbf{2}$ matrix (pronounced "two by two"). The matrix in Eq. (3.5) has 3 rows and 4 columns, and hence is called a $\mathbf{3} \times \mathbf{4}$ matrix. A very large proportion of matrices in science and engineering tend to be square in shape simply because they represent the solution of n equations in n unknowns.

The order, *row – column*, is inviolable and must always be followed since that is the universal convention. One possible mnemonic is to say that it is **R**eally **C**lever or maybe it may be remembered as the **R**ight **C**onvention. So it's perfect if you are called **R**hodri **C**harles but not if you are **C**eridwen **R**hys.

This R/C convention is particularly important when specifying matrices with generalised entries using subscripts, such as,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}. \tag{3.6}$$

Here the entry a_{ij} corresponds to **row** i and **column** j. Sometimes it is necessary to use a comma to ensure that the different subscripts are clear: $a_{i,j}$.

Matrices are either square or rectangular as we saw above, but they never take other shapes, such as triangular or hexagonal. However, square matrices can exhibit a very wide variety of patterns, some of which are due to what the application was that was translated into matrix form.

A square matrix may be described as being **lower triangular** if the entries above the **main diagonal** (i.e. top left to bottom right) are zero, such as in

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}. \tag{3.7}$$

A matrix which is **upper triangular** has nonzero entries below the main diagonal, such as in

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{pmatrix}. \tag{3.8}$$

Such matrices must be square.

Later, when we consider systematic methods for solving simultaneous equations, the first step will be to develop ways of transforming a fully-populated matrix to upper triangular form.

A **diagonal matrix** has zero entries both above and below the main diagonal. An example is

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{pmatrix}. \tag{3.9}$$

In a sense a diagonal matrix is simultaneously lower-triangular and upper-triangular.

Some of the entries on the main diagonal may be also be zero, but if all of them are, then we merely have the 4×4 **zero matrix**:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.10}$$

One special diagonal matrix is the **identity matrix** where all the main diagonal entries are equal to $\mathbf{1}$:

$$\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix}. \tag{3.11}$$

Often we adopt a shorthand notation, $\boldsymbol{I_n}$, for the $\boldsymbol{n \times n}$ identity matrix. Hence the above $\mathbf{4 \times 4}$ matrix may be written as $\boldsymbol{I_4}$.

The identity matrix is so-called because multiplication with a vector leaves the vector unchanged:

$$\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix}. \tag{3.12}$$

A **symmetric matrix** is a square matrix with the property that $\boldsymbol{a_{ij} = a_{ji}}$. An example is

$$\begin{pmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{2} & \mathbf{5} & \mathbf{10} & \mathbf{101} \\ \mathbf{3} & \mathbf{10} & \mathbf{-100} & \boldsymbol{\pi} \\ \mathbf{4} & \mathbf{101} & \boldsymbol{\pi} & \boldsymbol{\sqrt{2}} \end{pmatrix}, \tag{3.13}$$

where the red font shows that row 2 and column 2 are identical. For such matrices row $\boldsymbol{n}$ and column $\boldsymbol{n}$ are identical.

An **antisymmetric matrix** is a square matrix with the property that $\boldsymbol{a_{ij} = -a_{ji}}$ and the diagonal entries are therefore zero. An example is

$$\begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{-2} & \mathbf{9} \\ \mathbf{-1} & \mathbf{0} & \mathbf{3} & \mathbf{8} \\ \mathbf{2} & \mathbf{-3} & \mathbf{0} & \mathbf{7} \\ \mathbf{-9} & \mathbf{-8} & \mathbf{-7} & \mathbf{0} \end{pmatrix}, \tag{3.14}$$

where the values in row $\boldsymbol{n}$ correspond to the negative of the values in column $\boldsymbol{n}$. Row 2 and column 2 are coloured as an example of that. Terms on the diagonal are zero because $\boldsymbol{a_{ii} = -a_{ii}}$ implies that $\boldsymbol{a_{ii} = 0}$.

The **transpose** of a matrix, $\boldsymbol{A}$, is formed by interchanging its rows and columns, and it is denoted by $\boldsymbol{A^T}$. For example, if

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} \\ \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{9} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix} \quad \text{then} \quad \boldsymbol{A^T} = \begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{6} & \mathbf{1} \\ \mathbf{2} & \mathbf{4} & \mathbf{7} & \mathbf{1} \\ \mathbf{3} & \mathbf{5} & \mathbf{8} & \mathbf{1} \\ \mathbf{4} & \mathbf{6} & \mathbf{9} & \mathbf{1} \end{pmatrix}. \tag{3.15}$$

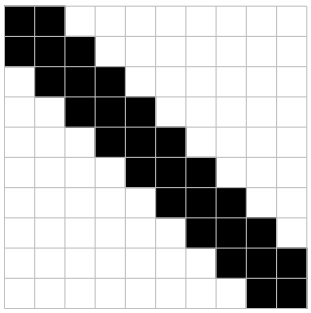
Therefore the transpose of an upper triangular matrix is a lower triangular matrix and vice versa. The transpose of a symmetric matrix is equal to the original matrix ($\boldsymbol{A = A^T}$). The transpose of an antisymmetric matrix satisfies $\boldsymbol{A^T = -A}$. And if a second transposition is undertaken then we return to the orginal matrix: $\boldsymbol{(A^T)^T = T}$.

The **tridiagonal matrix** has nonzero entries on its main diagonal, and also on its immediate **superdiagonal** and **subdiagonal**:

$$\begin{pmatrix} b_1 & c_1 & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & 0 \\ 0 & a_3 & b_3 & c_3 & 0 \\ 0 & 0 & a_4 & b_4 & c_4 \\ 0 & 0 & 0 & a_5 & b_5 \end{pmatrix}. \tag{3.16}$$

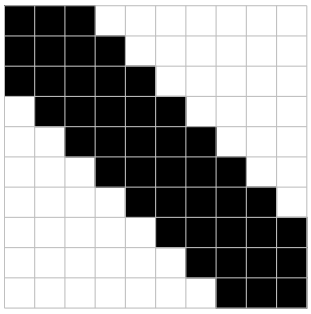
Here, the main diagonal elements are denoted by b_j, the superdiagonal elements by c_j and the subdiagonal elements by a_j. The subscript tells the row in which the element resides. Such matrices arise when using certain numerical methods to solve 2nd boundary value ODEs.

We may generalise this presentation a little further using the following diagram which represents the structure of a 10×10 matrix.



Schematic of a tridiagonal matrix

The nonzero entries are illustrated in black and the zero entries in white. Similarly, a pentadiagonal matrix takes the form,



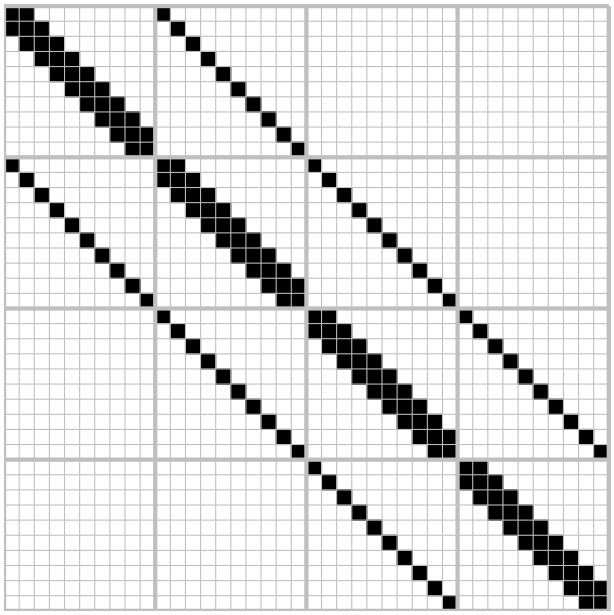
Schematic of a pentadiagonal matrix

A method which solves a fully-populated matrix/vector equation will often be very inefficient when faced with a tridiagonal matrix, and therefore more efficient schemes which take account of the pattern of coefficients are used. The numerical aspect of this will be covered in ME20021 Modelling Techniques 2 next year.

Finally, a **block tridiagonal matrix** takes the form,

$$\begin{pmatrix} \boldsymbol{B_1} & \boldsymbol{C_1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{A_2} & \boldsymbol{B_2} & \boldsymbol{C_2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A_3} & \boldsymbol{B_3} & \boldsymbol{C_3} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A_4} & \boldsymbol{B_4} & \boldsymbol{C_4} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A_5} & \boldsymbol{B_5} \end{pmatrix}. \tag{3.17}$$

Here, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ represent $M \times M$ square matrices. Such a pattern will be formed when solving certain 2nd order PDEs, again in ME20021 Modelling Techniques 2. For such applications the $\boldsymbol{B}$-submatrices tend to be tridiagonal, while the $\boldsymbol{A}$ and $\boldsymbol{C}$ submatrices are frequently diagonal or, at worst, tridiagonal. Here is a representation of such a matrix with a $\mathbf{4} \times \mathbf{4}$ arrangement of blocks each of which is a $\mathbf{10} \times \mathbf{10}$ square matrix, and it will therefore represent $\mathbf{40}$ equations in $\mathbf{40}$ unknowns:



Schematic of a block tridiagonal matrix

For these systems the number of nonzero entries will be incredibly small, and they are often termed **sparse matrices**. These cannot really be solved directly, but rather through efficient iterative methods. In such cases it is entirely impractical to store all of the elements of the matrix on the computer. An example would be where a Partial Differential Equation is to be solved on a grid of $\boldsymbol{N} \times \boldsymbol{M}$ points. The above matrix would then be an $\boldsymbol{NM} \times \boldsymbol{NM}$ matrix with $\boldsymbol{N^2M^2}$ elements. The number of nonzero elements may be shown to be only $\mathbf{5}\boldsymbol{NM} - \mathbf{2}(\boldsymbol{M} + \boldsymbol{N})$. To put that into perspective, if we set $\boldsymbol{N} = \boldsymbol{M} = \mathbf{100}$, then the matrix has a total of $\mathbf{10^8}$ elements of which only $\mathbf{39\,600}$ are nonzero.

Other patterns or types arise such as the periodic tridiagonal matrix, block pentadiagonal matrices, Hessenberg matrices, orthogonal matrices, singular matrices, the Hessian matrix, rotation matrices and the Wronskian.

Note: The fields of fluid mechanics and solid mechanics occasionally make use of what are called tensors. A matrix may be viewed as a two dimensional form of a tensor. But tensors are frequently higher dimensional versions of a matrix. For instance a third order tensor may be regarded as a cubic array of numbers, and a thorough derivation of the Navier-Stokes equations for fluid flows requires the manipulation of fourth order tensors. Very fortunately this aspect is well outside of the scope of your degree programme!

3.4 Matrix addition

Two matrices may be added by adding corresponding elements. For example,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{pmatrix}. \tag{3.18}$$

The reason why we might wish to do this is if we need to add two different sets of simultaneous equations together, such as $A\underline{x} + B\underline{x} = (A+B)\underline{x}$. We will have some examples later which arise naturally.

Clearly such a addition process works only if the matrices have the same number of rows and the same number of columns as one another. Thus these two $n \times m$ matrices may be added and therefore such matrices are called **compatible with respect to addition**. If not then they are **incompatible with respect to addition**.

Matrix subtraction follows the same procedure of manipulating the corresponding terms.

Example 3.1: Add the matrices $A = \begin{pmatrix} 1 & 0 \\ -3 & 5 \end{pmatrix}$ and $B = \begin{pmatrix} 2 & -5 \\ 3 & 1 \end{pmatrix}$ and also find $A - B$.

The sum of A and B is $A + B = \begin{pmatrix} 3 & -5 \\ 0 & 6 \end{pmatrix}$.

The difference is $A - B = \begin{pmatrix} -1 & 5 \\ -6 & 4 \end{pmatrix}$.

3.5 Matrix multiplication

If the manner in which two matrices are added is straightforward, the method by which they are multiplied is quite strange at first. Actually it follows in the same way as for matrix/vector multiplication which was introduced earlier, but we will now derive the manner in which it must be done, which also indicates one reason why one might need to multiply two matrices.

Suppose we have the following matrix/vector equation,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \tag{3.19}$$

where the vector of y values is given in terms of a vector of x values by the relation,

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{3.20}$$

then what is the equation for the three x values?

In terms of a shorthand matrix notation we could write the above analysis as,

$$A\underline{y} = \underline{r} \quad \text{and} \quad \underline{y} = B\underline{x} \qquad \Longrightarrow \qquad AB\underline{x} = \underline{r}. \tag{3.21}$$

So clearly we need to know how to multiply two matrices together.

This may be done in a longhand tedious way as follows. We may rewrite both systems of equations in longhand fashion to get

$$\begin{aligned} a_{11}y_1 + a_{12}y_2 + a_{13}y_3 &= r_1 \\ a_{21}y_1 + a_{22}y_2 + a_{23}y_3 &= r_2 \\ a_{31}y_1 + a_{32}y_2 + a_{33}y_3 &= r_3 \end{aligned} \tag{3.22}$$

$$\begin{aligned} y_1 &= b_{11}x_1 + b_{12}x_2 + b_{13}x_3 \\ y_2 &= b_{21}x_1 + b_{22}x_2 + b_{23}x_3 \\ y_3 &= b_{31}x_1 + b_{32}x_2 + b_{33}x_3 \end{aligned} \tag{3.23}$$

Now we shall substitute the expressions for y_j given in Eq. (3.23) into Eq. (3.22). After a fair while we get,

$$\begin{aligned} (a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31})x_1 + (a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32})x_2 + (a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33})x_3 &= r_1 \\ (a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31})x_1 + (a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32})x_2 + (a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33})x_3 &= r_2 \\ (a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31})x_1 + (a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32})x_2 + (a_{31}b_{13} + a_{32}b_{23} + a_{33}b_{33})x_3 &= r_3 \end{aligned} \tag{3.24}$$

or, in matrix notation,

$$\underbrace{\begin{pmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} & a_{11}b_{13}+a_{12}b_{23}+a_{13}b_{33} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} & a_{21}b_{13}+a_{22}b_{23}+a_{23}b_{33} \\ a_{31}b_{11}+a_{32}b_{21}+a_{33}b_{31} & a_{31}b_{12}+a_{32}b_{22}+a_{33}b_{32} & a_{31}b_{13}+a_{32}b_{23}+a_{33}b_{33} \end{pmatrix}}_{AB} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}. \tag{3.25}$$

Given Eq. (3.23), this equation is identical to,

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}}_{AB} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}. \tag{3.26}$$

We may extend this to larger square matrices: if A is the $n \times n$ matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{3.27}$$

and B is defined in a similar manner, then the product, called C where $C = AB$, when written fully is very lengthy, but the ij entry of the product (i.e. the entry in row i and column j) is

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}, \tag{3.28}$$

which may be written more compactly as

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}. \tag{3.29}$$

It now worth going back to see, for every product of a and b in Eq. (3.25), the roles played by the outer subscripts $a_{ik}b_{kj}$, and the inner subscripts, $a_{ik}b_{kj}$. This summation notation is especially useful if one were to encode a matrix multiplication subroutine in Fortran or C.

In the product AB we say that B is **premultiplied** by A, or, equivalently, that A is **postmultiplied** by B.

3.6 Commutivity, Associativity and Distributivity

These are three bits of jargon which describe concepts which are not only fairly simple, but also very important.

Commutivity is concerned with the order in which an arithmetical operation is undertaken. That normal addition is commutative is exemplified by the fact that $\mathbf{1}+\mathbf{2}=\mathbf{2}+\mathbf{1}$. Given the close relation between normal addition and matrix addition, namely that it involves corresponding terms in two matrices, matrix addition is also commutative, i.e. if two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are compatible in addition then $\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{B}+\boldsymbol{A}$ is always true.

If that seemed straightforward, it might be surprising to learn that matrix multiplication is not commutative in general. Here are two examples:

Example 3.2: Multiply the matrices $\boldsymbol{A}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}$ and $\boldsymbol{B}=\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}$, finding both $\boldsymbol{AB}$ and $\boldsymbol{BA}$.

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{0} & \mathbf{1}\end{pmatrix}, \qquad \boldsymbol{BA}=\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{-2} & \mathbf{-1}\end{pmatrix}. \tag{3.30}$$

Just to be secure, the following four equations highlight this vector scalar product way of calculating the product of two matrices:

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{0} & \mathbf{1}\end{pmatrix} \qquad \boldsymbol{R_1},\ \boldsymbol{C_1}$$

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{0} & \mathbf{1}\end{pmatrix} \qquad \boldsymbol{R_1},\ \boldsymbol{C_2}$$

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{0} & \mathbf{1}\end{pmatrix} \qquad \boldsymbol{R_2},\ \boldsymbol{C_1}$$

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{3} & \mathbf{2}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{-1} & \mathbf{1}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{0} & \mathbf{1}\end{pmatrix} \qquad \boldsymbol{R_2},\ \boldsymbol{C_2}$$

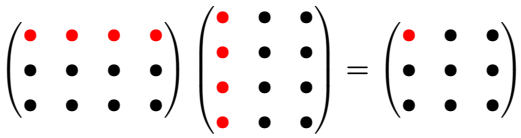
Example 3.3: Find $\boldsymbol{AB}$ and $\boldsymbol{BA}$ for $\boldsymbol{A}=\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{1}\\ \mathbf{1} & \mathbf{0} & \mathbf{-1}\end{pmatrix}$ and $\boldsymbol{B}=\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\end{pmatrix}$

$$\boldsymbol{AB}=\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{1}\\ \mathbf{1} & \mathbf{0} & \mathbf{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\end{pmatrix}=\begin{pmatrix}\mathbf{3} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{6} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{2}\end{pmatrix}, \tag{3.31}$$

$$\boldsymbol{BA}=\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2} & \mathbf{1}\\ \mathbf{1} & \mathbf{0} & \mathbf{-1}\end{pmatrix}=\begin{pmatrix}\mathbf{3} & \mathbf{-1} & \mathbf{1}\\ \mathbf{-1} & \mathbf{5} & \mathbf{-1}\\ \mathbf{1} & \mathbf{-1} & \mathbf{3}\end{pmatrix} \tag{3.32}$$

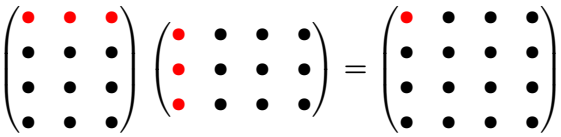
Normal scalar multiplication is commutative, of course, since $\boldsymbol{xy}=\boldsymbol{yx}$. The scalar product of two vectors is also commutative, but not the vector product of two vectors. Likewise we find that the result of mixing of the ingredients of a cake and then cooking them is not the same as when they are cooked first before being mixed.

A further complication with matrices is that, while the product $\boldsymbol{AB}$ may be calculated, it is not necessarily true that $\boldsymbol{BA}$ has the same dimensions or even that it exists! For example,



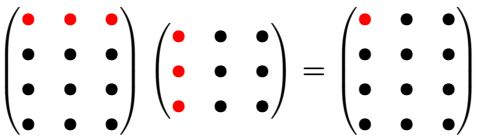
$$\underset{3\times4}{\boldsymbol{A}} \times \underset{4\times3}{\boldsymbol{B}} = \underset{3\times3}{\boldsymbol{AB}} \tag{3.33}$$

and



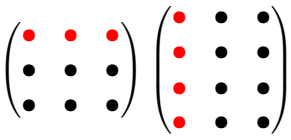
$$\underset{4\times3}{\boldsymbol{B}} \times \underset{3\times4}{\boldsymbol{A}} = \underset{4\times4}{\boldsymbol{BA}}. \tag{3.34}$$

show that these alternative products are different sizes. In the following cases,



$$\underset{4\times3}{\boldsymbol{B}} \times \underset{3\times3}{\boldsymbol{C}} = \underset{4\times3}{\boldsymbol{BC}} \tag{3.35}$$

and



$$= \text{incompatible} \qquad \underset{3\times3}{\boldsymbol{C}} \times \underset{4\times3}{\boldsymbol{B}} = \text{incompatible} \tag{3.36}$$

show that one product exists but the other one doesn't. However, if the matrices $\boldsymbol{D}$ and $\boldsymbol{E}$ are both square matrices of the same size, then both $\boldsymbol{DE}$ and $\boldsymbol{ED}$ will exist and have the same sizes as $\boldsymbol{D}$ and $\boldsymbol{E}$, but they are generally different.

One interesting consequence of this is that a rotation of a vector about the x-axis (which is equivalent to a suitable multiplication by a matrix) followed by a rotation about the y-axis (another matrix) does not lead to same resultant vector when the rotations are undertaken in the opposite order.:w

There are some special cases when $\boldsymbol{AB} = \boldsymbol{BA}$; these include (i) when $\boldsymbol{B} = \boldsymbol{I}$, the identity matrix; and (ii) when $\boldsymbol{AB} = \boldsymbol{I}$. In this last case $\boldsymbol{B}$ is the inverse matrix of $\boldsymbol{A}$ and it is denoted by $\boldsymbol{A^{-1}}$. We have $\boldsymbol{AA^{-1}} = \boldsymbol{A^{-1}A} = \boldsymbol{I}$; we will say a little more about this later.

Associativity is concerned with bracketing. For scalar addition and multiplication, associativity means that $(a + b) + c = a + (b + c)$ and $(ab)c = a(bc)$. Matrix addition and matrix multiplication are both associative, subject to the requirement of compatibility. Therefore, for compatible matrices, the following are always true:

$$(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}), \qquad (\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}). \tag{3.37}$$

Distributivity is concerned the swapping the order in which addition and multiplication take place. Matrix operations are also distributive, by which is meant that

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}. \tag{3.38}$$

Note: This has been a dry and dull-as-ditchwater selection of information transfer material. Unfortunately most of it is foundational for what follows. I have given only a small number of Examples; only one addition is really required, and there's going to be sufficient practice of the scalar-product-of-vectors approach to the multiplication of matrices on the problem sheets. Thus far, the chief danger will be arithmetical errors.

3.7 Determinants

The determinant of a matrix is a numerical value which is associated with that matrix, and it is defined only for square matrices. For those who already know a little about matrices and have some experience of determinants, it may seem a strange topic to introduce right now. Well, we will need it later when covering eigenvalues and eigenvectors, and so determinants must come before that. In addition, the matrix topic in the next subsection may, on some occasions, be usefully explained using the language of determinants. There is also a strong link with vectors last semester in terms of the geometric understanding of what determinants represent. So we're going for it now rather than later.

If the determinant of a matrix is zero, then a matrix/vector equation involving that matrix will have a unique solution if the determinant is nonzero. Otherwise it has either an infinite number of solutions or none at all. Later we will discuss how these observations fit into the geometry of planes in 3D space, and then a zero determinant can be understood in those terms.

3.7.1 Determinants of 2×2 matrices

Determinants arise naturally when solving matrix/vector equations. If we take the general case of a $\mathbf{2 \times 2}$ matrix/vector equation,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}, \tag{3.39}$$

or, in shorthand notation,

$$A\underline{x} = \underline{r}, \tag{3.40}$$

or in longhand algebraic notation,

$$\begin{aligned} ax + by &= r \\ cx + dy &= s, \end{aligned} \tag{3.41}$$

then the solution of the equivalent pair of simultaneous equations may be written in the form

$$x = \frac{dr - bs}{ad - bc}, \qquad y = \frac{-cr + as}{ad - bc}, \tag{3.42}$$

where one could have simply solved the pair of simultanous equations using one's favourite technique. This may be written in matrix/vector form as follows:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}. \tag{3.43}$$

Here $(ad - bc)$ is known as the **determinant of A**, and it may be written in many forms,

$$\det(A) \quad \text{or} \quad |A| \quad \text{or} \quad \begin{vmatrix} a & b \\ c & d \end{vmatrix} \quad \text{or} \quad \det \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{3.44}$$

Equation (3.43) may also be written as

$$\underline{x} = A^{-1}\underline{r} \qquad \text{where} \qquad A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \tag{3.45}$$

where $\boldsymbol{A^{-1}}$ is the **inverse matrix** of $\boldsymbol{A}$.

Example 3.4: Find the inverse of $\boldsymbol{A} = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$, and show that $\boldsymbol{AA^{-1} = A^{-1}A = I}$.

We will use Eq. (3.45) to provide an illustration. The determinant of $\boldsymbol{A}$ is $\boldsymbol{(3\times 3 - 5\times 2) = 1}$. Hence,

$$\boldsymbol{A^{-1}} = \begin{pmatrix} \boldsymbol{2} & \boldsymbol{-3} \\ \boldsymbol{-3} & \boldsymbol{5} \end{pmatrix}, \tag{3.46}$$

and so

$$\boldsymbol{AA^{-1}} = \begin{pmatrix} \boldsymbol{5} & \boldsymbol{3} \\ \boldsymbol{3} & \boldsymbol{2} \end{pmatrix}\begin{pmatrix} \boldsymbol{2} & \boldsymbol{-3} \\ \boldsymbol{-3} & \boldsymbol{5} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{A^{-1}A} = \begin{pmatrix} \boldsymbol{2} & \boldsymbol{-3} \\ \boldsymbol{-3} & \boldsymbol{5} \end{pmatrix}\begin{pmatrix} \boldsymbol{5} & \boldsymbol{3} \\ \boldsymbol{3} & \boldsymbol{2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1} \end{pmatrix}. \tag{3.47}$$

This illustrates the general case that, not only is the multiplication of a matrix by its inverse a commutative operation, but it always yields the identity matrix.

3.7.2 Geometric interpretation of the 2×2 determinant

We may rewrite the solution of Eq. (3.39) which is given in Eq. (3.43) in the form,

$$\begin{aligned} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} &= \frac{\boldsymbol{1}}{\boldsymbol{ad-bc}}\begin{pmatrix} \boldsymbol{d} & \boldsymbol{-b} \\ \boldsymbol{-c} & \boldsymbol{a} \end{pmatrix}\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix} \\ &= \frac{\boldsymbol{1}}{\boldsymbol{ad-bc}}\begin{pmatrix} \boldsymbol{rd-sb} \\ \boldsymbol{-cr+sa} \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{\boldsymbol{rd-sb}}{\boldsymbol{ad-bc}} \\ \\ \dfrac{\boldsymbol{-cr+sa}}{\boldsymbol{ad-bc}} \end{pmatrix} \end{aligned} \tag{3.48}$$

and hence we may write the solution is the form known as Cramer's rule,

$$\boldsymbol{x} = \frac{\begin{vmatrix} \boldsymbol{r} & \boldsymbol{b} \\ \boldsymbol{s} & \boldsymbol{d} \end{vmatrix}}{\begin{vmatrix} \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{c} & \boldsymbol{d} \end{vmatrix}} \qquad \boldsymbol{y} = \frac{\begin{vmatrix} \boldsymbol{a} & \boldsymbol{r} \\ \boldsymbol{c} & \boldsymbol{s} \end{vmatrix}}{\begin{vmatrix} \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{c} & \boldsymbol{d} \end{vmatrix}} \tag{3.49}$$

where each solution component is given by a ratio of determinants. With regard to the numerators, the red typesetting emphasizes the fact that, for the first component of the solution vector ($\boldsymbol{x}$ here), involves the first column of $\boldsymbol{A}$ being replaced by the right hand side vector, $\underline{\boldsymbol{r}}$. The second solution component then involves the second column in the same way.

Now we are in a position to understand the role played by the determinant of $\boldsymbol{A}$. First, we shall rearrange the two scalar equations which are represented by Eq. (3.39) in the classical '$\boldsymbol{y = mx + c}$' form:

$$\boldsymbol{y = (r - ax)/b}, \qquad \boldsymbol{y = (s - cx)/d}. \tag{3.50}$$

Figure 3.1 gives a typical sketch of such lines.

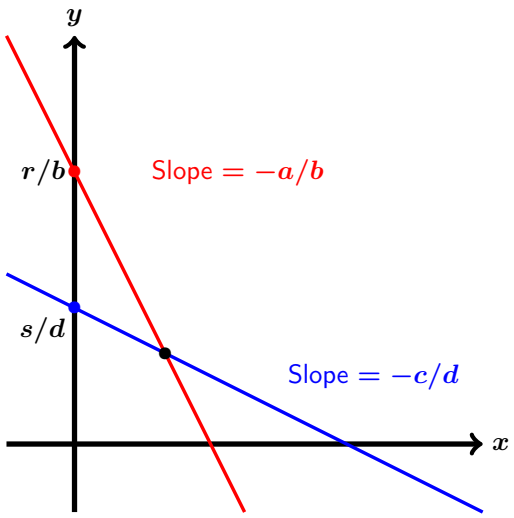


Figure 3.1. Illustrating two intersecting lines when $\mathbf{det}\boldsymbol{A} \neq \mathbf{0}$.

Geometrically, it is clear that this system of two equations in two unknowns will have a solution if the two slopes are different. Therefore, the condition for the existence of a solution is that,

$$a/b \neq c/d \quad \Longrightarrow \quad ad - bc \neq 0 \quad \Longrightarrow \quad \mathbf{det}\boldsymbol{A} \neq \mathbf{0}. \tag{3.51}$$

When $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ then either the lines are parallel, or else they are coincident, as shown in Fig. 3.2.

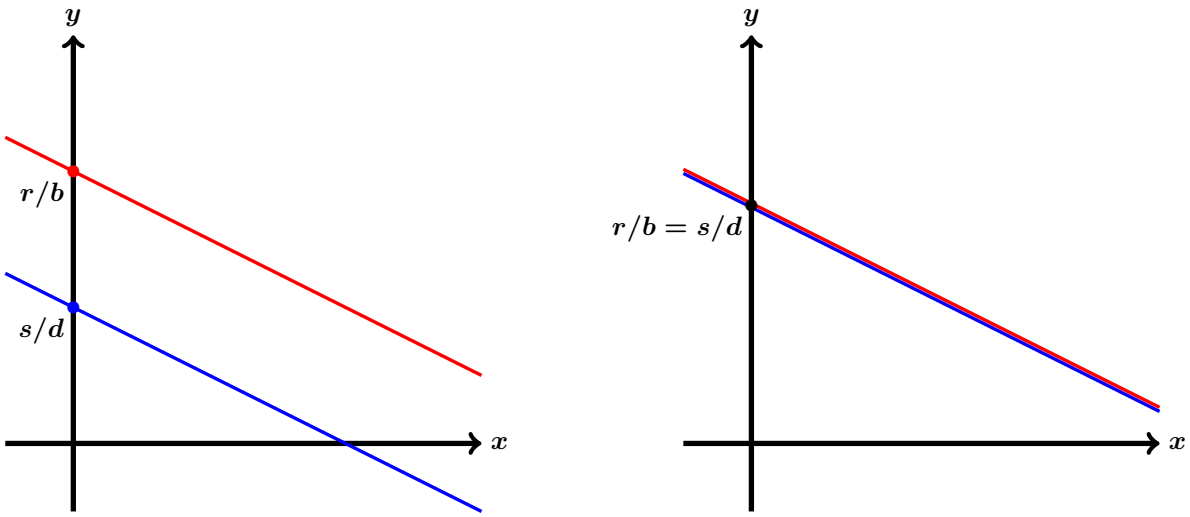


Figure 3.2. The two cases for which $\mathbf{det}\boldsymbol{A} = \mathbf{0}$.

When the lines are parallel the determinants which form the numerators in Eq. (3.49) are nonzero, meaning that Cramer's rule yields a formula which is effectively of a nonzero/zero form. When the lines are coincident, then the Cramer's rule formula gives zero/zero. The following examples show the full range of possibilities.

Example 3.5: Use Cramer's rule to solve the following three equations.

$$\text{(i)}\ \begin{pmatrix} \mathbf{3} & \mathbf{4} \\ \mathbf{1} & \mathbf{2} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \mathbf{8} \\ \mathbf{4} \end{pmatrix}; \qquad \text{(ii)}\ \begin{pmatrix} \mathbf{2} & \mathbf{4} \\ \mathbf{1} & \mathbf{2} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \mathbf{8} \\ \mathbf{5} \end{pmatrix}; \qquad \text{(iii)}\ \begin{pmatrix} \mathbf{2} & \mathbf{4} \\ \mathbf{1} & \mathbf{2} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \mathbf{8} \\ \mathbf{4} \end{pmatrix}.$$

These three systems look very similar indeed, but appearances can be deceptive. We'll call the matrix $\boldsymbol{A}$ in each case, for convenience.

Case (i): Here $\mathbf{det}\boldsymbol{A} = \mathbf{2}$, and Cramer's rule becomes

$$x = \tfrac{1}{2}\begin{vmatrix} \mathbf{8} & \mathbf{4} \\ \mathbf{4} & \mathbf{2} \end{vmatrix} = \mathbf{0}, \qquad y = \tfrac{1}{2}\begin{vmatrix} \mathbf{3} & \mathbf{8} \\ \mathbf{1} & \mathbf{4} \end{vmatrix} = \mathbf{2}.$$

The lines intersect and have a unique solution.

Case (ii): In this case $\mathbf{det}\boldsymbol{A} = \mathbf{0}$. Blindly following Cramer's rule we obtain,

$$x = \tfrac{1}{0}\begin{vmatrix} \mathbf{8} & \mathbf{4} \\ \mathbf{5} & \mathbf{2} \end{vmatrix} = -\tfrac{4}{0}, \qquad y = \tfrac{1}{0}\begin{vmatrix} \mathbf{2} & \mathbf{8} \\ \mathbf{1} & \mathbf{5} \end{vmatrix} = \tfrac{2}{0}.$$

The nonzero numerators tell us that the lines are parallel.

Case (iii): In this case $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ again. Another blind following of Cramer's rule now yields,

$$x = \tfrac{1}{0}\begin{vmatrix} \mathbf{8} & \mathbf{4} \\ \mathbf{4} & \mathbf{2} \end{vmatrix} = \tfrac{0}{0}, \qquad y = \tfrac{1}{0}\begin{vmatrix} \mathbf{2} & \mathbf{8} \\ \mathbf{1} & \mathbf{4} \end{vmatrix} = \tfrac{0}{0}.$$

Thus these two lines are coincident and therefore they 'intersect' along the whole line. We may write down a parametric vectorial form for this solution as follows. If we let $x = \alpha$, then $y = \mathbf{2} - \frac{1}{2}\alpha$. These translate into the vector form,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{2} \end{pmatrix} + \alpha \begin{pmatrix} \mathbf{1} \\ -1/2 \end{pmatrix}. \tag{3.52}$$

The vector which is multiplied by α is the direction of the line.

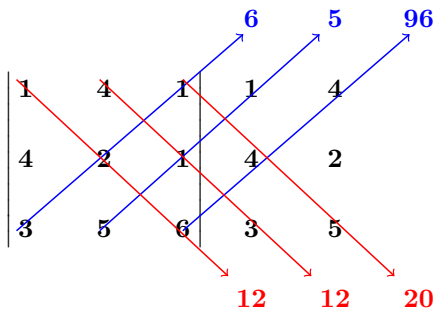
Note: The message of this story is that $\mathbf{det}\boldsymbol{A} \neq \mathbf{0}$ is the condition which guarantees that $\boldsymbol{A}\underline{x} = \underline{r}$ has a unique solution. This result also applies for larger determinants.

3.7.3 Determinants of 3×3 matrices

For a $\mathbf{3} \times \mathbf{3}$ matrix the corresponding process of solving the three simultaneous equations is much more tedious and is very lengthy. It eventually results in a solution which looks like Eq. (3.48) but where each of the three components consists of a ratio of sums of products of three terms. The denominators are identical and are the determinant of the matrix.

We'll look at this more generally in a moment, but there is an often-used quick way of evaluating $\mathbf{3} \times \mathbf{3}$ determinants and this is illustrated using the matrix,

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{1} & \mathbf{4} & \mathbf{1} \\ \mathbf{4} & \mathbf{2} & \mathbf{1} \\ \mathbf{3} & \mathbf{5} & \mathbf{6} \end{pmatrix}:$$



(3.53)

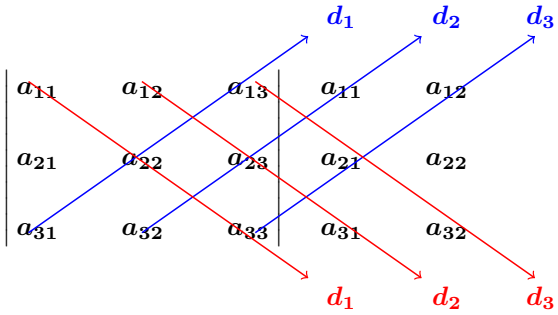
Hence the determinant is

$$(12 + 12 + 20) - (6 + 5 + 96) = -63.$$

Now we will use a more general notation for this 3×3 matrix, and define

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{3.54}$$

The determinant of $\boldsymbol{A}$ is given by



(3.55)

Hence the determinant is

$$\det(\boldsymbol{A}) = \underbrace{a_{11}a_{22}a_{33}}_{d_1} + \underbrace{a_{12}a_{23}a_{31}}_{d_2} + \underbrace{a_{13}a_{21}a_{32}}_{d_3} - \underbrace{a_{13}a_{22}a_{31}}_{d_1} - \underbrace{a_{11}a_{23}a_{32}}_{d_2} - \underbrace{a_{12}a_{21}a_{33}}_{d_3}. \tag{3.56}$$

This may be arranged into various convenient forms for evluation. There are six and they correspond to what is called an expansion about a named row or column. Two such are as follows.

Expansion about the 1st row:

$$\begin{aligned} \det(\boldsymbol{A}) &= a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ &= a_{11}(\text{minor of } a_{11}) - a_{12}(\text{minor of } a_{12}) + a_{13}(\text{minor of } a_{13}). \end{aligned}$$

Expansion about the 2nd column:

$$\begin{aligned} \det(\boldsymbol{A}) &= -a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{22}\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} - a_{32}\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \\ &= -a_{21}(\text{minor of } a_{21}) + a_{22}(\text{minor of } a_{22}) - a_{23}(\text{minor of } a_{23}). \end{aligned}$$

There are four further possibilities namely, expansions about the second and third rows and about the first and third columns. The term, **minor**, refers to the suitable $\mathbf{2} \times \mathbf{2}$ determinant. Specifically, the minor of a_{ij} is the determinant formed from the $\mathbf{2} \times \mathbf{2}$ matrix which arises when row i and column j are removed from the original $\mathbf{3} \times \mathbf{3}$ matrix. A more pictorial way of looking at this may be afforded by the following.

Expansion about the 1st row:

$$\det \boldsymbol{A} = \quad a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \quad - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} \quad + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{3.57}$$

$$\begin{matrix} \boldsymbol{a_{11}} & a_{12} & a_{13} \\ a_{21} & \boldsymbol{a_{22}} & \boldsymbol{a_{23}} \\ a_{31} & \boldsymbol{a_{32}} & \boldsymbol{a_{33}} \end{matrix} \qquad \begin{matrix} a_{11} & \boldsymbol{a_{12}} & a_{13} \\ \boldsymbol{a_{21}} & a_{22} & \boldsymbol{a_{23}} \\ \boldsymbol{a_{31}} & a_{32} & \boldsymbol{a_{33}} \end{matrix} \qquad \begin{matrix} a_{11} & a_{12} & \boldsymbol{a_{13}} \\ \boldsymbol{a_{21}} & \boldsymbol{a_{22}} & a_{23} \\ \boldsymbol{a_{31}} & \boldsymbol{a_{32}} & a_{33} \end{matrix}$$

where the blue components correspond to the minor, while the grey components are those which have been *removed*.

Note: that this formula is essentially the same as the one which we used to find the vector product of two vectors back in ME10304 Mathematics 1 where the first row consisted of the three unit vectors. I say *essentially* for although this formula gives exactly the same result, it has been organised slightly differently here. Back in ME10304 we made use of "ghostly" columns in order to maintain precisely the same formula throughout, but here we haven't done so. The discrepancy between the methods of evaluation is taken care of by the lone minus sign in Eq. (3.57).

Expansion about the 2nd column:

$$\det \boldsymbol{A} = \quad -a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} \quad + a_{22}\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \quad - a_{32}\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \tag{3.58}$$

$$\begin{matrix} a_{11} & \boldsymbol{a_{12}} & a_{13} \\ \boldsymbol{a_{21}} & a_{22} & \boldsymbol{a_{23}} \\ \boldsymbol{a_{31}} & a_{32} & \boldsymbol{a_{33}} \end{matrix} \qquad \begin{matrix} \boldsymbol{a_{11}} & a_{12} & \boldsymbol{a_{13}} \\ a_{21} & \boldsymbol{a_{22}} & a_{23} \\ \boldsymbol{a_{31}} & a_{32} & \boldsymbol{a_{33}} \end{matrix} \qquad \begin{matrix} \boldsymbol{a_{11}} & a_{12} & \boldsymbol{a_{13}} \\ \boldsymbol{a_{21}} & a_{22} & \boldsymbol{a_{23}} \\ a_{31} & \boldsymbol{a_{32}} & a_{33} \end{matrix}$$

The individual plus and minus signs in these last two formulae appear to have two different patterns, but they do follow a universal **checkerboard pattern**:

$$\begin{pmatrix} \boxed{+} & - & + & - & + & \cdots \\ - & + & - & + & - & \cdots \\ + & - & + & - & + & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix}, \tag{3.59}$$

where there is always a "+" anchored in the top left entry.

Note that each term of a square matrix also has what is called a cofactor. The minor is solely the $\mathbf{2} \times \mathbf{2}$ determinant, but the cofactor also accounts for the sign as given by the checkerboard pattern, Thus the cofactor of a_{ij} is given by $(-\mathbf{1})^{i+j}$ multiplied by the minor.

So there are potentially six different ways of writing out the determinant of a $\mathbf{3 \times 3}$ matrix. One other reason behind giving these various alternatives is that greater speed in evaluating determinants is obtained when there is a zero entry in the matrix. For example the following determinant may be expanded about the first column to give,

$$\begin{vmatrix} \mathbf{5} & \mathbf{0} & \mathbf{1} \\ \mathbf{3} & \mathbf{1} & \mathbf{5} \\ \mathbf{2} & \mathbf{1} & \mathbf{-1} \end{vmatrix} = \mathbf{5}\begin{vmatrix} \mathbf{1} & \mathbf{5} \\ \mathbf{1} & \mathbf{-1} \end{vmatrix} - \mathbf{3}\begin{vmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} \end{vmatrix} + \mathbf{2}\begin{vmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{5} \end{vmatrix} = \mathbf{5(-6) - 3(-1) + 2(-1) = -29.} \tag{3.60}$$

But if we expand about the second column (or, alternatively the first row) then we need to evaluate only two $\mathbf{2 \times 2}$ determinants:

$$\begin{vmatrix} \mathbf{5} & \mathbf{0} & \mathbf{1} \\ \mathbf{3} & \mathbf{1} & \mathbf{5} \\ \mathbf{2} & \mathbf{1} & \mathbf{-1} \end{vmatrix} = \mathbf{1}\begin{vmatrix} \mathbf{5} & \mathbf{1} \\ \mathbf{2} & \mathbf{-1} \end{vmatrix} - \mathbf{1}\begin{vmatrix} \mathbf{5} & \mathbf{1} \\ \mathbf{3} & \mathbf{5} \end{vmatrix} = \mathbf{1(-7) - 1(22)} = \mathbf{-29.} \tag{3.61}$$

So, less work....interesting....we'll squirrel that thought away for a moment.

3.7.4 Determinants of 4×4 matrices

As with $\mathbf{2 \times 2}$ and $\mathbf{3 \times 3}$ determinants, the determinant of a $\mathbf{4 \times 4}$ matrix may be expressed in terms of quotients of sums, but now these sums are of the products of four terms. The denominator of this version of Eq. (3.48) is the determinant of the matrix, but this is now far too lengthy to write down in full. But just as the general expression for the determinant of a $\mathbf{3 \times 3}$ matrix may be written as the sum of three $\mathbf{2 \times 2}$ determinant, so the general expression for the determinant of a $\mathbf{4 \times 4}$ matrix may be written as the sum of four $\mathbf{3 \times 3}$ determinants. Here are two examples.

Expansion about the first row:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$$

$$= +a_{11}\begin{vmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} & a_{24} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} - a_{14}\begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{vmatrix}$$

$$\begin{vmatrix} \mathbf{a_{11}} & a_{12} & a_{13} & a_{14} \\ a_{21} & \mathbf{a_{22}} & \mathbf{a_{23}} & \mathbf{a_{24}} \\ a_{31} & \mathbf{a_{32}} & \mathbf{a_{33}} & \mathbf{a_{34}} \\ a_{41} & \mathbf{a_{42}} & \mathbf{a_{43}} & \mathbf{a_{44}} \end{vmatrix} \quad \begin{vmatrix} a_{11} & \mathbf{a_{12}} & a_{13} & a_{14} \\ \mathbf{a_{21}} & a_{22} & \mathbf{a_{23}} & \mathbf{a_{24}} \\ \mathbf{a_{31}} & a_{32} & \mathbf{a_{33}} & \mathbf{a_{34}} \\ \mathbf{a_{41}} & a_{42} & \mathbf{a_{43}} & \mathbf{a_{44}} \end{vmatrix} \quad \begin{vmatrix} a_{11} & a_{12} & \mathbf{a_{13}} & a_{14} \\ \mathbf{a_{21}} & \mathbf{a_{22}} & a_{23} & \mathbf{a_{24}} \\ \mathbf{a_{31}} & \mathbf{a_{32}} & a_{33} & \mathbf{a_{34}} \\ \mathbf{a_{41}} & \mathbf{a_{42}} & a_{43} & \mathbf{a_{44}} \end{vmatrix} \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} & \mathbf{a_{14}} \\ \mathbf{a_{21}} & \mathbf{a_{22}} & \mathbf{a_{23}} & a_{24} \\ \mathbf{a_{31}} & \mathbf{a_{32}} & \mathbf{a_{33}} & a_{34} \\ \mathbf{a_{41}} & \mathbf{a_{42}} & \mathbf{a_{43}} & a_{44} \end{vmatrix}$$

Expansion about the second column:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$$

$$= -a_{12}\begin{vmatrix} a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} + a_{22}\begin{vmatrix} a_{11} & a_{13} & a_{14} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} - a_{32}\begin{vmatrix} a_{11} & a_{12} & a_{14} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} + a_{42}\begin{vmatrix} a_{11} & a_{13} & a_{14} \\ a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33} & a_{34} \end{vmatrix}$$

$$\begin{matrix} a_{11} & \mathbf{a_{12}} & a_{13} & a_{14} \\ \mathbf{a_{21}} & a_{22} & \mathbf{a_{23}} & \mathbf{a_{24}} \\ \mathbf{a_{31}} & a_{32} & \mathbf{a_{33}} & \mathbf{a_{34}} \\ \mathbf{a_{41}} & a_{42} & \mathbf{a_{43}} & \mathbf{a_{44}} \end{matrix} \qquad \begin{matrix} \mathbf{a_{11}} & a_{12} & \mathbf{a_{13}} & \mathbf{a_{14}} \\ a_{21} & \mathbf{a_{22}} & a_{23} & a_{24} \\ \mathbf{a_{31}} & a_{32} & \mathbf{a_{33}} & \mathbf{a_{34}} \\ \mathbf{a_{41}} & a_{42} & \mathbf{a_{43}} & \mathbf{a_{44}} \end{matrix} \qquad \begin{matrix} \mathbf{a_{11}} & a_{12} & \mathbf{a_{13}} & \mathbf{a_{14}} \\ \mathbf{a_{21}} & a_{22} & \mathbf{a_{23}} & \mathbf{a_{24}} \\ a_{31} & \mathbf{a_{32}} & a_{33} & a_{34} \\ \mathbf{a_{41}} & a_{42} & \mathbf{a_{43}} & \mathbf{a_{44}} \end{matrix} \qquad \begin{matrix} \mathbf{a_{11}} & a_{12} & \mathbf{a_{13}} & \mathbf{a_{14}} \\ \mathbf{a_{21}} & a_{22} & \mathbf{a_{23}} & \mathbf{a_{24}} \\ \mathbf{a_{31}} & a_{32} & \mathbf{a_{33}} & \mathbf{a_{34}} \\ a_{41} & \mathbf{a_{42}} & a_{43} & a_{44} \end{matrix}$$

These are but two examples of eight possible expansions. Again the option to be chosen will attempt to minimise the number of determinant evaluations. Similarly the determinant of a $\mathbf{5 \times 5}$ matrix may be expanded in terms of five $\mathbf{4 \times 4}$ determinants and so on. Again, and in all cases, one MUST take into account the checkboard pattern of signs given in Eq. (3.59).

3.7.5 A fast method for finding determinants

From the above it has to be clear that the number of determinant evaluations will increase alarmingly as the size of the matrix increases. If one attempts to evaluate how many *multiplications* need to be undertaken, then we get the following. A $\mathbf{2 \times 2}$ determinant requires only 2 multiplications, but a $\mathbf{3 \times 3}$ needs 9, a $\mathbf{4 \times 4}$ needs 40 and a $\mathbf{6 \times 6}$ needs 1236. Practically, this means that I can guarantee that you won't get a $\mathbf{6 \times 6}$ determinant in the examination! Here's a Table of how many multiplications are needed for differently-sized $\boldsymbol{N \times N}$ determinants (and you may wish to try to reproduce this for yourself as an exercise!).

$\boldsymbol{N}$	Multiplications
2	**2**
3	**9**
4	**40**
5	**205**
6	**1 236**
7	**8 659**
8	**69 280**
9	**623 529**
10	**6 235 300**

Believe it or not it is possible to show that $\boldsymbol{(e-1)n!}$ is a good approximation to the number of multiplications for an $\boldsymbol{N \times N}$ matrix when $\boldsymbol{N}$ is large. So clearly this method must be improved.

The solution to our woes may be illustrated by the following manipulation of a $\mathbf{2 \times 2}$ determinant. By way of a preamble, we have already seen that the presence of zero entries will increase the speed with which a determinant

may be evaluated, and therefore we will find a way to introduce extra zeros in the determinant in order to do this.

Note: The following is a derivation which forms a proof-of-concept; this proof won't be examined.

Given the general $\mathbf{2} \times \mathbf{2}$ determinant,

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \tag{3.62}$$

let us, seemingly at random, add γ times the values in the second column to the respective values in the first column. We get this:

$$\begin{vmatrix} a + \gamma b & b \\ c + \gamma d & d \end{vmatrix} = (a + \gamma b)d - (c + \gamma d)b = ad - bc. \tag{3.63}$$

In other words, the value of a determinant is unaltered by adding any multiple of one column to another column. This is also true when adding any multiple of one row to another row.

Here is a proof that this also works for $\mathbf{3} \times \mathbf{3}$ determinants.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} + \gamma a_{21} & a_{32} + \gamma a_{22} & a_{33} + \gamma a_{23} \end{vmatrix}$$

$$= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} + \gamma a_{22} & a_{33} + \gamma a_{23} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} + \gamma a_{21} & a_{33} + \gamma a_{23} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} + \gamma a_{21} & a_{32} + \gamma a_{22} \end{vmatrix}$$

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

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

It is more difficult to demonstrate, but it is also true, for $\boldsymbol{N} \times \boldsymbol{N}$ matrices. So the devious plan will be to undertake as many of these row and/or column manipulations as we need to reduce the overall number of multiplications that needs to be undertaken.

Warning: For the purposes of determinant evaluation, never multiply a column or a row by some constant because you will then alter the value of the determinant. This is why:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \qquad \text{but} \qquad \begin{vmatrix} \alpha a & b \\ \alpha c & d \end{vmatrix} = \alpha(ad - bc). \tag{3.64}$$

It is also wise never to think about swapping two rows or two columns. Odd-numbered rows may be swapped and even-numbered rows may be swapped with impunity, but if an odd-numbered one and an even-numbered one are swapped then the determinant is multiplied by -1. The same is true for columns. My view is that it is better not to undertake any swapping around.

Example. The following 4×4 determinant is a good example of why I have introduced this fast method. Note how the row and column manipulations are designed to get as many zeros as is possible.

$$\begin{vmatrix} \mathbf{1} & 2 & 3 & 4 \\ \mathbf{2} & 2 & -2 & 4 \\ \mathbf{1} & 2 & 1 & 4 \\ \mathbf{4} & 3 & 2 & 1 \end{vmatrix} \qquad \begin{matrix} C_2 \leftarrow C_2 - C_1 \\ C_3 \leftarrow C_3 + C_1 \\ C_4 \leftarrow C_4 - 2C_1 \end{matrix}$$

$$= \begin{vmatrix} 1 & 1 & 4 & 2 \\ \mathbf{2} & 0 & 0 & 0 \\ 1 & 1 & 2 & 2 \\ 4 & -1 & 6 & -7 \end{vmatrix} \qquad \begin{matrix} \text{now to expand about } \boldsymbol{R_2} \\ \text{the red } \mathbf{2} \text{ needs a minus} \end{matrix}$$

$$= -(\mathbf{2}) \begin{vmatrix} 1 & 4 & 2 \\ 1 & 2 & 2 \\ \mathbf{-1} & \mathbf{6} & \mathbf{-7} \end{vmatrix} \qquad \begin{matrix} R_1 \leftarrow R_1 + R_3 \\ R_2 \leftarrow R_2 + R_3 \end{matrix} \qquad (3.65)$$

$$= -2 \begin{vmatrix} 0 & 10 & -5 \\ 0 & 8 & -5 \\ \mathbf{-1} & 6 & -7 \end{vmatrix} \qquad \begin{matrix} \text{expand about } \boldsymbol{C_1} \\ \text{the blue "}\mathbf{-1}\text{" needs a plus} \end{matrix}$$

$$= +(-2) \times (\mathbf{-1}) \begin{vmatrix} 10 & -5 \\ 8 & -5 \end{vmatrix}$$

$$= 2 \times (-10) = -20.$$

Note: This is not the only way in which the row and column manipulations may be done. In the third line of this analysis, where we have just arrived at the 3×3 determinant, a close look at rows 1 and 2 show that there are two elements in common. So a slightly quicker route through the remaining analysis would involve subtracting row 1 from row 2 or vice versa, and then to eliminate another row and column. It is also worth attempting a few different ways from the start as a means of practice: the answer should always be the same!

3.8 Geometric interpretation of the 3×3 determinant

We have already seen Cramer's rule for 2×2 determinant, but it is repeated here with a subscripted notation, for convenience. If

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \qquad (3.66)$$

then Cramer's rule solution is given by,

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \qquad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}. \qquad (3.67)$$

Cramer's rule also applies for larger square matrices in $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$ systems. For $\mathbf{3} \times \mathbf{3}$ matrices the analogous formula is,

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}} \qquad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}} \qquad x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}. \tag{3.68}$$

Having seen this, it is quite clear how we may extend the formula to larger systems such $\mathbf{4} \times \mathbf{4}$ and so on.

Like the 2D case considered earlier, Cramer's rule will not work when $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ but what, geometrically, does this mean?

In ME10304 Mathematics 1, we saw that three vectors, $\underline{\boldsymbol{a}}$, $\underline{\boldsymbol{b}}$ and $\underline{\boldsymbol{c}}$, are independent when $\underline{\boldsymbol{a}} \cdot \underline{\boldsymbol{b}} \times \underline{\boldsymbol{c}} \neq \mathbf{0}$. Dependent vectors correspond to $\underline{\boldsymbol{a}} \cdot \underline{\boldsymbol{b}} \times \underline{\boldsymbol{c}} = \mathbf{0}$ and they lie in a common plane. Let us translate this condition into a determinantal form:

$$\begin{aligned}
\mathbf{0} &= \underline{\boldsymbol{a}} \cdot \underline{\boldsymbol{b}} \times \underline{\boldsymbol{c}} \\
&= \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \cdot \begin{vmatrix} \underline{\boldsymbol{i}} & \underline{\boldsymbol{j}} & \underline{\boldsymbol{k}} \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} && \text{by definition of the vector product} \\
&= (a_1\underline{\boldsymbol{i}} + a_2\underline{\boldsymbol{j}} + a_3\underline{\boldsymbol{k}}) \cdot \left[\underline{\boldsymbol{i}} \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - \underline{\boldsymbol{j}} \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + \underline{\boldsymbol{k}} \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix} \right] && \text{using Eq. (3.57)} \\
&= a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix} && \text{using the scalar products} \\
&= \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}.
\end{aligned} \tag{3.69}$$

We also recall from ME10304 Mathematics 1 that an equation of the form,

$$a_1x + a_2y + a_3z = \text{constant} \tag{3.70}$$

is the equation of a plane surface in 3D and that (a_1, a_2, a_3) is the direction which is normal to the plane. Therefore $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ when the normals to the three planes lie only in a 2D plane. Perhaps this is difficult to visualize! Hopefully the following will help.

When $\mathbf{det}\boldsymbol{A} \neq \mathbf{0}$, then any two of the planes will have a line of intersection (think of a piece of paper with a single fold, the fold acting as the line of intersection) and a third plane then intersects with the fold at one point. This forms the unique solution.

When $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ there are many ways in which this can happen:

- Two planes are parallel and the third intersects both.
- Two planes are identical and the third intersects both.
- All three planes are parallel.
- Two planes are identical and the third parallel.
- All three planes are identical.
- Like a Toblerone packet, i.e. all three intersect pairwise, but the intersections are parallel. See Fig 3.3, below.

Figure 3.3. Showing an instance of three planes which do not cross at a point.

Once we reach $\mathbf{4} \times \mathbf{4}$ determinants and larger, it becomes increasingly difficult to visualize what $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ means geometrically. The main message is that $\mathbf{det}\boldsymbol{A} \neq \mathbf{0}$ means that an equation of the form, $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$, will have a unique solution. On the other hand, when $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ then either there is no solution or there is an infinite number of solutions (see Figs. 3.1 and 3.2 for the analogous 2D case). This also means that the inverse matrix, $\boldsymbol{A}^{-1}$, does not exist when $\mathbf{det}\boldsymbol{A} = \mathbf{0}$ and in such cases $\boldsymbol{A}$ is called a **singular matrix**.

3.9 The Gaussian Elimination method

The amount of work involved in solving $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$ using Cramer's Rule is excessive. Even if each determinant were to use the row/column manipulations derived above then it would still remain awkward to apply, and indeed it is never used for large systems of simultaneous linear equations. Its main use here has really been to interpret the determinant geometrically.

The Gaussian Elimination method does not rely on the evaluation of determinants at all and is a systematic way of solving the equivalent system of simultaneous equations. Briefly, this method adds and subtracts multiples of one equation to another, the chief intermediate stage being when the matrix has achieved upper-triangular form.

The simplest thing to do is simply to launch into an example. The exposition will undertaken in both the simultaneous equations form and in what is called the **augmented matrix** form in which we manipulate the

matrix entries whilst still understanding that they represent coefficients in the original equations. We shall consider a $\mathbf{3} \times \mathbf{3}$ example.

Example 3.6: Solve the following $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$ system.

$$\begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{2} \\ \mathbf{2} & \mathbf{5} & \mathbf{-1} \\ \mathbf{3} & \mathbf{1} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \mathbf{6} \\ \mathbf{6} \\ \mathbf{5} \end{pmatrix}. \tag{3.71}$$

We proceed as follows.

Row manipulations	Equations							Augmented matrix form.			
	$\boldsymbol{x}$	$+$	$\mathbf{3}\boldsymbol{y}$	$+$	$\mathbf{2}\boldsymbol{z}$	$=$	$\mathbf{6}$	$\mathbf{1}$	$\mathbf{3}$	$\mathbf{2}$	$\mathbf{6}$
	$\mathbf{2}\boldsymbol{x}$	$+$	$\mathbf{5}\boldsymbol{y}$	$-$	$\boldsymbol{z}$	$=$	$\mathbf{6}$	$\mathbf{2}$	$\mathbf{5}$	$\mathbf{-1}$	$\mathbf{6}$
	$\mathbf{3}\boldsymbol{x}$	$+$	$\boldsymbol{y}$	$+$	$\boldsymbol{z}$	$=$	$\mathbf{5}$	$\mathbf{3}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{5}$
$\boldsymbol{R_1}$	$\boldsymbol{x}$	$+$	$\mathbf{3}\boldsymbol{y}$	$+$	$\mathbf{2}\boldsymbol{z}$	$=$	$\mathbf{6}$	$\mathbf{1}$	$\mathbf{3}$	$\mathbf{2}$	$\mathbf{6}$
$\boldsymbol{R_2 - 2R_1}$		$-$	$\boldsymbol{y}$	$-$	$\mathbf{5}\boldsymbol{z}$	$=$	$\mathbf{-6}$	$\mathbf{0}$	$\mathbf{-1}$	$\mathbf{-5}$	$\mathbf{-6}$
$\boldsymbol{R_3 - 3R_1}$		$-$	$\mathbf{8}\boldsymbol{y}$	$-$	$\mathbf{5}\boldsymbol{z}$	$=$	$\mathbf{-13}$	$\mathbf{0}$	$\mathbf{-8}$	$\mathbf{-5}$	$\mathbf{-13}$
$\boldsymbol{R_1}$	$\boldsymbol{x}$	$+$	$\mathbf{3}\boldsymbol{y}$	$+$	$\mathbf{2}\boldsymbol{z}$	$=$	$\mathbf{6}$	$\mathbf{1}$	$\mathbf{3}$	$\mathbf{2}$	$\mathbf{6}$
$\boldsymbol{R_2}$		$-$	$\boldsymbol{y}$	$-$	$\mathbf{5}\boldsymbol{z}$	$=$	$\mathbf{-6}$	$\mathbf{0}$	$\mathbf{-1}$	$\mathbf{-5}$	$\mathbf{-6}$
$\boldsymbol{R_3 - 8R_2}$				$+$	$\mathbf{35}\boldsymbol{z}$	$=$	$\mathbf{35}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{35}$	$\mathbf{35}$

The matrix has now been transformed into upper-triangular form and this marks the end of the **elimination phase** of the Gaussian Elimination method.

With regard to the red annotations on the left, these keep a record of what has been done, and either the row has been left untouched, or else it has had a multiple of the first row (in the first instance) or the second row (in the second instance) added to it in order to achieve zeros below the main diagonal.

Given that the matrix is in upper-triangular form, it is now straightfoward to continue with the Gaussian Elimination method by undertaking the **back substitution** phase. The final row gives $\mathbf{35}\boldsymbol{z} = \mathbf{35}$ and hence $\boldsymbol{z} = \mathbf{1}$. The penultimate row has $-\boldsymbol{y} - \mathbf{5}\boldsymbol{z} = \mathbf{-6}$ and so $\boldsymbol{y} = \mathbf{1}$. Finally, the first row gives $\boldsymbol{x} + \mathbf{3}\boldsymbol{y} + \mathbf{2}\boldsymbol{z} = \mathbf{6}$ and hence $\boldsymbol{x} = \mathbf{1}$. Therefore the final solution is,

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \tag{3.72}$$

Note: The general aim in practise should be to use solely the augmented matrix form, rather than to write out the simultaneous equations in parallel. But it must be understood that any row manipulations are, in effect, equivalent to adding or subtracting equations. Thus **Gaussian Elimination uses only row manipulations**, unlike when finding determinants. In addition, the multiplication of a whole row in the augmented matrix form by a constant (perhaps for numerical convenience to avoid fractions) is allowed because the whole equation is being multiplied.

Note: Reduction to upper-triangular form is a convention, and reduction to lower-triangular form also works as do other exotic versions. However, I wish us to maintain convention particularly in the exams.

Example 3.6 was a systematic approach to solving an $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$ system. However, if one required the solution of the same equation but with a different $\underline{\boldsymbol{r}}$, then much time can be saved by finding the two solutions simultaneously. The augmented matrix notation lends itself easily to this. We'll give it a go now.

Example 3.7: Solve the following two $\boldsymbol{A}\underline{\boldsymbol{x}} = \underline{\boldsymbol{r}}$ systems simultaneously.

$$\begin{pmatrix} 1 & 3 & 2 \\ 2 & 5 & -1 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 5 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 1 & 3 & 2 \\ 2 & 5 & -1 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 9 \\ 7 \end{pmatrix}.$$

We proceed as follows.

	x	$+$	$3y$	$+$	$2z$	$=$	6	5	1	3	2	6	5
	$2x$	$+$	$5y$	$-$	z	$=$	6	9	2	5	-1	6	9
	$3x$	$+$	y	$+$	z	$=$	5	7	3	1	1	5	7
R_1	x	$+$	$3y$	$+$	$2z$	$=$	$6,$	5	1	3	2	6	5
$R_2 - 2R_1$		$-$	y	$-$	$5z$	$=$	$-6,$	-1	0	-1	-5	-6	-1
$R_3 - 3R_1$		$-$	$8y$	$-$	$5z$	$=$	$-13,$	-8	0	-8	-5	-13	-8
R_1	x	$+$	$3y$	$+$	$2z$	$=$	$6,$	5	1	3	2	6	5
R_2		$-$	y	$-$	$5z$	$=$	$-6,$	-1	0	-1	-5	-6	-1
$R_3 - 8R_2$				$+$	$35z$	$=$	$35,$	0	0	0	35	35	0

Again the use of back-substitution yields the solution for the blue data:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \tag{3.73}$$

There is no reason why one couldn't have as many right hand sides as one wishes and be able to solve a large number of equations all of which have the same matrix. But one use of multiple right-hand sides is to find the inverse of a matrix. So if we set the equation, $\boldsymbol{AX} = \boldsymbol{I}$, where all three matrices are square, then $\boldsymbol{X} = \boldsymbol{A}^{-1}$.

Example 3.8: Use Gaussian Elimination to find the inverse of the matrix,

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix} \tag{3.74}$$

We proceed as follows.

	x	$-$	$2y$	$+$	z	$=$	$1,$	$0,$	0	1	-2	1	1	0	0
	x			$-$	z	$=$	$0,$	$1,$	0	1	0	-1	0	1	0
	x	$+$	y	$+$	z	$=$	$0,$	$0,$	1	1	1	1	0	0	1
R_1	x	$-$	$2y$	$+$	z	$=$	$1,$	$0,$	0	1	-2	1	1	0	0
$R_2 - R_1$		$+$	$2y$	$-$	$2z$	$=$	$-1,$	$1,$	0	0	2	-2	-1	1	0
$R_3 - R_1$		$+$	$3y$			$=$	$-1,$	$0,$	1	0	3	0	-1	0	1
R_1	x	$-$	$2y$	$+$	z	$=$	$1,$	$0,$	0	1	-2	1	1	0	0
R_2		$+$	$2y$	$-$	$2z$	$=$	$-1,$	$1,$	0	0	2	-2	-1	1	0
$R_3 - \frac{3}{2}R_2$				$+$	$3z$	$=$	$\frac{1}{2},$	$-\frac{3}{2},$	$1,$	0	0	3	$\frac{1}{2}$	$-\frac{3}{2}$	1

The three separated solutions, duly colour-coded, are

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \tfrac{1}{6}\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \tfrac{1}{6}\begin{pmatrix} 3 \\ 0 \\ -3 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \tfrac{1}{6}\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \tag{3.75}$$

and hence the final solution is,

$$\boldsymbol{A^{-1}} = \tfrac{1}{6}\begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{2} \\ \mathbf{-2} & \mathbf{0} & \mathbf{2} \\ \mathbf{1} & \mathbf{-3} & \mathbf{2} \end{pmatrix}. \tag{3.76}$$

It is good practice to check that this result is correct. It should be that both $\boldsymbol{AA^{-1}}$ and $\boldsymbol{A^{-1}A}$ are equal to $\boldsymbol{I}$. For the sake of avoiding fractions in the typesetting, we'll multiply $\boldsymbol{A}$ by $\mathbf{6}\boldsymbol{A^{-1}}$ both ways around. So we obtain,

$$\begin{pmatrix} \mathbf{1} & \mathbf{-2} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix}\begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{2} \\ \mathbf{-2} & \mathbf{0} & \mathbf{2} \\ \mathbf{1} & \mathbf{-3} & \mathbf{2} \end{pmatrix} = \begin{pmatrix} \mathbf{6} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{6} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{6} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{2} \\ \mathbf{-2} & \mathbf{0} & \mathbf{2} \\ \mathbf{1} & \mathbf{-3} & \mathbf{2} \end{pmatrix}\begin{pmatrix} \mathbf{1} & \mathbf{-2} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix}, \tag{3.77}$$

as required.

Note: There is a not-so-speedy cousin of the Gaussian Elimination method called the Gauss-Jordan method. This method follows the elimination phase but a second elimination phase after which the matrix has been reduced to a diagonal form.

Note: Gaussian Elimination is true workhorse and has been put to use for systems of simultaneous equations which can easily consist of matrices which may be as large as $\mathbf{1000} \times \mathbf{1000}$, but it is quite possible for them to be orders of magnitude larger than that.

3.10 Eigenvalues and eigenvectors

3.10.1 Applications

This is one of the most important parts of matrix theory especially for the engineer and scientist. A very large number of applications are associated with eigenvalues and eigenvectors. These include the determination of the modes and frequencies of vibration of structures (e.g. bridges, buildings and aircraft), acoustics (e.g. pressure waves in auditoriums, resonant frequencies of musical instruments), stability theory (how fast does a fluid need to move before the laminar motion becomes unstable, e.g. cloud rolls in the sky, ocean waves), Google search algorithms. These applications can often involve huge matrices and therefore computer methods are required. The present part of the unit will deal mostly with $\mathbf{2 \times 2}$ and $\mathbf{3 \times 3}$ matrices since they can be studied in a fairly short timescale but this does give some idea of what happens with larger systems.

Given the emphasis on the solution of linear constant-coefficient ODEs in this unit, I shall embed the rationale for finding eigenvalues and eigenvectors within the context of solving such ODEs or, rather, of solving systems of such ODEs.

3.10.2 Terminology.

The *eigen* in *eigenvalue* is a German word that Google Translate translates as *peculiar*. I reckon that this is the almost old-fashioned meaning of *belonging to* or *being associated with*, rather than the more modern/common usage where the object or behaviour is strange or odd in some way. This thought is reinforced by the fact that some textbooks call eigenvalues and eigenvectors, characteristic values and characteristic vectors, respectively.

3.10.3 Example 3.9 — preamble.

In a sense we have already found some eigenvalues and eigenvectors when we solved for complementary functions earlier but they were hidden. For example, if we wished to solve the ODE,

$$\boldsymbol{x'' - 4x' + 3x = 0,} \tag{3.78}$$

then we would substitute $\boldsymbol{x = e^{\lambda t}}$, then find that

$$\boldsymbol{\lambda^2 - 4\lambda + 3 = 0} \tag{3.79}$$

is the auxiliary (or indicial or *characteristic*) equation. Then we would go on to find that $\boldsymbol{\lambda = 1, 3}$ are the two values, while the corresponding functions are $\boldsymbol{e^t}$ and $\boldsymbol{e^{3t}}$. The final solution is then

$$\boldsymbol{x = Ae^t + Be^{3t},} \tag{3.80}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary constants. There are no vectors in sight here, none at all! However, and as mentioned above, the two values of $\boldsymbol{\lambda}$ are sometimes called characteristic values and this hints that our single ODE analysis fits into the eigenvalue idea.

Now I am going to be a little sneaky here because I would like to use the following pair of first order ODEs for our first foray into the finding of eigenvalues and eigenvectors:

$$\boldsymbol{x' = 2x + y, \qquad y' = x + 2y.} \tag{3.81}$$

With a little bit of hassle you might be able to manipulate these equations to remove either $\boldsymbol{y}$ or $\boldsymbol{x}$ from between them, and this will result in precisely Eq. (3.78) above. Therefore we could substitute the solution for $\boldsymbol{x}$ which is given by Eq. (3.80) into the ODE in Eq. (3.81) to find that

$$\boldsymbol{y = -Ae^t + Be^{3t}.} \tag{3.82}$$

In matrix notation the solution is $\begin{pmatrix} x \\ y \end{pmatrix} = \boldsymbol{A}e^{t}\begin{pmatrix} 1 \\ -1 \end{pmatrix} + \boldsymbol{B}e^{3t}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

So we have already solved the system of coupled first order equations given in Eq. (3.81) but we have done so in a way that is very inefficient when faced with much larger systems. This is one reason why matrix eigenvalue theory is important. For the following 'derivation' of the method of finding eigenvalues and eigenvectors, not only do we already know what the solution is, but I am also going to do it in two different ways in parallel. The left hand side doesn't use anything that resembles a matrix, while the right hand side gives the analysis using the language of matrices.

3.10.4 Example 3.9.

Now we shall solve the pair of equations given in Eq. (3.81) in two different ways simultaneously.

$\begin{aligned} x' &= 2x + y \\ y' &= x + 2y \end{aligned}$	$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$
Let $\begin{aligned} x &= Xe^{\lambda t} \\ y &= Ye^{\lambda t} \end{aligned}$ X and Y are arbitrary but they will turn out to be related to one another.	Let $\begin{pmatrix} x \\ y \end{pmatrix} = e^{\lambda t}\begin{pmatrix} X \\ Y \end{pmatrix}$ X and Y are arbitrary but they will turn out to be related to one another.
$\Longrightarrow \quad \begin{aligned} \lambda X &= 2X + Y \\ \lambda Y &= X + 2Y \end{aligned}$ Have cancelled $e^{\lambda t}$ both sides	$\Longrightarrow \quad \lambda\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$ Have cancelled $e^{\lambda t}$ both sides
$\Longrightarrow \quad \begin{aligned} Y &= (\lambda - 2)X \\ X &= (\lambda - 2)Y \end{aligned}$ — (E1) Now to eliminate Y	$\Longrightarrow \quad \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ — (E2) Now we need a nonzero solution for X and Y and hence...
$\begin{aligned} &\Longrightarrow \quad X = (\lambda - 2)^2 X \\ &\Longrightarrow \quad [(\lambda - 2)^2 - 1]X = 0 \end{aligned}$ We need a nonzero solution and hence...	$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0$ since otherwise $X = Y = 0$.
$\Longrightarrow \quad [(\lambda - 2)^2 - 1) = 0$ $\Longrightarrow \quad \lambda^2 - 4\lambda + 3 = 0$ we get the auxiliary equation.	$\Longrightarrow \quad [(\lambda - 2)^2 - 1) = 0$ $\Longrightarrow \quad \lambda^2 - 4\lambda + 3 = 0$

$\Longrightarrow \lambda = 1, 3$	$\Longrightarrow \lambda = 1, 3$ These are the eigenvalues of $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
When $\lambda = 1$ then Eq. (E1) gives $Y = -X$ $X = -Y$	When $\lambda = 1$ then Eq. (E2) gives $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
So let $X = A \Rightarrow Y = -A$ A is an arbitrary constant.	$\Longrightarrow \begin{pmatrix} X \\ Y \end{pmatrix} = A \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ A is an arbitrary constant.
When $\lambda = 3$ then Eq. (E1) gives $Y = X$ $X = Y$	When $\lambda = 3$ then Eq. (E2) gives $\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
So let $X = B \Rightarrow Y = B$ B is an arbitrary constant.	$\Longrightarrow \begin{pmatrix} X \\ Y \end{pmatrix} = B \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ B is an arbitrary constant.
Hence $\begin{matrix} x = Ae^t + Be^{3t} \\ y = -Ae^t + Be^{3t} \end{matrix}$ This is a 2nd order system $\Rightarrow$ two complementary functions and two arbitrary constants.	Hence $\begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^t + B \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{3t}$ $\uparrow$ Eigenvector corresponding to $\lambda = 1$ $\uparrow$ Eigenvector corresponding to $\lambda = 3$

It is worth going through each of the above columns in turn to see how one line moves to the next, and then it is worth comparing the two analyses because they clearly cover the same ground but do so from slightly different perspectives. As already mentioned, neither approach is better than the other when the matrix is a 2×2, but the matrix-based approach becomes better and more efficient for larger matrices.

3.10.5 General comment

The manner in which the eigenvalues and eigenvectors of a matrix are found is generally taught in a dry-as-dust manner and with no indication as to why this might be useful. The main purpose of Example 3.9 has been to show that they arise naturally when solving a system of two constant-coefficient first order ODEs.

Here is the recipe for finding the eigenvalues and eigenvectors of the square matrix, M, with the various

explanations along the way. We begin by insisting that there are nonzero solutions of the equation,

$$M\underline{x} = \lambda\underline{x}. \tag{3.83}$$

Given that $\underline{x} = I\underline{x}$, then this equation may also be written in the form,

$$M\underline{x} = \lambda I\underline{x} \qquad \text{or} \qquad (M - \lambda I)\underline{x} = \mathbf{0}. \tag{3.84}$$

If the determinant of $(M - \lambda I)$ is nonzero, then the solution is that $\underline{x}$ is the zero vector. Therefore nonzero solutions for $\underline{x}$ may only exist if the determinant of $(M - \lambda I)$ is zero:

$$|M - \lambda I| = 0. \tag{3.85}$$

If M is an $N \times N$ matrix, then the evaluation of this determinant yields an N^{th} order polynomial for λ, noting that we saw that the equation for λ was quadratic in Example 3.9 where the matrix is a 2×2 matrix. These values of λ are called the **eigenvalues** of the matrix, M.

Having obtained the different values of λ, one may then solve Eq. (3.84) to find the corresponding **eigenvectors**, $\underline{x}$. Each eigenvalue will have its associated eigenvector.

For the purposes of the exam for ME10305 I will not stray into the territory when one gets repeated values of λ, or where they are complex, although such things can happen in real life.

Finally, I would like to mention that computer packages which solve for eigenvalues and eigenvectors usually normalise each eigenvector and therefore they will have unit length. For Example 3.9 the vector form of our solution will be,

$$\begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} e^{t} + B \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} e^{3t}. \tag{3.86}$$

Of course, if two initial conditions are provided then the two forms of the general solution will yield identical results. We will see this in the next example.

3.10.6 Example 3.10.

Question. Find the eigenvalues and eigenvectors of $\begin{pmatrix} -3 & 4 \\ 1 & -3 \end{pmatrix}$ and hence solve

$$\underline{x}' = \begin{pmatrix} -3 & 4 \\ 1 & -3 \end{pmatrix} \underline{x} \qquad \underline{x}' = M\underline{x}$$

subject to the initial condition,

$$\underline{x} = \begin{pmatrix} 4 \\ 0 \end{pmatrix} \text{ at } t = 0.$$

Answer. We shall start by finding the eigenvalues of the given matrix. Hence,

$$0 = \begin{vmatrix} -3-\lambda & 4 \\ 1 & -3-\lambda \end{vmatrix} = (-3-\lambda)^2 - 4 = \lambda^2 + 6\lambda + 5 = (\lambda+1)(\lambda+5). \qquad 0 = |M - \lambda I|$$

Hence,

$$\lambda = -1, -5 \qquad \text{—— the eigenvalues.}$$

Now to find the eigenvectors corresponding to each eigenvalue. When $\lambda = -1$ we have

$$\begin{pmatrix} -2 & 4 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad (M - \lambda I)\underline{x} = 0 \quad \text{with } \lambda = -1$$

$$\implies \quad \begin{pmatrix} X \\ Y \end{pmatrix} = A \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{3.87}$$

Here A is an arbitrary constant. I decided to let $Y = A$ and then X was found to be $2A$. I could have done it the other way around by setting $X = A$ but I wanted to avoid fractions. One doesn't have to avoid fractions.

When $\lambda = -5$ we have

$$\begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad (M - \lambda I)\underline{x} = 0 \text{ again but with } \lambda = -5$$

$$\implies \quad \begin{pmatrix} X \\ Y \end{pmatrix} = B \begin{pmatrix} 2 \\ -1 \end{pmatrix}. \tag{3.88}$$

So the coefficients of A and B are the eigenvectors which correspond respectively to the eigenvalues, -1 and -5. Now we can go on to the ODE part of the question.

The solution of the given pair of 1st order ODEs may now be written immediately as,

$$\begin{pmatrix} x \\ y \end{pmatrix} = Ae^{-t} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + Be^{-5t} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

This is the general solution of system of ODEs. But when $t = 0$ we have,

$$\begin{pmatrix} 4 \\ 0 \end{pmatrix} = A \begin{pmatrix} 2 \\ 1 \end{pmatrix} + B \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Almost by inspection we can see that $A = B = 1$. Hence the final solution is

$$\begin{pmatrix} x \\ y \end{pmatrix} = e^{-t} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + e^{-5t} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

The problem is solved so maybe I should stop here. However, I did mention earlier about what would happen if the eigenvectors were normalised, so let us try it out. In the present case the general solution would be

$$\begin{pmatrix} x \\ y \end{pmatrix} = Ae^{-t} \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} + Be^{-5t} \begin{pmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{pmatrix}.$$

Application of the initial condition will yield

$$\begin{pmatrix} 4 \\ 0 \end{pmatrix} = A \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} + B \begin{pmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{pmatrix},$$

from which we find that $A = B = \sqrt{5}$, and therefore we obtain the same solution as before, namely that,

$$\begin{pmatrix} x \\ y \end{pmatrix} = e^{-t} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + e^{-5t} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

3.10.7 Example 3.11 and a dirty trick

The aim here is to solve

$$\underline{x}' = \begin{pmatrix} 3 & -4 \\ -1 & 3 \end{pmatrix} \underline{x}.$$

Ah, wait a moment, this matrix seems familiar. If we were denote by M the matrix in Example 3.10, then this matrix is $-M$. So the real aim is to find out what the effect is on the eigenvalues and eigenvectors of changing the sign of all of the entries in a matrix, and hence on the solution of the ODE system. We shall do this from scratch without reference to Example 3.10.

$$\text{Let } \underline{x} = e^{\lambda t}\underline{X} \implies \begin{pmatrix} 3 & -4 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \lambda \begin{pmatrix} X \\ Y \end{pmatrix} \implies \begin{pmatrix} 3-\lambda & -4 \\ -1 & 3-\lambda \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So we need

$$0 = \begin{vmatrix} 3-\lambda & -4 \\ -1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 4$$

$$\implies \lambda = \pm 2 + 3 = 1, 5.$$

Therefore the signs of the eigenvalues change.

Now to find the eigenvectors.

$$\text{When } \lambda = 1: \quad \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} X \\ Y \end{pmatrix} = A \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\text{When } \lambda = 5: \quad \begin{pmatrix} -2 & -4 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} X \\ Y \end{pmatrix} = A \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

But the eigenvectors remain the same.

So

$$\underline{x} = Ae^{t} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + Be^{5t} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

The message here is that, while the eigenvalues of M and of $-M$ have opposite signs, the eigenvectors remain the same.

3.10.8 Example 3.12 and a second dirty trick

Solve

$$\underline{x}'' = \begin{pmatrix} -3 & 4 \\ 1 & -3 \end{pmatrix} \underline{x}.$$

This is the same matrix as in Example 3.10, M, but these are now second order equations. Let us see what happens as we run the analysis through.

$$\text{Let } \underline{x} = e^{\lambda t}\underline{X} \implies \begin{pmatrix} -3 & 4 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \lambda^2 \begin{pmatrix} X \\ Y \end{pmatrix}$$

$$\implies \lambda^2 \underline{X} = \begin{pmatrix} -3 & 4 \\ 1 & -3 \end{pmatrix} \underline{X}$$

$$\begin{pmatrix} -3-\lambda^2 & 4 \\ 1 & -3-\lambda^2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So we need

$$0 = \begin{vmatrix} -3-\lambda^2 & 4 \\ 1 & -3-\lambda^2 \end{vmatrix} = (3+\lambda^2)^2 - 4$$

$$\implies \quad \lambda^2 \quad = \quad \pm 2 - 3 \quad = \quad -1, -5.$$

Because we are solving a system of second order equations the eigenvalues of M are now notated as λ^2.

Hence,

$$\lambda = \pm j, \pm\sqrt{5}j.$$

In this case we have four values of λ which may feel a little strange, given our previous examples, but we are solving a fourth order system which is composed of two second order equations. When we have solved a single fourth order equation in the past we have always found four values of λ, and hence four complementary functions. In the present context how do we proceed when the matrix has only two eigenvalues? The answer is as follows:

$$\underline{x} = \underbrace{(A\cos t + B\sin t)}_{\lambda^2=-1}\begin{pmatrix}2\\1\end{pmatrix} + \underbrace{(C\cos\sqrt{5}t + D\sin\sqrt{5}t)}_{\lambda^2=-5}\begin{pmatrix}2\\-1\end{pmatrix}.$$

For both values of λ when $\lambda^2 = -1$ we see the functions of time which we would normally expect (i.e. $\lambda = \pm j$ corresponds to $\sin t$ and $\cos t$) but they share the same eigenvector because that is the one which is associated with the *single* $\lambda^2 = -1$. The same observation may be made for $\lambda^2 = -5$. Again, a fourth order system requires four arbitrary constants.

3.10.9 Example 3.13

Solve

$$\underline{x}'' = \begin{pmatrix}3 & -4\\ -1 & 3\end{pmatrix}\underline{x}.$$

Given how Examples 3.10, 3.11 and 3.12 have played out, we may say immediately that $\lambda^2 = 1, 5$ and that the eigenvectors are the same as in those three examples. Therefore we may write the solutions immediately,

$$\underline{x} = \underbrace{(Ae^t + Be^{-t})}_{\lambda^2=1}\begin{pmatrix}2\\1\end{pmatrix} + \underbrace{(Ce^{\sqrt{5}t} + De^{-\sqrt{5}t})}_{\lambda^2=5}\begin{pmatrix}2\\-1\end{pmatrix}.$$

3.10.10 Example 3.14. Three masses on springs.

We will motivate the following 3×3 matrix example using an undamped mass/spring system which involves three second-order equations. The matrix which is obtained will then be used as part of three first order equations which are obtained by replacing the second derivatives by first derivatives. As far as I am aware this latter system of first order equations has no application, so we'll simply use it to gain experience of finding the eigenvalues and eigenvectors of 3×3 matrices. Afterwards we will solve the mass/spring system properly.

We have three masses, m, with four springs each with spring stiffness, k. The outer springs are attached to fixed points. This is illustrated in Fig. 3.4 on the next page.

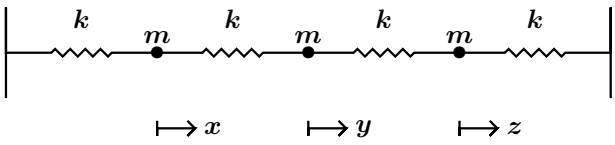


Figure 3.4. Showing the three masses and their displacements from equilibrium.

The equations are derived using Newton's law: $\boldsymbol{F} = \boldsymbol{ma}$, but the derivation of this is outside of the scope of this unit. The equations are

$$\begin{array}{rcccl} m\ddot{x} & = & k(y-x) - kx & = & k(-2x+y) \\ m\ddot{y} & = & k(z-y) + k(x-y) & = & k(x-2y+z) \\ m\ddot{z} & = & -kz + k(y-z) & = & k(y-2z). \end{array}$$

The terms in blue arise from that derivation and are included for interest only. These equations may now be written in matrix/vector form:

$$m\begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{pmatrix} = k\begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

This matrix is tridiagonal, a form which is retained should one wish to consider yet more masses and springs arranged in the same way as above. For the purposes of this example we shall simplify the problem by first setting $k = m$, and by replacing all second derivatives with first derivatives. **Therefore the present example consists of solving the following system of equations.**

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{3.89}$$

If we let $\underline{\boldsymbol{x}} = e^{\lambda t}\underline{\boldsymbol{X}}$, then

$$\lambda\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \implies \begin{pmatrix} -2-\lambda & 1 & 0 \\ 1 & -2-\lambda & 1 \\ 0 & 1 & -2-\lambda \end{pmatrix}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{3.90}$$

For nonzero solutions for $\underline{\boldsymbol{X}}$ we require that,

$$\begin{aligned} 0 &= \begin{vmatrix} -2-\lambda & 1 & 0 \\ 1 & -2-\lambda & 1 \\ 0 & 1 & -2-\lambda \end{vmatrix} \\ &= (-2-\lambda)\Big[(-2-\lambda)^2 - 1\Big] - (-2-\lambda) \qquad \text{...so many minus signs...} \\ &= -(2+\lambda)\Big[(2+\lambda)^2 - 2\Big] \qquad \text{after some careful arithmetic} \end{aligned}$$

is satisfied and hence $\boxed{\lambda = -2,\ -2 \pm \sqrt{2}}$ are the three eigenvalues.

Now to find the three eigenvectors. We'll take the middle eigenvalue first, so let $\lambda = -2$ in Eq. (3.90). This gives

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This matrix clearly has a zero determinant because the first and third rows are identical. On multiplying out the first row and the vector we get $Y = 0$. When multiplying out the second row we may set $X = B$ and therefore $Z = -B$. Hence the eigenvector is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = B \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

where B is an arbitrary constant.

When $\lambda = -2 - \sqrt{2}$ we have

$$\begin{pmatrix} \sqrt{2} & 1 & 0 \\ 1 & \sqrt{2} & 1 \\ 0 & 1 & \sqrt{2} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This matrix doesn't look as though it has a zero determinant, but if rows 1 and 2 are added then the result is $\sqrt{2}$ times row 2. Here is a way to solve this one:

1st equation: $\sqrt{2}X + Y = 0$. Let $X = C \Rightarrow Y = -\sqrt{2}C$.

3rd equation: $Y + \sqrt{2}Z = 0$. Hence $Z = C$.

2nd equation: This is already satisfied — must be, since the determinant is zero.

So

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = C \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}$$

is the eigenvector corresponding to $\lambda = -2 - \sqrt{2}$.

When $\lambda = -2 + \sqrt{2}$ exactly the same analysis ensues as for $\lambda = -2 - \sqrt{2}$ but with all instances of $\sqrt{2}$ replaced by $-\sqrt{2}$ — do check this to confirm my statement. Hence the third and final eigenvector is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = A \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}.$$

Now we are in a position to write down the solution of Eq. (3.89). It is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = Ae^{(-2+\sqrt{2})t} \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix} + Be^{-2t} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + Ce^{(-2-\sqrt{2})t} \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}.$$

3.10.11 Example 3.15.

Given the detailed analysis of Example 3.14 and given our earlier manipulations/tricks we are are in a position to solve for the mass/spring system shown in Fig. 3.4 but with $\boldsymbol{k = m}$ assumed. For completeness the system is,

$$\begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Our earlier experience with the $\mathbf{2 \times 2}$ examples tells us that an $e^{\lambda t}$ substitution will eventually yield $\boldsymbol{\lambda^2 = -2, -2 \pm \sqrt{2}}$ as the three eigenvalues and that the respective eigenvectors will be the same as in Example 3.14.

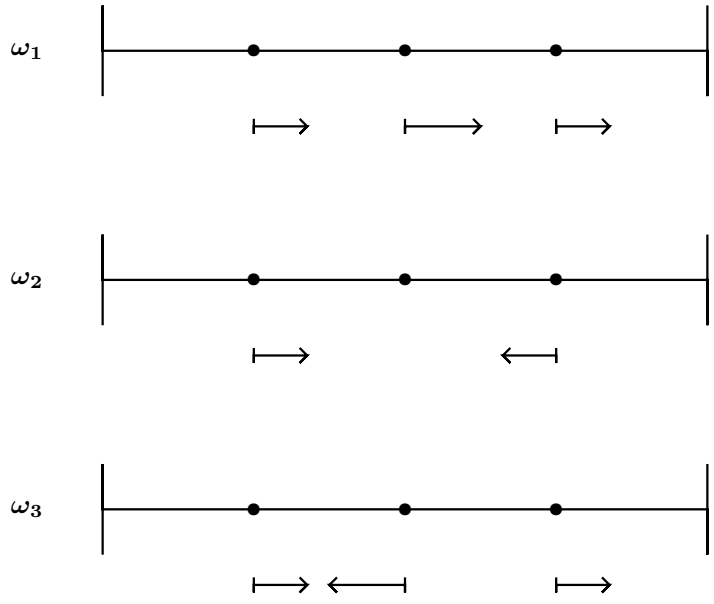
Given that all three values of $\boldsymbol{\lambda^2}$ are negative this means that all six values of $\boldsymbol{\lambda}$ are purely imaginary and hence the resulting motion will be composed of different sinusoidal terms. This is to be expected because we have modelled an undamped mass/spring system. For compactness of of presentation I shall define,

$$\omega_1^2 = 2 - \sqrt{2}, \qquad \omega_2^2 = 2, \qquad \omega_3^2 = 2 + \sqrt{2}.$$

Hence the solution is

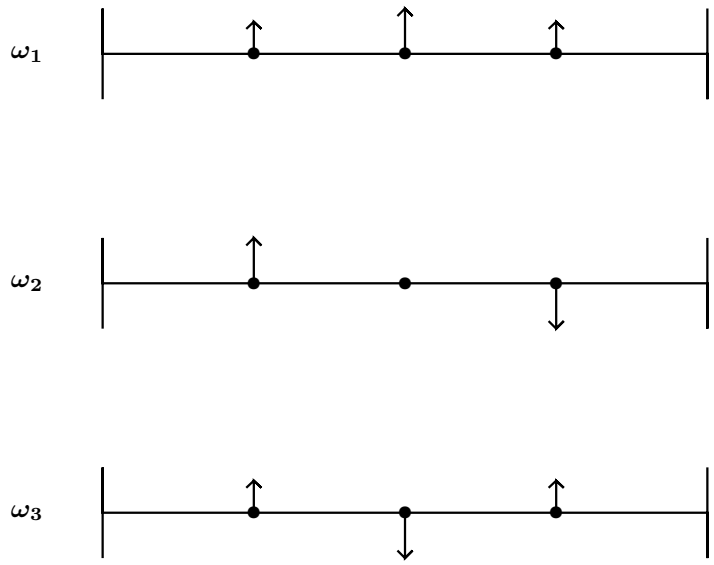
$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = (A\cos\omega_1 t + B\sin\omega_1 t) \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix} + (C\cos\omega_2 t + D\sin\omega_2 t) \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + (E\cos\omega_3 t + F\sin\omega_3 t) \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}.$$

We may visualise the motions of the masses correspond to each of the eigenvectors:



Here ω_1 is the lowest frequency and the three masses move in the same direction as one another. This is generally true of vibrating structures such as aircraft wings.

The equations we have solved here also apply to transverse vibrations although the manner in k is derived theoretically is different from the above longitudinal vibrations. In such an instance the mode shapes will look like the following.



3.10.12 A fifth order system

As already mentioned, systems which are composed of sets of mass and springs of the kind shown in Fig.3.4 always produce tridiagonal matrices. Thus a system of five equal masses and six identical springs will be modelled by the system,

$$m\begin{pmatrix} x_1'' \\ x_2'' \\ x_3'' \\ x_4'' \\ x_5'' \end{pmatrix} = k\begin{pmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}.$$

3.10.13 Example 3.16.

This one is the same as Example 3.15 except that are no restraining springs at either end. This is depicted in Figure 3.5:

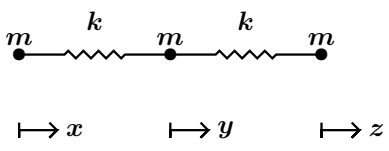


Figure 3.5. Showing the three masses and their displacements from equilibrium.

Again we shall assume that $k = m$ just to get some nice clean numbers.

$$\begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Let $\underline{x} = e^{\lambda t}\underline{X}$ and therefore we eventually get to the following determinantal equation for λ^2:

$$\begin{aligned} 0 &= \begin{vmatrix} -1-\lambda^2 & 1 & 0 \\ 1 & -2-\lambda^2 & 1 \\ 0 & 1 & -1-\lambda^2 \end{vmatrix} \qquad 0 = |M - \lambda I| \\ &= -(1+\lambda^2)\Big[(2+\lambda^2)(1+\lambda^2) - 1\Big] - (-1-\lambda^2) \\ &= -(1+\lambda^2)(\lambda^4 + 3\lambda^2) \qquad \text{Again much care is needed over minus signs.} \\ &= -\lambda^2(1+\lambda^2)(3+\lambda^2). \end{aligned}$$

Hence the three eigenvalues are $\lambda^2 = 0, -1, -3$.

When $\lambda^2 = 0$:

$$\begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad X = Y = Z \quad \Longrightarrow \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = A \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

In detail: let $X = A$, then row 1 of the matrix/vector equation gives $Y = A$ too. Then either row 2 or row 3 will give $Z = A$ as well.

When $\lambda^2 = -1$:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad Y = 0,\ X = -Z \quad \Longrightarrow \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = B \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

In detail: the first row gives $Y = 0$. If we then set $X = B$ then the second row yields $Z = -B$. The third row is identical to the first.

When $\lambda^2 = -3$:

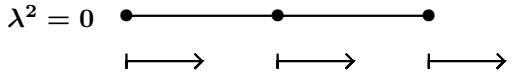
$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad X = Z,\ Y = -2X \quad \Longrightarrow \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = C \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

In detail: let $X = C$ and then the first row gives $Y = -2C$. Then either row 2 or row 3 yields $Z = C$.

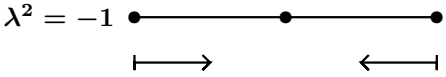
Now we may write down the solution. There will be six terms, two per eigenvalue/eigenvector, but given that $\lambda^2 = 0$ it means that $\lambda = 0$ is a repeated root of the characteristic equation. The solution is

$$\underline{x} = \underbrace{(A + Bt)}_{\lambda=0,0} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \underbrace{(C\cos t + D\sin t)}_{\lambda=\pm j} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + \underbrace{(E\cos\sqrt{3}t + F\sin\sqrt{3}t)}_{\lambda=\pm\sqrt{3}j} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

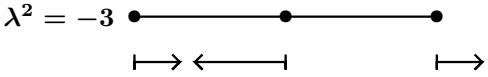
We may now check the mode shapes to see how these differ from those in Example 3.15.



The masses do not move relative to one another



The centre of gravity remains stationary



The centre of gravity remains stationary

We could imagine this system of springs to be placed in space with you as an observer at a fixed distance with zero relative motion. If the three masses were to be given a set of initial conditions, all of course being in the one direction depicted above, then the resulting motion would be a combination of the above shapes. If the total initial momentum were to be zero (i.e. $m(x_1' + x_2' + x_3') = 0$ in this case) then B would be zero, which means that the centre of mass would remain stationary. Otherwise, B would be nonzero and the mass/spring system would drift away at a constant speed, B.

3.10.14 Final remarks

In the above I have given quite a lot of space to the various mode shapes, and by doing so I have tried to draw conclusions about what the mathematics means in real life. I hope that this has made the topic of eigenvalues and eigenvectors interesting or, if not, at least convinced you of how useful they are in real world engineering.

In practice, the analysis of the vibrational characteristics of an an airframe or a building or bridge will involve the splitting up of these structures into small elements each of which is attached to its neighbours with some form of spring stiffness and damping (both extensional and rotational) connecting each pair. The resulting matrices can be huge.

In my early career (i.e. between my BSc and PhD) I spent a couple of years analysing the wing bending and torsion for both the BAC111 (the fundamental wing bending mode had a frequency of 4Hz, and the the corresponding torsional mode was 7Hz) and the Panavia Tornado, both of which have been retired since. Typically the wing was split into 10 sections mathematically, and each section was free to move in roll and pitch relative to its immediate neighbour and subect to bending and torsional stiffness and damping. This meant that I had a set of 20 second order equations to study. Feedback from accelerometers was also included — this analysis came from some background Laplace Transform work — and some crude approximations to the aerodynamic loading in both roll and pitch. Then the eigenvalues of this matrix were computed as a function of the air speed and the type of feedback (i.e. control law) included in the model. The aim was to find how fast the aircraft could be persuaded to fly without its wing falling off, or rather, to try to design the control law to maximise the safe operational speed. Therefore it was essential for all the λ-values to have a negative real part. Once one of them becomes positive then the wing would flutter (i.e. disturbances would oscillate in time but also grow exponentially) and, well, goodbye wing!

Another story from that time involved the job British Aerospace, as it was then, had converting some VC10s into air-to-air refuelling tankers. A massive hole was made in the top of the fuselage in order to get the fuel tank in. It was then shut up and resealed. However, this meant that the structure itself had changed and therefore it was essential to test this modified beast for its resonant modes and to check these against theory. I had no involvement in that apart from one bit of Saturday morning overtime. The vibration of the structure was monitored by means of a large number of accelerometers embedded in 1 inch cubes of wood which were then glued to the skin of the aircraft. My role was to clamber up the scaffolding armed with a hammer and chisel to hack them off the starboard tail plane. I don't recall health and safety being mentioned but I survived. I don't think that I completed my task that cleanly for some paint came off. However, the job was done quickly and I was casting around for what to do next so I decided to do some gentle squats, as one does, just to see what the effect might be and to my astonishment the port wing started moving! Clearly this must have been part of a fuselage torsion mode, but I eventually managed to find the first resonant frequency! Only very mild movements on my part built up to quite a large movement of the wing. And, of course, I was hidden from view!

If, based on this personal confession for educational purposes, you feel that you need to see me about questions on Control Theory next academic year, then do please bear in mind that the only thing I remember is that I did it then, in the early 1980s. I know nothing now apart from a vague sense of deja vu with a fair bit of nostalgia. I have very knowledgeable colleagues who will be only too pleased to feel superior to me!

As far as the exam is concerned, I will only examine you on the actual finding of the eigenvalues and eigenvectors, and on their application/use in solving a system of ODEs much as I have done above. I will not be asking about what the eigenvector means in practice. If in doubt, then have a quick look at some past exam questions. I generally will use $\mathbf{2} \times \mathbf{2}$, $\mathbf{3} \times \mathbf{3}$ and possibly tridiagonal $\mathbf{4} \times \mathbf{4}$ matrices in these questions.

4 NUMERICAL MATHEMATICS — Iteration schemes

4.1 Applications

The subject of root-finding is a classical area of mathematics which has a wide range of applications. The usual first example is to devise a method (typically the Newton-Raphson scheme) to solve

$$x^2 - 2 = 0.$$

Of course, it is quite easy to find the square root of 2 if you have a calculator with a square root button, but real life isn't that straightforward. One might need to find the roots of a quintic polynomial such as,

$$x^5 - 2x^4 - 3x^3 + 6x^2 + 2x - 4 = 0,$$

but there is no formula like that of the quadratic equation which may be applied. Although it may be possible to find at least one root *by inspection* for this quintic, it is almost never the case in a real life application. So we have to resort to purely numerical methods to find roots. (Incidentally, the above quintic factorises into $(x^2-1)(x^2-2)(x-2)=0$, and so its roots are $x = \pm 1$, $\pm\sqrt{2}$ and 2.)

A very common application of root-finding is in the numerical solution of ODEs which are Boundary Value Problems; this is covered in ME20014 Modelling Techniques 1 next semester. Suppose that you have all but one of the Initial Conditions (ICs) that are required at $x = x_{\min}$ and one final condition at $x = x_{\max}$ which also needs to be satisfied. This means that one IC is missing from the full list. Aaargh, what can we do?

The simplest methods for solving ODEs generally rely on (i) reducing the system to first order form, and then (ii) marching the solution forward one timestep at a time. When the problem is an Initial Value Problem with a full set of ICs then such a marching scheme works straightforwardly. But when one IC is missing, we can guess the unknown IC (call this value α), then march the solution forward numerically from $x = x_{\min}$ to $x = x_{\max}$ using those standard methods, and finally we find out how closely the computed final value is to the value given by the imposed final condition. Whatever that final condition is, the discrepancy is a function of the guessed value of the unknown IC. Therefore it may always be written in the form, $F(\alpha) = 0$, i.e. the discrepancy is a function of the value of the IC. So somewhere behind the scenes is a function, F, that we cannot write down but may evaluate by solving the ODE numerically. Then we need to iterate to find the correct value or values of α.

We shall look at two different ways of doing this. First, we'll consider what I shall call **ad hoc** schemes. These are quick to write down but tend to be a bit slow when they work. Moreover each scheme tends to work only about half the time. Second, we'll consider the **Newton-Raphson** scheme which performs very much better in many ways. For both schemes I shall also introduce you to the idea of a perturbation analysis. In the present context this will serve to demonstrate why the iteration schemes work as they do. But the general idea of a perturbation analysis is a central idea in stability theory where questions such as the following need to be answered: **How fast can this aeroplane fly before its wings fall off?** The idea is simple: introduce a small perturbation/disturbance and then find out if it grows or decays. If the disturbance evolves like a sinusoidal function multiplied by a growing exponential in time, then that is known as flutter.

4.2 Roots of Equations

Prior to considering the methods for finding roots, it is worth considering that equations may have any number of roots. We will look at three example cases here, namely

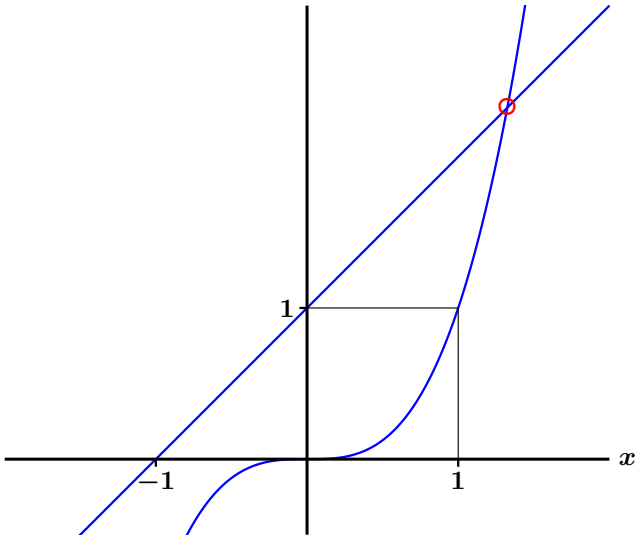
$$x^3 - x - 1 = 0, \qquad x^4 = \frac{1}{1+x^2} \qquad \text{and} \qquad x \sin x = 1. \tag{4.1}$$

As we shall see these have $\mathbf{1}$, $\mathbf{2}$ and an infinite number of roots, respectively. It is quite possible, of course, for an equation not to have any roots, such as for $\boldsymbol{x^2 + 2 = 0}$.

There is no general method for determining the number of roots that an equation might have, but a good sketch helps enormously. We shall consider the above three examples in turn in order to get a feeling for how to assess the number of roots.

4.2.1 Example 4.1: $x^3 - x - 1 = 0$.

A good rearrangement of this is the following: $\boldsymbol{x^3 = x + 1}$, and then we may sketch both sides of this equation and see where the intersection is or intersections are.



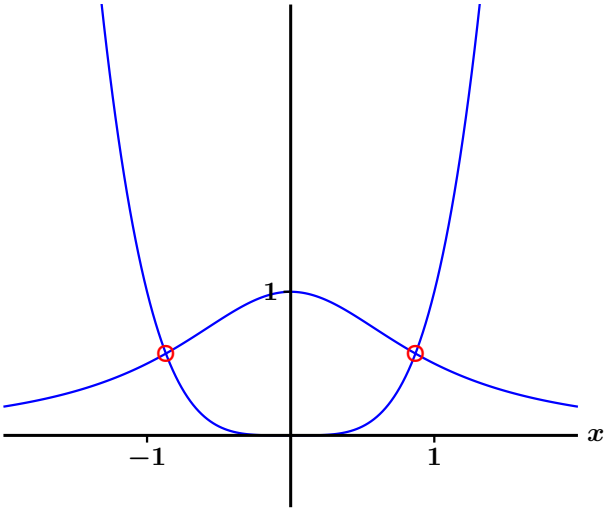
When $\boldsymbol{x = 1}$ the cubic has the value $\mathbf{1}$ but the straight line has the value, $\mathbf{2}$. However, the cubic is about to grow very rapidly and therefore we will expect an intersection not far above $\boldsymbol{x = 1}$; this is marked as the red circle.

When $\boldsymbol{x = -1}$, the straight line has the value $\mathbf{0}$, but the cubic is at $\boldsymbol{-1}$ and is already descending more rapidly than the line is. Therefore there will be no intersection on that side of the origin.

Our analysis above suggests that there will be only one root, and indeed that is the case.

4.2.2 Example 4.2: $x^4 = \dfrac{1}{1+x^2}$.

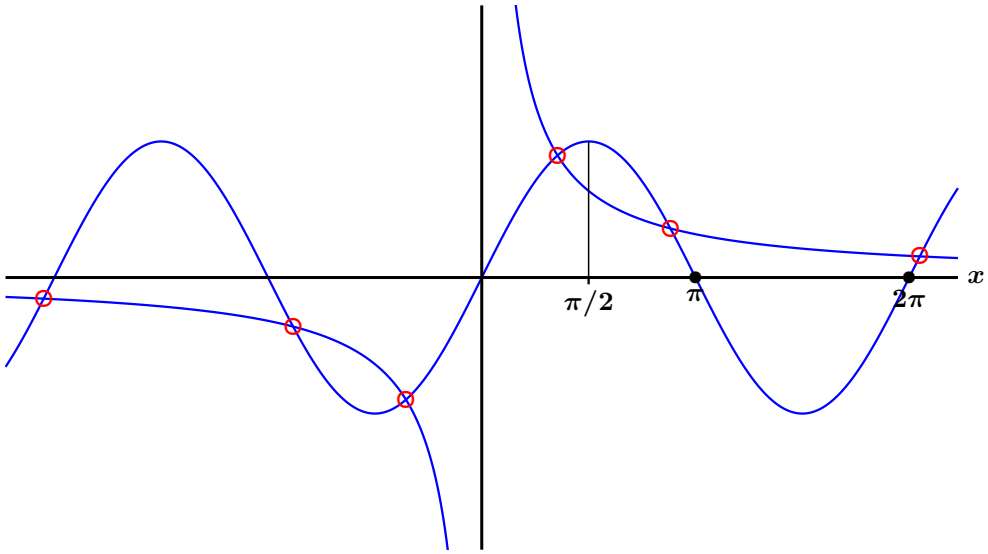
In this example we'll just sketch the two functions which are on either side of the equals sign.



We can state immediately that the minimum of x^4 is 0 but it then grows without bound as the magnitude of x increases, while maximum of $1/(1+x^2)$ is 1 but it then decreases to zero as the magnitude of x increases. Both functions are even. Therefore we can conclude that there will be two zeros, and these are placed symmetrically about the vertical axis.

4.2.3 Example 4.3: $x \sin x = 1$.

This equation may be rearranged into the form, $\sin x = 1/x$, and now we can sketch the two sides of this equation.



In this sketch I have singled out the value, $x = \pi/2$, for at that point the sine wave takes a unit value, while the hyperbola, $1/x$, takes the value $2/\pi$ which is less than 1. Therefore there must be two intersections either side of that location. This was done to determine if the two curves crossed there or if they just missed doing so.

Given the form taken by the sketch of $1/x$ (i.e. that it tends to zero as x increases), it may be also concluded that the roots of the original equation must be close to $x = n\pi$, where n is an integer because the hyperbola gets close to the x-axis. In fact, one could go a little further and say that the roots are just above $n\pi$ when n is even and just below $n\pi$ when n is odd.

It is also clear from the sketch that roots are also symmetrically placed about the origin.

It is possible to sketch both $x \sin x$ and 1 on the same graph and come to the same conclusions. However, I think that it is a little more tricky to do this. It might be worth trying that to see if you agree with me.

4.3 Ad hoc methods

The Latin, *ad hoc*, means **for this,** or maybe, **for its purpose,** or even **unplanned** or simply that **we're making it up as we go along.** Perhaps that last one is a little unfair.....

The plan is to find a root of the equation by first rewriting it in the form,

$$x = f(x), \tag{4.2}$$

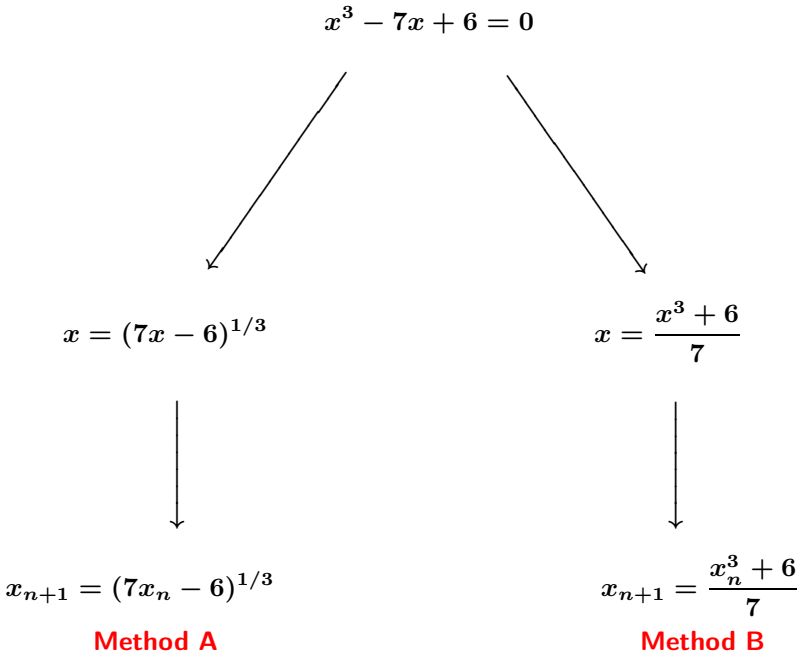
and then converting it to an iteration scheme,

$$x_{n+1} = f(x_n),$$

where n is the iteration number and where x_0 is the initial iterate.

4.3.1 Example 4.4: Solve $x^3 - 7x + 6 = 0$.

This cubic was created by multiplying out $(x-1)(x-2)(x+3) = 0$ and hence the roots are $x = 1, 2$ and -3. This may be written in the form given by Eq.(4.2) in two different ways.



Now let us try these with $x_0 = 1.1$:

Method A

$$x_{n+1} = (7x_n - 6)^{1/3}$$

$$\begin{aligned}
\text{Let} \quad x_0 &= 1.1 \\
\implies x_1 &= 1.193483 \\
\implies x_2 &= 1.330329 \\
\implies x_3 &= 1.490653 \\
\implies x_4 &= 1.642923 \\
&\vdots \\
\implies x_8 &= 1.946008
\end{aligned}$$

Method B

$$x_{n+1} = \frac{x_n^3 + 6}{7}$$

$$\begin{aligned}
\text{Let} \quad x_0 &= 1.1 \\
\implies x_1 &= 1.047286 \\
\implies x_2 &= 1.021239 \\
\implies x_3 &= 1.009297 \\
\implies x_4 &= 1.004022 \\
&\vdots \\
\implies x_8 &= 1.000137
\end{aligned}$$

Method A diverges from the root at $x = 1$, but Method B appears to be converging towards the root at $x = 1$.

Let us try a different starting iterate.

Method A

$$x_{n+1} = (7x_n - 6)^{1/3}$$

$$\begin{aligned}
\text{Let} \quad x_0 &= 2.1 \\
\implies x_1 &= 2.056710 \\
\implies x_2 &= 2.032548 \\
\implies x_3 &= 2.018809 \\
\implies x_4 &= 2.010912 \\
&\vdots \\
\implies x_8 &= 2.001255
\end{aligned}$$

Method B

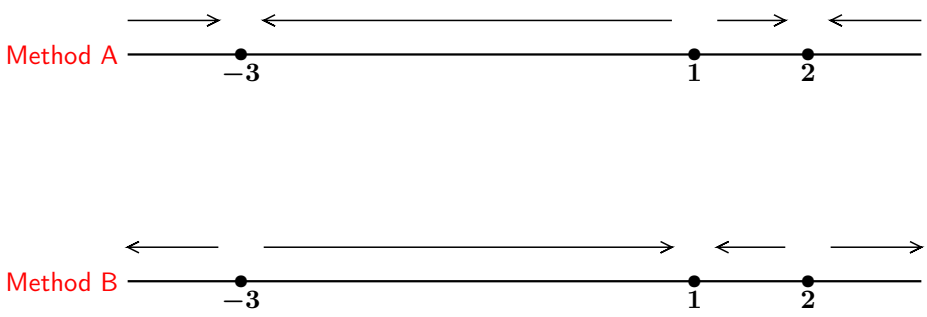
$$x_{n+1} = \frac{x_n^3 + 6}{7}$$

$$\begin{aligned}
\text{Let} \quad x_0 &= 2.1 \\
\implies x_1 &= 2.180143 \\
\implies x_2 &= 2.337467 \\
\implies x_3 &= 2.681620 \\
\implies x_4 &= 3.611966 \\
&\vdots \\
\implies x_8 &= 6.79 \times 10^{12}
\end{aligned}$$

In this case Method A converges towards $x = 2$ and Method B diverges away from it.

Clearly the two schemes are complementary and between them we are able to find all three roots. Note that Method A will also converge to the third root at $x = -3$; it is worth checking this.

The following is a graphical demonstration of the convergence properties of the two methods.



So for a chosen method the convergence properties alternate for each successive root.

4.4 Perturbation Analysis

Let us now analyse in detail what happens to the iterates when they are close to the desired root. This is done using a **perturbation analysis** which tells us what happens to small errors.

Let us consider a small perturbation from the root $x = 1$ for Method A. Let us also substitute $x_n = 1 + \epsilon$ into $x_{n+1} = (7x_n - 6)^{1/3}$, where the perturbation, ϵ, is assumed to be very small. Technically this is often written as $|\epsilon| \ll 1$. We get

$$\begin{aligned} x_{n+1} &= (7x_n - 6)^{1/3} && \text{Method A} \\ &= \Big[(7 + 7\epsilon) - 6\Big]^{1/3} \\ &= \Big[1 + 7\epsilon\Big]^{1/3} \\ &= 1 + \tfrac{7}{3}\epsilon + \cdots && \text{Two terms of a Binomial series.} \end{aligned}$$

In the above, the dots represents terms like ϵ^2, ϵ^3 and even higher powers all of which are negligible compared with the magnitude of ϵ.

Given that $\boxed{x_n = 1 + \epsilon}$ and that $\boxed{x_{n+1} \simeq 1 + \frac{7}{3}\epsilon}$, then the perturbation has grown in size and therefore the root at $x = 1$ cannot be found using Method A.

Let us consider how Method B behaves for the same root.

$$\begin{aligned} x_{n+1} &= \tfrac{1}{7}(x_n^3 + 6) && \text{Method B} \\ &= \tfrac{1}{7}\Big[(1 + \epsilon)^3 + 6\Big] \\ &= \tfrac{1}{7}\Big[(1 + 3\epsilon + \cdots) + 6\Big] \\ &= \tfrac{1}{7}\Big[7 + 3\epsilon + \cdots\Big] \\ &= 1 + \tfrac{3}{7}\epsilon + \cdots . \end{aligned}$$

In this case the error in x_{n+1} is smaller than the error in x_n. Hence the perturbation decreases in magnitude every iteration, and the iteration scheme will converge to the root, $x = 1$.

Given that magnitude of the perturbation for x_{n+1} is a multiple of the perturbation for x_n we can say that **Method A diverges linearly** and that **Method B converges linearly.**

Just the one more....let us consider Method A for the root, $x = 2$. In our table of iterates it is readily seen that the error/perturbation roughly halved for each iteration. Let us see what the perturbation analysis tells us; we let $x_n = 2 + \epsilon$:

$$\begin{aligned}
x_{n+1} &= \left[7x_n - 6\right]^{1/3} && \text{Method A} \\
x_{n+1} &= \left[7(2+\epsilon) - 6\right]^{1/3} \\
&= \left[8 + 7\epsilon\right]^{1/3} \\
&= 2\left[1 + \tfrac{7}{8}\epsilon\right]^{1/3} && \text{Get the } 2 \text{ outside for convenience} \\
&= 2\left[1 + \tfrac{7}{24}\epsilon + \cdots\right] && \text{Two terms of a Binomial series} \\
&= 2 + \tfrac{7}{12}\epsilon + \cdots .
\end{aligned}$$

So $\boxed{x_n = 2 + \epsilon}$ becomes $\boxed{x_{n+1} = 2 + \frac{7}{12}\epsilon + \cdots .}$

Therefore the error is close to being halved every iteration, as we observed earlier. Therefore Method A converges for the root $x = 2$.

Note: In summary, we may say the following.

If we have $x_n = x_{\text{exact}} + \epsilon$ and $x_{n+1} = x_{\text{exact}} + c\epsilon + \cdots$, then

$$|c| < 1 \implies \text{linear convergence}$$

and

$$|c| > 1 \implies \text{linear divergence}$$

The sign of c is irrelevant. If it were to be negative, then that only means that the errors change sign every iteration.

4.5 Repeated roots

We shall consider how these *ad hoc* methods behave when trying to find a twice-repeated root. To this end we will consider,

$$(x-1)^2(x+2) = 0 \qquad \text{or} \qquad x^3 - 3x + 2 = 0.$$

We'll define two methods in the same way as before. We shall begin with $x_0 = 1.01$ in the first Table below and then try $x_0 = 0.99$ in the second Table.

Method A

$$x_{n+1} = (3x_n - 2)^{1/3}$$

$$\begin{aligned}
\text{Let} \quad & x_0 = 1.01 \\
\implies \quad & x_1 = 1.009912 \\
\implies \quad & x_2 = 1.009805 \\
\implies \quad & x_3 = 1.009711 \\
& \vdots \\
\implies \quad & x_8 = 1.009264
\end{aligned}$$

Method B

$$x_{n+1} = \frac{x_n^3 + 2}{3}$$

$$\begin{aligned}
\text{Let} \quad & x_0 = 1.01 \\
\implies \quad & x_1 = 1.010100 \\
\implies \quad & x_2 = 1.010203 \\
\implies \quad & x_3 = 1.010307 \\
& \vdots \\
\implies \quad & x_8 = 1.010863
\end{aligned}$$

Method A

$$\begin{aligned}
\text{Let} \quad & x_0 = 0.99 \\
\implies \quad & x_1 = 0.989898 \\
\implies \quad & x_2 = 0.989795 \\
\implies \quad & x_3 = 0.989689 \\
& \vdots \\
\implies \quad & x_8 = 0.988880
\end{aligned}$$

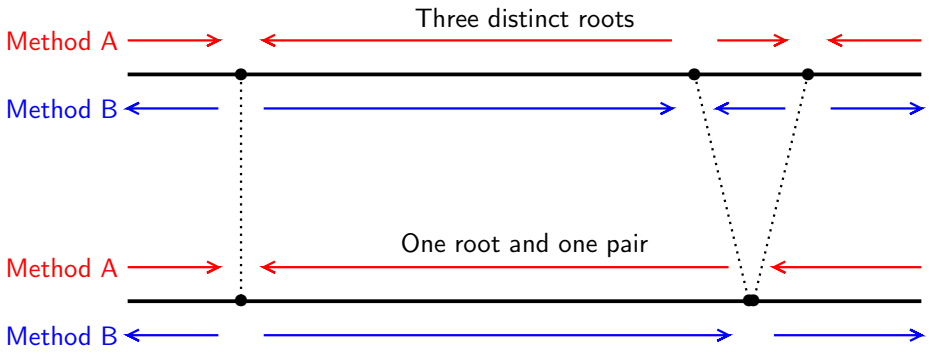
Method B

$$\begin{aligned}
\text{Let} \quad & x_0 = 0.99 \\
\implies \quad & x_1 = 0.990100 \\
\implies \quad & x_2 = 0.990197 \\
\implies \quad & x_3 = 0.990293 \\
& \vdots \\
\implies \quad & x_8 = 0.990745
\end{aligned}$$

Note 1: Convergence or divergence is **very very painfully slow**. In fact it takes 903 iterations to get to $\mathbf{1.001}$, 9903 iterations to get to $\mathbf{1.0001}$ and 99770 to get to $\mathbf{1.00001}$. This isn't good. It's appalling.

Note 2: For a chosen root, one of the Methods converges to it from above but diverges away from below, while the other one does the opposite. This is both strange and different from our experience earlier where the iterations either converge towards the root from both sides or else they diverge away from both sides. We need to check this out using the perturbation analysis.

4.5.1 Graphical explanation



Note how the arrows (which denote the direction of movement of successive iterates behave) as the distance between the upper two roots decreases.

4.5.2 Perturbation Analysis

Let us analyse Method A for the double root, $x = 1$, by setting $x_n = 1 + \epsilon$.

$$
\begin{aligned}
x_{n+1} &= (3x_n - 2)^{1/3} && \text{Method A} \\
x_{n+1} &= \Big[(3 + 3\epsilon) - 2\Big]^{1/3} \\
&= \Big[1 + 3\epsilon\Big]^{1/3} \\
&= 1 + \tfrac{1}{3}(3\epsilon) + \frac{(\frac{1}{3})(\frac{-2}{3})}{2}(3\epsilon)^2 + \cdots && \text{Three terms of a Binomial series.} \\
& && \text{because two terms gives } x_n \\
&= \underbrace{1 + \epsilon}_{=x_n} - \epsilon^2 + \cdots
\end{aligned}
$$

This analysis shows that (i) if $\epsilon > 0$, then the next iterate is very slightly closer to the root, and therefore a very slow convergence is obtained, and (ii) if $\epsilon < 0$, then the next iterate is slightly further away, and a slow divergence is obtained.

A similar Method B analysis shows that $x_{n+1} = 1 + \epsilon + \epsilon^2$, and therefore it acts in the opposite way to Method A.

A similar analysis for a triple root shows that $x_{n+1} = x_n + c\epsilon^3 + \cdots$, where c is a constant and hence this yields an even slower converence or divergence than for double roots.

4.6 Brief Summary of ad hoc methods

- For *ad hoc* methods we need to use two different Methods (i.e. different rearrangements of the equation being solved) in order to find all the roots of the equation.

- Single roots display linear convergence or divergence.

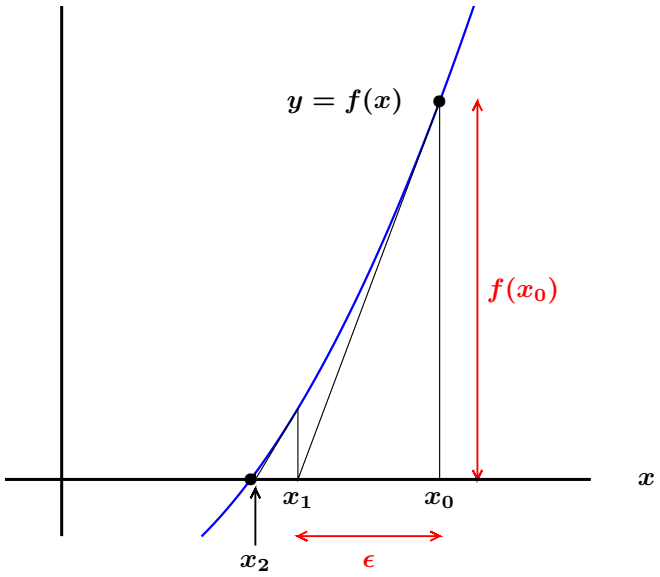
- Double roots display a much worse speed of convergence than for single roots. Convergence is disastrously slow.

4.7 The Newton-Raphson method

This is the usual workhorse not only for solving for the roots of a single equation but also for finding the roots of coupled system of nonlinear algebraic equations. In the present notes we shall cover only one equation in one unknown.

We shall motivate the formula in two different ways: graphical and by the use of Taylor's series.

4.7.1 Graphical Motivation



The aim is to find the root given by the bullet point on the x-axis. The initial iterate is $x = x_0$, as marked above, and the next iterate, x_1, is obtained by drawing the tangent to $y = f(x)$ at $x = x_0$ and then finding where it intersects the x-axis.

The slope of this tangent is $f'(x_0)$, while the slope is also given by $f(x_0)/\epsilon$, i.e. the height divided by the base of the triangle. According to this diagram we have $\epsilon = x_0 - x_1$, and therefore we may equate these two ways of writing down the slope:

$$\begin{aligned} \frac{f(x_0)}{\epsilon} &= f'(x_0) \\ \Longrightarrow \qquad \epsilon &= \frac{f(x_0)}{f'(x_0)} \\ \Longrightarrow \qquad x_0 - x_1 &= \frac{f(x_0)}{f'(x_0)} \\ \Longrightarrow \qquad x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)}. \end{aligned}$$

So this is how we obtain x_1 from x_0. Clearly the same approach will happen for the next iterate, and therefore the general formula for the method is,

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}} \qquad \text{The Newton-Raphson method}$$

.

4.7.2 Mathematical Motivation

We will derive the above formula by using two terms in a Taylor's series.

Let x be the exact solution. We will let x_n be the current iterate and we'll say that its error is precisely, ϵ. This means that,

$$x = x_n + \epsilon, \tag{4.3}$$

by definition. Hence,

$$\begin{aligned} 0 &= f(x) && \text{by definition} \\ &= f(x_n + \epsilon) && \text{also by definition} \\ &= f(x_n) + \epsilon f'(x_n) + \cdots && \text{two terms of the Taylor's series.} \end{aligned} \tag{4.4}$$

As before, the dots denote terms in ϵ^2, ϵ^3 and so on. The equation given by the last line above may be solved to give,

$$\epsilon = -\frac{f(x_n)}{f'(x_n)},$$

or, more properly, given that we have neglected all the terms denoted by the dots,

$$\epsilon \approx -\frac{f(x_n)}{f'(x_n)}.$$

Given that this is an approximation for ϵ, we need to modify Eq. (4.3) to the form,

$$x_{n+1} = x_n + \epsilon,$$

and therefore we obtain the Newton-Raphson formula:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

4.8 An example of the use of the Newton-Raphson method

We will look at the main example used for the *ad hoc* methods, namely to find the roots of $x^3 - 7x + 6 = 0$. Using Eq. (4.4) the formula for the Newton-Raphson method is

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^3 - 7x_n + 6}{3x_n^2 - 7} \\ &= \frac{2x_n^3 - 6}{3x_n^2 - 7}. \end{aligned} \tag{4.5}$$

We shall try three cases and these are shown in the following Table.

$$\begin{aligned} x_0 &= 1.1 \\ x_1 &= 0.990\,504\,451\,038\,5755 \\ x_2 &= 0.999\,933\,743\,233\,6637 \\ x_3 &= 0.999\,999\,996\,708\,0032 \\ x_4 &= 0.999\,999\,999\,999\,9999 \end{aligned}$$

$$\begin{aligned} x_0 &= 2.1 \\ x_1 &= 2.009\,951\,845\,906\,9023 \\ x_2 &= 2.000\,116\,453\,000\,0373 \\ x_3 &= 2.000\,000\,016\,269\,6456 \\ x_4 &= 2.000\,000\,000\,000\,0004 \end{aligned}$$

$$\begin{aligned} x_0 &= -2.9 \\ x_1 &= -3.004\,827\,207\,899\,0674 \\ x_2 &= -3.000\,010\,451\,675\,8714 \\ x_3 &= -3.000\,000\,000\,049\,1567 \\ x_4 &= -3.000\,000\,000\,000\,0000 \end{aligned}$$

There are various things to notice here. The first is that all three roots have been obtained using the method. The second is the *astonishing* speed of convergence, especially when compared with the somewhat weedy performance of the *ad hoc* methods.

When the error in an iterate is small, the error in the next iterate is proportial to the square of that, rather than being a constant multiplied by it. I have displayed 16 decimal places and coloured the correct decimal places in red in the above computations to enable the speed of convergence to be shown clearly.

We may use a perturbation analysis to model this.

4.9 Perturbation analysis

In the first instance let us consider the root, $x = 1$, by setting $x_n = 1 + \epsilon$ as before and where $|\epsilon| \ll 1$. We will use the lower formula in Eq. (4.5). Hence,

$$
\begin{aligned}
x_{n+1} &= \frac{2(1+\epsilon)^3 - 6}{3(1+\epsilon)^2 - 7} && \text{using (4.5)} \\
&= \frac{2 + 6\epsilon + 6\epsilon^2 + 2\epsilon^3 - 6}{3 + 6\epsilon + 3\epsilon^2 - 7} \\
&= \frac{-4 + 6\epsilon + 6\epsilon^2 + 2\epsilon^3}{-4 + 6\epsilon + 3\epsilon^2} && \text{tidying up} \\
&= \frac{-4 + 6\epsilon + (3\epsilon^2 + 3\epsilon^2) + 2\epsilon^3}{-4 + 6\epsilon + 3\epsilon^2} && \text{to reproduce the denominator} \\
&= 1 + \frac{3\epsilon^2 + 2\epsilon^3}{-4 + 6\epsilon + 3\epsilon^2} && \text{tidying up} \\
&= 1 - \tfrac{3}{4}\epsilon^2 + \cdots && \text{using leading order terms.}
\end{aligned}
$$

It is possible to get to this point by using the upper formula in Eq. (4.5), but the analysis takes much longer and it may also be necessary to to use the Binomial expansion. Again, I would try this approach once just to see what would be involved in doing it this way.

The final result shows that the error in x_{n+1} is proportional to the square of the error in x_n, as we had observed from the above Tables. This shows that the Newton-Raphson method converges quadratically.

For the root, $x = 2$ we'll follow the same sort of analysis using $x_n = 2 + \epsilon$ to start with. Hence

$$
\begin{aligned}
x_{n+1} &= \frac{2(2+\epsilon)^3 - 6}{3(2+\epsilon)^2 - 7} && \text{using (4.5)} \\
&= \frac{2(8 + 12\epsilon + 6\epsilon^2 + \epsilon^3) - 6}{3(4 + 4\epsilon + \epsilon^2) - 7} \\
&= \frac{10 + 24\epsilon + 12\epsilon^2 + 2\epsilon^3}{5 + 12\epsilon + 3\epsilon^2} \\
&= \frac{10 + 24\epsilon + (6\epsilon^2 + 6\epsilon^2) + 2\epsilon^3}{5 + 12\epsilon + 3\epsilon^2} && \text{looking for } 2\times \text{ the denominator} \\
&= \frac{2(5 + 12\epsilon + 3\epsilon^2) + 6\epsilon^2 + 2\epsilon^3}{5 + 12\epsilon + 3\epsilon^2} \\
&= 2 + \frac{6\epsilon^2 + 2\epsilon^3}{5 + 12\epsilon + 3\epsilon^2} \\
&= 2 + \tfrac{6}{5}\epsilon^2 + \cdots && \text{using leading order terms.}
\end{aligned}
$$

Again we have quadratic convergence, but don't be fooled that the coefficient of ϵ^2 is multiplied by a constant which is greater than 1 — this is still *quadratic* convergence

We shall also look at the third root, $x = -3$, for the sake of completeness, but setting $x_n = -3 + \epsilon$. Hence,

$$\begin{aligned}
x_{n+1} &= \frac{2(-3+\epsilon)^3 - 6}{3(-3+\epsilon)^2 - 7} && \text{using (4.5)} \\
&= \frac{2(-27 + 27\epsilon - 9\epsilon^2 + \epsilon^3) - 6}{3(9 - 6\epsilon + \epsilon^2) - 7} \\
&= \frac{-60 + 54\epsilon - 18\epsilon^2 + 2\epsilon^3}{20 - 18\epsilon + 3\epsilon^2} \\
&= \frac{-3(20 - 18\epsilon + 3\epsilon^2) - 9\epsilon^2 + 2\epsilon^3}{20 - 18\epsilon + 3\epsilon^2} \\
&= -3 + \frac{-9\epsilon^2 + 2\epsilon^3}{20 - 18\epsilon + 3\epsilon^2} \\
&= -3 - \tfrac{9}{20}\epsilon^2 + \cdots && \text{using leading order terms.}
\end{aligned}$$

For a third time we obtain quadratic convergence.

Note: In summary, we may say the following:

If we have $\boxed{x_n = x_{\text{exact}} + \epsilon}$, then $\boxed{x_{n+1} = x_{\text{exact}} + c\epsilon^2 + \cdots}$. Therefore if our initial iterate is close enough to the exact solution, then the Newton-Raphson method will converge quadratically to that root. The value of c here is irrelevant because it multiplies ϵ^2.

Therefore the Newton-Raphson method performs extremely well when the roots are single roots. The next task is to find out how well it performs for double roots.

4.10 Application of the Newton-Raphson method to a double root.

Let us consider how to find the roots of

$$x^3 - 12x + 16 = 0.$$

This cubic has been created by multiplying out the following, $(x-2)^2(x+4)$, and therefore the roots are 2 (twice) and -4 once.

The Newton-Raphson scheme is,

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^3 - 12x_n + 16}{3x_n^2 - 12} \\ &= \frac{2x_n^3 - 16}{3x_n^2 - 12} \\ &= \frac{2(x_n^3 - 8)}{3(x_n^2 - 4)}. \end{aligned} \tag{4.6}$$

The eagle-eyed will notice that the final fraction has $(x-2)$ as a common factor. We could cancel these out, but if a root-finding problem from another source happened to have a double root, then it generally wouldn't be obvious that a cancellation could take place. Therefore all of the computations below will use the formula above. But I will nevertheless state that, as we converge towards the root, $x = 2$, then both the numerator and the denominator will tend to zero and this may cause some numerical round-off error.

$$\begin{aligned} x_0 &= -4.1 \\ x_1 &= -4.003\,174\,603 \\ x_2 &= -4.000\,003\,354 && \text{Quadratic convergence} \\ x_3 &= -4.000\,000\,000 && \text{Single root} \end{aligned}$$

$$\begin{aligned} x_0 &= 2.1 \\ x_1 &= 2.050\,407 \\ x_2 &= 2.025\,308 \\ x_3 &= 2.012\,680 && \text{Linear convergence} \\ x_4 &= 2.006\,347 && \text{Double root} \end{aligned}$$

Therefore we retain quadratic convergence for $x = -4$, the single root, but for the double root Newton-Raphson converges linearly. Whilst this is slower than for a single root, it is much better than for a double root being found with an *ad hoc* method.

4.11 Perturbation analysis for a double root.

Let us consider what happens to small errors for this double root. We shall let $x_n = 2 + \epsilon$ in Eq. (4.6), and hence,

$$
\begin{aligned}
x_{n+1} &= \frac{2}{3} \times \frac{(2+\epsilon)^3 - 8}{(2+\epsilon)^2 - 4} && \text{using (4.6)} \\
&= \frac{2}{3} \times \frac{(8 + 12\epsilon + 6\epsilon^2 + \epsilon^3) - 8}{(4 + 4\epsilon + \epsilon^2) - 4} \\
&= \frac{2}{3} \times \frac{(8 + 12\epsilon + 6\epsilon^2 + \epsilon^3 - 8)}{(4 + 4\epsilon + \epsilon^2 - 4)} \\
&= \frac{2}{3} \times \frac{(12\epsilon + 6\epsilon^2 + \epsilon^3)}{(4\epsilon + \epsilon^2)} \\
&= \frac{2 + \epsilon + \frac{1}{6}\epsilon^2}{1 + \frac{1}{4}\epsilon} && \text{cancelling the common factor, } \epsilon \\
&= (2 + \epsilon + \cdots)(1 - \tfrac{1}{4}\epsilon + \cdots) && \text{Can neglect } \epsilon^2 \text{ terms} \\
&= 2 + \tfrac{1}{2}\epsilon + \cdots && \text{and have used the Binomial expansion}
\end{aligned}
$$

Therefore the Newton-Raphson method converges linearly to this double root. In this case we see that errors halve for every iteration; this halving always happens for a double root and the proof of this is given by the solution of a problem sheet question.

4.12 Brief Summary of the Newton-Raphson method.

- The Newton-Raphson method is used ubiquitously primarily because it is a nice compromise between the rate of convergence and the complexity of the formula. Ad hoc methods are easier, but converge much more slowly. On the other hand, there are faster methods than the Newton-Raphson such as Halley's method, which has cubic convergence, and yes, this is *the* Edmund Halley of comet fame!
- For single roots the Newton-Raphson method displays quadratic convergence for all roots. A caveat is that one needs to be sufficiently close to the root for this to work; see later.
- For double roots, the performance for the Newton-Raphson method has been degraded, but nevertheless it still converges linearly which isn't hopeless!
- Although it hasn't been demonstrated here, the Newton-Raphson method also converges linearly (though a little slower) for triple roots and so on.
- **Outside of this unit:** Newton-Raphson is commonly used for solving systems of nonlinear algebraic equations, i.e. N equations in N unknowns. The derivation of this involves Taylor's series of more than one function, partial differentiation and matrices. Sounds terrible but bizarrely one gets used to it really quickly.

4.13 Some practical considerations.

In the summary on the previous page it was stated that the Newton-Raphson method will converge when the initial iterate is sufficiently close to the root. This is true, but there are other practical issues which I feel would be good for the new practitioner to know. So I'll discuss a few cases.

4.13.1 The root of a linear function.

Let us suppose that we wish to find the root of $x - a = 0$ numerically. Obviously this is $x = a$ by simple rearrangement, but my excuse for introducing this very very simple case is to illustrate a general idea. So please run with this for now....

We have $f(x) = x - a$. The Newton-Raphson scheme is,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

in general, but for this function it is,

$$x_{n+1} = x_n - \frac{x_n - a}{1} = a.$$

So we obtain the exact solution after just one iteration, and this hasn't even used the value of x_n! All initial guesses will give us the correct root after one iteration. Perhaps this shouldn't be surprising given the graphical motivation for the Newton-Raphson method we saw above.

An interesting implication from this arises when solving a Boundary Value Problem ODE numerically when that ODE is linear and we have one initial condition missing. If we were to use the Newton-Raphson method to iterate for that missing initial condition, then it too would yield the correct value after only one iteration.

4.13.2 A globally convergent case.

Suppose we really do wish to find the square root of 2 without using the square root button on your calculator, then we could use the Newton-Raphson method to do it using

$$f(x) = x^2 - 2 = 0.$$

This is an interesting case because any initial guess for $\sqrt{2}$ will yield a converging sequence of iterates. The method is

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n}$$

which simplifies to

$$x_{n+1} = \frac{x_n^2 + 2}{2x_n}.$$

In fact, we can see almost immediately that if x_n is monstrously huge then,

$$\begin{aligned} x_{n+1} &= \frac{x_n^2 + 2}{2x_n} \\ &\approx \frac{x_n^2}{2x_n} \qquad \text{because } x_n^2 \gg 2 \\ &= \tfrac{1}{2}x_n. \end{aligned}$$

Here, I have assumed that x_n^2 is so much larger than 2 that the latter can be neglected. Therefore, when $x_n \gg 1$, we expect the successive iterations to halve successively at first. That's the prediction so let us try it out with $x_0 = 100$:

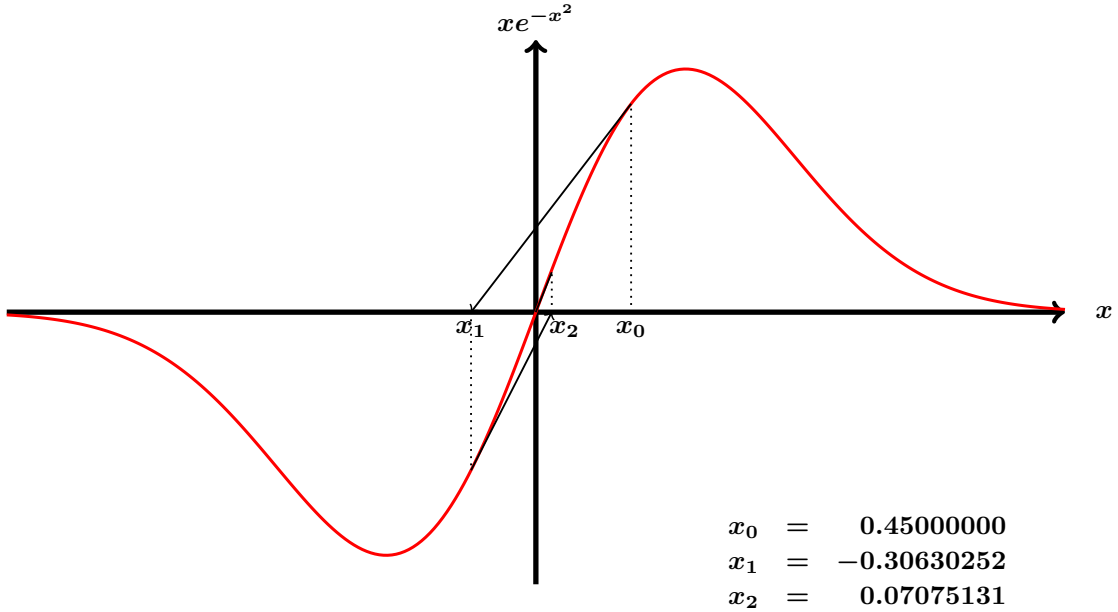
$$
\begin{aligned}
x_0 &= 100 \\
x_1 &= 50.009999999999998 && \text{the linear phase} \\
x_2 &= 25.024996000799838 \\
x_3 &= 12.552458046745901 \\
x_4 &= 6.3558946949311395 \\
x_5 &= 3.3352816092804338 \\
x_6 &= 1.9674655622311490 \\
x_7 &= 1.4920008896897232 \\
x_8 &= 1.4162413320389438 \\
x_9 &= 1.4142150140500531 && \text{the quadratic phase} \\
x_{10} &= 1.4142135623738401 \\
x_{11} &= 1.4142135623730951
\end{aligned}
$$

Those digits which have been coloured in red are correct.

4.13.3 A typical case.

The convergence characteristics depend vry much on the shape of the functions whose roots are being sought. In this final subsubsection we'll consider $f(x) = xe^{-x^2}$ which has only the one root, at $x = 0$, and where convergence proceeds differently depending on one's initial guess.

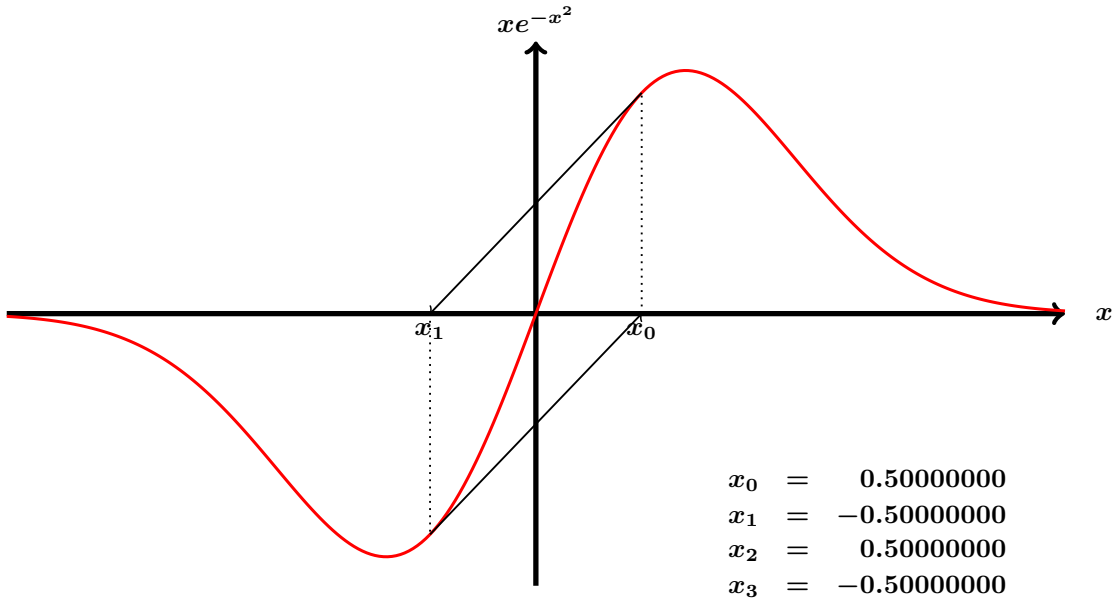
An example of convergence where $x_0 = 0.45$.



$$
\begin{aligned}
x_0 &= 0.45000000 \\
x_1 &= -0.30630252 \\
x_2 &= 0.07075131 \\
x_3 &= -0.00071549 \\
x_4 &= 0.00000000 \\
x_5 &= 0.00000000
\end{aligned}
$$

The initial iterate is within a zone of convergence.

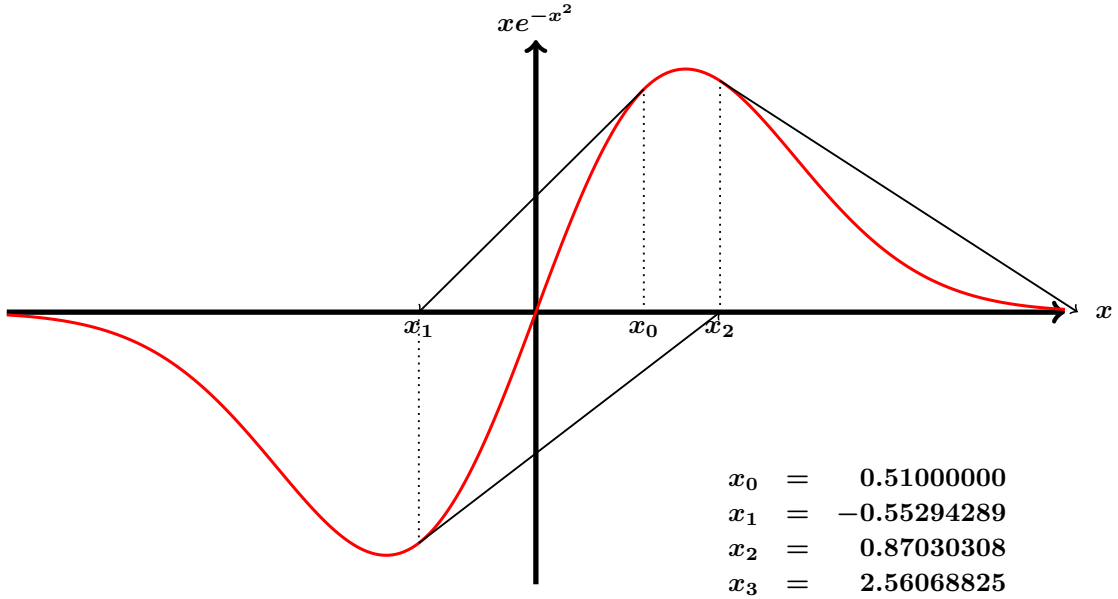
An example of a lack of convergence where $x_0 = 0.5$.



$$
\begin{aligned}
x_0 &= 0.50000000 \\
x_1 &= -0.50000000 \\
x_2 &= 0.50000000 \\
x_3 &= -0.50000000 \\
x_4 &= 0.50000000 \\
x_5 &= -0.50000000
\end{aligned}
$$

The initial iterate is on the borderline between convergence and lack of convergence. The resulting iterates merely oscillate between two values.

An example of divergence where $x_0 = 0.51$.



$$
\begin{aligned}
x_0 &= 0.51000000 \\
x_1 &= -0.55294289 \\
x_2 &= 0.87030308 \\
x_3 &= 2.56068825 \\
x_4 &= 2.77206647 \\
x_5 &= 2.96499036
\end{aligned}
$$

Successive iterations diverge.

5 FOURIER SERIES

5.1 Introduction

In ME10304 Mathematics 1 we saw how the Taylor Series technique was used to represent functions as power series. One example is

$$e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{(-t)^n}{n!}, \tag{5.1}$$

and another is

$$\cos t = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n}}{(2n)!}. \tag{5.2}$$

Fourier Series is a different type of infinite series where a periodic function is represented in terms of a suitable series of sines and/or cosines. One such Fourier series is,

$$\begin{aligned} g(t) &= \frac{1}{6} - \sum_{n=1}^{\infty} \frac{\cos 2\pi n t}{n^2\pi^2} \\ &= \frac{1}{6} - \frac{1}{\pi^2}\left[\cos 2\pi t + \frac{\cos 4\pi t}{2^2} + \frac{\cos 6\pi t}{3^2} + \frac{\cos 8\pi t}{4^2} + \cdots\right], \end{aligned}$$

which represents a function with a unit period.

Another series is the following,

$$f(t) = \frac{3}{8} + \sum_{n=1}^{\infty} \left[1 - 2\cos\left(\frac{n\pi}{2}\right) + 2\cos\left(\frac{3n\pi}{4}\right) - \cos n\pi\right] \frac{2}{n^2\pi^2} \cos n\pi t. \tag{5.3}$$

This is quite a complicated expression, but then the function which it represents is also quite complicated. One thing which may be said is that the function has a period of 2; the $n = 1$ term in (5.3) is $\cos \pi t$ which has a period of 2. The function is also even because every term is even.

Perhaps it is worth seeing how different partial sums of the above expression converge towards the function. [Note, by *partial sum* I mean the sum of the first N terms, say, of the infinite sum.]

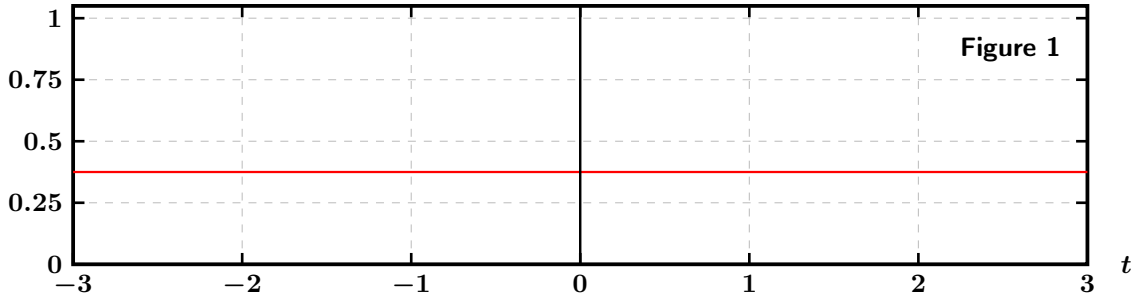
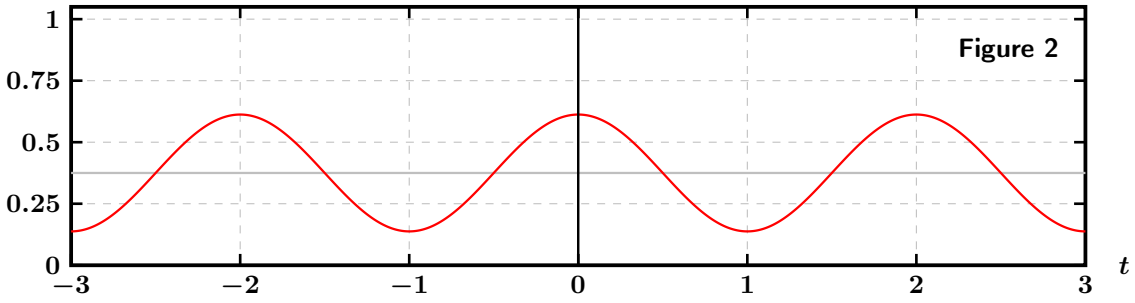
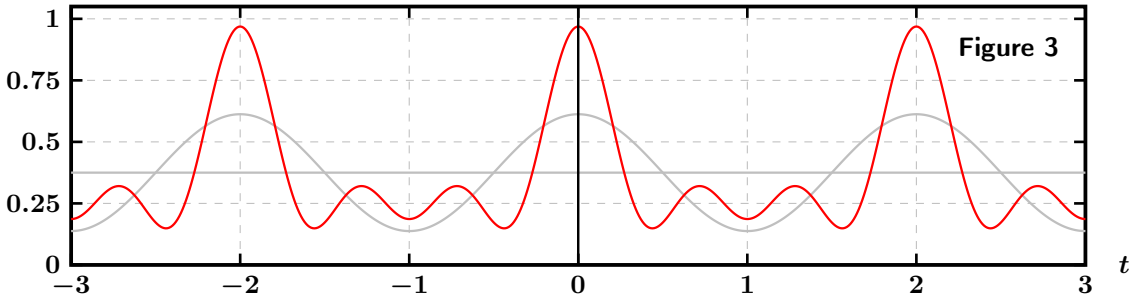


Figure 1

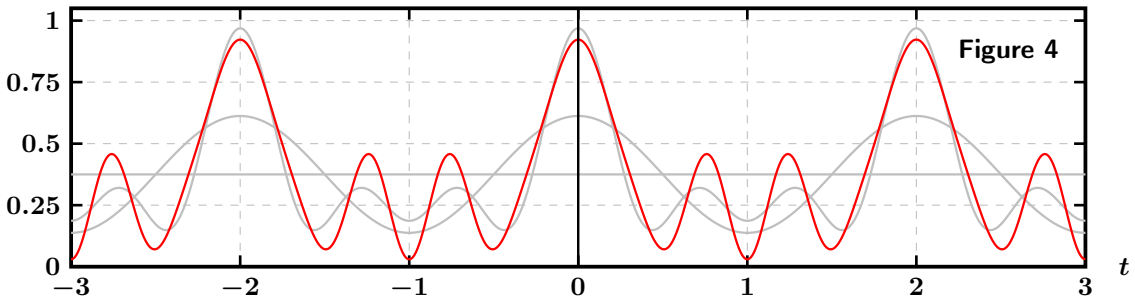
This is zeroth partial sum, i.e. just the constant in Eq. (5.3). Clearly this doesn't tell us much at all about what the function looks like, although this value is the mean of the function because each of the cosine terms in (5.3) has a mean of zero.



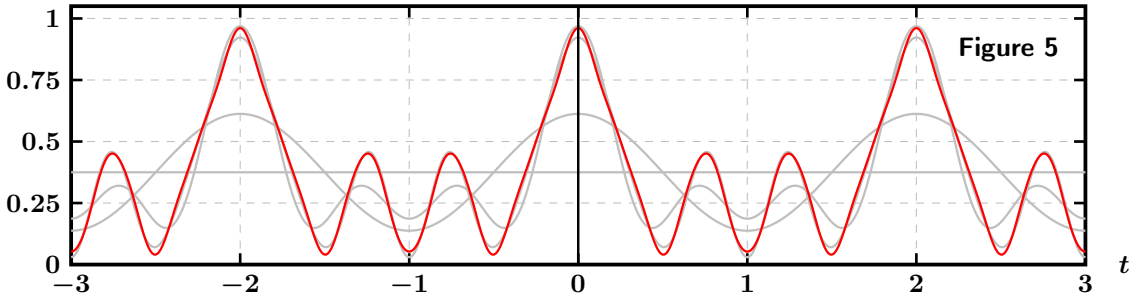
This is the $N = 1$ partial sum, i.e. the sum of the constant and the first cosine, and it is displayed in red. The grey line shows the $N = 0$ partial sum for comparison. Again there is a strong feeling that little may be said about what $f(t)$ looks like.



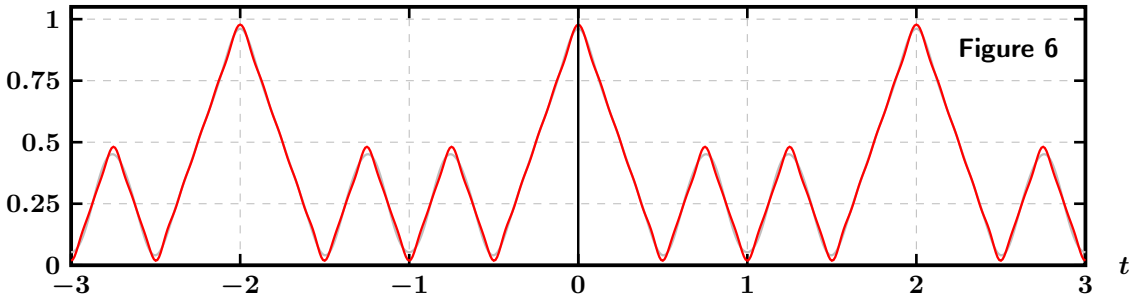
This represents the $N = 3$ partial sum, where the grey lines represent the $N = 0$ and $N = 1$ partial sums. This latest curve now feels as though we are getting some semblance of an idea of what $f(t)$ looks like, i.e. the function has well-defined peaks.



This represents the $N = 5$ partial sum, with the grey curves representing the $N = 0$, $N = 1$ and $N = 3$ partial sums.

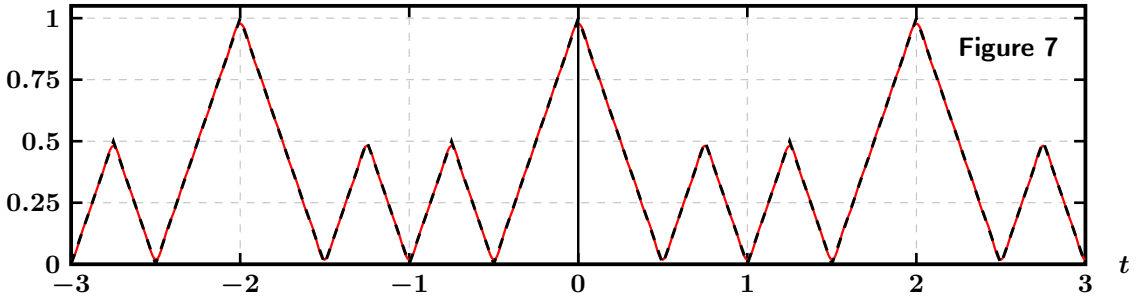


This is the $N = 10$ partial sum. Some parts of the red curve look fairly straight...



This is the $N = 20$ partial sum. Some more parts of the red curve now look straight too. The grey curve, which may only just be seen, is the $N = 10$ partial sum.

OK, so let us compare the $N = 20$ partial sum with $f(t)$ itself:



Here the dotted black curve is $f(t)$ and it is indeed composed solely of straight line segments. The period of this function is 2, and I have plotted three periods of the function itself.

Now I have to come clean and state explicitly what $f(t)$ is. In the range $-1 \leq t \leq 1$ it is

$$f(t) = \begin{cases} 2t + 2 & -1 \leq t \leq -\frac{3}{4}, \\ -2t - 1 & -\frac{3}{4} \leq t \leq -\frac{1}{2}, \\ 2t + 1 & -\frac{1}{2} \leq t \leq 0, \\ -2t + 1 & 0 \leq t \leq \frac{1}{2}, \\ 2t - 1 & \frac{1}{2} \leq t \leq \frac{3}{4}, \\ -2t + 2 & \frac{3}{4} \leq t \leq 1. \end{cases} \tag{5.4}$$

This represents the behaviour of $f(t)$ in the range, $-1 \leq t \leq 1$; this shape is then reproduced to the right (viz. $1 \leq t \leq 3$, $3 \leq t \leq 5$, etc) and to the left (viz. $-3 \leq t \leq -1$, $-5 \leq t \leq -3$, etc) in a fashion that makes the function periodic.

This complexity shown in Eq. (5.4) is what yielded the quite complicated coefficients in the cosine terms in Eq. (5.3). The reason I chose this example was to demonstrate that even something as bonkers as this function is may still be represented as a summation of cosines.

5.2 Fourier's Theorem — some preliminary concepts

The various examples of Fourier Series presented above were confined to ones involving only a constant and a series of cosine terms, i.e. to even functions. In practice we will also need sines to complete the general picture.

So the act of finding the Fourier Series of a periodic function is essentially the finding of the constant term and the coefficients of those sinusoids. The detailed formulae for this will be given later, but for now I would like to present a few essential preliminary ideas.

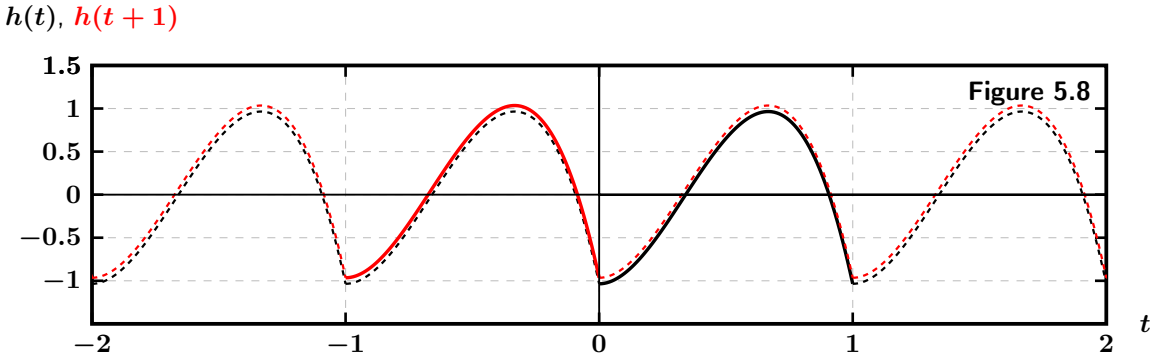
5.2.1 What is a periodic function?

This would seem to be a simple question but.....

The function $\cos t$ has a period of 2π and the function $\sin 2\pi t$ has a period of 1, as almost everyone would agree. The reason is simple — these functions repeat every interval of 2π and 1, respectively! However, $\cos t$ also repeats every interval of 4π, 6π and so on. So we could, therefore, be a little more precise and say that 2π is the *fundamental period* although almost everyone says the 2π is the period. But the fundamental period is the smallest of all the possible values that one could choose.

By constrast, functions such e^t, $t\sin t$ and t^2 are not periodic.

A rather strange notation is given by the following: If $h(t) = h(t+P)$, then the function $h(t)$ has the period, P. This is illustrated below:



The black curve is

$$h(t) = \tfrac{27}{4}(t^2 - t^3) - 1 \qquad \text{in the range } 0 \le t \le 1 \tag{5.5}$$

and where

$$h(t) = h(t+1) \quad \text{for all values of } t. \tag{5.6}$$

That part of the function which is defined in (5.5) corresponds to the continuous black line, whereas its periodic extension to other values of t which is given in (5.6) corresponds to the dashed black line. The red line (both continuous and dashed) then corresponds to $h(t+1)$ and shows the backward phase shift by -1.

5.2.2 Symmetries

We shall consider briefly the effect of symmetries on the properties of functions.

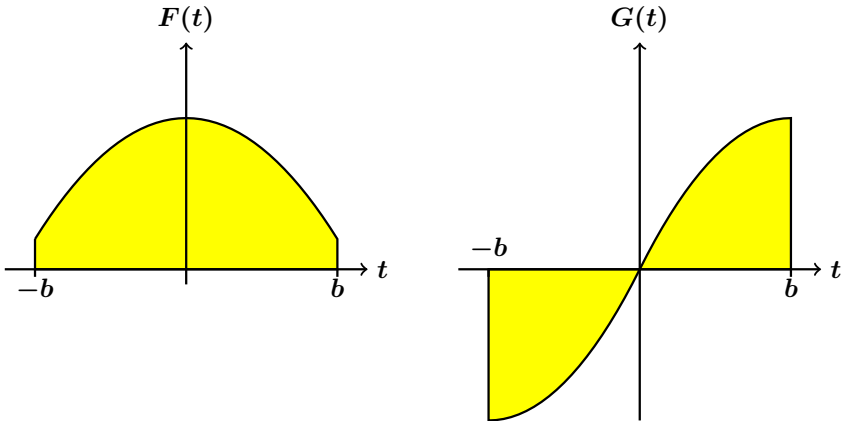


Figure 5.9. Showing an even function, $F(t)$, and an odd function, $G(t)$.

The left hand sketch in Fig. 5.9 shows an even function. This is also known as a *symmetric function* about $t = 0$ and it could also be described as having a mirror symmetry about the vertical axis. It satisfies the functional relationship, $F(t) = F(-t)$. Examples include $\cos t$, $\sin^2 t$, t^4, $|t|$, $1/(1 + t^2)$.

Given that the area under the curve for $F(t)$ to the right or the vertical axis is the same as that to the left of the axis, we may write down the following property of even functions,

$$\int_{-b}^{b} F(t)\, dt = 2 \int_{0}^{b} F(t)\, dt. \tag{5.7}$$

This is useful since fewer arithmetical mistakes are made when the lower limit is zero than when it is negative.

The right hand sketch in Fig. 5.9 shows an odd function. This is also known as an *antisymmetric function* about $t = 0$, and it is sometimes described as having a rotational symmetry about the origin. It satisfies the functional relationship, $F(t) = -F(-t)$. This one always feels a little odd, but we're used to the specific case, $\sin(t) = -\sin(-t)$. Examples include $\sin t$, t^3, $t \cos t$, $t/(1 + t^2)$.

The actual amounts of yellow which are displayed are equal to the left and to the right, but the area to the left of the vertical axis is negative, mathematically. Therefore the area to the left and the area to the right are of equal magnitude but have opposite signs. This leads to the following property for odd functions,

$$\int_{-b}^{b} G(t)\, dt = 0. \tag{5.8}$$

The properties given in Eqs. (5.7) and (5.8) will be used quite often in the following text.

We also have the following symmetry relationships for functions:

$$\texttt{even} \times \texttt{even} = \text{even} \qquad \texttt{even} \times \texttt{odd} = \text{odd} \qquad \texttt{odd} \times \texttt{odd} = \text{even}. \tag{5.9}$$

Finally, functions which are neither even nor odd are called *asymmetric*, the 'a' prefix meaning "not' or without', as in *aperiodic*, *atheistic*, *atypical*, *achromatic*, *amoral*, *anoxic*, *asymptomatic* and *amusement*. Examples include e^t, $\sin t + \cos t$, $H(t)$ (the unit step function), $1/(1 + t + t^2)$.

5.2.3 The Fourier Series

If a function, $f(t)$, has fundamental period, P, i.e. that $f(t) = f(t+P)$ for all values of t, then we may rewrite the function in the form,

$$f(t) = \tfrac{1}{2}A_0 + \sum_{n=1}^{\infty}\left[A_n \cos\frac{2\pi nt}{P} + B_n \sin\frac{2\pi nt}{P}\right]. \tag{5.10}$$

These expressions show explicitly that $f(t)$ will typically contain an infinite number of sinusoidal components. As mentioned before, there is a constant term, which is written as $\frac{1}{2}A_0$ by convention, which may be interpreted as the average of $f(t)$ since both the sines and the cosines average out to zero over one period.

The coefficients, A_n and B_n, are given by

$$A_n = \frac{2}{P}\int_{t_0}^{t_0+P} f(t)\,\cos\frac{2\pi nt}{P}\,dt, \qquad \text{for } n = 0, 1, 2, \cdots, \tag{5.11}$$

$$B_n = \frac{2}{P}\int_{t_0}^{t_0+P} f(t)\,\sin\frac{2\pi nt}{P}\,dt, \qquad \text{for } n = 1, 2, 3, \cdots. \tag{5.12}$$

Note that the A_n coefficients are conventionally associated with the cosines and B_n with the sine terms. Note also that there is no B_0 term because the corresponding sine would be zero!

A third note is concerned with the presence of the mysterious t_0 in the limits of the integrals. This merely represents an arbitrary number. The most important thing is that the range of integration is P, the given period of $f(t)$. In practice, the value of t_0 will be quite obvious given how the function is defined. Referring to the function given in Eq. (5.4) the period is $P = 2$, but given the precise range of values over which $f(t)$ has been defined explicitly, one would use $t_0 = -1$ in (5.11) and (5.12). A second example is Eq. (5.5); here the period is $P = 1$ and we would use $t_0 = 0$ in (5.10) because this is the lower end of the given range of t-values. These two examples illustrate the two most common ranges of integration which are $\int_{-P/2}^{P/2}$ and $\int_0^P$.

5.2.4 How to derive the formulae for the Fourier coefficients

This subsection is for information only — this analysis will not be examined.

It may be wondered where the expressions given in Eqs. (5.11) and (5.12) come from. We may start by assuming that Eq. (5.10) is correct. Then it may be integrated over one period where, for convenience, I will take $t_0 = -P/2$. In this way Eq. (5.10) becomes,

$$\begin{aligned}\int_{-P/2}^{P/2} f(t)\,dt &= \int_{-P/2}^{P/2} \tfrac{1}{2}A_0\,dt + \sum_{n=1}^{\infty}\left[A_n\int_{-P/2}^{P/2}\cos\frac{2\pi nt}{P}\,dt + B_n\int_{-P/2}^{P/2}\sin\frac{2\pi nt}{P}\,dt\right]\\ &= \tfrac{1}{2}A_0 P.\end{aligned}$$

All of the integrals which involve sines and cosines are integrals over a whole number of periods of each of these functions and hence each of the integrals is zero. The only one which is left is the constant term. Hence,

$$A_0 = \frac{2}{P}\int_{-P/2}^{P/2} f(t)\,dt,$$

as given by (5.11) when $n = 0$.

To find the A_n values we may first multiply (5.10) by $\cos(2\pi mt/P)$ (note the m in the cosine) and integrate in the same way. We get

$$\begin{aligned}
&\int_{-P/2}^{P/2} f(t)\cos\frac{2\pi mt}{P}\,dt \\
&= \int_{-P/2}^{P/2} \tfrac{1}{2}A_0\cos\frac{2\pi mt}{P}\,dt + \sum_{n=1}^{\infty}\left[A_n\int_{-P/2}^{P/2}\cos\frac{2\pi nt}{P}\cos\frac{2\pi mt}{P}\,dt + B_n\int_{-P/2}^{P/2}\sin\frac{2\pi nt}{P}\cos\frac{2\pi mt}{P}\,dt\right] \\
&= \tfrac{1}{2}A_mP.
\end{aligned}$$

This needs to be explained! The term involving $\boxed{A_0}$ involves an integral of a cosine over a whole number of periods, and therefore that term is zero.

The terms involving $\boxed{B_n}$ involve a sine multiplied by a cosine; the use of the multiple angle formulae means that these integrands may be changed to the sum of two sines:

$$\sin\frac{2\pi nt}{P}\cos\frac{2\pi mt}{P}\,dt = \frac{1}{2}\left[\sin\frac{2\pi(n+m)t}{P} + \sin\frac{2\pi(n-m)t}{P}\right],$$

and again, integration over one period renders these terms to be zero. They are also odd functions, so this is a second argument for a zero integral!

The terms involving $\boxed{A_n}$ may also be simplified using a multiple angle formulae:

$$\cos\frac{2\pi nt}{P}\cos\frac{2\pi mt}{P} = \frac{1}{2}\left[\cos\frac{2\pi(n-m)t}{P} - \cos\frac{2\pi(n+m)t}{P}\right].$$

All the terms, $\cos(2\pi(n+m)t/P)$, integrate to zero, as do all of the $\cos(2\pi(n-m)t/P)$ terms except for the one where $n=m$. In this case we have

$$\cos\frac{2\pi(n-m)t}{P} = 1$$

and therefore it is this term which has a nonzero integral and it is what provides the above $\frac{1}{2}A_mP$ result.

To find the B_n values we may first multiply (5.10) by $\sin(2\pi mt/P)$ and integrate again. We get

$$\begin{aligned}
&\int_{-P/2}^{P/2} f(t)\sin\frac{2\pi mt}{P}\,dt \\
&= \int_{-P/2}^{P/2} \tfrac{1}{2}A_0\sin\frac{2\pi mt}{P}\,dt + \sum_{n=1}^{\infty}\left[A_n\int_{-P/2}^{P/2}\cos\frac{2\pi nt}{P}\sin\frac{2\pi mt}{P}\,dt + B_n\int_{-P/2}^{P/2}\sin\frac{2\pi nt}{P}\sin\frac{2\pi mt}{P}\,dt\right] \\
&= \tfrac{1}{2}B_mP.
\end{aligned}$$

Once more, all of the integrals except the one for which $n=m$ are zero.

Comment 1:

There is something interesting going on behind the scenes here. Perhaps I could illustrate it this way: let c_n denote $\cos(2\pi nt/P)$. Then,

$$\frac{2}{P}\int_{-P/2}^{P/2} c_n c_m \, dt = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$$

This is reminscent of what happens when we take the dot product of any two of the three unit vectors in 3D Cartesian space. If we define each of the unit vectors like this:

$$\underline{v}_1 = \hat{\underline{i}}, \qquad \underline{v}_2 = \hat{\underline{j}}, \qquad \underline{v}_3 = \hat{\underline{k}}$$

then

$$\underline{v}_n . \underline{v}_m = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$$

And just as each point in 3D may be written as the sum of multiples of those three unit vectors, the multiples being called *coordinates*, so the Fourier coefficients *may* be regarded as being coordinates in an infinite dimensional space, where each cosine (and sine) may be regarded as the equivalent of the unit vectors. So each periodic function corresponds to one point in an infinite dimensional space! I doubt if we'll use that in good hard engineering practice, but I still think that it is interesting.

Comment 2:

Another way of viewing the act of finding the Fourier coefficients is that the process of multiplying by the cosine or the sine and then integrating over one period may be regarded as being a *filter*. We have a periodic function, $f(t)$, and the process, **multiplication followed by integration**, filters out all of those terms which we are not interested in. This is similar to having a piece of, say, yellow cellophane placed in front of a coloured picture where only the light from the yellow component of that picture is able to pass through. We can then measure the amount of yellow there is in the picture by this removal of other frequencies.

Figure 5.10. An illustration of the filtering out of non-yellow frequencies.

5.3 Some examples

We will consider four different examples of Fourier Series.

5.3.1 Example 5.1.

Let us apply the formulae given in (5.11) and (5.12) to the function, $f(t)$, which is defined as follows,

$$f(t) = t - t^2 \quad \text{in} \quad 0 \le t \le 1, \qquad f(t) = f(t+1) \quad \text{for all values of } t. \tag{5.13}$$

This function is defined explicitly over one period, and then the rest of the function is determined by insisting that it has a period equal to 1, as shown in Fig. 5.11, below.

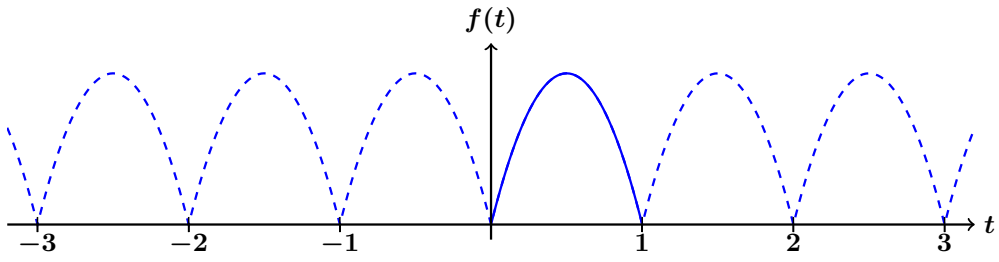


Figure 5.11. The function $f(t)$ given by equation (5.13). The solid line depicts the function as defined in the interval $0 \le t \le 1$. The dashed line shows the periodic extension of that interval.

When $P = 1$ the Fourier series and the formulae for the Fourier coefficients which are given in Eqs. (5.10) to (5.12) become,

$$f(t) = \tfrac{1}{2}A_0 + \sum_{n=1}^{\infty} \Big[A_n \cos\, 2\pi nt + B_n \sin\, 2\pi nt \Big], \tag{5.14}$$

where

$$A_n = 2\int_0^1 f(t)\, \cos\, 2\pi nt\, dt, \qquad \text{for } \; n = 0, 1, 2, \cdots, \tag{5.15}$$

$$B_n = 2\int_0^1 f(t)\, \sin\, 2\pi nt\, dt, \qquad \text{for } \; n = 1, 2, 3, \cdots, \tag{5.16}$$

where we note that the lower limit is chosen to be consistent with the definition of $f(t)$ in (5.13).

For A_0 we obtain,

$$A_0 = 2\int_0^1 (t - t^2)\, dt = \frac{1}{3}.$$

For $\boldsymbol{A_n}$ and $\boldsymbol{B_n}$ we find that,

$$\begin{aligned}
\boldsymbol{A_n} &= \mathbf{2}\int_0^1 (\boldsymbol{t}-\boldsymbol{t^2})\cos \mathbf{2}\pi \boldsymbol{n t}\, \boldsymbol{dt},\\
&= \mathbf{2}\Big[(\boldsymbol{t}-\boldsymbol{t^2})\Big(\frac{\sin \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{2}\pi \boldsymbol{n}}\Big) - (\mathbf{1}-\mathbf{2}\boldsymbol{t})\Big(\frac{-\cos \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{4}\pi^{\mathbf{2}}\boldsymbol{n^2}}\Big) + (\mathbf{-2})\Big(\frac{-\sin \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{8}\pi^{\mathbf{3}}\boldsymbol{n^3}}\Big)\Big]_{\mathbf{0}}^{\mathbf{1}},\\
&= \Big[\frac{(\mathbf{1}-\mathbf{2}\boldsymbol{t})\cos \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{2}\pi^{\mathbf{2}}\boldsymbol{n^2}}\Big]_{\mathbf{0}}^{\mathbf{1}} \quad = \quad \Big[\frac{-\mathbf{1}-(\mathbf{1})}{\mathbf{2}\pi^{\mathbf{2}}\boldsymbol{n^2}}\Big] \quad = \quad -\frac{\mathbf{1}}{\pi^{\mathbf{2}}\boldsymbol{n^2}},\\
\boldsymbol{B_n} &= \mathbf{2}\int_0^1 (\boldsymbol{t}-\boldsymbol{t^2})\sin \mathbf{2}\pi \boldsymbol{n t}\, \boldsymbol{dt},\\
&= \mathbf{2}\Big[(\boldsymbol{t}-\boldsymbol{t^2})\Big(\frac{-\cos \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{2}\pi \boldsymbol{n}}\Big) - (\mathbf{1}-\mathbf{2}\boldsymbol{t})\Big(\frac{-\sin \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{4}\pi^{\mathbf{2}}\boldsymbol{n^2}}\Big) + (\mathbf{-2})\Big(\frac{\cos \mathbf{2}\pi \boldsymbol{n t}}{\mathbf{8}\pi^{\mathbf{3}}\boldsymbol{n^3}}\Big)\Big]_{\mathbf{0}}^{\mathbf{1}},\\
&= \mathbf{0}.
\end{aligned}$$

Having obtained the coefficients it remains to substitute them into the definition of the Fourier Series:

$$\boldsymbol{f(t)} = \frac{\mathbf{1}}{\mathbf{6}} + \sum_{\boldsymbol{n}=\mathbf{1}}^{\infty}\Big(-\frac{\mathbf{1}}{\pi^{\mathbf{2}}\boldsymbol{n^2}}\Big)\cos \mathbf{2}\pi \boldsymbol{n t}.$$

For this particular definition of $\boldsymbol{f(t)}$ the $\boldsymbol{B_n}$ coefficients are precisely zero. The integration by parts was unnecessary because the sketch of $\boldsymbol{f(t)}$ shows that it is an even function, and therefore there will not be any sine terms in the Fourier Series because each one is an odd function.

5.3.2 Example 5.2.

Let us find the Fourier Series of

$$\boldsymbol{f(t)} = \boldsymbol{t^2} \quad \text{in} \quad \boldsymbol{-1 \le t \le 1}, \qquad \boldsymbol{f(t)} = \boldsymbol{f(t+2)} \quad \text{for all values of } \boldsymbol{t}. \tag{5.17}$$

The sketch of the function is given below.

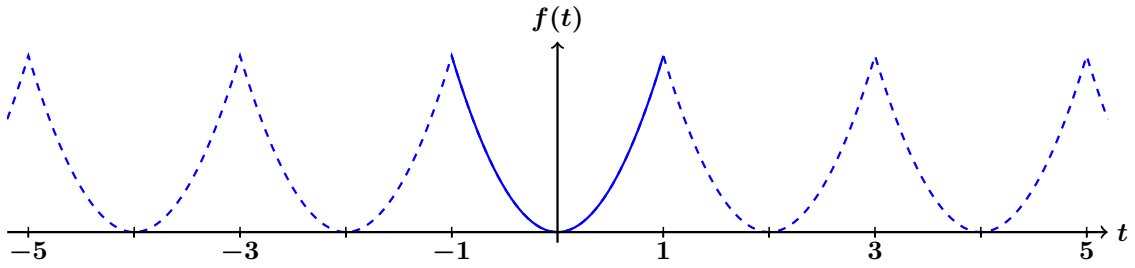


Figure 5.12. The function $\boldsymbol{f(t)}$ given by equation (5.17). The solid line depicts the function as defined in the interval $\boldsymbol{-1 \le t \le 1}$. The dashed line shows the periodic extension of that interval.

This function is also even as seen above, and therefore we may state immediately that $\boldsymbol{B_n = 0}$ for all $\boldsymbol{n}$-values since the $\boldsymbol{B_n}$ coefficients multiply sines which are odd.

The period is $P = 2$ and, given the range over which $f(t)$ is defined, we will integrate between $t = -1$ and $t = 1$. Therefore Eqs. (5.10) to (5.12) become,

$$f(t) = \tfrac{1}{2}A_0 + \sum_{n=1}^{\infty}\Big[A_n \cos n\pi t + B_n \sin n\pi t\Big], \tag{5.18}$$

$$A_n = \int_{-1}^{1} f(t)\, \cos n\pi t\, dt, \qquad \text{for } n = 0, 1, 2, \cdots, \tag{5.19}$$

$$B_n = \int_{-1}^{1} f(t)\, \sin n\pi t\, dt, \qquad \text{for } n = 1, 2, 3, \cdots. \tag{5.20}$$

Using (5.18) and (5.19) we get,

$$A_0 = \int_{-1}^{1} t^2\, dt = 2\int_{0}^{1} t^2\, dt = \tfrac{2}{3}$$

and

$$\begin{aligned} A_n &= \int_{-1}^{1} t^2 \cos n\pi t\, dt && \text{using (5.19)} \\ &= 2\int_{0}^{1} t^2 \cos n\pi t\, dt && \text{using the even symmetry} \\ &= \frac{4\cos n\pi}{\pi^2 n^2} && \text{after integration by parts} \\ &= \frac{4(-1)^n}{\pi^2 n^2} && \text{using } \cos n\pi = (-1)^n. \end{aligned}$$

There is no need to use Eq. (5.20) given that this $f(t)$ is even, and hence $B_n = 0$. Therefore the Fourier series is

$$\begin{aligned} f(t) &= \tfrac{1}{3} + \sum_{n=1}^{\infty} \frac{4(-1)^n \cos n\pi t}{\pi^2 n^2} \\ &= \tfrac{1}{3} + \frac{4}{\pi^2}\Big[-\cos \pi t + \frac{\cos 2\pi t}{2^2} - \frac{\cos 3\pi t}{3^2} + \frac{\cos 4\pi t}{4^2} \cdots\Big]. \qquad \text{showing the first few terms.} \end{aligned} \tag{5.21}$$

Note: It is possible to use expressions like the present Fourier Series to obtain results involving the summation of series. For example, if we set $t = 0$ in (5.21), then $f(t) = 0$ and therefore (5.21) becomes

$$0 = \tfrac{1}{3} + \frac{4}{\pi^2}\Big[-1 + \frac{1}{2^2} - \frac{1}{3^2} + \frac{1}{4^2} - \frac{1}{5^2} + \cdots\Big],$$

which, upon rearrangement, becomes

$$1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots = \frac{\pi^2}{12}.$$

Similarly, the setting of $t = 1$ gives

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}.$$

If one had been presented with either of these series and asked if they converge, then d'Alembert's test would be inconclusive. But here we have shown that they are convergent and, even better, what they converge to.

5.3.3 Example 5.3.

We shall find the Fourier Series of

$$f(t) = t^2 \quad \text{in} \quad 0 \le t \le 2, \qquad f(t) = f(t+2) \quad \text{for all values of } t. \tag{5.22}$$

The sketch of the function is given below.

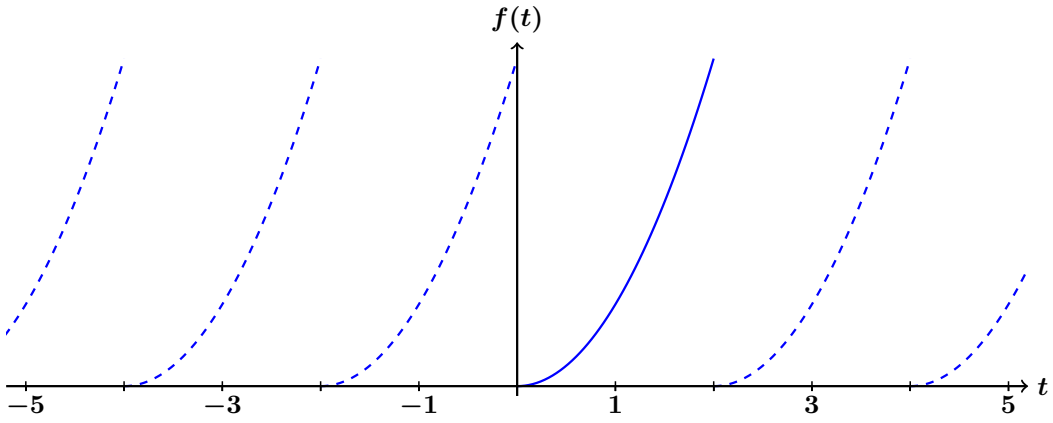


Figure 5.13. The function $f(t)$ given by equation (5.22). The solid line depicts the function as defined in the interval $0 \le t \le 2$. The dashed line shows the periodic extension of that interval.

Superficially, if one glanced at the mathematical definition of this function, it would be easy to think that it is identical to the one given in Example 5.2 because both are defined as being $f(t) = t^2$ and both have a period equal to 2. However, the explicit definitions of the functions in these two examples correspond to different ranges of values of t and this makes a HUGE difference (i) to how they look, (ii) their respective symmetries and (iii) to the resulting Fourier series. The present one is shown in the above Figure, and it is clearly neither even nor odd, which means that there will be both A_n and B_n values.

We should use Eqs. (5.10) to (5.12) with $P = 2$ and $t_0 = 0$, i.e.,

$$f(t) = \tfrac{1}{2}A_0 + \sum_{n=1}^{\infty}\Big[A_n \cos n\pi t + B_n \sin n\pi t\Big], \tag{5.23}$$

$$A_n = \int_0^2 f(t)\, \cos n\pi t\, dt, \qquad \text{for } \; n = 0, 1, 2, \cdots, \tag{5.24}$$

$$B_n = \int_0^2 f(t)\, \sin n\pi t\, dt, \qquad \text{for } \; n = 1, 2, 3, \cdots. \tag{5.25}$$

The Fourier coefficients are as follows:

$$A_0 = \int_0^2 t^2\, dt = \frac{8}{3}, \qquad A_n = \int_0^2 t^2 \cos n\pi t\, dt = \frac{4}{n^2\pi^2}, \qquad B_n = \int_0^2 t^2 \sin n\pi t\, dt = -\frac{4}{n\pi}.$$

Hence the Fourier series of the above function is,

$$f(t) = \frac{4}{3} + \sum_{n=1}^{\infty}\Big[\Big(\frac{4}{n^2\pi^2}\Big)\cos n\pi t - \Big(\frac{4}{n\pi}\Big)\sin n\pi t\Big]. \tag{5.26}$$

5.3.4 Example 5.4.

We shall find the Fourier Series of

$$f(t) = t - t^3 \quad \text{in} \quad -1 \leq t \leq 1, \qquad f(t) = f(t+2) \quad \text{for all values of } t. \tag{5.27}$$

The sketch of the function is given below.

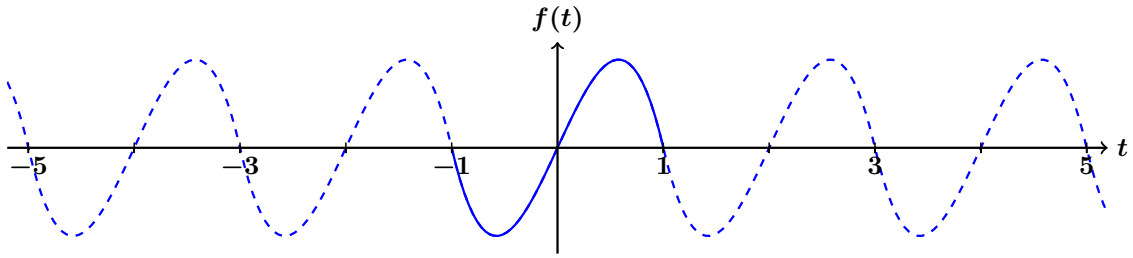


Figure 5.14. The function $f(t)$ given by equation (5.27). The solid line depicts the function as defined in the interval $0 \leq t \leq 2$. The dashed line shows the periodic extension of that interval.

The function is clearly odd and therefore $A_0 = 0$ (a constant is formally an even function) and $A_n = 0$. The values of B_n are now given by,

$$\begin{aligned} B_n &= \int_{-1}^{1} (t - t^3) \sin n\pi t \, dt && \text{from (5.20)} \\ &= 2\int_{0}^{1} (t - t^3) \sin n\pi t \, dt && \text{odd} \times \text{odd} = \text{even} \\ &= \frac{12(-1)^{n+1}}{n^3\pi^3}. && \text{after three integrations by parts} \end{aligned}$$

Hence the Fourier Series is

$$f(t) = \sum_{n=1}^{\infty} \frac{12(-1)^{n+1}}{n^3\pi^3} \sin n\pi t.$$

5.4 On rates of convergence

In the four examples above it may be seen that there are different speeds of convergence of the series. The Fourier coefficients in Examples 5.1 and 5.2 decay like n^{-2} while the slower ones in Example 5.3 decay like n^{-1} and those in Example 5.4 decay like n^{-3}. Why is this?

The answer is that the rate of decay depends on the degree of continuity of the function.

We will illustrate how the speed of convergence of the Fourier coefficients depends on the degree of continuity of the function by considering the three functions, $f_1(t)$, $f_2(t)$ and $f_3(t)$, which are defined in the interval $-1 < t < 1$ according to

$$f_1(t) = t \qquad f_2(t) = 1 - t^2 \qquad f_3(t) = t - t^3, \tag{5.28}$$

and each has a period equal to 2.

The following Figure shows the variation with t of $f_1(t)$.

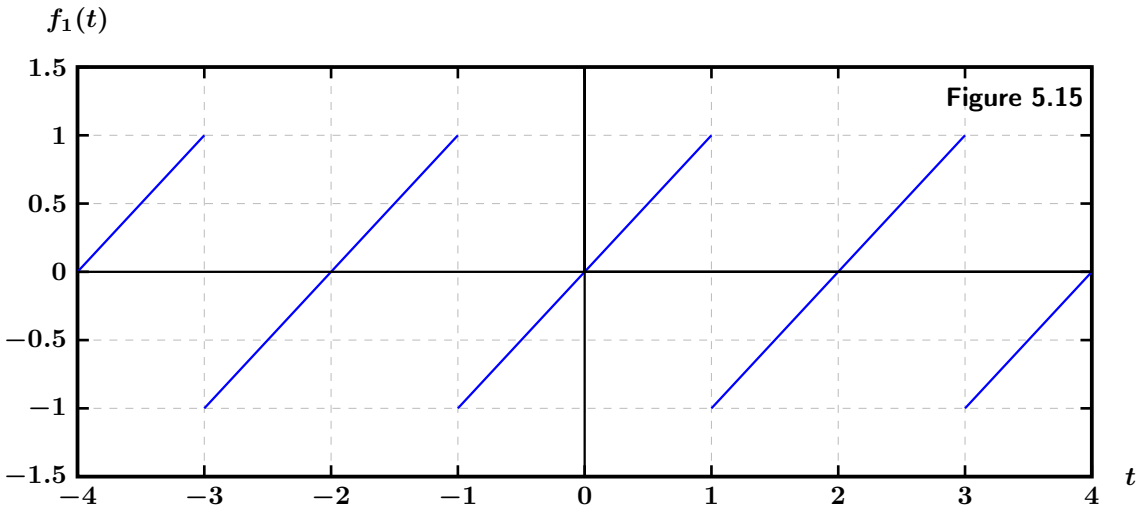


Figure 5.15

This function is continuous except at $t = \pm 1, \pm 3, \pm 5, \cdots$. At $t = 1$, the limit as t tends towards 1 from below is $f_1 = 1$ while the limit as t tends towards 1 from above is $f_1 = -1$. This second limit is, in fact, identical to saying that the limit as t tends towards -1 from above is $f_1 = -1$. If one is happy with that statement, then a quicker way of determining if $f_1(t)$ is continuous at these end-points is to look at the values of f_1 at the two end-points.

Referring to Eq. (5.28) we see that $\boxed{f_1(-1) \neq f_1(1)}$ and therefore the function is discontinuous.

For the function, $f_2(t)$, we see that f_2 is continuous but f_2' is discontinuous since

$$\boxed{f_2(1) = f_2(-1)} \text{ but } \boxed{f_2'(1) \neq f_2'(-1)}.$$

For the function, $f_3(t)$, we see that f_3 and f_3' are continuous but f_3'' is discontinuous since

$$\boxed{f_3(1) = f_3(-1)} \text{ and } \boxed{f_3'(1) = f_3'(-1)} \text{ but } \boxed{f_3''(1) \neq f_3''(-1)}.$$

How do these observations about the continuity of the functions affect the speed of convergence of their respective Fourier Series? Well, here are the series.....

$$f_1 = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin n\pi t, \qquad f_1 \text{ is discontinuous}$$

$$f_2 = \frac{2}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} \cos n\pi t, \qquad f_2 \text{ is continuous, } f_2' \text{ is discontinuous}$$

$$f_3 = \frac{12}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^3} \sin n\pi t, \qquad f_3 \text{ and } f_3' \text{ are continuous, } f_3'' \text{ is discontinuous}$$

and so on. Therefore if we were to have a function, $F(t)$, for which F, F', F'' and F''' are continuous but F'''' is discontinuous then the Fourier coefficients will decay as n^{-5}.

5.5 What happens to the Fourier Series at a discontinuity?

When we find the Fourier Series of a function we are replacing it by a set of sines and cosines all of which are continuous functions. So what happens when those continuous functions try to model a discontinuous function? We will look at this using the following function:

$$\mathcal{F}(t) = e^t \quad \text{in} \quad -1 \le t \le 1, \qquad \mathcal{F}(t) = \mathcal{F}(t+2) \quad \text{for all values of } t. \tag{5.29}$$

We can predict in advance of any sketch or plot that $\mathcal{F}(t)$ has a discontinuity because $\mathcal{F}(-1) \neq \mathcal{F}(1)$ since $\mathcal{F}(-1) = e^{-1}$ and $\mathcal{F}(1) = e$.

Its Fourier Series is

$$\mathcal{F}(t) = \sinh 1 + \sum_{n=1}^{\infty} \left(\frac{2(-1)^n \sinh 1}{1 + n^2\pi^2} \right) \Big(\cos n\pi t - n\pi \sin n\pi t \Big). \tag{5.30}$$

It is a good exercise to check that this Fourier Series is correct for the simplification at the end is awkward.

The coefficient of the sine terms is $-n\pi/(1 + n^2\pi^2)$ which is essentially an n^{-1} behaviour when n is large, and this reflects the discontinuity in $\mathcal{F}$.

The next four Figures show both $\mathcal{F}$ and 5, 10, 25 and 100 terms in the Fourier Series.

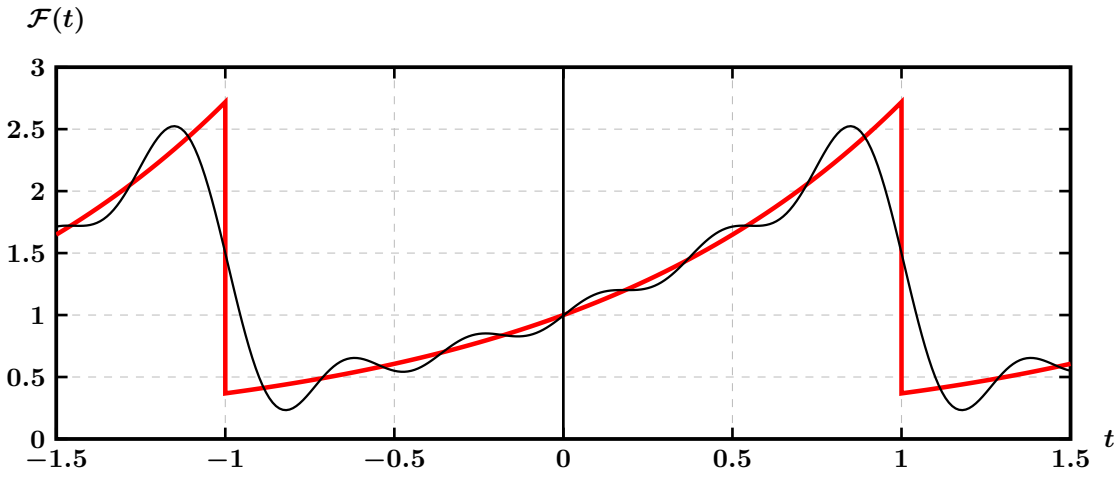


Figure 5.16. The function $\mathcal{F}(t)$ given by equation (5.29) and shown in red. Also displayed is the 5-term partial sum of its Fourier Series as given by Eq. (5.30) and displayed as the black curve.

Here we see that the Fourier Series approximates fairly closely the function, $\mathcal{F}$, except for near to the discontinuity — an act of mathematical democracy. At the discontinuity itself, the Fourier Series takes a value which is roughly the mean of two possible values that $\mathcal{F}$ can take at these points.

As the number of terms in the partial sums increases, the function is approimated increasingly well. The 'wavelength' of the 'wiggles' in the Fourier Series curve decreases. Very close to any discontinuity, the Fourier Series overshoots the maximum in $\mathcal{F}$ (and undershoots the minimum) by an amount which is roughly 9% of drop between the ends of the discontinuity — this is known as the Gibbs Phenomenon. As the number of terms increase, this region of overshoot is compressed increasingly and for practical purposes the overshoot becomes insignificant.

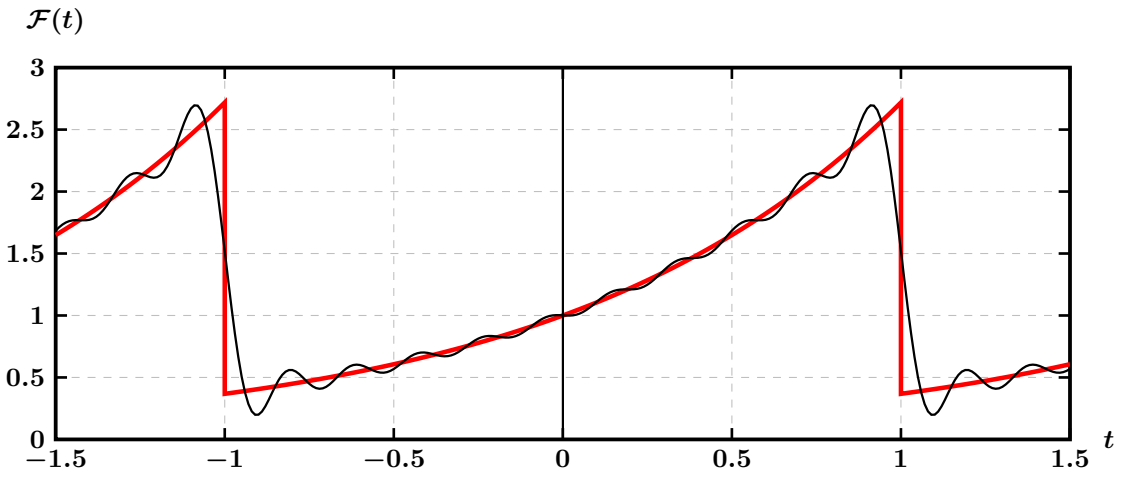


Figure 5.17. The function $\mathcal{F}(t)$ given by equation (5.29) and shown in red. Also displayed is the $\mathbf{10}$-term partial sum of its Fourier Series.

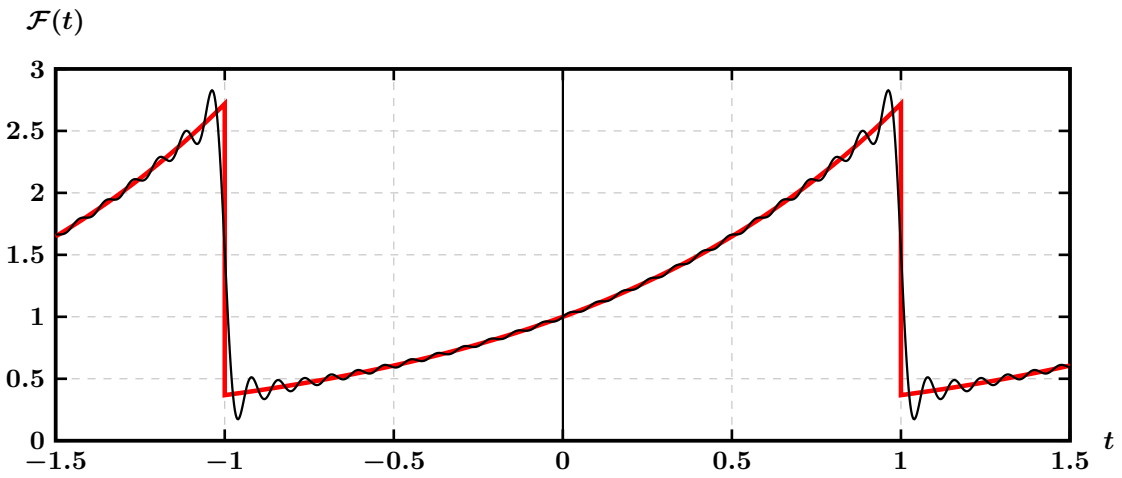


Figure 5.18. The function $\mathcal{F}(t)$ given by equation (5.29) and shown in red. Also displayed is the $\mathbf{25}$-term partial sum of its Fourier Series.

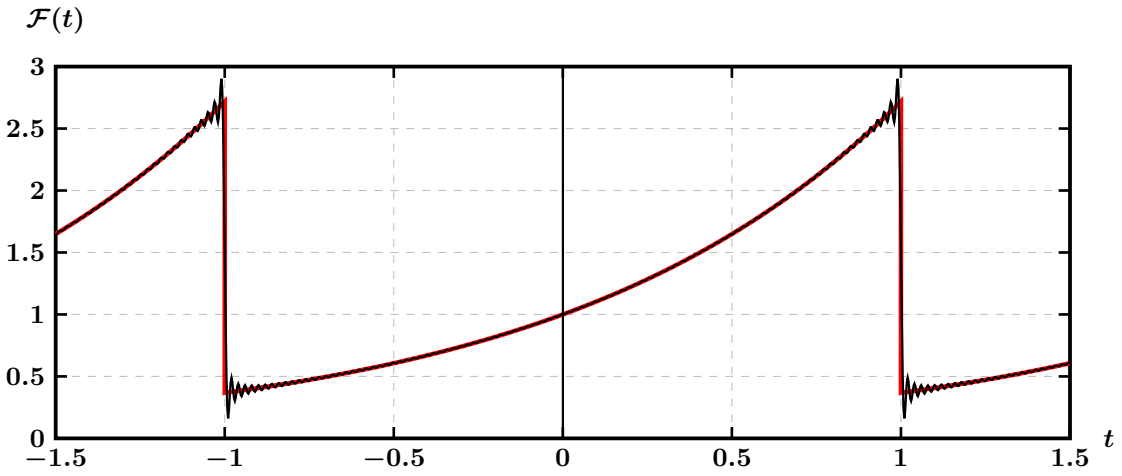


Figure 5.19. The function $\mathcal{F}(t)$ given by equation (5.29) and shown in red. Also displayed is the $\mathbf{100}$-term partial sum of its Fourier Series.

Finally, we note that the numerical value taken by the Fourier Series at the discontinuity is the average of the two possible limits as t tends towards the point of discontinuity. For the present case this means that when $t = 1$, the value of the Fourier Series is equal to,

$$\tfrac{1}{2}\Big[\lim_{t\to 1^-} \mathcal{F}(t) + \lim_{t\to 1^+} \mathcal{F}(t)\Big].$$

In this case the value is $\frac{1}{2}(e + e^{-1})$, i.e. $\cosh 1$.

5.6 Solving ODEs with periodic forcing.

We shall now consider two examples of linear constant-coefficient ODEs which are subject to a forcing term which is periodic in time. The Complementary Function is, of course, independent of this forcing, but the Particular Integral may be split into an infinite sum of Particular Integrals each of which corresponds to one of the forcing terms.

5.6.1 Example 5.5.

Consider the following equation,

$$\frac{d^2y}{dt^2} + K^2 y = \frac{1}{3} + \sum_{n=1}^{\infty} \frac{4(-1)^n \cos n\pi t}{\pi^2 n^2}, \tag{5.31}$$

where the right hand side is the periodic function given in Example 5.2 above and also in Eqs. (5.17) and (5.21). The Complementary Function is clearly,

$$y_{\rm cf} = A\cos Kt + B\sin Kt,$$

where A and B are arbitrary.

As the equation is linear, we can find the Particular Integral for every term on the right hand side of (5.31) and then we may add each of these to get the overall Particular Integral. So the **Particular Integral of the sum is equal to the sum of the Particular Integrals**.

For the constant term we have

$$\frac{d^2y}{dt^2} + K^2 y = \tfrac{1}{3} \qquad \Rightarrow \qquad y_{\rm pi} = \frac{1}{3K^2}. \tag{5.32}$$

For the $n^{\rm th}$ cosine term we have

$$\frac{d^2y}{dt^2} + K^2 y = \frac{4(-1)^n}{\pi^2 n^2}\cos n\pi t \qquad \Rightarrow \qquad y_{\rm pi} = \frac{4(-1)^n}{\pi^2 n^2 (K^2 - n^2\pi^2)}\cos n\pi t. \tag{5.33}$$

Hence the overall Particular Integral is

$$y_{\rm pi} = \frac{1}{3K^2} + \sum_{n=1}^{\infty} \frac{4(-1)^n}{\pi^2 n^2 (K^2 - n^2\pi^2)}\cos n\pi t. \tag{5.34}$$

This function may be plotted out by evaluating a sufficient number of terms in the summation for each desired value of t.

Note: The forcing term in Eq. (5.31) is a Fourier Series which decays as n^{-2}, and therefore its function is one which is continuous but has a discontinuous derivative. The solution, on the other hand, is a Fourier Series which decays as n^{-4}, and therefore the function, y_{pi}, and its first two derivatives are continuous, but its third derivative is discontinuous.

Note also that the solution given by Eq. (5.34) is valid only when $K \neq n\pi$, i.e. K is not an integer multiple of π. So let us see what happens when K happens to be equal to $m\pi$, where m is an integer. Then we can say first is that all the individual terms in (5.34) apply perfectly when $n \neq m$. But when $n = m$ we will need to solve,

$$\frac{d^2y}{dt^2} + m^2\pi^2 y = \frac{4(-1)^m \cos m\pi t}{\pi^2 m^2}, \tag{5.35}$$

where I have used $K = m\pi$ explicitly on the left hand side to show the nature of the difficulty. In terms of 'λ-values' we have a repeated pair, $\pm m\pi j$, $\pm m\pi j$.

After a little bit of work we obtain the following component of the Particular Integral,

$$y_{\mathrm{pi}} = \frac{1}{2m\pi}\left(\frac{4(-1)^m}{\pi^2 m^2}\right) t \sin m\pi t,$$

which corresponds to the $n = m$ term when $K = m\pi$.

Therefore the full Particular Integral is,

$$y_{\mathrm{pi}} = \frac{1}{3K^2} + \frac{1}{2m\pi}\left(\frac{4(-1)^m}{\pi^2 m^2}\right) t \sin m\pi t + \sum_{\substack{n=1 \\ n\neq m}}^{\infty} \frac{4(-1)^n}{\pi^2 n^2 (K^2 - n^2\pi^2)} \cos n\pi t. \tag{5.36}$$

This solution displays resonance because the applied forcing, $f(t)$, contains a Fourier component which is part of the Complementary Function, the physical response to which is a wave which grows linearly with time ($t \sin m\pi t$).

5.6.2 Example 5.6

Please treat the following as *for information only*. There will not be a question on this.

Consider the solution of the equation

$$\frac{d^2y}{dt^2} + c\frac{dy}{dt} + K^2 y = f(t), \tag{5.37}$$

where $f(t)$ is as in Example 5.5 and where $c > 0$. The Particular Integral corresponding to the constant component of $f(t)$ is as in Example 5.5. However, the PI corresponding to the general term in the infinite series involves both cosines and sines. Therefore let us consider the representative problem,

$$\frac{d^2y}{dt^2} + c\frac{dy}{dt} + K^2 y = A \cos \omega t, \tag{5.38}$$

where we may substitute the correct values of A and ω later. We let $y = C \cos \omega t + D \sin \omega t$ be the PI and need to find C and D. Substitution of this into (5.37) gives, eventually,

$$\left[-\omega^2 C + \omega c D + K^2 C\right] \cos \omega t + \left[-\omega^2 D - \omega c C + K^2 D\right] \sin \omega t = A \cos \omega t.$$

Equating coefficients gives

cosine terms: $$(K^2-\omega^2)C+\omega cD=A,$$

sine terms: $$-\omega cC+(K^2-\omega^2)D=0.$$

We eventually obtain,

$$C=A\frac{K^2-\omega^2}{(K^2-\omega^2)^2+(\omega c)^2},\qquad D=A\frac{\omega c}{(K^2-\omega^2)^2+(\omega c)^2}.$$

Hence the PI for the representative system is,

$$y_{\rm pi}=A\frac{(K^2-\omega^2)\cos\omega t+(\omega c)\sin\omega t}{(K^2-\omega^2)^2+(\omega c)^2}.$$

Now, since we are solving the equation,

$$\frac{d^2y}{dt^2}+c\frac{dy}{dt}+K^2y=\tfrac{1}{3}+\sum_{n=1}^{\infty}\frac{4(-1)^n\cos n\pi t}{\pi^2n^2}$$

then the solution is

$$y_{\rm pi}=\frac{1}{3K^2}+\sum_{n=1}^{\infty}\frac{4(-1)^n}{n^2\pi^2}\times\frac{(K^2-n^2\pi^2)\cos n\pi t+(n\pi c)\sin n\pi t}{(K^2-n^2\pi^2)^2+(n\pi c)^2}. \tag{5.39}$$

For these damped systems (i.e. cases where $c>0$) the denominator in (5.39) is never zero, and therefore we will never get an infinite particular integral. Thus there is no resonance in a damped system in the sense that there exist solutions which grow indefinitely with time. That said, if K is precisely equal to an integer multiple of π (say $m\pi$), and if c is very small, then there will be a very large periodic response of the form, $\sin m\pi t$. Specifically, the offending term in Eq. (5.39) is,

$$\frac{4(-1)^m}{m^2\pi^2}\times\frac{\sin m\pi t}{m\pi c},$$

and this term will dominate all of the others.

5.7 Some final comments

- The material contained in this section of the unit draws on quite a large amount of information. Do not despair! Have a look at past exam questions for guidance on what I intend to set as an exam question.
- As a rough guide you will be asked to find a Fourier Series and to solve an ODE. It is always good to sketch the given function even if I don't ask you to do so. It will give you information about whether the function is even or odd or neither. In fact, it is occasionally absolutely indispensible. Here are four examples; try to guess their symmetries quickly and then sketch them carefully to see if you were right.

$$f(t) = \sin \pi t \quad (0 \leq t \leq 1), \qquad \text{with } f(t) = f(t+1)$$

$$f(t) = t - \tfrac{1}{2} \quad (0 \leq t \leq 1), \qquad \text{with } f(t) = f(t+1)$$

$$f(t) = t^2 \quad (0 \leq t < 1), \qquad f(t) = -(t-2)^2 \quad (1 < t \leq 2) \qquad \text{with } f(t) = f(t+2)$$

$$f(t) = t \quad (0 \leq t < 1), \qquad f(t) = 2 - t \quad (1 < t \leq 2) \qquad \text{with } f(t) = f(t+2)$$

- Fourier Series will be used again in the context of solving certain Partial Differential Equations in year 2 semester 2. The unit is Modelling Techniques 2 (ME20021).

6 The Method of Least Squares

6.1 Introduction

Whenever experimental work is performed and measurements are taken as a control parameter is varied, it is well-known that errors of various kinds will ensue; you have experience of this with the Errors Lab last semester. One example might be that the drag force on a car needs to be known as a function of its speed. A wind tunnel is set up, and the force on the car is measured as the wind tunnel speed is increased. In all of these cases many measurements are taken and then the data is plotted with the result that one may well obtain what looks like a cloud of dust with a slight hint of purpose just like the following.

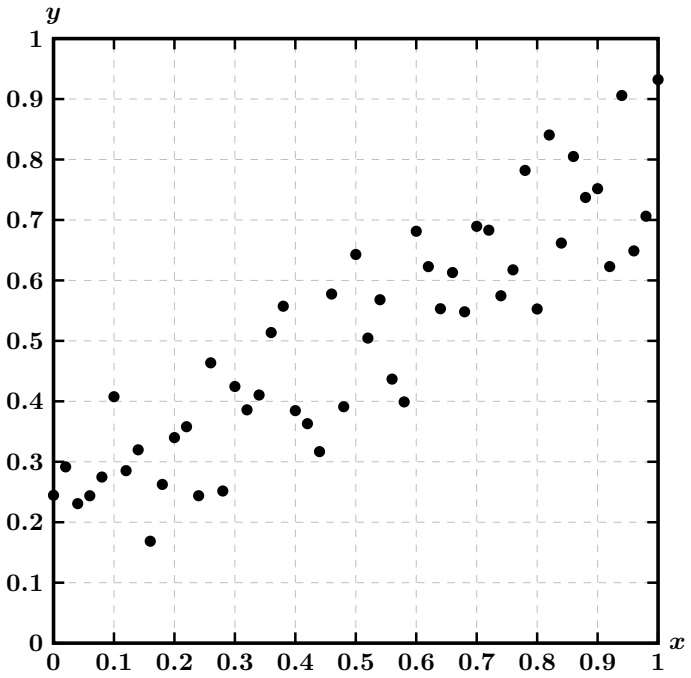


Figure 6.1. Showing measured values of y against the control parameter, x.

Clearly y increases as x increases, but it is a very noisy set of data with large errors.

Clearly too there must be some underlying truth — y varies with x, but how do we fit a line to this data? This is the realm of the theory of least squares.

Or shall we simply guess what that relationship might be? Let's have a go....

How about this?

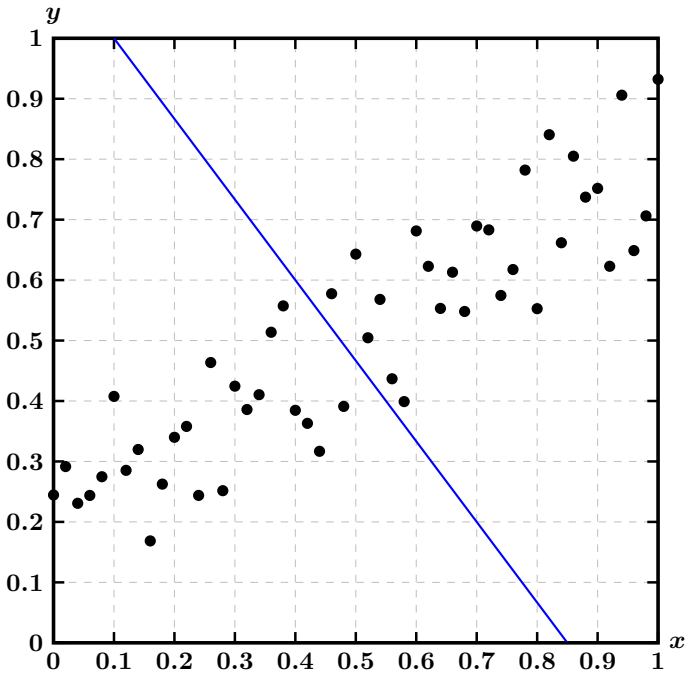


Figure 6.2. Showing a crazy guess for the fitted line.

No, clearly that is stupid! But what about the following guesses?

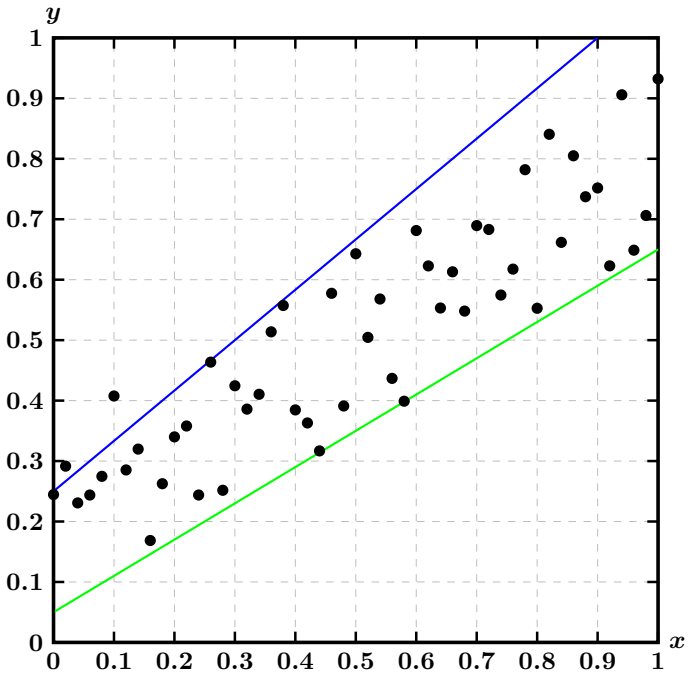


Figure 6.3. Some slightly better guesses for the fitted line.

The blue line is clearly much too high (almost all the data points are below it) while the green line is much too low. Obviously it should be somewhere in the middle.

What about this one, then, which one could have achieved by eye with a ruler and pencil?

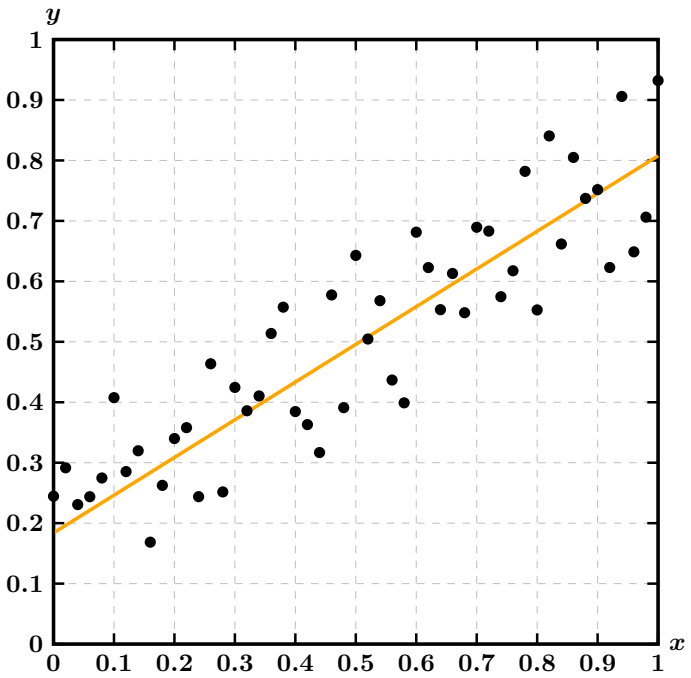


Figure 6.4. Showing a good guess for the fitted line.

Well, this does look plausible, and I doubt if many would look at it and say that it is incorrect. However, the next Figure displays the mathematically correct straight line.

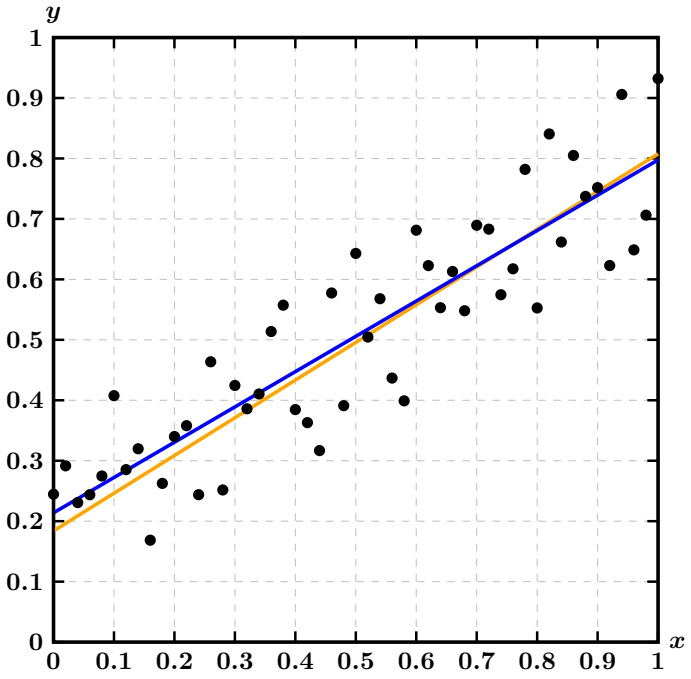


Figure 6.5. Showing the plausible guess and the correct fitted line.

In this Figure our "plausible" line is shown in orange while the blue line is the line of best fit. The plausible line has slightly different values for the slope and intercept.

In reality, I took the line $y = 0.2 + 0.6x$ and added a uniformly distributed random noise to it with a maximum amplitude of 0.25. Therefore the following Figure shows how our least squares fit compares with the original data.

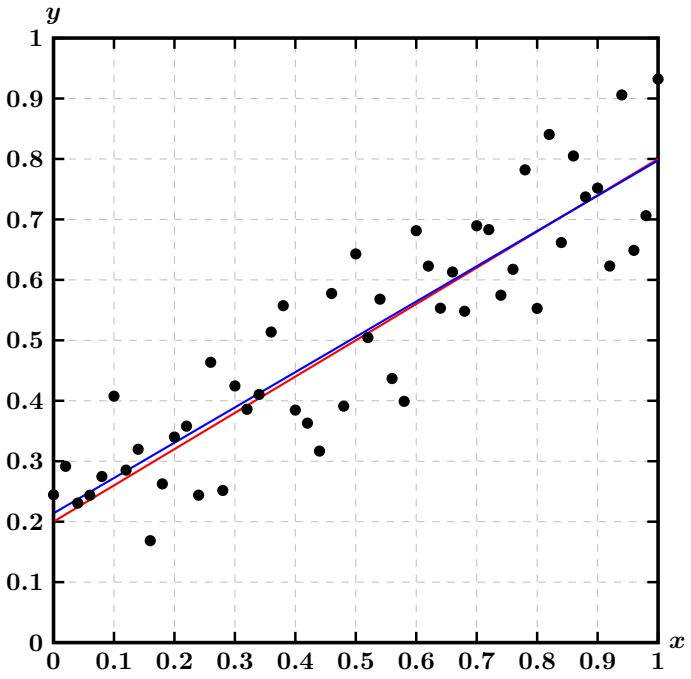


Figure 6.6. Showing the underlying line and the correct fitted line.

Here the red line is the intended behaviour, $y = 0.2 + 0.6x$, while the blue line is the least squares fit for comparison. It is generally the case that one will have more confidence in the fit as the number of data points increases.

6.2 The initial theory

The blue straight line above is called the best fit, but the question is how one may define 'best'. To answer this we begin with what is called the **residual.** It is illustrated in the following Figure.

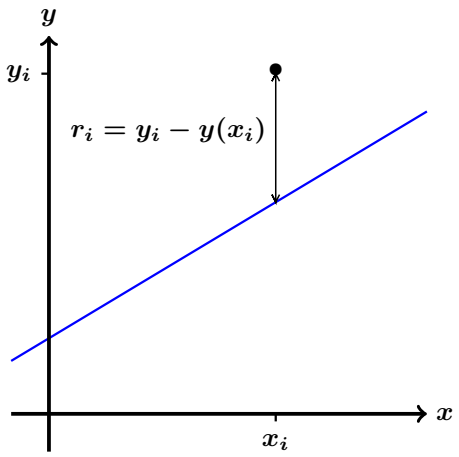


Figure 6.7. Definition of the residual, r_i.

The simplest way of defining the residual verbally is to say that it is the vertical distance between the data point, y_i, and the line as represented by $y(x_i)$. We use the vertical distance because we choose the value of x and then measured the system's response, y. The sign is not of importance here because we will be summing the squares of the residuals, so we could, if we wished, define the residual the other way around, i.e. as $r_i = y(x_i) - y_i$. The subscript, i, denotes a general data point.

We will illustrate the general method by fitting data to a straight line though the origin, i.e. $y = mx$. The residual corresponding to the point, (x_i, y_i), and the line, $y = mx$, is therefore,

$$r_i \quad = \quad y_i - y(x_i) \quad = \quad y_i - mx_i. \tag{6.1}$$

The right hand side here is just a rearrangement of the equation we are fitting, $y - mx = 0$, but evaluated at the given data point. So if we have a perfect set of data and if we have chosen the correct slope, m, then this residual and all of the others will be precisely zero. This doesn't happen in practice and therefore a compromise is needed.

The method of Least Squares is precisely what it says that it is, namely it seeks to minimise the sum of the squares of the residuals. We will define S to be that sum:

$$\begin{aligned} S &= \sum_{i=1}^{N} r_i^2 \\ &= \sum_{i=1}^{N} (y_i - mx_i)^2 \qquad \text{by definition} \\ &= \sum_{i=1}^{N} y_i^2 - 2m\sum_{i=1}^{N} x_i y_i + m^2 \sum_{i=1}^{N} x_i^2. \end{aligned}$$

As the coefficient of m^2 is positive, this means that S must have a minimum value as m varies, as opposed to having a maximum. Therefore it is possible to find this minimum by differentiating S with respect to m and by setting the outcome to zero. Hence,

$$0 \quad = \quad \frac{dS}{dm} \quad = \quad 2m\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i y_i, \tag{6.2}$$

and this yields,

$$m \quad = \quad \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \tag{6.3}$$

When we apply this formula to the data used in Fig. 6.1 we get the following:

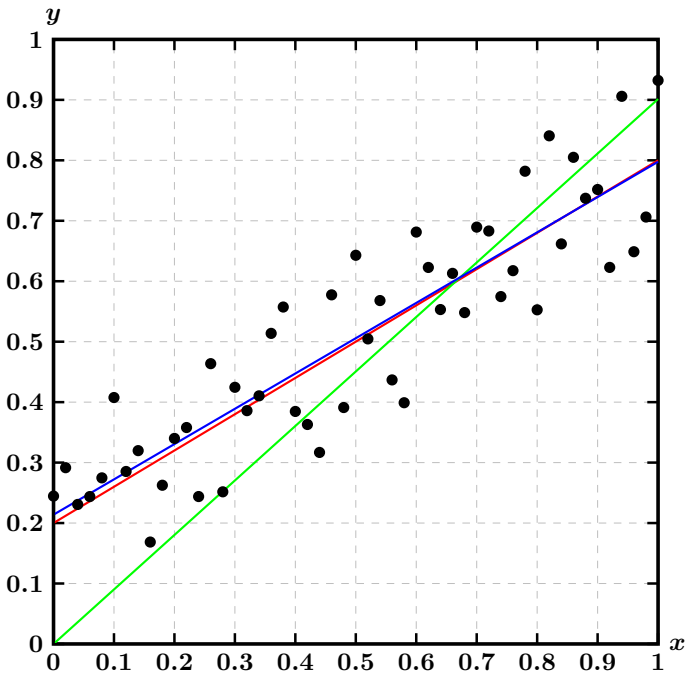


Figure 6.8. Fitting a straight line through the origin. The other lines are as in Fig. 6.6.

Clearly this isn't very good because we have assumed that the straight line passes through the origin, whereas it should pass close to $y = 0.2$. This has the effect of rotating the curve as one can see.

Now we shall relax the restriction of having the line pass through the origin.

6.3 Fitting a general straight line

Now we consider a general line, $y = mx + c$. The residual for any chosen data point has the form,

$$r_i = y_i - y(x_i) = y_i - mx_i - c. \tag{6.4}$$

Therefore the sum of the squares of the residuals is,

$$S = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - mx_i - c)^2. \tag{6.5}$$

I won't expand this because it produces six terms, but it is important to note that one of those six terms is,

$$\sum_{i=1}^{N} c^2 = N\,c^2. \tag{6.6}$$

I've mentioned this because the left hand side is equivalent to adding together N copies of c^2; this is often forgotten in the heat of an exam.

The value S is a function of two variables, m and c. Both m^2 and c^2 in this function have positive coefficients and therefore S will also have a well-defined minimum. The location of this minimum corresponds to when both

$$\frac{\partial S}{\partial m} = 0 \quad \text{and} \quad \frac{\partial S}{\partial c} = 0, \tag{6.7}$$

simultaneously. Using the chain rule we have,

$$\mathbf{0} = \frac{\partial S}{\partial m} = -2m\sum_{i=1}^{N} x_i^2 - 2c\sum_{i=1}^{N} x_i + 2\sum_{i=1}^{N} x_i y_i, \tag{6.8}$$

$$\mathbf{0} = \frac{\partial S}{\partial c} = -2m\sum_{i=1}^{N} x_i - 2cN + 2\sum_{i=1}^{N} y_i. \tag{6.9}$$

These two equations for m and c may be rearranged into the form,

$$m\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i, \tag{6.10}$$

$$m\sum_{i=1}^{N} x_i + cN = \sum_{i=1}^{N} y_i, \tag{6.11}$$

or, more compactly, into the following matrix/vector form:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}. \tag{6.12}$$

This may be solved for m and c quite easily when one has evaluated the summations.

The data which was used to plot Fig. 6.1 was also used to evaluate all the summations in Eq. (6.12), and the resulting least squares line is the **blue** line in Fig. 6.6.

6.4 Fitting a horizontal line.

I can think of no good reason why one would wish to do this, but there is a quick way of fitting a horizontal line, given the analysis of §1.3. A horizontal line corresponds to $m = 0$, and therefore the act of declaring the value of m means that we cannot use the result of setting $\partial S/\partial m = 0$ in that analysis. This means that Eq. (6.8) cannot be used, and therefore we are left with Eq. (6.9) but with $m = 0$. This is

$$-2cN + 2\sum_{i=1}^{N} y_i = 0,$$

or

$$c = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{6.13}$$

which is just the average value of all of the y-values. This feels right to me.

6.5 Fitting a quadratic equation

Now we will fit the quadratic, $y = ax^2 + bx + c$, to suitable data. We have

$$S = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - ax_i^2 - bx_i - c)^2. \tag{6.14}$$

This is to be minimised with respect to $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. Therefore we need to set all three first partial derivatives of $\boldsymbol{S}$ to zero. We obtain,

$$\mathbf{0} = \frac{\boldsymbol{\partial S}}{\boldsymbol{\partial a}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{2}(\boldsymbol{y_i} - \boldsymbol{a x_i^2} - \boldsymbol{b x_i} - \boldsymbol{c})(-\boldsymbol{x_i^2}), \tag{6.15}$$

$$\mathbf{0} = \frac{\boldsymbol{\partial S}}{\boldsymbol{\partial b}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{2}(\boldsymbol{y_i} - \boldsymbol{a x_i^2} - \boldsymbol{b x_i} - \boldsymbol{c})(-\boldsymbol{x_i}), \tag{6.16}$$

$$\mathbf{0} = \frac{\boldsymbol{\partial S}}{\boldsymbol{\partial c}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{2}(\boldsymbol{y_i} - \boldsymbol{a x_i^2} - \boldsymbol{b x_i} - \boldsymbol{c})(-\mathbf{1}), \tag{6.17}$$

where the chain rule has been used. After a little more work we get,

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i x_i^2 \\ \sum_{i=1}^{N} y_i x_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}. \tag{6.18}$$

Having done this analysis, and also if one notes carefully the pattern of the summations in the matrix, it should become clear how one would proceed when fitting cubic equations and so on.

However, it is generally regarded as not being particularly safe to employ polynomials of high order, and especially so if there are not many data points. The least squares matrix becomes increasingly singular as the order of the polynomial increases, and this can lead to wild swings in the computed coefficents of the powers of $\boldsymbol{x}$ when the data exhibit small changes. So the general practical policy is to fit fairly low order polynomials and to maximise the amount of data used in order to have a safe least squares fit.

Now we may mimic the trick we played in §6.4 by using Eq. (6.18) to write down straightaway how to fit the quadratic, $\boldsymbol{y = a x^2 + c}$. The linear term, $\boldsymbol{bx}$, isn't needed since we are setting $\boldsymbol{b = 0}$ instead of applying, $\boldsymbol{\partial S/\partial b = 0}$. This means that the second equation in Eq. (6.18) is no longer valid. Therefore we can write down the first and third rows of the matrix with $\boldsymbol{b = 0}$. This yields the following system:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^2 & N \end{pmatrix} \begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i x_i^2 \\ \sum_{i=1}^{N} y_i \end{pmatrix}. \tag{6.19}$$

In a similar fashion we could have made the alternative assumption that the intercept is zero and then fit the quadratic, $\boldsymbol{y = a x^2 + b x}$ by setting $\boldsymbol{c = 0}$ in Eq. (6.18) and by not using the third row of the matrix (which

came from applying $\partial S/\partial c = 0$):

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i x_i^2 \\ \sum_{i=1}^{N} y_i x_i \end{pmatrix}. \tag{6.20}$$

6.6 A general analysis

So far we have looked at cases where we have tried to fit one, two or three functions all of which are polynomials. There may well be circumstances where other functions might be used. Examples could include exponentials, sinusoids and non-integer powers of x. It is possible and, indeed, quite straightforward to extend our analysis to general functions.

Let us attempt to fit the following functions to a set of data:

$$y = af(x) + bg(x), \tag{6.21}$$

where $f(x)$ and $g(x)$ are unspecified functions. The expression for the residual when $x = x_i$ is,

$$r_i = y_i - y(x_i) = y_i - af_i - bg_i,$$

where I have used $f_i = f(x_i)$ and $g_i = g(x_i)$ as shorthand notation. Then the sum of the squares of the residuals is,

$$S = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - af_i - bg_i)^2, \tag{6.22}$$

where our aim is find those values of a and b which minimise S. On setting both the a-derivative and the b-derivative of S to zero, we find that,

$$0 = \frac{\partial S}{\partial a} = -2a\sum_{i=1}^{N} f_i^2 - 2b\sum_{i=1}^{N} f_i g_i + 2\sum_{i=1}^{N} f_i y_i, \tag{6.23}$$

$$0 = \frac{\partial S}{\partial b} = -2a\sum_{i=1}^{N} f_i g_i - 2b\sum_{i=1}^{N} g_i^2 + 2\sum_{i=1}^{N} g_i y_i. \tag{6.24}$$

In matrix/vector form this becomes,

$$\begin{pmatrix} \sum_{i=1}^{N} f_i^2 & \sum_{i=1}^{N} f_i g_i \\ \sum_{i=1}^{N} f_i g_i & \sum_{i=1}^{N} g_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} f_i y_i \\ \sum_{i=1}^{N} g_i y_i \end{pmatrix}. \tag{6.25}$$

This general example not only shows that the matrix will always be symmetric, but it also suggests immediately how a larger number of general functions might be fitted. So if one wishes to use,

$$y = a\,f(x) + b\,g(x) + c\,h(x,)$$

then a, b and c are given by the solutions of the following equation,

$$\begin{pmatrix} \sum_{i=1}^{N} f_i^2 & \sum_{i=1}^{N} f_i g_i & \sum_{i=1}^{N} f_i h_i \\ \sum_{i=1}^{N} f_i g_i & \sum_{i=1}^{N} g_i^2 & \sum_{i=1}^{N} g_i h_i \\ \sum_{i=1}^{N} f_i h_i & \sum_{i=1}^{N} g_i h_i & \sum_{i=1}^{N} h_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} f_i y_i \\ \sum_{i=1}^{N} g_i y_i \\ \sum_{i=1}^{N} h_i y_i \end{pmatrix}, \tag{6.26}$$

and so on.

Equations (6.25) and (6.26) form a generalisation of all the above matrix/vector systems which we have derived for polynomials. Thus they may be applied to curves such as the following:

$$y = a\cos x + b\sin x, \tag{6.27}$$

$$y = a + bx^{-1}, \tag{6.28}$$

$$y = ae^{bx}, \tag{6.29}$$

$$z = a + bx + cy. \tag{6.30}$$

The detailed derivations of these will not be presented here because similar questions are featured on the problem sheet and the associated solutions. However, the general analysis of this subsection applies directly. I will mention here that Eq. (6.29) cannot be done immediately because the constant, b, is in the exponent. But this equation may be processed by taking logs of both sides to turn it into a linear fit of $\ln y$ in terms of x where the coefficients, $\ln a$ and b are the unknowns. Equation (6.30) represents the fitting of a plane to a set of three-dimensional data; again this may be done using the general analysis of this subsection.

It is worth mentioning, for the sake of completeness, that there are other types of curve-fitting which cannot be done in the way we have seen. One example of this is

$$y = ae^{bx} + ce^{dx}, \tag{6.31}$$

where the four constants, a, b, c and d are sought. This is an example of a nonlinear least squares fit where the resulting equations for the four constants are simultaneous nonlinear equations. This has to be solved using a four-dimensional Newton-Raphson method, an iterative scheme, details of which are well outside of the scope of this unit.

6.7 Practical considerations

In this last subsection we'll consider the role played by visual observation on the goodness of fit. The basic data set that has been used involves the function, $y = x - 0.4x^2$, where uniformly distributed random errors have been superimposed. We have used 101 points.

In Fig. 6.9 the noise has a maximum amplitude of 0.25 while it is 0.05 in Fig. 6.10. The RMS (root mean square) of the residuals has also been computed and stated as a rough guide to the goodness of fit. Changes in this RMS will generally be reflected in how good the fit will appear to us.

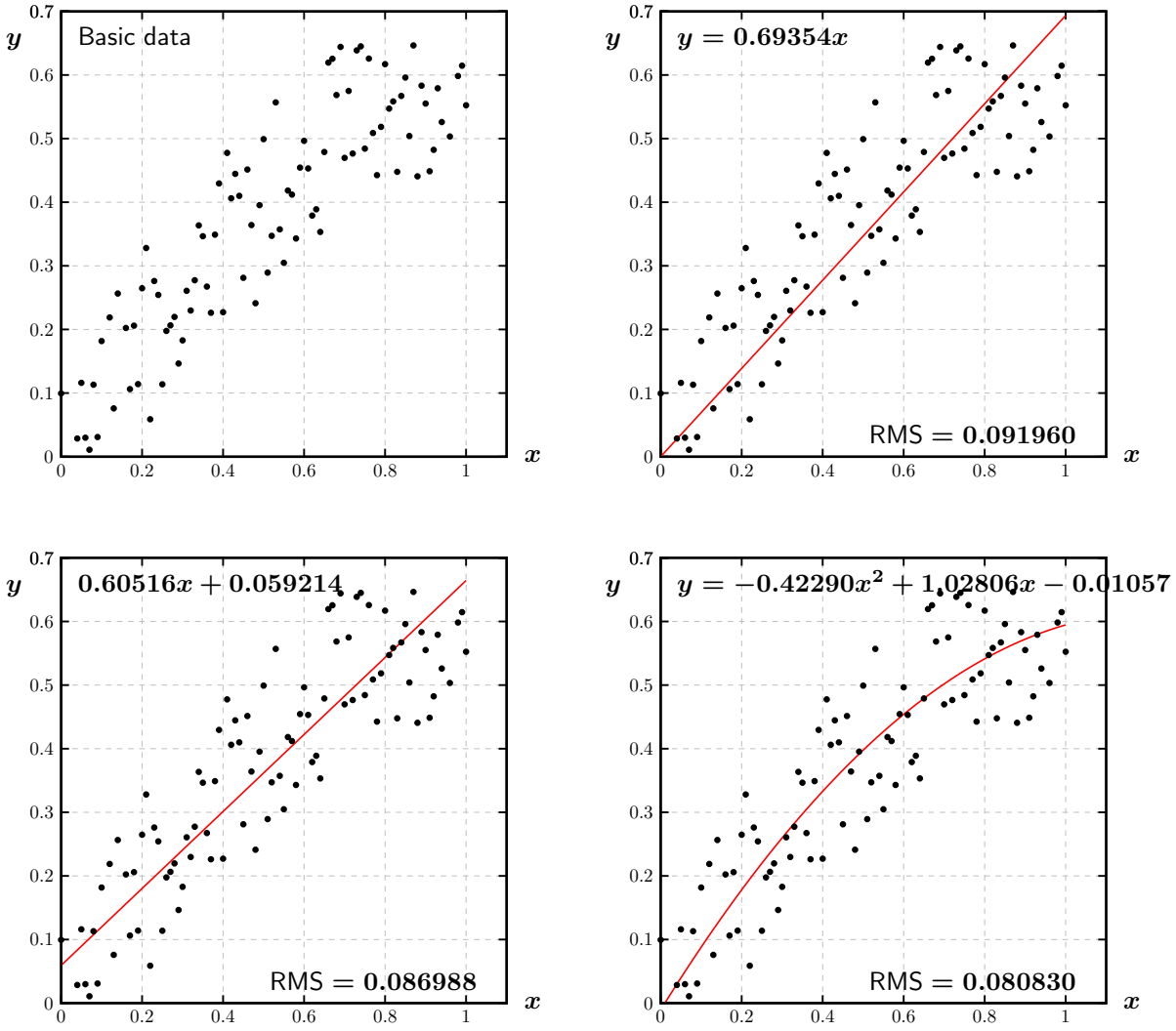


Figure 6.9. Three different least squares fits to the function, $y = x - 0.4x^2$, subject to a uniformly distributed noise with maximum amplitude, 0.25. Using 101 points.

The basic data suggests some large errors. One can see a gradual rise as if it were a linear function, so we will first fit such a line through the origin.

The use of $y = cx$ results in a fit where there is a relatively large number of data points below the line near to $x = 1$, and correspondingly large number above the line near to $x = 0$. This suggests that we should include a constant so that the vertical axis has a nonzero intercept, and this will allow the slope to decrease a little.

Fitting $y = mx + c$ does reduce the RMS value a little. In fact, it is probably best to start with this fit unless one knows for sure (e.g. drag forces on a car) that the correlation must pass through the origin.

In this case the errors are so large that it is not clear if we should try for a quadratic fit, but we have anyway. This reduces the RMS again but only by a small amount. That said, though, the fit certainly looks better but it doesn't seem as though it is worth trying a cubic..

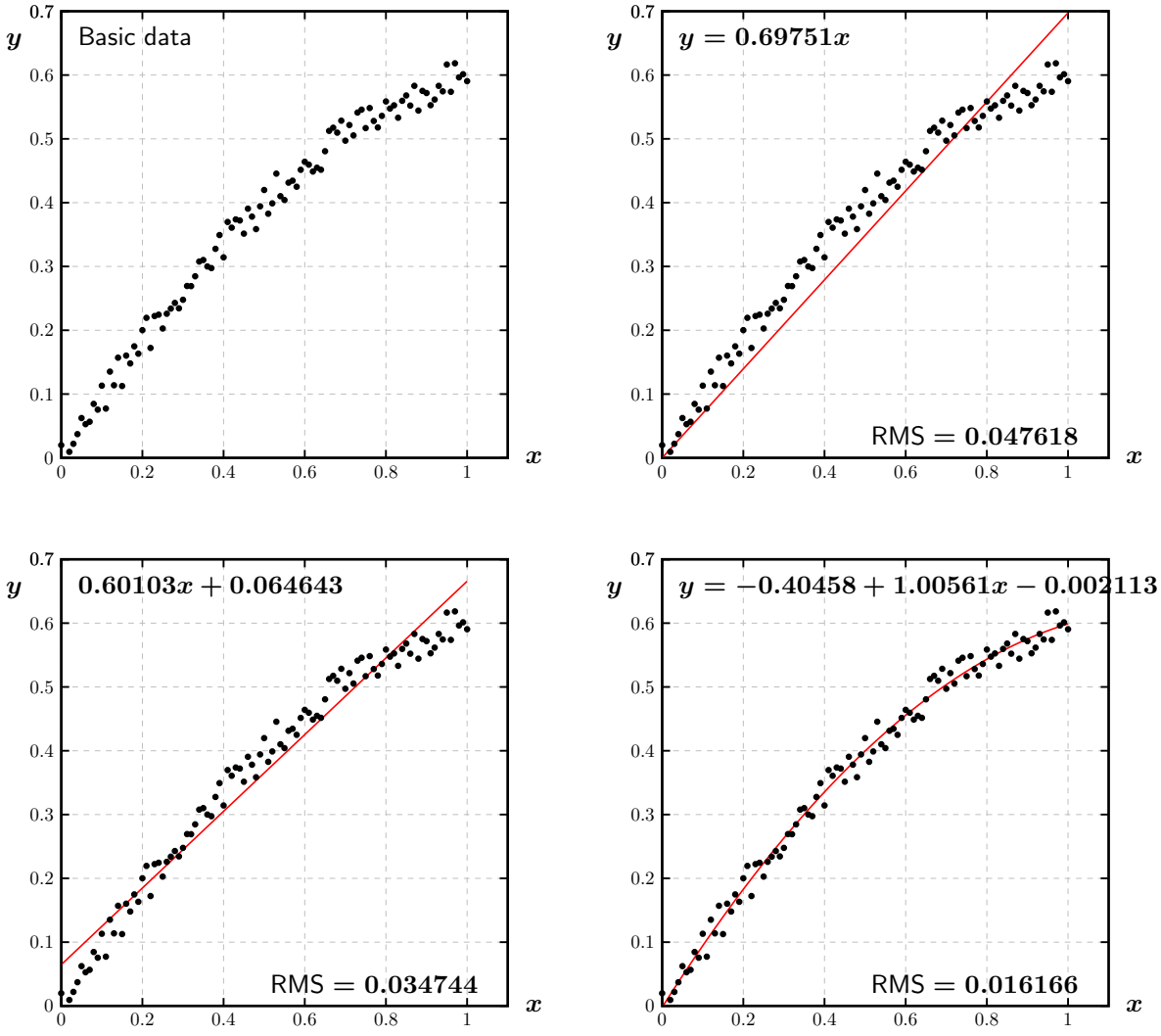


Figure 6.10. Three different least squares fits to the function, $y = x - 0.4x^2$, subject to a uniformly distributed noise with maximum amplitude, 0.05. Using 101 points.

The basic data already looks quadratic because the spread due to the presence of errors is quite small.

On fitting $y = mx$ it is clear that half of the data points are below the line and half above, roughly. So we'll add the constant.

When we fit $y = mx + c$ the RMS has reduced quite well. However, the data points are above the line in the middle of the range and below elsewhere. That qualitative aspect is definitely unsatisfactory and therefore we really do have to add a quadratic term.

The quadratic fit more than halves the RMS value from that of the general linear fit. It is also very evident, satisfyingly good if you like, that the quadratic works very well indeed and has captured the correct shape.

6.8 Final comments

A typical exam question on this topic will involve a derivation of the kind given here. If I happen to ask for the derivation of the formula for the straight line, then I do not expect a quotation of a memorised formula — I have often seen this on exam scripts but I cannot give marks without the asked-for derivation. There will also be some data points to analyse which will typically be five or six points so that there isn't too much numerical calculation.

This is as far as I wish to go with Least Squares theory. I haven't mentioned the more statistical aspects of this topic, such as discussions of the $\boldsymbol{R^2}$ value and its meaning.

Department of Mechanical Engineering, University of Bath

Mathematics 2 ME10305

Problem Sheet 1 — ODEs

1. What is the order of the following equations or systems of equations?

In each case rewrite them in first order form. **Do not** try to solve them!

Are these equations/systems linear or nonlinear, and do they constitute Initial Value Problems or Boundary Value Problems? (Primes denote derivatives with respect to t.)

(a) $y'' + ty = 0$ subject to $y(0) = 0, \quad y'(0) = 1.$

(b) $y''' + y'' - 2yz = 0, \qquad z' = ty$ subject to $y(0) = 1, \quad y'(0) = 0, \quad y'(\infty) = 0, \quad z(0) = 0.$

(c) $y'''' + 2(y + y'')^3 y' + y^5 = 1,$ subject to $y = y' = y'' = y''' - 1 = 0$ at $t = 0.$

(d) $x'' + 2x - y = 0, \qquad y'' - x + 2y - z = 0, \qquad z'' - 3y + 2z = 0,$
subject to $x(0) = 1, \quad x'(0) = 0, \quad y(0) = y'(0) = 0, \quad z(0) = z'(0) = 0.$

(e) $f' = g, \qquad g'' + fg' + f'g = 0,$ subject to $f(0) = 0, \quad g(0) = 1, \quad g(\infty) = 0.$

2. Solve the following equations by direct integration of both sides.

(a) $y' = \cos t$ subject to $y(0) = 1,$ (b) $y' = e^{2t} + 1$ subject to $y(1) = 1.$

3. Use separation of variables to find the solutions to the following ODEs.

(a) $\dfrac{dy}{dt} = \dfrac{4t}{y}, \quad y(0) = 1,$ (b) $\dfrac{dy}{dt} = 3t^2 y, \quad y(0) = 1,$

(c) $\dfrac{dy}{dt} = t(1 + y^2) \quad y(0) = 1,$ (d) $\dfrac{dy}{dt} = t^2(1 - y^2), \quad y(0) = 2,$

(e) $\dfrac{dy}{dt} = y - y^2, \quad y(0) = 2,$ (f) $t^2 \dfrac{dy}{dt} = y - t^3 y, \quad y(1) = 1,$

(g) $\dfrac{d^2y}{dt^2} = \dfrac{1}{t}\dfrac{dy}{dt}, \quad y(0) = 1, \, y'(1) = 2.$

4. Find the Integrating Factor and hence solve the following 1st order equations.

(a) $\dfrac{dy}{dt} + \dfrac{y}{t} = 1,$ (b) $\dfrac{dy}{dt} - \dfrac{y}{t} = 1,$ (c) $\dfrac{dy}{dt} + \dfrac{3y}{t} = t^{-2},$ (d) $\dfrac{dy}{dt} + 2ty = 2t,$

(e) $\dfrac{dy}{dt} + y \cot t = 1,$ (f) $\dfrac{dy}{dt} + \dfrac{1 + 2t}{t} y = \dfrac{1}{t},$ (g) $\dfrac{dy}{dt} + 4t^3 y = t^3$

(h) $t\dfrac{dy}{dt} + (t + 1)y = t^2.$

5. The following differential equation

$$\frac{dy}{dt} = y^3 - y$$

falls into two different categories. First, it is of variables-separable type, and second it is an example of what is known as a Bernoulli equation.

(i) Use separation of variables, followed by partial fractions to find the solution subject to the initial condition that $y = 1/\sqrt{2}$ when $t = 0$.

(ii) Solve the equation for y again by first using the substitution, $y = z^{-1/2}$ where $z = z(t)$ is a new dependent variable — you will need to use the chain rule for this to find a formula for dz/dt in terms of dy/dt. This substitution should then give you a linear equation for z which may be solved.

6. The general form for Bernoulli's equation is

$$\frac{dy}{dt} + P(t)y^n + Q(t)y = 0,$$

where $P(t)$ and $Q(t)$ are given functions.

(i) Use the substitution $y = z^\alpha$, where z is a function of time and where α is constant to be found. After substitution, determine that value of α which reduces the equation for z into one of first order linear form.

(ii) You are now required to solve the ODE,

$$\frac{dy}{dt} + y - y^{-1} = 0, \qquad \text{subject to} \qquad y = 2 \text{ at } t = 0.$$

Use your general Bernoulli result to reduce the ODE to first order linear form and solve it. Check your answer for y by substituting it back into the above equation.

(iii) The above equation may also be solved using separation of variables. Please do it this way as well.

7. Another category of ODE could be called equidimensional. This is an example:

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

The method of solution is to substitute $y(x) = x\,v(x)$ to form an ODE for $v(x)$. The resulting equation should then be solvable using separation of variables. Solve the above equation subject to the initial condition, $y = 1$ when $x = 1$. Then check that your solution satisfies the original ODE. [You may also attempt Q4a using the same idea.]

Please note that all of the above could form at least part of an exam question. I will not ask for the derivation required in Q6i. For equations of Bernoulli type and for those in equidimensional form I will give the required substitutions (as I have in Q5ii and Q7).

Department of Mechanical Engineering, University of Bath

Mathematics ME10305

Problem Sheet 2 — ODE solutions

Questions 1 and 2 contain equations most of which are of exam standard. Questions 3 and 4 are longer, and while they use some ideas (e.g. l'Hôpital's rule) which won't be examined, questions of this type may arise and have arisen in past exams. It is best to be guided by past exam papers in this regard.

1. First find the general solution of the following homogeneous equations. Then find the solution which satisfies $y(0) = 1$ and $y'(0) = 0$ (additionally $y''(0) = 0$ for third and fourth order equations and $y'''(0) = 0$ for fourth order equations).

(a) $\dfrac{d^2y}{dt^2} + 5\dfrac{dy}{dt} + 4y = 0$; (b) $\dfrac{d^2y}{dt^2} + 4\dfrac{dy}{dt} + 4y = 0$; (c) $\dfrac{d^2y}{dt^2} + 2\dfrac{dy}{dt} + 5y = 0$;

(d) $\dfrac{d^2y}{dt^2} - 4\dfrac{dy}{dt} + 29y = 0$; (e) $\dfrac{d^3y}{dt^3} + 2\dfrac{d^2y}{dt^2} + \dfrac{dy}{dt} + 2y = 0$;

(f) $\dfrac{d^3y}{dt^3} + \dfrac{d^2y}{dt^2} - 2y = 0$; (g) $\dfrac{d^3y}{dt^3} - 3\dfrac{d^2y}{dt^2} + 3\dfrac{dy}{dt} - y = 0$; (h) $\dfrac{d^4y}{dt^4} + 4y = 0$;

(i) $\dfrac{d^4y}{dt^4} + 5\dfrac{d^2y}{dt^2} + 4y = 0$; (j) $\dfrac{d^4y}{dt^4} + 2\dfrac{d^2y}{dt^2} + y = 0$.

2. Find the general solution of the following inhomogeneous equations.

(a) $\dfrac{d^2y}{dt^2} + 9y = f(t)$ where $f(t)$ takes the following forms: (i) e^{at}, (ii) t^3, (iii) $\cos at$, (iv) $\cos 3t$.

(b) $\dfrac{d^2y}{dt^2} + 2\dfrac{dy}{dt} + 2y = f(t)$ where $f(t)$ takes the following forms: (i) e^{at}, (ii) t^2, (iii) $\cos at$.

(c) $\dfrac{d^2y}{dt^2} - 7\dfrac{dy}{dt} + 12y = f(t)$ where $f(t)$ takes the following forms: (i) e^{2t}, (ii) e^{3t}, (iii) t^2 (iv) $\cos at$.

(d) $\dfrac{d^3y}{dt^3} + 3\dfrac{d^2y}{dt^2} + 3\dfrac{dy}{dt} + y = t^3e^{-t}$. (Use the standard way first, then use the substitution, $y(t) = z(t)e^{-t}$.)

3. Solve the equation

$$\frac{d^2y}{dt^2} + \frac{dy}{dt} - 6y = e^{2t}, \qquad y(0) = 0, \quad \left.\frac{dy}{dt}\right|_{t=0} = 0,$$

using standard CF/PI methods.

Now we will attempt to solve the same equation using a slightly different method. First solve,

$$\frac{d^2y}{dt^2} + \frac{dy}{dt} - 6y = e^{at}, \qquad y(0) = 0, \quad \left.\frac{dy}{dt}\right|_{t=0} = 0,$$

where $a \neq 2$. Now let $a \to 2$ in the answer, and use L'Hôpital's rule to recover the solution when $a = 2$.

4. In this question the equation,

$$\frac{d^2y}{dt^2} + 5\frac{dy}{dt} + 6y = te^{-2t},$$

will be solved in two different ways.

(a) Use the Complementary Function/Particular Integral approach.

(b) Use the substitution $y = z(t)e^{-2t}$ to simplify the equation. You should then be able to integrate the resulting equation once with respect to t. The final first order equation for z may then be solved using the CF/PI approach.

Department of Mechanical Engineering, University of Bath

Mathematics ME10305

Problem Sheet 2 (Extension) — ODE solutions

This problem sheet contains questions all of which are over and above what is required in the exams. You may therefore treat this sheet as **purely optional**. Nevertheless, everything that is given here may be completed using what has been taught in Maths 1 and Maths 2, with a few hints and nudges along the way.

1. One notation for dy/dt which is sometimes used in textbooks and research papers is Dy. In essence, d/dt and D are directly equivalent to one another and are simply alternative ways of writing down the same thing. Given this, one may try to determine the inverse of D in the following way. Given that

$$\frac{dy}{dt} = f(t) \qquad \Rightarrow \qquad y = c + \int f(t)\,dt,$$

then we may define D^{-1} as follows,

$$Dy = f \qquad \Rightarrow \qquad y = \frac{1}{D} f(t) \qquad = \qquad c + \int f(t)\,dt.$$

In other words, D^{-1} is equivalent to an indefinite integral plus an arbitrary constant.

(a) Now consider the differential equation, $(D+a)y = f(t)$. Rewrite this in the usual way (i.e. $dy/dt + ay = f(t)$) and use the integrating factor approach to find y, not forgetting the arbitrary constant. When this is done, identify which part of your solution forms the Complementary function and which the Particular Integral. What you have written is then the equivalent of

$$y = \frac{1}{D+a} f(t),$$

and it defines the meaning of $(D+a)^{-1}$.

(b) Let us extend the result of Q1a to the following differential equation,

$$\frac{d^2y}{dt^2} + (a+b)\frac{dy}{dt} + aby = f(t).$$

This may also be written as

$$D^2y + (a+b)Dy + aby = f(t), \qquad \text{or} \qquad (D+a)(D+b)y = f(t).$$

If we now set $z = (D+b)y$ then $(D+a)z = f(t)$.

First solve $(D+a)z = f(t)$ for z by applying the result of Q1a directly. Then solve $(D+b)y = z$ to find y. Keep your wits about you on this one — the final answer will involve a double integral.

(c) Now we will modify slightly the answer given in Q1b for the case when $a = b$, which (in the terminology of the lectures) is a repeated-λ case. You should find that some integrals will simplify slightly.

(d) Apply the formula found in Q1b to solve the two equations,

$$y'' + 3y' + 2y = e^t \qquad \text{and} \qquad y'' + 3y' + 2y = e^{-t}.$$

(e) Suppose that we are solving a third order ODE with $f(t)$ on the right hand side. If it is written in the form,

$$(D+a)(D+b)(D+c)y = f(t),$$

and given the form of the answer Q1b, can you guess what the solution is?

2. The aim for this question is to solve $y' + ay = 1$ subject to $y(0) = 0$ using Taylor's series. First, write down a general expression for the Taylor's series about $t = 0$ for the function $y(t)$ — this is *not* the solution because we don't yet know the value of all of the derivatives of y at $t = 0$. However, we may substitute the initial value of y into the governing equation to find $y'(0)$. Now differentiate the governing equation once; this will allow us to find $y''(0)$. Differentate again and hence find $y'''(0)$. The pattern should now be clear. Hence write down the Taylor's series of the solution. Can you identify it?

3. This question was devised while I was watching the film, Gravity, en route to India, with only a thin skin of aluminium between me and a quarter of an atmosphere of air pressure at -50°C while travelling ay 500mph six miles above the ground. I am not sure that I like disaster movies while flying! Suppose that Sandra Bullock and George Clooney are stranded in space, 20m apart and stationary relative to each another, i.e. 10m from their centre of gravity (I am assuming that they have the same mass!). How long will it take for gravitational attraction to cause the couple get close enough together that they may grasp each other's hand? So if $x(t)$ is the distance of one of them from the other, how long will it take to reduce $x = 20$m to $x = 1$m as gravitational attraction draws them together? (Of course, this is being typeset on Valentine's day.) The governing equation is

$$m_1 \frac{d^2x}{dt^2} = -\frac{m_1 m_2 G}{x^2},$$

where $m_1 = m_2 = 60$kg are their masses, and $G = 6.67408 \times 10^{-11}$N m^{2}kg^{-2} is the gravitational constant. This is a nonlinear second order equation!

(a) Nothing in our lectures hints about how to solve this! However, d^2x/dt^2 is the same as dv/dt where $v = dx/dt$. Use the chain rule to show that

$$\frac{dv}{dt} = v\frac{dv}{dx}.$$

Use this substitution to solve for v in terms of x. Apply the initial condition, namely that at $t = 0$, $x = x_0$ and $v = 0$ (we'll keep the initial separation general for now).

(b) Now that we have v in terms of x, it is possible to solve this by first using the substitution, $x = x_0 \cos^2\theta$, to obtain an equation for θ in terms of t. This equation may be solved to find t in terms of θ. Don't let this worry you, for the whole point is that you need to find the time corresponding to a given distance. Now use $x_0 = 20$ and let $x = 1$ (it's probably best to find the corresponding value of θ here); what is the time? So how many days does it take for them to be reunited? (Cue suitable sad violin music...)

4. The Cauchy-Euler equation is a different class of linear ODE, and technically it is known as an equi-dimensional equation. The most general second order version is

$$x^2\frac{d^2y}{dx^2} + ax\frac{dy}{dx} + by = 0.$$

There are two ways of solving this equation, the first being to let $y = x^n$ (and then one will eventually be led to an indicial/auxiliary/characteristic equation for n) while the second is to change variables from x to ξ using $x = e^\xi$. Try to solve the equation

$$x^2\frac{d^2y}{dx^2} + 4x\frac{dy}{dx} + 2y = 0$$

using each of these two methods. [Note, when attempting the second, we are changing from dy/dx to $dy/d\xi$, and the chain rule will need to be used. Take care with the transformation of the second derivative — the product rule will be needed!]

[Continued overleaf]

Suppose now that we wish to solve

$$x^2\frac{d^2y}{dx^2} + 5x\frac{dy}{dx} + 4y = 0.$$

The first method given above leads to a repeated value of n and then it isn't obvious how to proceed in this context. So adopt the second method, solve the equation, and this will show how one should proceed when using the otherwise quicker and simpler first method.

Department of Mechanical Engineering, University of Bath

ME10305 Mathematics 2

Laplace Transforms Sheet 0

This problem sheet is an experiment to see if you can solve an ordinary differential equation using Laplace Transforms before I even begin lecturing on the topic. It's your choice if you wish to rise to the challenge. I'll try to walk you through it gently.

Here's the definition. If we have a function of time, $\boldsymbol{f(t)}$, then its Laplace Transform is defined to be $\boldsymbol{F(s)}$ where

$$\boldsymbol{F(s) = \mathcal{L}\big[f(t)\big] = \int_0^\infty f(t)e^{-st}\,dt,}$$

where $\boldsymbol{\mathcal{L}\big[\ \ \big]}$ is just a mathematical shorthand for saying the words, 'Laplace Transform of...'. Don't worry about what all of this means — it's really weird — just go with the flow for now.

Q1. Use the definition of the Laplace Transform to find $\boldsymbol{\mathcal{L}\big[e^{-at}\big]}$. This should be a nice quick integral, and your answer should come out to be $\boldsymbol{1/(s+a)}$.

Q2. As the aim is to solve a differential equation we had better find an expression for $\boldsymbol{\mathcal{L}\big[dy/dt\big]}$.

So let $\boldsymbol{Y(s) = \mathcal{L}\big[y(t)\big]}$, and apply the Laplace Transform formula to $\boldsymbol{dy/dt}$.

You will need one integration by parts to do this, and the final answer will involve both $\boldsymbol{Y(s)}$ and the value of $\boldsymbol{y}$ at $\boldsymbol{t=0}$, i.e. $\boldsymbol{y(0)}$. In the context of ODEs the value of $\boldsymbol{y(0)}$ represents the initial condition.

Q3. Believe it or not, we are now in a position to solve an ODE. So use both of the above results to find the Laplace Transform of the ordinary differential equation,

$$\boldsymbol{\frac{dy}{dt} + 2y = e^{-t},} \qquad \text{subject to } \boldsymbol{y(0) = 2}.$$

Once the equation (with the boundary condition) has been transformed, first rearrange it to find $\boldsymbol{Y(s)}$ explicitly, then use partial fractions to simplify that expression, and then finally use the result in Q1 to find the function, $\boldsymbol{y(t)}$, which corresponds to your $\boldsymbol{Y(s)}$.

Q4. Perhaps you should solve the equation using the CF/PI method (i) to ensure that you can after such a long break(!) and (ii) to check that the Laplace Transform solution is correct.

Department of Mechanical Engineering, University of Bath

ME10305 Mathematics 2

Laplace Transforms Sheet 1

It is normal in questions on Laplace Transforms to have ready access to the LTs of functions like sinusoids, exponentials and powers. Here, though, I will need you to derive these results.

1. Find the Laplace Transforms of the following functions using the definition of the Laplace Transform (rather than by looking up the result in a table):

(a) e^{3t} (b) e^{-3t} (c) $\cos \omega t$ (d) te^{-3t} (e) t^3 (f) $t \cos \omega t$ (g) $f'''(t)$

(h) The unit pulse: $f(t) = 1$ for $t < 1$, $f(t) = 0$ otherwise (i) $\cosh \omega t$ (j) $t^2 e^{-t}$

(k) $t^{-1/2}$ [Hint: set $x = (st)^{1/2}$ to transform the integral and use the result $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$.]

2. Use the Laplace Transform to solve the following equations:

$$\text{(a)} \quad \frac{dy}{dt} + 4y = 6, \qquad y(0) = 2.$$

$$\text{(b)} \quad \frac{d^2y}{dt^2} + 16y = 0, \qquad y(0) = 0, \quad \frac{dy}{dt}(0) = 1.$$

$$\text{(c)} \quad \frac{d^2y}{dt^2} + 4y = 29e^{-5t}, \qquad y(0) = 0, \quad \frac{dy}{dt}(0) = -3.$$

$$\text{(d)} \quad y''' + y'' + 4y' + 4y = 0, \qquad y(0) = 0, \quad y'(0) = 3, \quad y''(0) = -5.$$

[You may also practice on any of the linear constant coefficient equations from the ODEs section of the unit, but note that there may be some awkwardnesses due to the fact that (i) the questions weren't designed for nice LT solutions, (ii) many don't have initial conditions specified, (iii) some of the results derived in the 3rd and 4th Laplace Transform lectures may be of considerable use.]

3. Find the Laplace Transform of $z(t) = \int_0^t y(\tau)\, d\tau$. [Hint: recall that $z'(t) = y(t)$ here.]

4. Find the solution of the ODE, $y'' + 2y' + y = 2e^{-t}$, subject to $y(0) = y'(0) = 0$. [Hint: you may need to consult the solution to Q1j.]

5. Factorise the denominator of the following fractions into complex factors, and use partial fractions to find their Inverse Laplace Transforms: [**Note:** that I won't expect such complex factorisation in the exam.]

(a) $\frac{1}{s^2+b^2}$ (b) $\frac{s}{s^2+b^2}$ (c) $\dfrac{1}{s^2 + 2cs + c^2 + d^2}$ (d) $\dfrac{s+c}{s^2 + 2cs + c^2 + d^2}$.

These results may be used to solve the following equations:

(e) $y'' + 4y' + 5y = 0, \quad y(0) = 0, \quad y'(0) = 1.$

(f) $y'' + 2y' + 2y = e^{-t}, \quad y(0) = 0, \quad y'(0) = 0.$

6. Write down the values of the following integrals.

$$\int_{-\infty}^{\infty} \delta(t)e^{2t}\,dt, \qquad \int_{-\infty}^{\infty} \delta(t-1)e^{-t^2}\,dt, \qquad \int_{-\infty}^{\infty} \delta(t-2)\sin \pi t\,dt, \qquad \int_{0}^{\infty} \delta(t+2)t^3\,dt.$$

7. Find the Laplace Transforms of the following functions:

(a) $e^{e^t}\delta(t-1)$, (b) $\displaystyle\sum_{n=0}^{\infty} \delta(t-n) = \delta(t) + \delta(t-1) + \delta(t-2) + \delta(t-3) + \ldots.$

[Look out for the geometric series....]

8. Use the Laplace Transform to solve the following equations:

(a) $\dfrac{dy}{dt} + 3y = \delta(t), \qquad y(0) = 1.$

(b) $\dfrac{d^2y}{dt^2} + 3\dfrac{dy}{dt} + 2y = \delta(t), \qquad y(0) = 1, \quad \dfrac{dy}{dt}(0) = b,$ where b is a constant.

(c) $\dfrac{d^3y}{dt^3} - \dfrac{dy}{dt} = 3\delta(t), \qquad y(0) = 1, \quad \dfrac{dy}{dt}(0) = 0, \quad y''(0) = -1.$

9. Laplace Transforms are perfectly set up to solve Initial Value Problems, but let us try them out on a Boundary Value Problem. The aim, then, is to solve $y'' + y = 0$, subject to $y(0) = 1$ and $y(\frac{1}{2}\pi) = 1$. At the outset, let $y'(0) = c$ and carry out the analysis using this unknown constant. Eventually you will have the opportunity to find c.

Department of Mechanical Engineering, University of Bath

ME10305 Mathematics 2

Laplace Transforms Sheet 2

10. First sketch the following functions, and then Find their Laplace Transforms:

(a) $H(t-a)t^3$ (b) $\sum_{n=0}^{\infty}(-1)^n H(t-n) = H(t) - H(t-1) + H(t-2) - H(t-3) + \ldots$

(c) $H(a-t)$,

[In one case it may be possible to simplify the final answer...]

11. Use the s-shift theorem to find the Inverse Laplace Transform of:

(a) $\frac{1}{s+a}$ (b) $\frac{2}{(s+a)^3}$ (c) $\frac{b}{(s+a)^2+b^2}$ (d) $\frac{s}{(s+a)^2+b^2}$.

12. [This is an exam-style question.] Find the Laplace Transforms of both $\cos bt$ and $\sin bt$. Then use the s-shift theorem to write down the Laplace Transforms of $e^{-at}\cos bt$ and $e^{-at}\sin bt$. Hence solve the ODE,

$$y'' + 6y' + 25y = 0$$

subject to $y(0) = 1$ and $y'(0) = 5$.

13. Use the t-Shift Theorem to find the Inverse Laplace Transform of:

(a) $\frac{e^{-as}}{s^3}$ (b) $\frac{e^{-as}}{s+b}$ (c) $\frac{e^{-as}}{s^2+b^2}$ (d) $\frac{e^{-as}}{(s+c)^2+b^2}$.

14. Use the convolution theorem to find the Inverse Laplace Transform of:

(a) $\frac{1}{(s+a)^2}$ (b) $\frac{1}{(s+a)(s^2+b^2)}$ (c) $\frac{1}{(s^2+a^2)^2}$ (d) $e^{-as} \times \frac{1}{s^2}$.

15. Use the convolution theorem to find the solutions to the following equations:

(a) $y' + 3y = e^{-2t}, \quad y(0) = 0$;

(b) $y'' + 5y' + 6y = 0, \quad y(0) = 0, \quad y'(0) = 1$;

(c) $y'' + y = e^{-t}, \quad y(0) = 0, \quad y'(0) = 0$.

16. Solve the system of equations,

$$x' = 2x - 4y + \delta(t),$$
$$y' = 3x - 5y,$$

subject to the initial conditions, $x(0) = y(0) = 0$.

17. Solve the system of equations,

$$x'' + 2x - 2y = \delta(t),$$
$$y'' - x + 3y = 0,$$

subject to the initial conditions, $x(0) = x'(0) = y(0) = y'(0) = 0$. When the final solution has been obtained, determine which of the four initial conditions has been violated, but can you guess this in advance?

18. [This is a project-like question which combines quite a large number of mathematical results.]

The overall aim for this question is to solve the ODE,

$$y' + y = \text{Ш}(t) = \sum_{n=-\infty}^{\infty} \delta(t-n).$$

The unusual symbol, Ш, which I cannot typeset properly(!), is known as the Shah function, and the symbol itself is the Cyrillic character, sha, which mimics the shape of the function. In various contexts it is also known as (i) the Dirac comb, (ii) more picturesquely as the bed of nails function, and (iii) more functionally as an impulse train.

The solution is a periodic function which has a period of 1, but this can't be found simply using the Laplace Transform because that is an integral from $t = 0$ onwards, whereas the Shah function consists of impulses at all integer values of t, both positive and negative. Therefore we will do this by determining the eventual 'steady periodic state' that is achieved when t becomes large. Enjoy the ride!

(a) Find the Laplace Transform of e^{-t}, and then apply the t-shift theorem to find the inverse Laplace Transform of $e^{-ns}/(s+1)$.

(b) Use the Laplace Transform on the ODE,

$$y' + y = \sum_{n=0}^{\infty} \delta(t-n), \qquad y(0) = 0,$$

to find an expression for $Y(s)$, the transform of $y(t)$. Do not simplify this expression for Y by, say, summing the geometric series!

(c) Now use the result of part (a) to write down an expression for $y(t)$ in terms of a sum of terms involving unit step functions.

(d) The sum we have obtained for $y(t)$ is infinite in length; do make sure that you're happy with this idea! Now we let $t = n + \epsilon$ in your expression for y, where n is the first positive integer below t, and where $0 \le \epsilon < 1$. You should then be able to factor $e^{-\epsilon}$ out of the resulting mess(!), and then be able to sum the resulting geometric series to obtain a compact formula for y.

(e) Now find $\lim_{n\to\infty} y$. This will give the long-term formula for $y(t)$, but written in terms of ϵ — this will be a valid formula for $y(t)$ between two neighbouring integer values of t. In the ultimate steady periodic state, what are the maximum and minimum values of y?

(f) See if you can guess what $y(t)$ looks like.

(g) An easier way to solve the main problem is to concentrate on the interval of time, $0 \le t < 1$, and to solve for

$$y' + y = \delta(t), \qquad y(0) = c,$$

using Laplace Transforms. The value of c may be found by insisting that $y(1) = y(0)$.

University of Bath, Department of Mechanical Engineering

ME10305 Mathematics 2.

Matrices Sheet 1 — Multiplication.

The aim of this problem sheet is to get used to performing matrix multiplication and to know what *compatibility with respect to multiplication* means. There are also some properties which involve matrix transposes which are useful to know, and also the concepts of commutivity and distributivity.

1. The matrices, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$, are defined as follows,

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{-1} & \mathbf{1} \\ \mathbf{3} & \mathbf{5} \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{-1} \\ \mathbf{2} & \mathbf{-1} & \mathbf{2} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \mathbf{2} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{2} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \mathbf{2} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{2} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{2} \end{pmatrix}.$$

Classify all these matrices in terms of numbers of rows and columns. Now make a list of which pairs may be multiplied together (i.e. are *compatible* with respect to multiplication) — for example, both $\boldsymbol{AB}$ and $\boldsymbol{BA}$ belong to this list. Now find all the permissible products of two matrices.

2. Having now determined $\boldsymbol{AB}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are as given in Q1, write down $(\boldsymbol{AB})^{\boldsymbol{T}}$, the transpose of $\boldsymbol{AB}$. Now calculate $\boldsymbol{B^T A^T}$. Is $(\boldsymbol{AB})^{\boldsymbol{T}} = \boldsymbol{B^T A^T}$? Is it obvious whether this last answer is true in general?

3. The matrix $\boldsymbol{A}$ is defined by

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{2} & \mathbf{1} & \mathbf{2} \\ \mathbf{0} & \mathbf{3} & \mathbf{-3} \\ \mathbf{1} & \mathbf{2} & \mathbf{-1} \end{pmatrix}.$$

Find $\boldsymbol{A^T}$. Now form the sums $\boldsymbol{A + A^T}$ and $\boldsymbol{A - A^T}$. What do you notice about these new matrices? Find the products $\boldsymbol{AA^T}$ and $\boldsymbol{A^T A}$. What do you conclude from these results?

4. We have seen that matrix multiplication, where the matrices are compatible, yields $\boldsymbol{AB \neq BA}$ in general, i.e. matrix multiplication is non-commutative. But I would like you to show that matrix multiplication is associative, that is, $\boldsymbol{A(BC) = (AB)C}$, where the term in brackets is computed first. Check one specific case:

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{2} & \mathbf{1} \\ \mathbf{1} & \mathbf{2} \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} \mathbf{1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} \end{pmatrix} \qquad \boldsymbol{C} = \begin{pmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{3} & \mathbf{4} \end{pmatrix}.$$

Now check the general case for $\mathbf{2} \times \mathbf{2}$ matrices:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a_{11}} & \boldsymbol{a_{12}} \\ \boldsymbol{a_{21}} & \boldsymbol{a_{22}} \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{b_{11}} & \boldsymbol{b_{12}} \\ \boldsymbol{b_{21}} & \boldsymbol{b_{22}} \end{pmatrix} \qquad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{c_{11}} & \boldsymbol{c_{12}} \\ \boldsymbol{c_{21}} & \boldsymbol{c_{22}} \end{pmatrix}.$$

Try to think of a way of generalising this result to any three square matrices, and then to any set of compatible matrices.

5. Not really part of the syllabus, but I needed something to fill the gap at the bottom of this page! If you have two tridiagonal matrices which are compatible with respect to multiplication and are subsequently multiplied together, then is there is a general statement that can be made about the pattern of the components in that product?

6. Rotation matrices are important for many applications and are especially so in robotics. I am not going to teach this formally, but I would like to introduce them and to play around with them a little.

We may define the following three rotation matrices:

$$\boldsymbol{R_x(\alpha)} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\cos\alpha} & \mathbf{-\sin\alpha} \\ \mathbf{0} & \mathbf{\sin\alpha} & \mathbf{\cos\alpha} \end{pmatrix} \qquad \text{(Rotation by an angle } \boldsymbol{\alpha} \text{ about the } \boldsymbol{x}\text{-axis.)}$$

$$\boldsymbol{R_y(\beta)} = \begin{pmatrix} \mathbf{\cos\beta} & \mathbf{0} & \mathbf{\sin\beta} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{-\sin\beta} & \mathbf{0} & \mathbf{\cos\beta} \end{pmatrix} \qquad \text{(Rotation by an angle } \boldsymbol{\beta} \text{ about the } \boldsymbol{y}\text{-axis.)}$$

$$\boldsymbol{R_z(\gamma)} = \begin{pmatrix} \mathbf{\cos\gamma} & \mathbf{-\sin\gamma} & \mathbf{0} \\ \mathbf{\sin\gamma} & \mathbf{\cos\gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix} \qquad \text{(Rotation by an angle } \boldsymbol{\gamma} \text{ about the } \boldsymbol{z}\text{-axis.)}$$

Therefore if the position vector of a point is $\underline{\boldsymbol{r}}$, and if that point is rotated by an angle, $\boldsymbol{\alpha}$, about the $\boldsymbol{x}$-axis, then the new location of the point is given by the matrix/vector product, $\boldsymbol{R_x(\alpha)}\,\underline{\boldsymbol{r}}$. If this new point is subsequently rotated by $\boldsymbol{\gamma}$ about the $\boldsymbol{z}$-axis, then its new location is given by, $\boldsymbol{R_z(\gamma)\,R_x(\alpha)}\,\underline{\boldsymbol{r}}$.

Thus a rotation about the $\boldsymbol{x}$-axis followed by a rotation about the $\boldsymbol{z}$-axis is $\boldsymbol{R_z(\gamma)\,R_x(\alpha)}$, where the rotation matrices only appear to have been written down in the wrong order!

(i) Perhaps it is not surprising that the inverse matrix of $\boldsymbol{R_x(\alpha)}$ is $\boldsymbol{R_x(-\alpha)}$, given what this notation means. Check that $\boldsymbol{R_x(\alpha)\,R_x(-\alpha) = I}$, the $\mathbf{3\times 3}$ identity matrix.

(ii) Find both $\boldsymbol{R_z(\gamma)\,R_x(\alpha)}$ and $\boldsymbol{R_x(\alpha)\,R_z(\gamma)}$. Are they equal? What is the implication of this general result? What about when $\boldsymbol{\alpha = \gamma = \frac{1}{4}\pi}$? What about when $\boldsymbol{\alpha = \gamma = \frac{1}{2}\pi}$?

(iii) If you really have time spare, then you could try the following. A point suffers the grave indignity of the following sequence of rotations: $\boldsymbol{R_x(\alpha)}$ then $\boldsymbol{R_z(\gamma)}$ then $\boldsymbol{R_x(-\alpha)}$ then $\boldsymbol{R_z(-\gamma)}$. This expresses a possibly naive thought that an arbitrarily chosen point will return to where it started after this sequence; do you think that it will? If not, what are the correct third and fourth rotations to cause the point to return?

7. A question for interest, perhaps. Fermat's last theorem is well-known: when $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{n}$ take positive integer values, the equation $\boldsymbol{a^n + b^n = c^n}$ has solutions only when $\boldsymbol{n = 2}$. However, a Michael Penn youtube video alerted me to the fact that this theorem doesn't apply when $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are matrices! So here's a straightforward exercise in matrix multiplication to check if Prof. Penn is correct:

$$\begin{pmatrix} \mathbf{1} & \mathbf{3} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}^{\mathbf{3}} + \begin{pmatrix} \mathbf{-1} & \mathbf{0} \\ \mathbf{1} & \mathbf{-1} \end{pmatrix}^{\mathbf{3}} = \begin{pmatrix} \mathbf{0} & \mathbf{3} \\ \mathbf{1} & \mathbf{0} \end{pmatrix}^{\mathbf{3}}.$$

University of Bath, Department of Mechanical Engineering

ME10305 Mathematics 2.

Matrices Sheet 2 — Determinants, Cramer's rule and Gaussian Elimination.

1. Find the determinant of the following matrices. Which matrices are singular (i.e. have a zero determinant)? For (d) and (g) attempt the evaluation of the determinant in more than one way just to practice the skill.

(a) $\begin{pmatrix} 6 & 2 \\ 8 & 3 \end{pmatrix}$ (b) $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (c) $\begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$ (d) $\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & -3 \\ 1 & 2 & -1 \end{pmatrix}$

(e) $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ (f) $\begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$ (g) $\begin{pmatrix} 2 & 1 & 1 & 1 \\ 0 & 3 & -3 & 1 \\ 1 & 2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$ (h) $\begin{pmatrix} b & a & a & a \\ a & b & a & a \\ a & a & b & a \\ a & a & a & b \end{pmatrix}$

2. The matrix $\boldsymbol{J_n}$ is an $\boldsymbol{n \times n}$ matrix where the diagonal entries have the value $\boldsymbol{-2}$, the superdiagonal and subdiagonal entries the value $\boldsymbol{1}$, and $\boldsymbol{0}$ elsewhere. For example, $\boldsymbol{J_1}$, $\boldsymbol{J_2}$ and $\boldsymbol{J_5}$ are

$$\boldsymbol{J_1 = (-2)} \qquad \boldsymbol{J_2} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \qquad \boldsymbol{J_5} = \begin{pmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}.$$

Such matrices arise in the numerical solution of second order ordinary differential equations.

Assume that $|\boldsymbol{J_1}| = \boldsymbol{-2}$, and then evaluate $|\boldsymbol{J_2}|$, $|\boldsymbol{J_3}|$, $|\boldsymbol{J_4}|$ and $|\boldsymbol{J_5}|$ directly from the matrix definitions. This should show you how to derive the recurrence relation,

$$\boldsymbol{|J_n| = -2|J_{n-1}| - |J_{n-2}|.}$$

Finally, what is the explicit value of $|\boldsymbol{J_n}|$?

3. Use Cramer's rule to solve the following systems of equations.

(a) $\begin{aligned} 2x + 5y &= -1 \\ -3x + 2y &= 2 \end{aligned}$ (b) $\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 1 \\ 6x_1 - 2x_2 - x_3 &= 2 \\ x_1 - x_2 + x_3 &= 2 \end{aligned}$ (c) $\begin{aligned} x_1 + 3x_2 - x_3 &= 3 \\ x_2 - 7x_3 &= 2 \\ 2x_1 - 5x_3 &= 1 \end{aligned}$

4. Use Gaussian Elimination to solve the following systems of equations.

(a) $\begin{aligned} 2x + 5y &= -1 \\ -3x + 2y &= 2 \end{aligned}$ (b) $\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 1 \\ 6x_1 - 2x_2 - x_3 &= 2 \\ x_1 - x_2 + x_3 &= 2 \end{aligned}$ (c) $\begin{aligned} x_1 + 3x_2 - x_3 &= 3 \\ x_2 - 7x_3 &= 2 \\ 2x_1 - 5x_3 &= 1 \end{aligned}$

Note that the above three systems of equations are identical to those in which were solved in Q3 using Cramer's rule.

(d) $\begin{pmatrix} 1 & 2 & -1 & 1 \\ 1 & 1 & -2 & 6 \\ 3 & 0 & 1 & 1 \\ -2 & 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ 2 \\ -1 \end{pmatrix}$ (e) $\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -2 \\ 7 \end{pmatrix}$

5. The matrix given in Q1e is singular, by which is meant that it has a zero determinant. Therefore a matrix/vector equation involving that matrix either has no solution or else it has an infinite number of them. The aim of this question is to see how Gaussian Elimination behaves in such a situation.

Try to solve the matrix/vector equation

$$\begin{pmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \mathbf{4} & \mathbf{5} & \mathbf{6} \\ \mathbf{7} & \mathbf{8} & \mathbf{9} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix}$$

using Gaussian Elimination to see the manner in which the procedure fails when the matrix is singular. However, it is possible to write down solutions for this case. I haven't covered this in the lectures and therefore you'll need to work out how to do it.

Now try to find the solution when the right hand side vector is $(\mathbf{-2}, \mathbf{1}, \mathbf{5})^{\boldsymbol{T}}$. Can you explain why two separate equations involving the same matrix has solutions in one case but not in another? (Hint: use $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})^{\boldsymbol{T}}$ as the right hand side as a third case.)

6. The aim here is to find the inverse of some matrices using Gaussian Elimination starting with the identity matrix as part of the augmented matrix scheme. Other general properties of inverses will arise along the way. Treat this question as practice in Gaussian elimination; the computation of inverses takes too long in the examination context (with the possible exception of a tridiagonal $\mathbf{3} \times \mathbf{3}$ matrix). Find the inverses of the following matrices.

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{-2} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} \mathbf{1} & \mathbf{3} & \mathbf{-1} \\ \mathbf{0} & \mathbf{4} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \mathbf{2} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{2} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{2} \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} \mathbf{2} & \mathbf{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{-1} & \mathbf{2} & \mathbf{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{-1} & \mathbf{2} & \mathbf{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{2} \end{pmatrix}, \quad \boldsymbol{E} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{2} \\ \mathbf{-1} & \mathbf{0} & \mathbf{3} \\ \mathbf{-2} & \mathbf{-3} & \mathbf{0} \end{pmatrix}, \quad \boldsymbol{F} = \begin{pmatrix} \mathbf{1/2} & \mathbf{1/3} & \mathbf{1/4} \\ \mathbf{1/3} & \mathbf{1/4} & \mathbf{1/5} \\ \mathbf{1/4} & \mathbf{1/5} & \mathbf{1/6} \end{pmatrix}$$

You may check your answer either by forming the product $\boldsymbol{M^{-1}M}$ or the product $\boldsymbol{MM^{-1}}$ or by consulting the web page: `https://matrix.reshish.com/inverse.php`.

What conclusion can you draw about the inverses of matrices which are symmetric, antisymmetric or tridiagonal?

University of Bath, Department of Mechanical Engineering

ME10305 Mathematics 2.

Matrices Sheet 3 — Eigenvalues, eigenvectors and solutions of ODE systems.

NOTE: that Q5 and Q6 are project-like questions involving eigenvectors and eigenvalues, and are beyond the remit of the unit.

1. Find the eigenvalues and eigenvectors of the following matrices;

$$\boldsymbol{A}=\begin{pmatrix}\mathbf{2} & \mathbf{-1}\\ \mathbf{-1} & \mathbf{2}\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}\mathbf{-3} & \mathbf{1}\\ \mathbf{1} & \mathbf{-3}\end{pmatrix},\quad \boldsymbol{C}=\begin{pmatrix}\mathbf{1} & \mathbf{3} & \mathbf{-1}\\ \mathbf{0} & \mathbf{4} & \mathbf{-1}\\ \mathbf{1} & \mathbf{1} & \mathbf{1}\end{pmatrix},\quad \boldsymbol{D}=\begin{pmatrix}\mathbf{2} & \mathbf{1} & \mathbf{1}\\ \mathbf{1} & \mathbf{2} & \mathbf{1}\\ \mathbf{1} & \mathbf{1} & \mathbf{2}\end{pmatrix},\quad \boldsymbol{E}=\begin{pmatrix}\mathbf{1} & \mathbf{1} & \mathbf{1}\\ \mathbf{0} & \mathbf{2} & \mathbf{2}\\ \mathbf{0} & \mathbf{0} & \mathbf{3}\end{pmatrix},$$

$$\boldsymbol{F}=\begin{pmatrix}\boldsymbol{b} & \boldsymbol{a} & \mathbf{0}\\ \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{a}\\ \mathbf{0} & \boldsymbol{c} & \boldsymbol{b}\end{pmatrix},\qquad \boldsymbol{G}=\begin{pmatrix}\boldsymbol{a} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{b} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \boldsymbol{c}\end{pmatrix},\qquad \boldsymbol{H}=\begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{1}\\ \mathbf{0} & \mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{0} & \mathbf{0}\end{pmatrix},\qquad \boldsymbol{J}=\begin{pmatrix}\mathbf{4} & \mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\\ \mathbf{0} & \mathbf{3} & \mathbf{4}\end{pmatrix}.$$

2. Now we apply the eigenvalue theory to solving ODE systems. Solve:

(a) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\mathbf{2} & \mathbf{-1}\\ \mathbf{-1} & \mathbf{2}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$ subject to $\boldsymbol{x(0)=1,\ y(0)=0.}$

(b) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\mathbf{-3} & \mathbf{6}\\ \mathbf{1} & \mathbf{-4}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$ subject to $\boldsymbol{x(0)=1,\ y(0)=0.}$

(c) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\mathbf{-1} & \mathbf{3}\\ \mathbf{4} & \mathbf{-5}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$ subject to $\boldsymbol{x(0)=4,\ y(0)=0.}$

(d) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{3} & \mathbf{-1}\\ \mathbf{0} & \mathbf{4} & \mathbf{-1}\\ \mathbf{1} & \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}$ subject to $\boldsymbol{x(0)=2,\ y(0)=2,\ z(0)=3.}$

In cases (a) and (d) you will be able to use the results of part of Q1 to lighten your load!

3. Solve the following two systems of equations. Some of the work of Q2a may be used.

(a) $\dfrac{\boldsymbol{d^2}}{\boldsymbol{dt^2}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\mathbf{2} & \mathbf{-1}\\ \mathbf{-1} & \mathbf{2}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$ subject to $\boldsymbol{x(0)=1,\ x'(0)=0,\ y(0)=0,\ y'(0)=0.}$

(b) $\dfrac{\boldsymbol{d^2}}{\boldsymbol{dt^2}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\mathbf{-2} & \mathbf{1}\\ \mathbf{1} & \mathbf{-2}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}$ subject to $\boldsymbol{x(0)=1,\ x'(0)=0,\ y(0)=0,\ y'(0)=0.}$

4. Solve the following systems of ODEs.

(a) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}=\begin{pmatrix}\mathbf{4} & \mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\\ \mathbf{0} & \mathbf{3} & \mathbf{4}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}$ (b) $\dfrac{\boldsymbol{d}}{\boldsymbol{dt}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}=-\begin{pmatrix}\mathbf{4} & \mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\\ \mathbf{0} & \mathbf{3} & \mathbf{4}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}$

(c) $\dfrac{\boldsymbol{d^2}}{\boldsymbol{dt^2}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}=\begin{pmatrix}\mathbf{4} & \mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{1} & \mathbf{-1}\\ \mathbf{0} & \mathbf{3} & \mathbf{4}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}.$

If you have already found the eigenvalues and eigenvectors of $\boldsymbol{J}$ in Q1, then this should be very quick to answer.

5. One of the applications of matrix theory is in the area of Markov chains, which is about probabilities that are associated with sequences of events. This is an example adapted from the textbook by Glyn James.

It is stated that dry days follow dry days with a probability of $\mathbf{0.5}$ while wet days follow wet days with probability, $\mathbf{0.6}$. The notation, $\boldsymbol{P(D_n)}$, is the probability that day $\boldsymbol{n}$ is dry, and clearly $\boldsymbol{P(W_n) = 1 - P(D_n)}$ is that it is wet on the same day, where no other choices of weather are available! All of this may be written as follows,

$$\begin{pmatrix} \boldsymbol{P(D_{n+1})} \\ \boldsymbol{P(W_{n+1})} \end{pmatrix} = \begin{pmatrix} \mathbf{0.5} & \mathbf{0.4} \\ \mathbf{0.5} & \mathbf{0.6} \end{pmatrix} \begin{pmatrix} \boldsymbol{P(D_n)} \\ \boldsymbol{P(W_n)} \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} \boldsymbol{P(D_0)} \\ \boldsymbol{P(W_0)} \end{pmatrix} = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix}.$$

This matrix is an example of a probability transition matrix and the initial condition (namely, that it is definitely dry!) is given. Although not necessary, see if you can understand how the matrix/vector equation has been constructed.

(a) Given the initial condition, find $\boldsymbol{P(D_1)}$ and $\boldsymbol{P(W_1)}$ using $\boldsymbol{n = 0}$ in the above equation to find the prediction for the weather on day 1. Carry on like this to, say, day 4 to see if you can guess what the long-term trend is.

(b) Now find the eigenvalues and eigenvectors of the probability transition matrix, and see if you can relate these to your predicted long-term trend found in part (a).

(c) Given that $\boldsymbol{A\underline{v} = \lambda\underline{v}}$ for eigenvectors and eigenvalues, the following is derived,

$$\boldsymbol{A^2\underline{v} = A(A\underline{v}) = A(\lambda\underline{v}) = \lambda\, A\underline{v} = \lambda^2\underline{v},}$$

and so on for $\boldsymbol{A^3}$, $\boldsymbol{A^4}$ and so on. Now rewrite the original initial condition in the form,

$$\begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} = \boldsymbol{A\underline{v}_1 + B\underline{v}_2,}$$

i.e. find the constants $\boldsymbol{A}$ and $\boldsymbol{B}$; here $\boldsymbol{\underline{v}_1}$ and $\boldsymbol{\underline{v}_2}$ are the eigenvectors. Now find out what happens as successive days fly by, but always keep the result in terms of a sum of multiples of the eigenvectors.

(d) The property you have just uncovered is a feature of probability transition matrices: one eigenvalue is equal to $\mathbf{1}$ and the corresponding eigenvector is the long-term trend. All of the other eigenvalues are smaller in magnitude. This happens because the elements in each column of the probability transition matrix adds to $\mathbf{1}$. Check all of these statements by finding the eigenvalues and eigenvectors of

$$\begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{1-a} & \boldsymbol{1-b} \end{pmatrix}.$$

(e) Finally we wish to design our probability so that we can have dry weekends and wet weekdays, so the long-term behaviour is that it will be dry $\mathbf{2/7}$ths of the time and wet $\mathbf{5/7}$ths of the time. Let $\boldsymbol{a = 0.5}$ (as it was at the start) and find the value of $\boldsymbol{b}$ which will ensure this outcome.

6. This final question is lengthy and well above the standard required for the exam. However, it may be done if one is led carefully through it. The aim is to find the eigenvalues of tridiagonal matrices such as the following, by deriving a recurrence relation:

$$\boldsymbol{J_2} = \begin{pmatrix} \mathbf{-2} & \mathbf{1} \\ \mathbf{1} & \mathbf{-2} \end{pmatrix} \quad \boldsymbol{J_3} = \begin{pmatrix} \mathbf{-2} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{-2} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{-2} \end{pmatrix} \quad \boldsymbol{J_4} = \begin{pmatrix} \mathbf{-2} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{-2} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{-2} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{-2} \end{pmatrix}.$$

You have already met these matrices in the problem sheet on determinants.

(i) Find the eigenvalues of $\boldsymbol{J_2}$, $\boldsymbol{J_3}$, $\boldsymbol{J_4}$ and $\boldsymbol{J_5}$ by using the standard techniques. I am not interested in finding the eigenvectors for now. Note that the factor $\boldsymbol{(\lambda + 2)}$ plays a vital role, and therefore *do not* expand your expressions for the determinants; keep them in terms of powers of $\boldsymbol{(\lambda + 2)}$. You should be able to find simple analytical values for the eigenvalues, even for $\boldsymbol{J_5}$ (for which three of the five eigenvalues are negative integers).

(ii) While undertaking part (i), you should have noticed how $\det(\boldsymbol{J_n})$ may be written in terms of $\det(\boldsymbol{J_{n-1}})$ and $\det(\boldsymbol{J_{n-2}})$ in the same manner as we found in Q2 of the determinants problem sheet. Show that,

$$\det(\boldsymbol{J_n}) = \boldsymbol{-(2+\lambda)}\det(\boldsymbol{J_{n-1}}) - \det(\boldsymbol{J_{n-2}}).$$

Now use your expressions for $\det(\boldsymbol{J_2})$ and $\det(\boldsymbol{J_3})$ obtained in part (i) to show that your expression for $\det(\boldsymbol{J_4})$ is correct. Likewise show that your expression for $\det(\boldsymbol{J_5})$ is correct. What is the polynomial which represents $\det(\boldsymbol{J_6}) = \mathbf{0}$?

(iii) Now we will go into further detail with $\boldsymbol{J_5}$ to find a general way of finding the eigenvalues for $\boldsymbol{J_n}$ and the eigenvectors. Assume that the eigenvector for $\boldsymbol{J_5}$ has the following form,

$$\begin{pmatrix} \mathbf{-2-\lambda} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{-2-\lambda} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{-2-\lambda} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{-2-\lambda} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{-2-\lambda} \end{pmatrix} \begin{pmatrix} \boldsymbol{\sin\alpha} \\ \boldsymbol{\sin 2\alpha} \\ \boldsymbol{\sin 3\alpha} \\ \boldsymbol{\sin 4\alpha} \\ \boldsymbol{\sin 5\alpha} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{\alpha}$ is currently unknown. Now write out the equation corresponding to, say, row 3 of this equation, rewrite this equation using multiple angle formulae (i.e. let $\boldsymbol{\sin 4\alpha = \sin(3\alpha + \alpha)}$ and $\boldsymbol{\sin 2\alpha = \sin(3\alpha - \alpha)}$), and hence show that,

$$\boldsymbol{\lambda = -2 + 2\cos\alpha.}$$

Row 1 of the matrix equation is a special case as it has only two coefficients; check that it too gives the same expression for $\boldsymbol{\lambda}$. Row 5 is also a special case, but this should yield $\boldsymbol{\sin 6\alpha = 0}$. Now you are in a position to write down a simple formula for the eigenvalues, $\boldsymbol{\lambda}$, for $\boldsymbol{J_5}$ which should fit with your original calculations. What are the eigenvectors? What is the implication for $\boldsymbol{J_n}$ in general?

Department of Mechanical Engineering, University of Bath

Mathematics 2 ME10305

Problem Sheet — Root finding and iteration schemes.

1. The aim for this question is to repeat some of the techniques used in the lectures to find roots of equations.

(a) Use a sketch to find the number of roots there are likely to be of the equation, $f(x) = x^3 - 2x + 1 = 0$.

(b) Use two *ad hoc* iteration schemes to determine the roots of the equation.

(c) Use the Newton–Raphson scheme with $x_0 = 1.1$ as the initial iterate to find one of those roots.

(d) Taking this root, use the perturbation method to determine how quickly each method used converges to that root.

2. Find the only real root of the cubic $x^3 - x^2 - x - 1 = 0$ correct to six significant figures. Use any method you like.

3. Use a suitable sketch to show that $f(x) = e^{-x} - x = 0$ has only one root. Use both the possible ad hoc schemes and the Newton-Raphson method to find that root. Analyze the approach to the solution for all three methods by setting $x_n = X + \epsilon_n$ where X is the solution of $f(X) = 0$, i.e. it satisfies $e^{-X} = X$.

4. Use the Newton-Raphson method to find the first 4 positive roots of $f(x) = x \sin x - 1 = 0$. Rough locations of the roots may be obtained using a suitable sketch.

5. So let us create a general perturbation analysis of the convergence of the Newton-Raphson method towards a double root. We'll fix the roots to be at $x = 0$ reflects a general situation perfectly, and therefore we will consider $f(x) = x^2 g(x)$ where $g(0) \neq 0$. Write down the Newton-Raphson formula for this $f(x)$, and use a perturbation analysis to determine how quickly the iteration scheme will converge to $x = 0$. What happens when we have $f(x) = x^m g(x)$ where m is a positive integer?

6. [This question is best tackled using some suitable software to undertake the computations.]

The objective is to find the zeros of the function, $f(x) = x^{1/3} - \ln x$, where it is no secret that any such zeros must be positive. Use both of the possible *ad hoc* methods and the Newton-Raphson method to find these zeros. I am not sure that it will be useful to sketch this function, but trialling a few tentative values of x is a good start.

7. [If you have access to a machine/software which can compute with complex numbers, then you may undertake this question, should you wish.]

Write down the Newton-Raphson scheme for $f(x) = x^2 + 1$. Now use $x_0 = 0.5 + 0.5j$ as the initial iterate. To what value does the Newton-Raphson scheme converge?

Use the same method for finding the square root of $2j$. Use $x_0 = 1 + 0j$ in this case.

8. [This is a project-style of question. It is lengthy and intricate, but it ends up with an algebraic equation to solve for which the Newton-Raphson method is well-suited. The background application is on the vibrations of a beam.]

First, an introduction to Ordinary Differential Eigenvalue problems. I'll summarise the process first with a 2nd order ODE, and your job will be to apply the same ideas to a 4th order ODE.

The vibrations of a taut string are described by the wave equation, and eventually one obtains the ODE,

$$\frac{d^2y}{dx^2} + \omega^2 y = 0, \qquad \text{subject to } \; y(0) = 0, \; y(1) = 0.$$

The value, y, is a displacement, like that of a violin string, and the boundary conditions represent a zero displacement at both ends, which is what one expects of a violin. Clearly $y = 0$ satisfies the ODE and boundary conditions, but we've heard violins and therefore we need nonzero solutions. The value, ω, is related to the frequency of vibration of the string, and nonzero solutions (eigensolutions!) arise for certain frequencies only, and it is these values which we seek (eigenvalues!). The analysis proceeds as follows.

The general solution is $y = A\cos\omega x + B\sin\omega x$. Given that $y(0) = 0$ we must therefore have $A = 0$, which means that we now have $y = B\sin\omega x$. Application of $y(1) = 0$ yields,

$$B\sin\omega = 0.$$

We can't have $B = 0$ because that means that string has no displacement, and that defeats the purpose of the analysis. So we must have $\omega = n\pi$, where n is a positive integer; these values of ω are called the eigenvalues of the ODE. For a chosen value of n, the associated disturbance shape is $y = B\sin n\pi x$ where B is arbitary; these are the eigensolutions.

Your task, should you wish to take it on, is to use a similar analysis of the corresponding equation for a beam, namely,

$$\frac{d^4y}{dx^4} - \omega^4 y = 0 \qquad \text{subject to } \; y(0) = y'(0) = 0, \; y(1) = y'(1) = 0.$$

The boundary conditions are consistent with those of a cantilever: zero displacement and zero slope.

(a) Use the substitution, $y = e^{\lambda x}$, to write down the general solution in terms of four functions and with four arbitrary constants. Where you have to choose between exponentials and hyperbolic functions, I would advise the hyperbolics on this occasion. Sorry.

(b) Now apply the boundary conditions to get four algebraic equations. The following will be a somewhat arduous trek. The aim is to try to eliminate three of the arbitrary constants in order to have an equation involving the last arbitrary constant and an expression involing ω. If this has worked correctly you should get

$$\cos\omega \, \cosh\omega = 1. \tag{1}$$

(c) Sketch both $\cos\omega$ and $1/\cosh\omega$ to estimate where the first root of Eq. (1) might be. (Ignore the obvious one at $x = 0$ which actually yields nothing of any use!)

(d) Apply Newton-Raphson to find this first value of ω. Again, ω is the freqency of vibration of the beam and, given how much more constrained the beam is compared with the string, you should obtain a higher lowest frequency here. In the solutions I will provide the first four values of ω and the corresponding shapes of vibration.

Department of Mechanical Engineering, University of Bath

Mathematics 2 ME10305

Problem Sheet — Fourier Series

Note: For the purposes of exam revision, questions 2, 4 and 5 are the important ones. Question 1 has a general importance while questions 3 and 6 provide some good background knowledge.

1. Which of the following functions are even, odd or neither about $x = 0$? Of those which are periodic, find the fundamental period.

(i) $\sin t$ (ii) $\sin^2 t$ (iii) $\sqrt{1-t^2}$ $(-1 \le t \le 1)$ (iv) te^{-t} (v) e^{-t^2}

(vi) te^{-t^2} (vii) $\sin t + \sin 3t$ (viii) $\sin t \sin 3t$ (ix) $\sin t \sin \sqrt{2}t$

(x) $f(t) = t + 1$ for $-1 < t < 1$, $f(t) = f(t+2)$

The final two functions have one definition in part of the period and another in the remaining part:

(xi) $f(t) = t$ for $0 \le t \le 1$, $f(t) = 2 - t$ for $1 \le t \le 2$, $f(t) = f(t+2)$

(xii) $f(t) = 1$ for $0 < t \le 1$, $f(t) = 2 - t$ for $1 \le t < 2$, $f(t) = f(t+2)$

2. Find the Fourier Series representations of the following functions, bearing in mind that quicker results may be obtained when symmetries are accounted for. In all cases, (i) sketch the function, (ii) try to predict in advance how fast the Fourier coefficients decay by checking the continuity of each function **before** attempting to find the Fourier Series.

(a) $f(t) = t^2$ $-\pi \le t \le \pi$ with $f(t) = f(t + 2\pi)$.

(b) $f(t) = t - t^2$ $0 \le t \le 1$ with $f(t) = f(t+1)$.

(c) $f(t) = \pi^2 t - t^3$ $-\pi \le t \le \pi$ with $f(t) = f(t + 2\pi)$.

(d) $f(t) = t - t^3$ $-1 \le t \le 1$ with $f(t) = f(t+2)$.

(e) $f(t) = \cos \alpha t$ $-1 \le t \le 1$ with $f(t) = f(t+2)$.

(f) $f(t) = \cosh \alpha t$ $-1 \le t \le 1$ with $f(t) = f(t+2)$.

(g) $f(t) = 1$ for $0 < t < 1$, $f(t) = -1$ for $1 < t < 2$, with $f(t) = f(t+2)$.

(h) $f(t) = 3t^5 - 10t^3 + 7t$ for $-1 \le t \le 1$ with $f(t) = f(t+2)$.

(i) $f(t) = |\sin t|$.

(j) $f(t) = t^2$ for $0 < t < 1$ with $f(t) = f(t+1)$.

3. The aim of this question is two-fold, to derive the formulae for the Fourier coefficients and to prove Parseval's theorem. For simplicity we will consider functions of period 2π. **(This question is over and above what would be expected in an exam question, but it is included to show where the formulae for the Fourier coefficients come from.)**

If m and n are nonzero integers, first show that

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \cos nt \, \cos mt \, dt = \begin{cases} 0 & \text{when } n \neq m, \\ 1 & \text{when } n = m. \end{cases}$$

Do the same for the integral of the product of two sines. Finally, show that the integral of $\sin nt \cos mt$ over the same range is zero.

Hence use these results and the standard definition of the Fourier series,

$$f(t) = \tfrac{1}{2}A_0 + \sum_{n=1}^{\infty}(A_n \cos nt + B_n \sin nt),$$

to find expressions for the Fourier coefficients.

For a function of period 2π, Parseval's theorem is

$$\frac{1}{\pi}\int_{-\pi}^{\pi}[f(t)]^2 dt = \tfrac{1}{2}A_0^2 + \sum_{n=1}^{\infty}[A_n^2 + B_n^2];$$

prove this using the results you have already derived. This result is related to the energy content of a periodic signal.

4. If $g(t) = t^2$ in the range $-\pi \leq t \leq \pi$, and $g(t)$ has a period equal to 2π, find its Fourier series. Hence find the Particular Integral of the ordinary differential equation,

$$\frac{dy}{dt} + cy = g(t).$$

5. Consider the response of the following undamped mass/spring system to a rectified sine wave signal:

$$\frac{d^2y}{dt^2} + K^2 y = |\sin t|.$$

By sketching the signal confirm that its period is π and determine its Fourier series. Hence find the response $y(t)$. For which values of K is there resonance?

6. **[Long and difficult.]**‘ Find the Fourier series of the response to the following damped system to the rectified sine wave:

$$y'' + cy' + K^2 y = |\sin t|.$$

For which values of K is there resonance?

Department of Mechanical Engineering, University of Bath

Mathematics 2 ME10305

Problem sheet — Least Squares Fitting of data

Note: Past exam papers are a good resource of questions of the length and type I will give in the forthcoming exam period. I would even recommend attempting these first before the following.

1. You are given the following data:

x :	0	$1/8$	$2/8$	$3/8$	$4/8$	$5/8$	$6/8$	$7/8$	1
y :	0.06625	0.27985	0.24641	0.28299	0.41285	0.41753	0.57629	0.56542	0.47777

(i) Assume that this data should lie on a straight line through the origin. Find that line.
(ii) Now assume that the data should lie on the straight line $y = mx + c$. What is the line?
(iii) Extend (ii) to a quadratic curve.

In all the cases determine the RMS of the residuals, $\sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2}$, of the line/curve obtained. Which fit is best?

[Note: the detailed summations required for this might be done more easily using a package such as Excel.]

2. In this question we are going to play a different game in the sense that the data to be fitted has a different type of randomness. I would like to see how much accuracy one can gain from a set of data which has been rounded quite severely. We'll use the conversion from miles to kilometres for this purpose. The following is a Table of data which is subject to different degrees of rounding off, namely to zero, $\mathbf{1}$, $\mathbf{2}$ and $\mathbf{3}$ decimal places; x denotes miles and yn denotes kilometres with n decimal places.

For each of these sets of data, fit a straight line through the origin to determine the least squares version of the conversion factor between the units. How accurate are they? You may compare with the exact value, $\mathbf{1.609344}$.

x :	1	2	3	4	5	6	7	8	9	10
$y0$:	2	3	5	6	8	10	11	13	14	16
$y1$:	1.6	3.2	4.8	6.4	8.0	9.7	11.3	12.9	14.5	16.1
$y2$:	1.61	3.22	4.83	6.44	8.05	9.66	11.27	12.87	14.48	16.09
$y3$:	1.609	3.219	4.828	6.437	8.047	9.656	11.265	12.875	14.484	16.093

If this has piqued your interest, then try the same with the conversion between ounces and grams, for which 1 ounce is equal to 28.349523 grams. Use 16 sets of data from 1oz to 16oz, and use the same number of decimal places as in the above miles/kilometres example. The raw data may be found at

`http://staff.bath.ac.uk/ensdasr/ME10305.bho/ounces-to-grams.txt`

although there is also a link to it at the unit webpage. Given that there are 16 data points, you might be able to coerce Excel into doing your calculations for you.

3. Suppose that you were given a set of experimental data where it is suspected that the data should satisfy an equation of the form

$$y = a + b/x.$$

How could the data be manipulated in order to use standard Least Squares theory? [Note: I can think of at least two different ways of doing this.]

What about the equation,

$$y = \frac{a}{x+b}?$$

4. In many experimental situations the observable, y, is a power law function of the parameter, x. In other words it takes the form,

$$y = ax^b,$$

where a and b need to be found. [For example, the rate of heat transfer from a hot vertical surface is proportional to the $\frac{1}{4}$ power of the temperature difference between the heated surface and the ambient conditions.] How would you convert this power-law relationship into a straight line relationship?

5. Experimental measurements have been taken of z, which is a function of both x and y. It is suspected that z is a linear function of x and y, and therefore it represents a plane in 3D space. Use least squares theory to determine the three unknown coefficients in the following equation for the plane,

$$z = ax + by + c.$$

6. An obsessive cyclist has a comprehensive set of data for his/her ride-times over the same route for a period of several years. Naturally the cyclist's journey times are slower when the weather is colder, and faster when it is warmer. The cyclist wishes to determine (i) what the long term general trend is in terms of speed, (ii) what seasonal effect should be expected given the time of year. To this end, the cyclist proposes a least squares fit of the form,

$$T = a + bt + c\cos(2\pi t) + d\sin(2\pi t)$$

where T is the ride-time and t is time measured in years. Develop the least squares theory which will allow the cyclist to achieve his/her twin objectives.

7. Let us generalize things a little. Suppose that we have to fit the following curve to measured data:

$$y = a\,f(x) + b\,g(x),$$

where $f(x)$ and $g(x)$ are chosen functions and a and b are to be found. It may help you to think of $f(x)$ and $g(x)$ as being 1 and x (as in Q1(ii)), or 1 and $1/x$ (as in Q3), or even $\cos(2\pi x)$ and $\sin(2\pi x)$ (the last two components in Q6). How does one modify least squares theory to this generalization?

8. An experiment has two measurables, $y(x)$ and $z(x)$, as the control parameter, x, is varied. Both y and z should be linear with different slopes, but should have the same intercept on the vertical axis. That is, we wish to fit the following to the data, where there are three constants to find:

$$y = ax + c, \qquad z = bx + c.$$

How is this done? [This is a simplified version of a problem a couple of third year students brought to me where they had to fit a straight line to five measureables all of which had the same intercept. So this question isn't a product of my wild imaginings! What was the answer for their problem?]