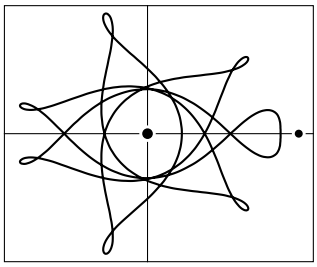

1

Differential and Difference Equations

10 Differential Equation Problems

100 Introduction to differential equations

As essential tools in scientific modelling, differential equations are familiar to every educated person. In this introductory discussion we do not attempt to restate what is already known, but rather to express commonly understood ideas in the style that will be used for the rest of this book.

The aim will always be to understand, as much as possible, what we expect to happen to a quantity that satisfies a differential equation. At the most obvious level, this means predicting the value this quantity will have at some future time. However, we are also interested in more general questions such as the adherence to possible conservation laws or perhaps stability of the long-term solution. Since we emphasize numerical methods, we often discuss problems with known solutions mainly to illustrate qualitative and numerical behaviour.

Even though we sometimes refer to 'time' as the independent variable, that is, as the variable on which the value of the 'solution' depends, there is no reason for insisting on this interpretation. However, we generally use x to denote the 'independent' or 'time' variable and y to denote the 'dependent variable'. Hence, differential equations will typically be written in the form

$$y'(x) = f(x, y(x)), \tag{100a}$$

where

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x}.$$

Numerical Methods for Ordinary Differential Equations, Third Edition. J. C. Butcher.

Sometimes, for convenience, we omit the x in $y(x)$.

The terminology used in (100a) is misleadingly simple, because y could be a vector-valued function. Thus, if we are working in $\mathbb{R}^N$, and x is permitted to take on any real value, then the domain and range of the function f which defines a differential equation and the solution to this equation are given by

$$\begin{aligned} f &: \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N, \\ y &: \mathbb{R} \to \mathbb{R}^N. \end{aligned}$$

Since we might be interested in time values that lie only in some interval $[a, b]$, we sometimes consider problems in which $y : [a, b] \to \mathbb{R}^N$, and $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$. When dealing with specific problems, it is often convenient to focus, not on the vector-valued functions f and y, but on individual components. Thus, instead of writing a differential equation system in the form of (100a), we can write coupled equations for the individual components:

$$\begin{aligned} y_1'(x) &= f_1(x, y_1, y_2, \ldots, y_N), \\ y_2'(x) &= f_2(x, y_1, y_2, \ldots, y_N), \\ \vdots &\qquad\qquad \vdots \\ y_N'(x) &= f_N(x, y_1, y_2, \ldots, y_N). \end{aligned} \tag{100b}$$

Autonomous differential equations

A differential equation for which f is a function not of x, but of y only, is said to be 'autonomous'. Some equations arising in physical modelling are more naturally expressed in one form or the other, but we emphasize that it is always possible to write a non-autonomous equation in an equivalent autonomous form. All we need to do to change the formulation is to introduce an additional component y_{N+1} into the y vector, and ensure that this can always maintain the same value as x, by associating it with the differential equation $y_{N+1}' = 1$. Thus, the modified system is

$$\begin{aligned} y_1'(x) &= f_1(y_{N+1}, y_1, y_2, \ldots, y_N), \\ y_2'(x) &= f_2(y_{N+1}, y_1, y_2, \ldots, y_N), \\ \vdots &\qquad\qquad \vdots \\ y_N'(x) &= f_N(y_{N+1}, y_1, y_2, \ldots, y_N), \\ y_{N+1}'(x) &= 1. \end{aligned} \tag{100c}$$

A system of differential equations alone does not generally define a unique solution, and it is necessary to add to the formulation of the problem a number of additional conditions. These are either 'boundary conditions', if further information is given at two or more values of x, or 'initial conditions', if all components of y are specified at a single value of x.

Initial value problems

If the value of $y(x_0) = y_0$ is given, then the pair of equations

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0,$$

is known as an 'initial value problem'. Our main interest in this book is with exactly this problem, where the aim is to obtain approximate values of $y(x)$ for specific values of x, usually with $x > x_0$, corresponding to the prediction of the future states of a differential equation system.

Note that for an N-dimensional system, the individual components of an initial value vector need to be given specific values. Thus, we might write

$$y_0 = [\,\eta_1 \quad \eta_2 \quad \cdots \quad \eta_N\,]^{\mathsf{T}}.$$

When the problem is formally converted to autonomous form (100c), the value of η_{N+1} must be identical to x_0, otherwise the requirement that $y_{N+1}(x)$ should always equal x would not be satisfied.

For many naturally occurring phenomena, the most appropriate form in which to express a differential equation is as a high order system. For example, an equation might be of the form

$$y^{(n)} = \phi\big(x, y, y', y'', \ldots, y^{(n-1)}\big),$$

with initial values given for $y(x_0), y'(x_0), y''(x_0), \ldots, y^{(n-1)}(x_0)$. Especially important in the modelling of the motion of physical systems subject to forces are equation systems of the form

$$\begin{aligned} y_1''(x) &= f_1(y_1, y_2, \ldots, y_N), \\ y_2''(x) &= f_2(y_1, y_2, \ldots, y_N), \\ &\vdots \qquad\qquad \vdots \\ y_N''(x) &= f_N(y_1, y_2, \ldots, y_N), \end{aligned} \tag{100d}$$

where the equations, though second order, do have the advantages of being autonomous and without $y_1', y_2', \ldots, y_N'$ occurring amongst the arguments of $f_1, f_2, \ldots, f_N$.

To write (100d) in what will become our standard first order system form, we can introduce additional components $y_{N+1}, y_{N+2}, \ldots, y_{2N}$. The differential equation system (100d) can now be written as the first order system

$$\begin{aligned} y_1'(x) &= y_{N+1}, \\ y_2'(x) &= y_{N+2}, \\ &\vdots \quad \vdots \\ y_N'(x) &= y_{2N}, \\ y_{N+1}'(x) &= f_1(y_1, y_2, \ldots, y_N), \\ y_{N+2}'(x) &= f_2(y_1, y_2, \ldots, y_N), \\ &\vdots \qquad\qquad \vdots \\ y_{2N}'(x) &= f_N(y_1, y_2, \ldots, y_N). \end{aligned}$$

101 The Kepler problem

The problems discussed in this section are selected from the enormous range of possible scientific applications. The first example problem describes the motion of a single planet about a heavy sun. By this we mean that, although the sun exerts a gravitational attraction on the planet, we regard the corresponding attraction of the planet on the sun as negligible, and that the sun will be treated as being stationary. This approximation to the physical system can be interpreted in another way: even though both bodies are in motion about their centre of mass, the motion of the planet relative to the sun can be modelled using the simplification we have described. We also make a further assumption, that the motion of the planet is confined to a plane.

Let $y_1(x)$ and $y_2(x)$ denote rectangular coordinates centred at the sun, specifying at time x the position of the planet. Also let $y_3(x)$ and $y_4(x)$ denote the components of velocity in the y_1 and y_2 directions, respectively. If M denotes the mass of the sun, γ the gravitational constant and m the mass of the planet, then the attractive force on the planet will have magnitude

$$\frac{\gamma M m}{y_1^2 + y_2^2}.$$

Resolving this force in the coordinate directions, we find that the components of acceleration of the planet, due to this attraction, are $-\gamma M y_1 (y_1^2 + y_2^2)^{-3/2}$ and $-\gamma M y_2 (y_1^2 + y_2^2)^{-3/2}$, where the negative sign denotes the inward direction of the acceleration.

We can now write the equations of motion:

$$\begin{aligned}
\frac{\mathrm{d}y_1}{\mathrm{d}x} &= y_3, \\
\frac{\mathrm{d}y_2}{\mathrm{d}x} &= y_4, \\
\frac{\mathrm{d}y_3}{\mathrm{d}x} &= -\frac{\gamma M y_1}{(y_1^2 + y_2^2)^{3/2}}, \\
\frac{\mathrm{d}y_4}{\mathrm{d}x} &= -\frac{\gamma M y_2}{(y_1^2 + y_2^2)^{3/2}}.
\end{aligned}$$

By adjusting the scales of the variables, the factor γM can be removed from the formulation, and we arrive at the equations

$$\frac{\mathrm{d}y_1}{\mathrm{d}x} = y_3, \tag{101a}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}x} = y_4, \tag{101b}$$

$$\frac{\mathrm{d}y_3}{\mathrm{d}x} = -\frac{y_1}{(y_1^2 + y_2^2)^{3/2}}, \tag{101c}$$

$$\frac{\mathrm{d}y_4}{\mathrm{d}x} = -\frac{y_2}{(y_1^2 + y_2^2)^{3/2}}. \tag{101d}$$

The solutions of this system are known to be conic sections, that is, ellipses, parabolas or hyperbolas, if we ignore the possibility that the trajectory is a straight line directed either towards or away from the sun. We investigate this further after we have shown that two 'first integrals', or invariants, of the solution exist.

Conservation of Hamiltonian and angular momentum

Theorem 101A *The quantities*

$$H = \tfrac{1}{2}\left(y_3^2 + y_4^2\right) - (y_1^2 + y_2^2)^{-1/2},$$
$$A = y_1 y_4 - y_2 y_3$$

are constant.

Proof. We verify that the values of $\mathrm{d}H/\mathrm{d}x$ and $\mathrm{d}A/\mathrm{d}x$ are zero if y satisfies (101a)–(101d). We have

$$\begin{aligned}\frac{\mathrm{d}H}{\mathrm{d}x} &= y_3\frac{\mathrm{d}y_3}{\mathrm{d}x} + y_4\frac{\mathrm{d}y_4}{\mathrm{d}x} + y_1\frac{\mathrm{d}y_1}{\mathrm{d}x}(y_1^2+y_2^2)^{-3/2} + y_2\frac{\mathrm{d}y_2}{\mathrm{d}x}(y_1^2+y_2^2)^{-3/2}\\ &= -\frac{y_1y_3}{(y_1^2+y_2^2)^{3/2}} - \frac{y_2y_4}{(y_1^2+y_2^2)^{3/2}} + \frac{y_1y_3}{(y_1^2+y_2^2)^{3/2}} + \frac{y_2y_4}{(y_1^2+y_2^2)^{3/2}}\\ &= 0\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{d}A}{\mathrm{d}x} &= y_1\frac{\mathrm{d}y_4}{\mathrm{d}x} + \frac{\mathrm{d}y_1}{\mathrm{d}x}y_4 - y_2\frac{\mathrm{d}y_3}{\mathrm{d}x} - \frac{\mathrm{d}y_2}{\mathrm{d}x}y_3\\ &= -\frac{y_1y_2}{(y_1^2+y_2^2)^{3/2}} + y_3y_4 + \frac{y_2y_1}{(y_1^2+y_2^2)^{3/2}} - y_4y_3\\ &= 0.\end{aligned}$$

□

The quantities H and A are the 'Hamiltonian' and 'angular momentum', respectively. Note that $H = T + V$, where $T = \frac{1}{2}\left(y_3^2 + y_4^2\right)$ is the kinetic energy and $V = -(y_1^2 + y_2^2)^{-1/2}$ is the potential energy.

A further property of this problem is its invariance under changes of scale of the variables:

$$\begin{aligned}y_1 &= \alpha^{-2}\overline{y}_1,\\ y_2 &= \alpha^{-2}\overline{y}_2,\\ y_3 &= \alpha\overline{y}_3,\\ y_4 &= \alpha\overline{y}_4,\\ x &= \alpha^{-3}\overline{x}.\end{aligned}$$

The Hamiltonian and angular momentum get scaled to

$$\begin{aligned}\overline{H} &= \tfrac{1}{2}\left(\overline{y}_3^2 + \overline{y}_4^2\right) - (\overline{y}_1^2 + \overline{y}_2^2)^{-1/2} &&= \alpha^{-2}H,\\ \overline{A} &= \overline{y}_1\overline{y}_4 - \overline{y}_2\overline{y}_3 &&= \alpha A.\end{aligned}$$

Invariance under orthogonal transformations

A second type of transformation is based on a two-dimensional orthogonal transformation (that is, a rotation or a reflection or a composition of these) Q, where $Q^{-1} = Q^{\mathsf{T}}$. The time variable x is invariant, and the position and velocity variables get transformed to

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \\ \overline{y}_4 \end{bmatrix}.$$

It is easy to see that $A = 0$ implies that the trajectory lies entirely in a subspace defined by $\cos(\theta)y_1 = \sin(\theta)y_2$, $\cos(\theta)y_3 = \sin(\theta)y_4$ for some fixed angle θ. We move on from this simple case and assume that $A \neq 0$. The sign of H is of crucial importance: if $H \geq 0$ then it is possible to obtain arbitrarily high values of $y_1^2 + y_2^2$ without $y_3^2 + y_4^2$ vanishing. We exclude this case for the present discussion and assume that $H < 0$. Scale H so that it has a value $-\frac{1}{2}$ and at the same time A takes on a positive value. This value cannot exceed 1 because we can easily verify an identity involving the derivative of $r = \sqrt{y_1^2 + y_2^2}$. This identity is

$$\left(r\frac{\mathrm{d}r}{\mathrm{d}x}\right)^2 = 2Hr^2 + 2r - A^2 = -r^2 + 2r - A^2. \tag{101e}$$

Since the left-hand side cannot be negative, the quadratic function in r on the right-hand side must have real roots. This implies that $A \leq 1$. Write $A = \sqrt{1 - e^2}$, for $e \geq 0$, where we see that e is the eccentricity of an ellipse on which the orbit lies. The minimum and maximum values of r are found to be $1 - e$ and $1 + e$, respectively. Rotate axes so that when $r = 1 - e$, $y_1 = 1 - e$ and $y_2 = 0$. At this point we find that $y_3 = 0$ and $y_4 = \sqrt{(1 + e)/(1 - e)}$. We will use these as initial values at $x = 0$.

Change to polar coordinates by writing $y_1 = r\cos(\theta)$, $y_2 = r\sin(\theta)$. It is found that

$$y_3 = \frac{\mathrm{d}y_1}{\mathrm{d}x} = \frac{\mathrm{d}r}{\mathrm{d}x}\cos(\theta) - r\frac{\mathrm{d}\theta}{\mathrm{d}x}\sin(\theta),$$
$$y_4 = \frac{\mathrm{d}y_2}{\mathrm{d}x} = \frac{\mathrm{d}r}{\mathrm{d}x}\sin(\theta) + r\frac{\mathrm{d}\theta}{\mathrm{d}x}\cos(\theta),$$

so that, because $y_1 y_4 - y_2 y_3 = \sqrt{1 - e^2}$, we find that

$$r^2\frac{\mathrm{d}\theta}{\mathrm{d}x} = \sqrt{1 - e^2}. \tag{101f}$$

From (101e) and (101f) we find a differential equation for the path traced out by the orbit

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2 = \frac{1}{1 - e^2}r^2\left(e^2 - (1 - r)^2\right),$$

and we can verify that this is satisfied by

$$\frac{1 - e^2}{r} = 1 + e\cos(\theta).$$

If we change back to Cartesian coordinates, we find that all points on the trajectory lie on the ellipse

$$(y_1 + e)^2 + \frac{y_2^2}{1 - e^2} = 1,$$

with centre $(-e, 0)$, eccentricity e, and major and minor axis lengths 1 and $\sqrt{1 - e^2}$ respectively.

As we have seen, a great deal is known about this problem. However, much less is known about the motion of a many-body gravitational system.

One of the aims of modern numerical analysis is to understand the behaviour of various geometrical properties. In some cases it is possible to preserve the value of quantities that are invariant in the exact solution. In other situations, such as problems where the Hamiltonian is theoretically conserved, it may be preferable to conserve other properties, such as what is known as 'symplectic behaviour'.

We consider further gravitational problems in Subsection 120.

102 A problem arising from the method of lines

The second initial value problem we consider is based on an approximation to a partial differential equation. Consider the parabolic system

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad (x, t) \in [0, 1] \times [0, \infty), \tag{102a}$$

where we have used t to represent time, x to represent distance and $u(x, t)$ to represent some quantity, such as temperature, which diffuses with time. For this problem it is necessary to impose conditions on the boundaries $x = 0$ and $x = 1$ as well as at the initial time $t = 0$. We may interpret the solution as the distribution of the temperature at points in a conducting rod, given that the temperature is specified at the ends of the rod. In this case the boundary conditions would be of the form $u(0, t) = \alpha(t)$ and $u(1, t) = \beta(t)$. Equation (102a) is known as the heat or diffusion equation, and the conditions given at $x = 0$ and $x = 1$ are known as Dirichlet conditions. This is in contrast to Neumann conditions, in which the values of $\partial u/\partial x$ are given at the ends of the x interval.

Space discretization

To convert this problem into an ordinary differential equation system, which mimics the behaviour of the parabolic equation, let $y_1(t), y_2(t), \ldots, y_N(t)$ denote the values of $u(\frac{1}{N+1}, t), u(\frac{2}{N+1}, t), \ldots, u(\frac{N}{N+1}, t)$, respectively. That is,

$$y_j(t) = u\Big(\frac{j}{N + 1}, t\Big), \qquad j = 0, 1, 2, \ldots, N + 1,$$

where we have included $y_0(t) = u(0, t)$, $y_{N+1}(t) = u(1, t)$ for convenience.

For $j = 1, 2, \ldots, N$, $\partial^2 u/\partial x^2$, evaluated at $x = j/(N + 1)$, is approximately equal to $(N + 1)^2(y_{j-1} - 2y_j + y_{j+1})$. Hence, the vector of derivatives of $y_1, y_2, \ldots, y_N$ is given by

$$
\begin{aligned}
\frac{\mathrm{d}y_1(t)}{\mathrm{d}t} &= (N+1)^2\big(\alpha(t) - 2y_1(t) + y_2(t)\big),\\
\frac{\mathrm{d}y_2(t)}{\mathrm{d}t} &= (N+1)^2\big(y_1(t) - 2y_2(t) + y_3(t)\big),\\
\frac{\mathrm{d}y_3(t)}{\mathrm{d}t} &= (N+1)^2\big(y_2(t) - 2y_3(t) + y_4(t)\big),\\
\vdots &\qquad\qquad \vdots\\
\frac{\mathrm{d}y_{N-1}(t)}{\mathrm{d}t} &= (N+1)^2\big(y_{N-2}(t) - 2y_{N-1}(t) + y_N(t)\big),\\
\frac{\mathrm{d}y_N(t)}{\mathrm{d}t} &= (N+1)^2\big(y_{N-1}(t) - 2y_N(t) + \beta(t)\big).
\end{aligned}
$$

This system can be written in vector–matrix form as

$$y'(t) = Ay(t) + v(t), \tag{102b}$$

where

$$
A = (N+1)^2\begin{bmatrix}
-2 & 1 & 0 & \cdots & 0 & 0 & 0\\
1 & -2 & 1 & \cdots & 0 & 0 & 0\\
0 & 1 & -2 & \cdots & 0 & 0 & 0\\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & -2 & 1 & 0\\
0 & 0 & 0 & \cdots & 1 & -2 & 1\\
0 & 0 & 0 & \cdots & 0 & 1 & -2
\end{bmatrix}, \quad
v = (N+1)^2\begin{bmatrix}
\alpha(t)\\ 0\\ 0\\ \vdots\\ 0\\ 0\\ \beta(t)
\end{bmatrix}.
$$

The original problem is 'dissipative' in the sense that, if u and v are each solutions to the diffusion equation, which have identical boundary values but different initial values, then

$$W(t) = \frac{1}{2}\int_0^1 \big(u(x,t) - v(x,t)\big)^2\,\mathrm{d}x$$

is non-increasing as t increases. We can verify this by differentiating with respect to t and by showing, using integration by parts, that the result found cannot be positive. We have

$$
\begin{aligned}
\frac{\mathrm{d}W}{\mathrm{d}t} &= \int_0^1 \Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial u(x,t)}{\partial t} - \frac{\partial v(x,t)}{\partial t}\right)\,\mathrm{d}x\\
&= \int_0^1 \Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial^2 u(x,t)}{\partial x^2} - \frac{\partial^2 v(x,t)}{\partial x^2}\right)\,\mathrm{d}x\\
&= \left[\Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)\right]_0^1\\
&\qquad\qquad - \int_0^1 \left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)^2\,\mathrm{d}x\\
&= -\int_0^1 \left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)^2\,\mathrm{d}x\\
&\le 0.
\end{aligned}
$$

Even though the approximation of (102a) by (102b) is not exact, it is an advantage of the discretization we have used, that the qualitative property is still present. Let y and z be two solutions to the ordinary differential equation system. Consider the nature of

$$\widehat{W}(t) = \frac{1}{2}\sum_{j=1}^{N}(y_j - z_j)^2.$$

We have

$$\begin{aligned}
\frac{\mathrm{d}\widehat{W}}{\mathrm{d}t} &= \sum_{i=1}^{N}(y_j - z_j)\left(\frac{\mathrm{d}y_j}{dt} - \frac{\mathrm{d}z_j}{\mathrm{d}t}\right) \\
&= (N+1)^2\sum_{j=1}^{N}(y_j - z_j)\left(y_{j-1} - 2y_j + y_{j+1} - z_{j-1} + 2z_j - z_{j+1}\right) \\
&= 2(N+1)^2\sum_{j=1}^{N-1}(y_j - z_j)(y_{j+1} - z_{j+1}) - 2(N+1)^2\sum_{j=1}^{N}(y_j - z_j)^2 \\
&= -(N+1)^2\sum_{j=0}^{N}(y_{j+1} - y_j - z_{j+1} + z_j)^2 \\
&\leq 0.
\end{aligned}$$

Sprectum of discretization

Another aspect of the discretization that might be explored is the spectrum of the matrix A, in comparison with the spectrum of the linear operator $u \mapsto \mathrm{d}^2u/\mathrm{d}x^2$ on the space of C^2 functions on $[0,1]$ for which $u(0) = u(1) = 0$. The eigenfunctions for the continuous problem are of the form $\sin(k\pi x)$, for $k = 1, 2, 3, \ldots$, and the corresponding eigenvalues are $-k^2\pi^2$. For the discrete problem, we need to find the solutions to the problem

$$(A - \lambda I)\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = 0, \tag{102c}$$

where $v_1, v_2, \ldots, v_N$ are not all zero. Introducing also $v_0 = v_{N+1} = 0$, we find that it is possible to write (102c) in the form

$$v_{j-1} - qv_j + v_{j+1} = 0, \qquad j = 1, 2, \ldots, N, \tag{102d}$$

where $q = 2 + \lambda/(N+1)^2$. The difference equation (102d) has solutions of the form

$$v_i = C(\mu^i - \mu^{-i}), \tag{102e}$$

where $\mu + \mu^{-1} = q$, unless $q = \pm 2$ (which is easily seen to be impossible). Because $v_{N+1} = 0$, it follows that $\mu^{2N+2} = 1$ and hence that

$$\mu = \exp\left(\frac{k\pi\mathrm{i}}{N+1}\right), \qquad k = 1, 2, \ldots, N,$$

with $\mathrm{i} = \sqrt{-1}$. Hence,

$$\lambda = -2(N+1)^2\left(1 - \cos\left(\frac{k\pi}{N+1}\right)\right) = -4(N+1)^2 \sin^2\left(\frac{k\pi}{2N+2}\right).$$

Denote this by λ_k.

The eigenvector corresponding to $\lambda = \lambda_k$ is found from (102e), with C chosen so that the eigenvectors are orthonormal. The result is

$$v_i = \sqrt{\frac{2}{N}} \sin\left(\frac{ik\pi}{2N+2}\right).$$

By spectral decomposition

$$A = \sum_{k=1}^{N} \lambda_k v v^{\mathsf{T}}$$

and furthermore

$$\phi(A) = \sum_{k=1}^{N} \phi(\lambda_k) v v^{\mathsf{T}}$$

for ϕ a suitable function. In particular, if $\alpha(t) = \beta(t) = 0$, the solution to (102b) over an interval $[0, h]$ is

$$\exp(Ah) y_0 = \sum_{k=1}^{N} \exp(\lambda_k h) v v^{\mathsf{T}} y_0. \tag{102f}$$

It is interesting to consider the relative contributions of the k terms on the right-hand side of (102f). For k small compared with N, we can use the approximation $\sin(\xi) \approx \xi$, which holds for small ξ, to give eigenvalue number k as $\lambda_k \approx -k^2\pi^2$. On the other hand, for k close to N, $\lambda_k \approx -4(N+1)^2 + (N+1-k)^2\pi^2$. This means that for moderate values of h, the terms in (102f) decay at a slow rate for low values of k but more rapidly for values of k close to N.

In the special case $N = 10$ and $h = 0.1$, we illustrate this effect by observing the behaviour of

$$\exp(Ah) - \sum_{k=1}^{K} \exp(\lambda_k h) v v^{\mathsf{T}}.$$

The element of this matrix with greatest magnitude is 3.8×10^{-3} for $K = 1$ and for $K = 2, 3, 4, 5$ the value reduces rapidly through the values 4.2×10^{-5}, 1.3×10^{-7}, 1.7×10^{-10} and 1.8×10^{-13}. For the remaining values of K, the value is close to zero.

These observations have an important consequence for numerical approximations. To model the behaviour of solutions to this problem, it is the low values of k which are significant. However, in many numerical approximations, the values of k close to N have an overwhelming influence on the stability of the computation. Problems like this are said to be 'stiff' and have an important role in later chapters of this book.

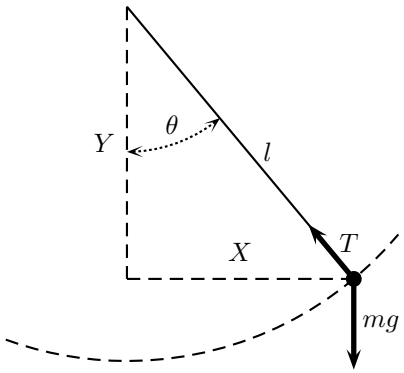

Figure 103(i) Simple pendulum

103 The simple pendulum

Formulation as a differential-algebraic equation

Consider a small mass m attached to a light inelastic string of length l, with the other end attached to the origin of coordinates, which can swing back and forth in a vertical plane. Let X, measured in a rightwards direction, and Y, measured in a downwards direction, be the coordinates. Because the string is inelastic, the tension T in the string always matches other forces resolved in the direction of the string so as to guarantee that the length does not change. The way these forces act on the mass is shown in Figure 103(i). Also shown is the angle θ defined by $X = \sin(\theta)$, $Y = \cos(\theta)$.

We denote by U and V, respectively, the velocity components in the X and Y directions. The motion of the pendulum is governed by the equations

$$\frac{\mathrm{d}X}{\mathrm{d}x} = U, \tag{103a}$$

$$\frac{\mathrm{d}Y}{\mathrm{d}x} = V, \tag{103b}$$

$$m\frac{\mathrm{d}U}{\mathrm{d}x} = -\frac{TX}{l}, \tag{103c}$$

$$m\frac{\mathrm{d}V}{\mathrm{d}x} = -\frac{TY}{l} + mg, \tag{103d}$$

$$X^2 + Y^2 = l^2, \tag{103e}$$

where, in addition to four differential equations (103a)–(103d), the constraint (103e) expresses the constancy of the length of the string. The tension T acts as a control variable, forcing this constraint to remain satisfied. By rescaling variables in a

suitable way, the 'differential-algebraic' equation system (103a)–(103e) can be rewritten with the constants m, g and l replaced by 1 in each case. In the rescaled formulation write $y_1 = X$, $y_2 = Y$, $y_3 = U$, $y_4 = V$ and $y_5 = T$, and we arrive at the system

$$\frac{\mathrm{d}y_1}{\mathrm{d}x} = y_3, \tag{103f}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}x} = y_4, \tag{103g}$$

$$\frac{\mathrm{d}y_3}{\mathrm{d}x} = -y_1 y_5, \tag{103h}$$

$$\frac{\mathrm{d}y_4}{\mathrm{d}x} = -y_2 y_5 + 1, \tag{103i}$$

$$y_1^2 + y_2^2 = 1. \tag{103j}$$

It will be convenient to choose initial values defined in terms of $\theta = \Theta$, with the velocity equal to zero. That is,

$$y_1(0) = \sin(\Theta), \quad y_2(0) = \cos(\Theta), \quad y_3(0) = y_4(0) = 0, \quad y_5(0) = \cos(\Theta).$$

The five variables are governed by four differential equations (103f)–(103i), together with the single algebraic constraint (103j). We will say more about this below, but first we consider the classical way of simplifying the problem.

Formulation as a single second order equation

Make the substitutions $y_1 = \sin(\theta)$, $y_2 = \cos(\theta)$. Because (103j) is automatically satisfied, the value of y_5 loses its interest and we eliminate this by taking a linear combination of (103h) and (103i). This gives the equation system

$$\cos(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}x} = y_3, \tag{103k}$$

$$-\sin(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}x} = y_4, \tag{103l}$$

$$-\cos(\theta)\frac{\mathrm{d}y_3}{\mathrm{d}x} + \sin(\theta)\frac{\mathrm{d}y_4}{\mathrm{d}x} = \sin(\theta). \tag{103m}$$

Differentiate (103k) and (103l) and substitute into (103m) and we obtain the well-known single-equation formulation of the simple pendulum:

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}x^2} + \sin(\theta) = 0, \tag{103n}$$

with initial values

$$\theta(0) = \Theta, \qquad \theta'(0) = 0.$$

It can be shown that the period of the pendulum is given by

$$T = 4\int_0^{\pi/2} \frac{\mathrm{d}\phi}{\sqrt{1 - \sin^2\phi \sin^2\frac{\Theta}{2}}}$$

Table 103(I) Period of simple pendulum for various amplitudes

Θ	T
$0°$	6.2831853072
$3°$	6.2842620831
$6°$	6.2874944421
$9°$	6.2928884880
$12°$	6.3004544311
$15°$	6.3102066431
$18°$	6.3221637356
$21°$	6.3363486630
$24°$	6.3527888501
$27°$	6.3715163462
$30°$	6.3925680085

and some values are given in Table 103(I).

The value for $0°$ can be interpreted as the period for small amplitudes. The fact that T increases slowly as Θ increases is the characteristic property of a simple pendulum which makes it of practical value in measuring time.

Formulation as a Hamiltonian problem

In the formulation (103n), write the H as the 'Hamiltonian'

$$H(p,q) = \frac{1}{2}p^2 - \cos(q),$$

where $q = \theta$ and $p = \mathrm{d}\theta/\mathrm{d}x$. The second order equation (103n) is now equivalent to the first order system

$$\begin{bmatrix} p' \\ q' \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial p} \\ \frac{\partial H}{\partial q} \end{bmatrix}.$$

Differential index and index reduction

Carry out three steps, of which the first is to differentiate (103j) and substitute from (103f) and (103g) to give the result

$$y_1 y_3 + y_2 y_4 = 0. \tag{103o}$$

The second step is to differentiate (103o) and to make various substitutions from (103f)–(103i) to arrive at the equation

$$y_2 + y_3^2 + y_4^2 - y_5 = 0. \tag{103p}$$

The third and final step is to differentiate (103p) and make various substitutions to arrive at the result

$$\frac{\mathrm{d}y_5}{\mathrm{d}x} = \frac{\mathrm{d}y_2}{\mathrm{d}x} + 2y_3\frac{\mathrm{d}y_3}{\mathrm{d}x} + 2y_4\frac{\mathrm{d}y_4}{\mathrm{d}x} = y_4 + 2y_3(-y_1y_5) + 2y_4(-y_2y_5 + 1),$$

which simplifies to

$$\frac{\mathrm{d}y_5}{\mathrm{d}x} = 3y_4. \tag{103q}$$

Given that consistent initial values are used, it seems that the equations (103f)–(103i) together with any of (103j), (103o), (103p) or (103q) give identical solutions.

Which of the possible formulations *should* be used? From the point of view of physical modelling, it seems to be essential to require that the length constraint (103j) should hold exactly. On the other hand, when it comes to numerical approximations to solutions, it is found that the use of this constraint in the problem description creates serious computational difficulties. It also seems desirable from a modelling point of view to insist that (103o) should hold exactly, since this simply states that the direction of motion is tangential to the arc on which it is constrained to lie.

104 A chemical kinetics problem

We next consider a model of a chemical process consisting of three species, which we denote by A, B and C. The three reactions are

$$A \to B, \tag{104a}$$

$$B + C \to A + C, \tag{104b}$$

$$B + B \to B + C. \tag{104c}$$

Let y_1, y_2 and y_3 denote the concentrations of A, B and C, respectively. We assume these are scaled so that the total of the three concentrations is 1, and that each of three constituent reactions will add to the concentration of any of the species exactly at the expense of corresponding amounts of the reactants. The reaction rate of (104a) will be denoted by k_1. This means that the rate at which y_1 decreases, and at which y_2 increases, because of this reaction, will be equal to k_1y_1. In the second reaction (104b), C acts as a catalyst in the production of A from B and the reaction rate will be written as k_2, meaning that the increase of y_1, and the decrease of y_3, in this reaction will have a rate equal to $k_2y_2y_3$. Finally, the production of C from B, (104c), will have a rate constant equal to k_3, meaning that the rate at which this reaction takes place will be $k_3y_2^2$. Putting all these elements of the process together, we find the system of differential equations for the variation with time of the three concentrations to be

$$\frac{\mathrm{d}y_1}{\mathrm{d}x} = -k_1y_1 + k_2y_2y_3, \tag{104d}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}x} = k_1y_1 - k_2y_2y_3 - k_3y_2^2, \tag{104e}$$

$$\frac{\mathrm{d}y_3}{\mathrm{d}x} = k_3y_2^2. \tag{104f}$$

If the three reaction rates are moderately small numbers, and not greatly different in magnitude, then this is a straightforward problem. However, vastly different magnitudes amongst k_1, k_2 and k_3 can make this problem complicated to understand as a chemical model. Also, as we shall see, the problem then becomes difficult to solve numerically. This problem was popularized by Robertson (1966), who used the reaction rates

$$k_1 = 0.04, \qquad k_2 = 10^4, \qquad k_3 = 3 \times 10^7.$$

Before looking at the problem further we note that, even though it is written as a three-dimensional system, it would be a simple matter to rewrite it in two dimensions, because $y_1 + y_2 + y_3$ is an invariant and is usually set to a value of 1, by an appropriate choice of the initial values. We always assume this value for $y_1 + y_2 + y_3$. Furthermore, if the initial value has non-negative values for each of the three components, then this situation is maintained for all positive times. To see why this is the case, write (104d), (104e) and (104f) in the forms

$$\begin{aligned} \frac{\mathrm{d}(\exp(k_1 x) y_1)}{\mathrm{d}x} &= \exp(k_1 x) k_2 y_2 y_3, \\ \frac{\mathrm{d}\big(\exp\big(\max(k_2, k_3)x\big) y_2\big)}{\mathrm{d}x} &= \exp\big(\max(k_2, k_3)x\big) F, \\ \frac{\mathrm{d}y_3}{\mathrm{d}x} &= k_3 y_2^2, \end{aligned}$$

where

$$\begin{aligned} F = k_1 y_1 + &\max(k_2, k_3) y_1 y_2 \\ &+ \big(\max(k_2, k_3) - k_2\big) y_2 y_3 + \big(\max(k_2, k_3) - k_3\big) y_2^2, \end{aligned}$$

so that each of $\exp(k_1 x) y_1$, $\exp(\max(k_2, k_3)x) y_2$ and y_3 is non-decreasing.

An interesting feature of this problem is that a small perturbation, that does not disturb the invariance of $y_1 + y_2 + y_3$, is damped out rapidly. To see why this is the case, eliminate y_1 so that the differential equation system in the remaining two components becomes

$$\frac{\mathrm{d}y_2}{\mathrm{d}x} = k_1(1 - y_2 - y_3) - k_2 y_2 y_3 - k_3 y_2^2, \tag{104g}$$

$$\frac{\mathrm{d}y_3}{\mathrm{d}x} = k_3 y_2^2. \tag{104h}$$

The Jacobian matrix, the matrix of partial derivatives, is given by

$$J(x) = \begin{bmatrix} -k_1 - k_2 y_3 - 2k_3 y_2 & -k_1 - k_2 y_2 \\ 2k_3 y_2 & 0 \end{bmatrix},$$

and the characteristic polynomial is

$$\lambda^2 + (k_1 + k_2 y_3 + 2k_3 y_2)\lambda + 2k_3 y_2 (k_1 + k_2 y_2). \tag{104i}$$

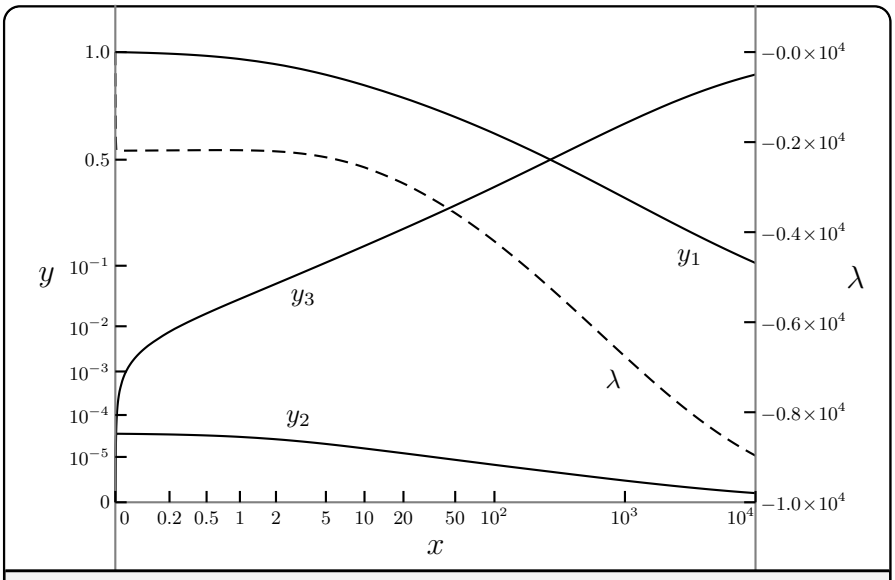

Figure 104(i) Solution and most negative eigenvalue for the Robertson problem

An analysis of the discriminant of (104i) indicates that for $y_2, y_3 \in (0, 1]$, both zeros are real and negative. Along the actual trajectory, one of the eigenvalues of $J(x)$, denoted by λ, rapidly jumps to a very negative value, with the second eigenvalue retaining a small negative value. Consider a small perturbation z to the solution, so that the solution becomes $y + z$. Because the two components of z are small we can approximate $f(y + z)$ by $f(y) + (\partial f/\partial y)z$. Hence, the perturbation itself satisfies the equation

$$\begin{bmatrix} \dfrac{\mathrm{d}z_2}{\mathrm{d}x} \\ \dfrac{\mathrm{d}z_3}{\mathrm{d}x} \end{bmatrix} = J(x) \begin{bmatrix} z_2 \\ z_3 \end{bmatrix}$$

and the negative eigenvalues of $J(x)$ guarantee the decay of the components of z.

The solution to this problem, together with the value of λ, is shown in Figure 104(i).

105 The Van der Pol equation and limit cycles

The simple pendulum, which we considered in Subsection 103, is a non-linear variant of the 'harmonic oscillator' problem $y'' = -y$. We now consider another non-linear generalization of this problem, by adding a term $\mu(1 - y^2)y'$, where μ is a positive constant, to obtain the 'Van der Pol equation'

$$y''(x) = \mu(1 - y(x)^2)y'(x) - y(x).$$

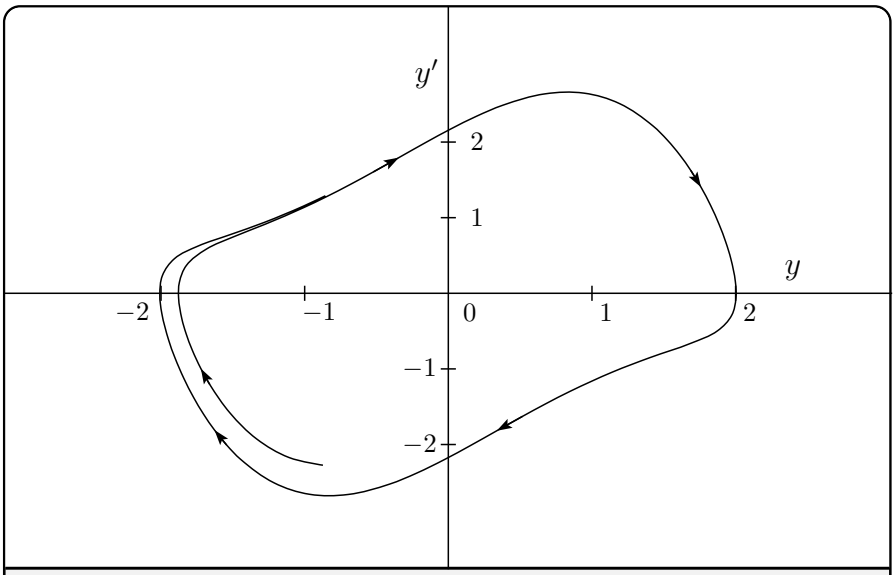

Figure 105(i) Van der Pol problem with $\mu = 1$

This problem was originally introduced by Van der Pol (1926) in the study of electronic circuits. If μ is small and the initial values correspond to what would be oscillations of amplitude less than 1, if μ had in fact been zero, it might be expected that the values of $y(x)$ would remain small for all time. However, the non-linear term has the effect of injecting more 'energy' into the system, as we see by calculating the rate of change of $E = \frac{1}{2}y'(x)^2 + \frac{1}{2}y(x)^2$. This is found to be

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\tfrac{1}{2}y'(x)^2 + \tfrac{1}{2}y(x)^2\right) = \mu(1 - y(x)^2)y'(x)^2 > 0,$$

as long as $|y| < 1$.

Similarly, if $|y|$ starts with a high value, then E will decrease until $|y| = 1$. It is possible to show that the path, traced out in the (y, y') plane, loops round the origin in a clockwise direction forever, and that it converges to a 'limit cycle' – a periodic orbit. In Figure 105(i), this is illustrated for $\mu = 1$. The path traced out in the (y, y') plane moves rapidly towards the limit cycle and is soon imperceptibly close to it. In Figure 105(ii), the case $\mu = 3$ is presented.

Of special interest in this problem, especially for large values of μ, is the fact that numerical methods attempting to solve this problem need to adjust their behaviour to take account of varying conditions, as the value of $1 - |y(x)|^2$ changes. The sharp change of direction of the path traced out near $(y, y') = (\pm 2, 0)$ for the $\mu = 3$ case, a phenomenon which becomes more pronounced as μ is further increased, is part of the numerical difficulty associated with this problem.

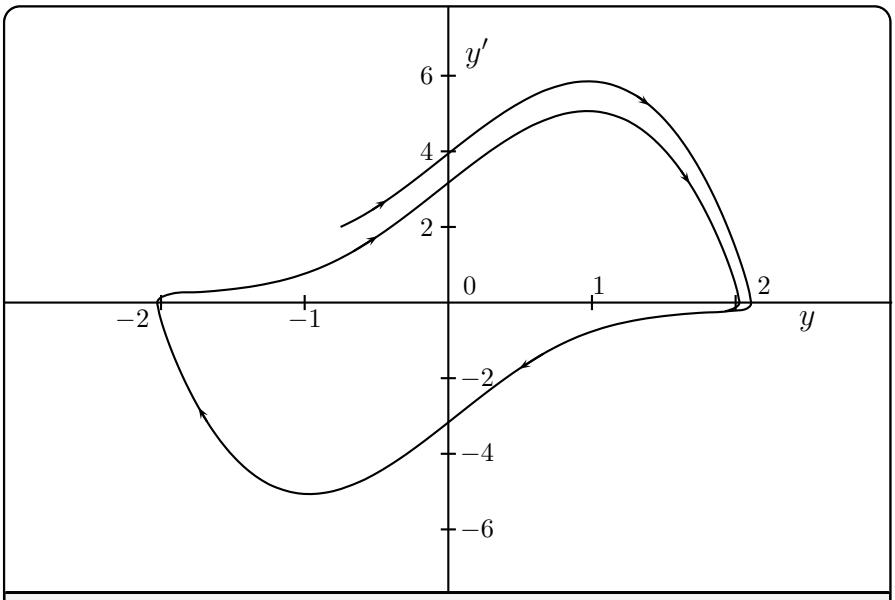

Figure 105(ii) Van der Pol problem with $\mu = 3$

106 The Lotka–Volterra problem and periodic orbits

In the modelling of the two-species 'predator–prey' problem, differential equation systems of the following type arise:

$$u' = u(2 - v), \tag{106a}$$

$$v' = v(u - 1), \tag{106b}$$

where the factors $2 - v$ and $u - 1$ can be generalized in various ways. This model was proposed independently by Lotka (1925) and Volterra (1926).

The two variables represent the time-dependent populations, of which v is the population of predators that feed on prey whose population is denoted by u. It is assumed that u would have been able to grow exponentially without limit, if the predator had not been present, and that the factor $2 - v$ represents the modification to its growth rate because of harvesting by the predator. The predator in turn, in the absence of prey, would die out exponentially, and requires at least a prey population of $u = 1$ to feed upon to be able to grow. Of the two stationary solutions, $(u, v) = (0, 0)$ and $(u, v) = (1, 2)$, the second is more interesting because small perturbations from this point will lead to periodic orbits around the stationary point. By dividing (106a) by (106b), we obtain a differential equation for the path traced out by (u, v). The solution is that $I(u, v)$ is constant, where

$$I(u, v) = \log(u) + 2\log(v) - u - v.$$

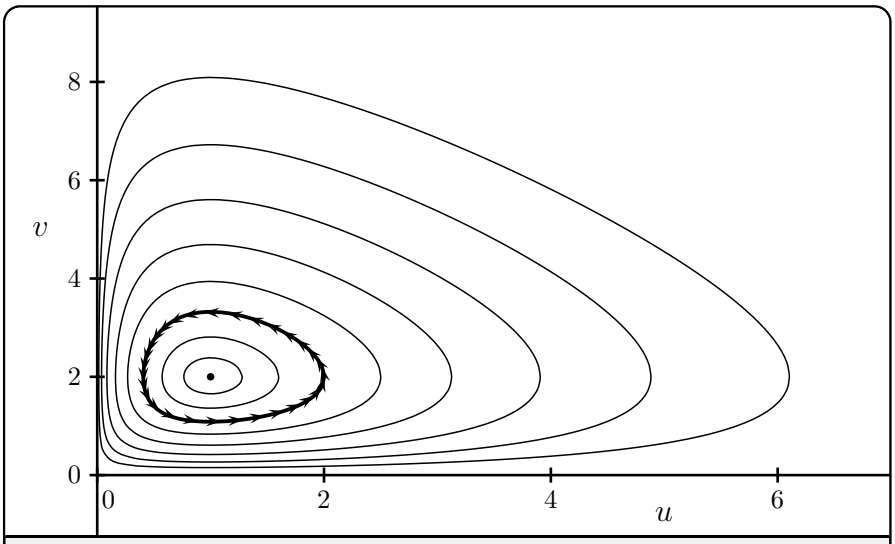

Figure 106(i) Phase diagram for Lotka–Volterra solution with $(u_0, v_0) = (2, 2)$, together with seven alternative orbits

It is interesting to try to calculate values of the period T, for a given starting point (u_0, v_0). To calculate T, change to polar coordinates centred at the stationary point

$$u = 1 + r\cos(\theta), \qquad v = 2 + r\sin(\theta)$$

and calculate the integral $\int_0^{2\pi} \phi(\theta)\mathrm{d}\theta$, where

$$\phi(\theta) = \frac{1}{v\cos^2(\theta) + u\sin^2(\theta)}. \tag{106c}$$

Starting values $(u_0, v_0) = (2, 2)$ lead to the orbit featured in Figure 106(i). Orbits with various other starting values are also shown.

The period, based on the integral of (106c), has been calculated with a varying number n of equally spaced values of $\theta \in [0, 2\pi]$, using the trapezoidal rule. It is known that for certain smooth functions, the error of this type of calculation will behave, not like a power of n^{-1}, but like $\exp(-\alpha n)$, for some problem-specific parameter α. This super-convergence is evidently realized for the present problem, where the observed approximations

$$T = \int_0^{2\pi} \phi(\theta)\mathrm{d}\theta \approx \frac{2\pi}{n}\sum_{k=0}^{n-1} \phi\left(\frac{2\pi k}{n}\right) \tag{106d}$$

are shown in Table 106(I) for $n = 10, 20, 40, \ldots, 320$. Evidently, to full machine accuracy, the approximations have converged to $T = 4.61487051945103$. An explanation of the phenomenon of rapid convergence of the trapezoidal rule for

```
1  function [period, points] = lotkavolterra(n , u0 , v0)
2     theta = linspace(0 ,2*pi , n+1) ;
3     co = cos(theta) ;
4     si = sin(theta) ;
5     C = u0 * v0^2*exp(-u0-v0) ;
6     r = ones(size(theta)) ;
7     u = 1 + r .* co ;
8     v = 2 + r .* si ;
9     carryon = 1 ;
10    while carryon
11       f = u .* v .^ 2 - C * exp(u + v) ;
12       df = -v .* r .* (v .* co .^ 2 + u .* si .^ 2) ;
13       dr = f ./df ;
14       r = r - dr ;
15       u = 1 + r .* co ;
16       v = 2 + r .* si ;
17       carryon = norm(dr , inf) > 1e - 9 ;
18    end
19    phi = 1 ./(v .* co .^ 2 + u .* si .^ 2) ;
20    period = (2 * pi/n) * sum(phi(1 : n)) ;
21    points = [u', v'] ;
22 end
```

Algorithm 106α Computation of orbit and period for the Lotka–Volterra problem

periodic functions can be found in Davis and Rabinowitz (1984) and in papers referenced in that book.

In Algorithm 106α, statements are presented to carry out the computations to generate Figure 106(i) and Table 106(I). To compute the value of r for each θ, the equation $f(r) = 0$ is solved, where

$$f(r) = (\exp(I(u, v)) - C)\exp(u + v) = uv^2 - C\exp(u + v),$$

with $C = u_0 v_0^2 \exp(-u_0 - v_0)$. Note that the statement in line 11 evaluates a vector with element number i equal to $u_i v_i^2 - C\exp(u_i + v_i)$, and that the statement in line 2 generates a vector with $n + 1$ components, equally spaced in $[0, 2\pi]$.

107 The Euler equations of rigid body rotation

For a rigid body on which no moments are acting, the three components of angular velocity, in terms of the principal directions of inertia fixed in the body, satisfy the

Table 106(I) Approximations to the period T, given by (106d) for $(u_0, v_0) = (2, 2)$

n	Approximate integral
10	4.62974838287860
20	4.61430252126987
40	4.61487057379480
80	4.61487051945097
160	4.61487051945103
320	4.61487051945103

Euler equations:

$$\begin{aligned} I_1 \frac{\mathrm{d}w_1}{\mathrm{d}t} &= (I_2 - I_3) w_2 w_3, \\ I_2 \frac{\mathrm{d}w_2}{\mathrm{d}t} &= (I_3 - I_1) w_3 w_1, \\ I_3 \frac{\mathrm{d}w_3}{\mathrm{d}t} &= (I_1 - I_2) w_1 w_2, \end{aligned} \tag{107a}$$

where the 'principal moments of inertia' I_1, I_2 and I_3 are positive. Denote the kinetic energy by $\frac{1}{2}E$ and the squared norm of the angular momentum by F. That is,

$$\begin{aligned} E &= I_1 w_1^2 + I_2 w_2^2 + I_3 w_3^2, \\ F &= I_1^2 w_1^2 + I_2^2 w_2^2 + I_3^2 w_3^2. \end{aligned}$$

Differentiate these expressions and substitute the expressions for $\mathrm{d}w_i/\mathrm{d}t$, $i = 1, 2, 3$, to obtain a zero result in each case. Hence, E and F are invariants of the solution to (107a). This observation provides useful tests on numerical methods for this problem because there is in general no reason why these invariants should be maintained in a numerical approximation.

Exercises 10

10.1 You are given the initial value problem

$$u'''(x) - 3u''(x) + 2u(x)u'(x) = 0, \quad u(1) = 2, \;\; u'(1) = -1, \;\; u''(1) = 4.$$

Show how to reformulate this problem in the form

$$y'(x) = f(y(x)), \qquad y(x_0) = y_0,$$

where $f : \mathbb{R}^3 \to \mathbb{R}^3$.

10.2 The matrix

$$A = (N-1)^2 \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$$

arises in the numerical solution of the heat equation, but with Neumann boundary conditions. Find the eigenvalues of A.

10.3 Calculate the period of an orbit of the Lotka–Volterra problem which passes through the point $(3, 2)$.

10.4 You are given the non-autonomous initial value problem

$$\begin{aligned} u' &= xu + x^2 v, & u(0) &= 3 \\ v' &= u - v + 2xw, & v(0) &= 2 \\ w' &= u + \frac{v}{1+x}, & w(0) &= 5. \end{aligned}$$

Show how to write this as an autonomous problem.

11 Differential Equation Theory

110 Existence and uniqueness of solutions

A fundamental question that arises in scientific modelling is whether a given differential equation, together with initial conditions, can be reliably used to predict the behaviour of the trajectory at later times. We loosely use the expression 'well-posed' to describe a problem that is acceptable from this point of view. The three attributes of an initial value problem that have to be taken into account are whether there actually exists a solution, whether the solution, if it exists, is unique, and how sensitive the solution is to small perturbations to the initial information. Even though there are many alternative criteria for answering these questions in a satisfactory manner, we focus here on the existence of a Lipschitz condition. This is especially convenient because the same type of condition can be used to study the behaviour of numerical approximations.

Definition 110A *The function $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is said to satisfy a 'Lipschitz condition in its second variable' if there exists a constant L, known as a 'Lipschitz constant', such that for any $x \in [a, b]$ and $Y, Z \in \mathbb{R}^N$, $\|f(x, Y) - f(x, Z)\| \leq L\|Y - Z\|$.*

We need a basic lemma on metric spaces known as the 'contraction mapping principle'. We present this without proof.

Lemma 110B *Let M denote a complete metric space with metric ρ and let $\phi : M \to M$ denote a mapping which is a contraction, in the sense that there exists a number k, satisfying $0 \leq k < 1$, such that, for any $\eta, \zeta \in M$, $\rho(\phi(\eta), \phi(\zeta)) \leq k\rho(\eta, \zeta)$. Then there exists a unique $\xi \in M$ such that $\phi(\xi) = \xi$.*

We can now state our main result.

Theorem 110C *Consider an initial value problem*

$$y'(x) = f(x, y(x)), \tag{110a}$$

$$y(a) = y_0, \tag{110b}$$

where $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is continuous in its first variable and satisfies a Lipschitz condition in its second variable. Then there exists a unique solution to this problem.

Proof. Let M denote the complete metric space of continuous functions $y : [a, b] \to \mathbb{R}^N$, such that $y(a) = y_0$. The metric is defined by

$$\rho(y, z) = \sup_{x \in [a,b]} \exp(-K(x - a))\|y(x) - z(x)\|,$$

where $K > L$. For given $y \in M$, define $\phi(y)$ as the solution Y on $[a, b]$ to the initial value problem

$$y'(x) = f(x, Y(x)),$$
$$Y(a) = y_0.$$

This problem is solvable by integration as

$$\phi(y)(x) = y_0 + \int_a^x f(s, y(s))ds.$$

This is a contraction because for any two $y, z \in M$, we have

$$\begin{aligned}\rho(\phi(y), \phi(z)) &\leq \sup_{x \in [a,b]} \exp(-K(x - a)) \left\| \int_a^x \big(f(s, y(s)) - f(s, z(s))\big)ds \right\| \\ &\leq \sup_{x \in [a,b]} \exp(-K(x - a)) \int_a^x \|f(s, y(s)) - f(s, z(s))\|\, ds\end{aligned}$$

$$\begin{aligned}
&\leq L \sup_{x\in[a,b]} \exp(-K(x-a)) \int_a^x \|y(s)-z(s)\|\,ds \\
&\leq L\rho(y,z) \sup_{x\in[a,b]} \exp(-K(x-a)) \int_a^x \exp(K(s-a))ds \\
&\leq \frac{L}{K}\rho(y,z).
\end{aligned}$$

The unique function y that therefore exists satisfying $\phi(y) = y$, is evidently the unique solution to the initial value problem given by (110a), (110b). □

The third requirement for being well-posed, that the solution is not overly sensitive to the initial condition, can be readily assessed for problems satisfying a Lipschitz condition. If y and z each satisfy (110a) with $y(a) = y_0$ and $z(a) = z_0$, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\|y(x)-z(x)\| \leq L\|y(x)-z(x)\|.$$

Multiply both sides by $\exp(-Lx)$ and deduce that

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(\exp(-Lx)\|y(x)-z(x)\|\big) \leq 0,$$

implying that

$$\|y(x)-z(x)\| \leq \|y_0-z_0\|\exp\big(L(x-a)\big). \tag{110c}$$

This bound on the growth of initial perturbations may be too pessimistic in particular circumstances. Sometimes it can be improved upon by the use of the 'one-sided Lipschitz condition'. This will be discussed in Subsection 112.

Local Lipschitz condition

Definition 110A is too restrictive to apply to many important practical problems. We can obtain a weaker version of Theorem 110C if we assume a local version of the Lipschitz condition.

Definition 110D *The function $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ is said to satisfy a 'local Lipschitz condition in its second variable' if for each sphere S in $\mathbb{R}^N$, there exists a constant $L(S)$, known as a 'local Lipschitz constant', such that for any $x \in [a,b]$ and $Y, Z \in S$, $\|f(x,Y) - f(x,Z)\| \leq L\|Y - Z\|$.*

111 Linear systems of differential equations

Linear differential equations are important because of the availability of a superposition principle. That is, it is possible for a linear differential equation system to combine known solutions to construct new solutions. The standard form of a linear system is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = A(x)y + \phi(x), \tag{111a}$$

where $A(x)$ is a possibly time-dependent linear operator. The corresponding 'homogeneous' system is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = A(x)y. \tag{111b}$$

The superposition principle, which is trivial to verify, states that:

Theorem 111A *If $\widehat{y}$ is a solution to* (111a) *and $y_1, y_2, \ldots, y_k$ are solutions to* (111b)*, then for any constants $\alpha_1, \alpha_2, \ldots, \alpha_k$, the function y given by*

$$y(x) = \widehat{y}(x) + \sum_{i=1}^{k} \alpha_i y_i(x),$$

is a solution to (111a)*.*

The way this result is used is to attempt to find the solution that matches a given initial value, by combining known solutions.

Many linear problems are naturally formulated in the form of a single high order differential equation

$$Y^{(m)}(x) - C_1(x)Y^{(m-1)}(x) - C_2(x)Y^{(m-2)}(x) - \cdots - C_m(x)Y(x) = g(x). \tag{111c}$$

By identifying $Y(x) = y_1(x), Y'(x) = y_2(x), \ldots, Y^{(m-1)} = y_m(x)$, we can rewrite the system in the form

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{bmatrix} = A(x)\begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{bmatrix} + \phi(x),$$

where the 'companion matrix' $A(x)$ and the 'inhomogeneous term' $\phi(x)$ are given by

$$A(x) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ C_m(x) & C_{m-1}(x) & C_{m-2}(x) & \cdots & C_1(x) \end{bmatrix}, \quad \phi(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ g(x) \end{bmatrix}.$$

When $A(x) = A$ in (111b) is constant, then to each eigenvalue λ of A, with corresponding eigenvector v, there exists a solution given by

$$y(x) = \exp(\lambda x)v. \tag{111d}$$

When a complete set of eigenvectors does not exist, but corresponding to λ there is a chain of generalized eigenvectors

$$Av_1 = \lambda v_1 + v, \quad Av_2 = \lambda v_2 + v_1, \quad \ldots, \quad Av_{k-1} = \lambda v_{k-1} + v_{k-2},$$

then there is a chain of additional independent solutions to append to (111d):

$$y_1 = x\exp(\lambda x)v_1, \quad y_2 = x^2\exp(\lambda x)v_2, \quad \ldots, \quad y_{k-1} = x^{k-1}\exp(\lambda x)v_{k-1}.$$

In the special case in which A is a companion matrix, so that the system is equivalent to a high order equation in a single variable, as in (111c), with $C_1(x) = C_1, C_2(x) = C_2, \ldots, C_m(x) = C_m$, each a constant, the characteristic polynomial of A is

$$P(\lambda) = \lambda^m - C_1\lambda^{m-1} - C_2\lambda^{m-2} - \cdots - C_m = 0.$$

For this special case, $P(\lambda)$ is also the *minimal* polynomial, and repeated zeros *always* correspond to incomplete eigenvector spaces and the need to use generalized eigenvectors. Also, in this special case, the eigenvector corresponding to λ, together with the generalized eigenvectors if they exist, are

$$v = \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{bmatrix}, \quad v_1 = \begin{bmatrix} 0 \\ 1 \\ 2\lambda \\ \vdots \\ (m-1)\lambda^{m-2} \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ \frac{(m-1)(m-2)}{2}\lambda^{m-3} \end{bmatrix}, \quad \ldots.$$

112 Stiff differential equations

Many differential equation systems of practical importance in scientific modelling exhibit a distressing behaviour when solved by classical numerical methods. This behaviour is distressing because these systems are characterized by very high stability, which can turn into very high *instability* when approximated by standard numerical methods. We have already seen examples of stiff problems, in Subsections 102 and 104, and of course there are many more such examples. The concept of the 'one-sided Lipschitz condition' was mentioned in Subsection 110 without any explanation. Stiff problems typically have large Lipschitz constants, but many have more manageable one-sided Lipschitz constants, and this can be an aid in obtaining realistic growth estimates for the effect of perturbations.

We confine ourselves to problems posed on an inner product space. Thus we assume that there exists an inner product on $\mathbb{R}^N$ denoted by $\langle u, v\rangle$, and that the norm is defined by $\|u\|^2 = \langle u, u\rangle$.

Definition 112A *The function f satisfies a 'one-sided Lipschitz condition', with 'one-sided Lipschitz constant' l if for all $x \in [a, b]$ and all $u, v \in \mathbb{R}^N$,*

$$\langle f(x, u) - f(x, v), u - v\rangle \leq l\|u - v\|^2.$$

It is possible that the function f could have a very large Lipschitz constant but a moderately sized, or even negative, one-sided Lipschitz constant. The advantage of this is seen in the following result.

Theorem 112B *If f satisfies a one-sided Lipschitz condition with constant l, and y and z are each solutions of*

$$y'(x) = f(x, y(x)),$$

then for all $x \geq x_0$,

$$\|y(x) - z(x)\| \leq \exp(l(x - x_0))\|y(x_0) - z(x_0)\|.$$

Proof. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\|y(x) - z(x)\|^2 &= \frac{\mathrm{d}}{\mathrm{d}x}\langle y(x) - z(x), y(x) - z(x)\rangle \\ &= 2\langle f(x, y(x)) - f(x, z(x)), y(x) - z(x)\rangle \\ &\leq 2l\|y(x) - z(x)\|^2.\end{aligned}$$

Multiply by $\exp\big(-2l(x - x_0)\big)$ and it follows that

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(\exp\big(-2l(x - x_0)\big)\|y(x) - z(x)\|^2\big) \leq 0,$$

so that $\exp\big(-2l(x - x_0)\big)\|y(x) - z(x)\|^2$ is non-increasing. □

Note that the problem described in Subsection 102 possesses the one-sided Lipschitz condition with $l = 0$.

Even though stiff differential equation systems are typically non-linear, there is a natural way in which a linear system arises from a given non-linear system. Since stiffness is associated with the behaviour of perturbations to a given solution, we suppose that there is a small perturbation $\epsilon Y(x)$ to a solution $y(x)$. The parameter ϵ is small, in the sense that we are interested only in asymptotic behaviour of the perturbed solution as this quantity approaches zero. If $y(x)$ is replaced by $y(x) + \epsilon Y(x)$ in the differential equation

$$y'(x) = f(x, y(x)), \tag{112a}$$

and the solution expanded in a series in powers of ϵ, with ϵ^2 and higher powers replaced by zero, we obtain the system

$$y'(x) + \epsilon Y'(x) = f(x, y(x)) + \epsilon\frac{\partial f}{\partial y}Y(x). \tag{112b}$$

Subtract (112a) from (112b) and cancel out ϵ, and we arrive at the equation governing the behaviour of the perturbation,

$$Y'(x) = \frac{\partial f}{\partial y}Y(x) = J(x)Y(x),$$

say. The 'Jacobian matrix' $J(x)$ has a crucial role in the understanding of problems of this type; in fact its spectrum is sometimes used to characterize stiffness. In a time interval Δx, chosen so that there is a moderate change in the value of the solution to (112a), and very little change in $J(x)$, the eigenvalues of $J(x)$ determine the growth rate of components of the perturbation. The existence of one or more large and negative values of $\lambda\Delta x$, for $\lambda \in \sigma(J(x))$, the spectrum of $J(x)$, is a sign that stiffness is almost certainly present. If $J(x)$ possesses complex eigenvalues, then we interpret this test for stiffness as the existence of a $\lambda = \mathrm{Re}\lambda + \mathrm{i}\mathrm{Im}\lambda \in \sigma(J(x))$ such that $\mathrm{Re}\lambda\Delta x$ is negative with large magnitude.

Exercises 11

11.1 Show how to modify Theorem 110C so that the Lipschitz condition holds only in a neighbourhood of y_0 and where the solution is only required to exist on $[a, \widetilde{b}]$, where $\widetilde{b}$ satisfies $a < \widetilde{b} \leq b$.

11.2 By finding two vectors α and β so that the system

$$y'(x) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} y(x) + \begin{bmatrix} \sin(x) \\ 0 \\ \cos(x) \end{bmatrix},$$

has a solution of the form $\widehat{y}(x) = \sin(x)\alpha + \cos(x)\beta$, find the general solution to this problem.

12 Further Evolutionary Problems

120 Many-body gravitational problems

We consider a more general gravitational problem involving n mutually attracting masses $M_1, M_2, \ldots, M_n$ at position vectors $y_1(x), y_2(x), \ldots, y_n(x)$, satisfying the $3n$-dimensional second order differential equation system

$$y_i''(x) = -\sum_{j \neq i} \frac{\gamma M_j (y_i - y_j)}{\|y_i - y_j\|^3}, \qquad i = 1, 2, \ldots, n.$$

Reformulated as a first order system, the problem is $6n$-dimensional because each of the y_i has three components and the velocity vectors y_i' also have three components.

To reduce this problem to a manageable level in situations of practical interest, some simplifications can be made. For example, in models of the solar system, the most massive planets, Jupiter, Uranus, Neptune and Saturn, are typically regarded as the only bodies capable of influencing the motion of the sun and of each other. The four small planets closest to the sun, Mercury, Venus, Earth and Mars, are, in this model, regarded as part of the sun in the sense that they add to its mass in attracting the heavy outer planets towards the centre of the solar system. To study the motion of the small planets or of asteroids, they can be regarded as massless particles, moving in the gravitation fields of the sun and the four large planets, but not at the same time influencing their motion.

The restricted three–body problem

Another model, involving only three bodies, is useful for studying the motion of an Earth–Moon satellite or of an asteroid close enough to the Earth to be strongly influenced by it as well as by the Sun. This system, known as the restricted three–body problem, regards the two heavy bodies as revolving in fixed orbits about their common centre of mass and the small body as attracted by the two larger bodies but not affecting their motion in any way. If it is possible to approximate the large-body orbits as circles, then a further simplification can be made by working in a frame of reference that moves with them. Thus, we would regard the two large bodies as being fixed in space with their rotation in the original frame of reference translated into a modification of the equations of gravitational motion.

To simplify this discussion, we use units scaled to reduce a number of constants to unit value. We scale the masses of the two larger bodies to $1-\mu$ and μ and their positions relative to the moving reference frame by the vectors $(\mu-1)e_1$ and μe_1, so that their centre of mass is at the origin of coordinates. Write y_1, y_2 and y_3 as the scalar variables representing the position coordinates of the small body and y_4, y_5 and y_6 as the corresponding velocity coordinates. Under these assumptions, the equations of motion become

$$
\begin{aligned}
y_1' &= y_4,\\
y_2' &= y_5,\\
y_3' &= y_6,\\
y_4' &= 2y_5 + y_1 - \frac{\mu(y_1+\mu-1)}{\left(y_2^2+y_3^2+(y_1+\mu-1)^2\right)^{3/2}} - \frac{(1-\mu)(y_1+\mu)}{\left(y_2^2+y_3^2+(y_1+\mu)^2\right)^{3/2}},\\
y_5' &= -2y_4 + y_2 - \frac{\mu y_2}{\left(y_2^2+y_3^2+(y_1+\mu-1)^2\right)^{3/2}} - \frac{(1-\mu)y_2}{\left(y_2^2+y_3^2+(y_1+\mu)^2\right)^{3/2}},\\
y_6' &= -\frac{\mu y_3}{\left(y_2^2+y_3^2+(y_1+\mu-1)^2\right)^{3/2}} - \frac{(1-\mu)y_3}{\left(y_2^2+y_3^2+(y_1+\mu)^2\right)^{3/2}}.
\end{aligned}
$$

Planar motion is possible; that is, solutions which satisfy $y_3 = y_6 = 0$ at all times. One of these is shown in Figure 120(i), with the values of (y_1, y_2) plotted as the orbit evolves. The heavier mass is at the point $(\mu, 0)$ and the lighter mass is at $(1-\mu, 0)$, where $(0,0)$ is marked 0 and $(1,0)$ is marked 1. For this calculation the value of $\mu = 1/81.45$ was selected, corresponding to the Earth–Moon system. The initial values for this computation were $(y_1, y_2, y_3, y_4, y_5, y_6) = (0.994, 0, 0, 0, -2.0015851063790825224, 0)$ and the period was 17.06521656015796.

A second solution, identical except for the initial value $(y_1, y_2, y_3, y_4, y_5, y_6) = (0.87978, 0, 0, 0, -0.3797, 0)$ and a period 19.14045706162071, is shown in Figure 120(ii).

Figure eight orbit

If the three masses are comparable in value, then the restriction to a simpler system that we have considered is not available. However, in the case of a number of equal masses, other symmetries are possible. We consider just a single example, in which three equal, mutually attracting masses move in a figure eight orbit. This is shown in Figure 120(iii).

121 Delay problems and discontinuous solutions

A functional differential equation is one in which the rate of change of $y(x)$ depends not just on the values of y for the same time value, but also on time values less than x. In the simplest case, this has the form

$$y'(x) = f(x, y(x), y(x - \tau)), \tag{121a}$$

where τ is a constant delay. Note that this cannot be cast as an initial value problem with the hope of actually defining a unique solution, because at an initial point x_0, the derivative depends on the value of $y(x_0 - \tau)$. What we will need to do in the case of (121a) is to specify the value of y on an initial interval $[x_0 - \tau, x_0]$.

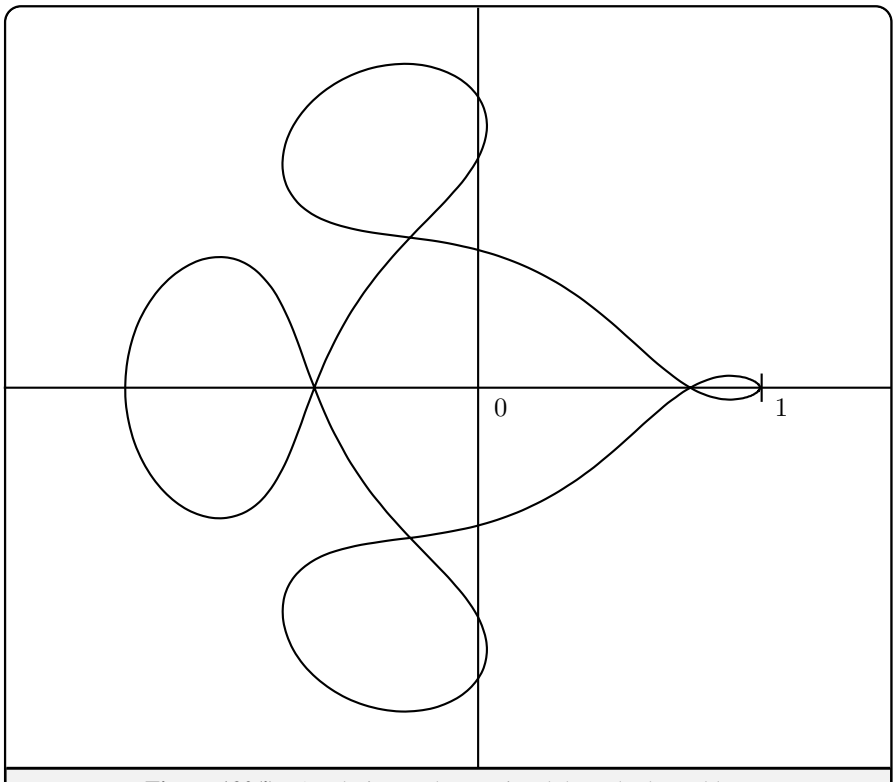

Figure 120(i) A solution to the restricted three–body problem

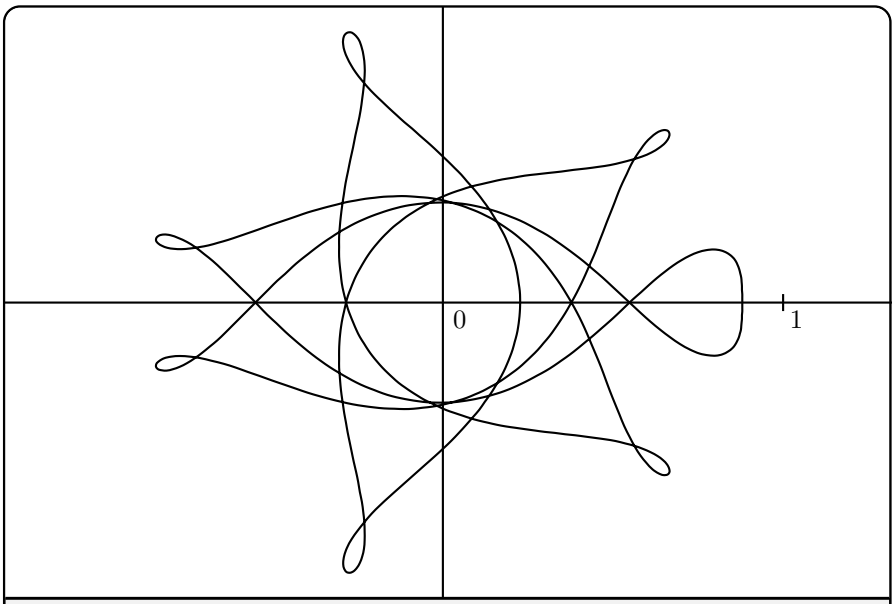

Figure 120(ii) A second solution to the restricted three–body problem

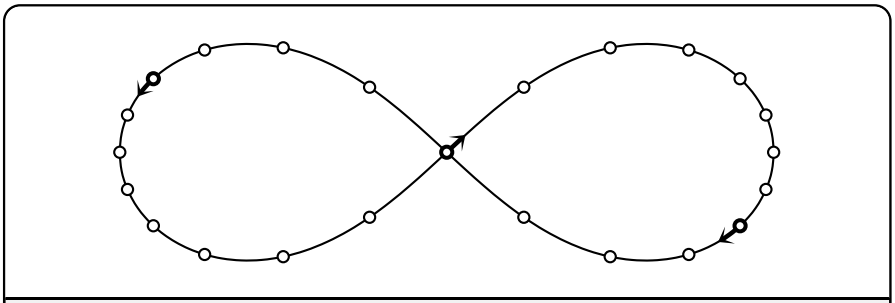

Figure 120(iii) A figure eight orbit for three equal masses

A linear delay differential equation

We consider the problem given by

$$y'(x) = -y(x - \tfrac{\pi}{2}), \quad x > 0, \qquad y(x) = x, \quad x \in [-\tfrac{\pi}{2}, 0]. \tag{121b}$$

For x in the interval $[0, \frac{\pi}{2}]$ we find

$$y(x) = -\int_0^x (x - \tfrac{\pi}{2})\, \mathrm{d}x = \tfrac{1}{2}x(\pi - x),$$

with $y(\frac{\pi}{2}) = \frac{1}{8}\pi^2$. This process can be repeated over the sequence of intervals $[\frac{\pi}{2}, \pi]$, $[\pi, \frac{3\pi}{2}], \ldots$ to obtain values of $y(x)$ shown in Figure 121(i) for $x \leq 4\pi$.

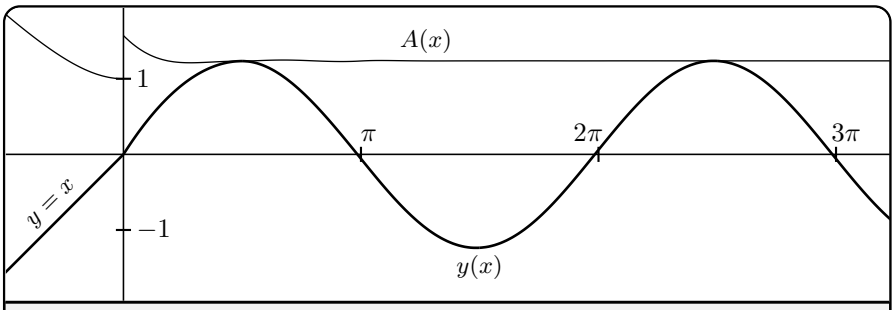

Figure 121(i) Solution to delay differential equation (121b)

It appears that the solution is attempting to approximate sinusoidal behaviour as time increases. We can verify this by estimating a local amplitude defined by

$$A(x) = (y(x)^2 + y'(x)^2)^{\frac{1}{2}}.$$

This function is also shown in Figure 121(i) and we note the discontinuity at $x = 0$, corresponding to the discontinuity in the value of $y'(x)$. Such discontinuities are to be expected because the right-derivative is given by the formula for $y'(x)$ for x positive and the left-derivative is found from the derivative of the initial function. For each positive integral multiple of $\frac{1}{2}\pi$, there will always be an inherited non-smooth behaviour but this will be represented by a discontinuity in increasingly higher derivatives.

We will now consider a problem with two delays.

An example with persistent discontinuities

A delay differential equation of 'neutral type' is one in which delayed values of y' also occur in the formulation. An example of this type of problem is

$$y'(x) = \tfrac{1}{2}y'(x-1) + ay(x-\sqrt{2}), \qquad x > 0, \tag{121c}$$

$$y(x) = 1, \qquad x \in [-\sqrt{2}, 0], \tag{121d}$$

where the constant is given by $a = \exp(\sqrt{2}) - \frac{1}{2}\exp(\sqrt{2}-1)$ and was contrived to ensure that $\exp(x)$ would have been a solution, if the initial information had been defined in terms of that function.

The solution is shown in Figure 121(ii) and we see that it seems to be approximating exponential behaviour more and more closely as x increases. However, there is a discontinuity in $y'(x)$ at every positive integer value of x. Specifically, for each n there is a jump given by

$$\lim_{x \to n+} y'(x) - \lim_{x \to n-} y'(x) = 2^{-n}a.$$

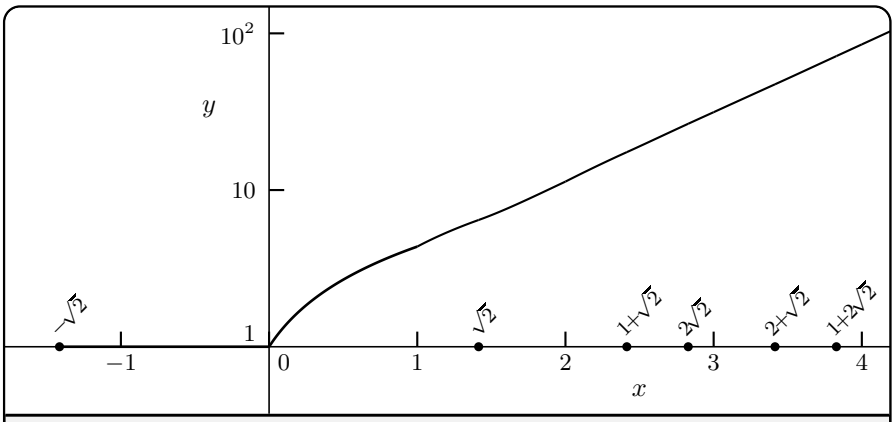

Figure 121(ii) Solution to neutral delay differential equation (121c)

122 Problems evolving on a sphere

Given a function $H(y)$, we will explore situations in which solutions to $y'(x) = f(y)$ preserve the value of $H(y(x))$. In the special case in which $H(y) = \frac{1}{2}\|y\|^2$, this will correspond to motion on a sphere. We recall the standard notation

$$\nabla(H) = \begin{bmatrix} \frac{\partial H}{\partial y_1} \\ \frac{\partial H}{\partial y_2} \\ \vdots \\ \frac{\partial H}{\partial y_N} \end{bmatrix}$$

and consider problems of the 'Poisson' form

$$y' = L(x, y)\nabla(H), \tag{122a}$$

where $L(x, y)$ is always a skew-symmetric matrix. For such problems $H(y(x))$ is invariant. To verify this, calculate

$$\frac{\mathrm{d}}{\mathrm{d}x}H(y(x)) = \sum_{i=1}^{N} \frac{\partial H}{\partial y_i} y_i'(x) = \nabla(H)^{\mathsf{T}} L(x, y)\nabla(y) = 0,$$

because of the skew-symmetry of L.

The Euler equations discussed in Subsection 107, provide two examples of this. To show that $E(w)$ is invariant write $H(w) = \frac{1}{2}E(w)$, and to show that $F(w)$ is invariant write $H(w) = \frac{1}{2}F(w)$. The problem reverts to the form of (122a), with y replaced by w, where $L(x, w)$ is given by

$$\begin{bmatrix} 0 & \frac{I_3 w_3}{I_1 I_2} & -\frac{I_2 w_2}{I_1 I_3} \\ -\frac{I_3 w_3}{I_1 I_2} & 0 & \frac{I_1 w_1}{I_2 I_3} \\ \frac{I_2 w_2}{I_1 I_3} & -\frac{I_1 w_1}{I_2 I_3} & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & -\frac{w_3}{I_1 I_2} & \frac{w_2}{I_1 I_3} \\ \frac{w_3}{I_1 I_2} & 0 & -\frac{w_1}{I_2 I_3} \\ -\frac{w_2}{I_1 I_3} & \frac{w_1}{I_2 I_3} & 0 \end{bmatrix},$$

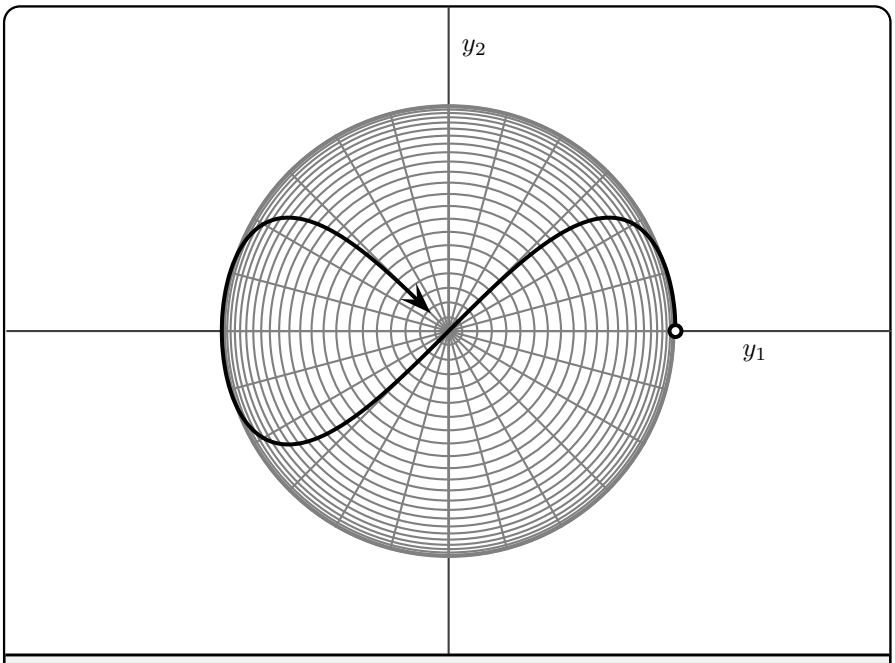

Figure 122(i) Solution to problem (122c) with y_3 pointing out of the page

respectively.

We now revert to the special case $H(x) = \frac{1}{2}y^{\mathsf{T}}y$, for which (122a) becomes

$$y' = L(x, y)y. \tag{122b}$$

An example is the contrived problem

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 0 & -y_1 & -\sin(x) \\ y_1 & 0 & -1 \\ \sin(x) & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \tag{122c}$$

with solution $y_1(x) = \cos(x)$, $y_2(x) = \cos(x)\sin(x)$, $y_3(x) = \sin^2(x)$. The solution values for $x \in [0, 1.4\pi]$ are shown in Figure 122(i).

Problems of the form (122b) are a special case of

$$Y' = L(x, Y)Y,$$

where Y has a number of columns. In this case the inner product of two specific columns will be invariant. In particular, if $Y(x)$ is a square matrix, initially orthogonal, and $L(x, Y)$ is always skew-symmetric, then $Y(x)$ will remain

orthogonal. Denote the elements of Y by y_{ij}. An example problem of this type is

$$Y'(x) = \begin{bmatrix} 0 & -1 & \mu y_{21} \\ 1 & 0 & -\mu y_{11} \\ -\mu y_{21} & \mu y_{11} & 0 \end{bmatrix} Y, \qquad Y(0) = I, \tag{122d}$$

with μ a real parameter. The solution to (122d) is

$$Y(x) = \begin{bmatrix} \cos(x) & -\sin(x)\cos(\mu x) & \sin(x)\sin(\mu x) \\ \sin(x) & \cos(x)\cos(\mu x) & -\cos(x)\sin(\mu x) \\ 0 & \sin(\mu x) & \cos(\mu x) \end{bmatrix}.$$

123 Further Hamiltonian problems

In the Hamiltonian formulation of classical mechanics, generalized coordinates q_1, q_2, ..., q_N and generalized momenta p_1, p_2, ..., p_N are used to represent the state of a mechanical system. The equations of motion are defined in terms of a 'Hamiltonian' function $H(p_1, p_2, \ldots, p_N, q_1, q_2, \ldots, q_N)$ by the equations

$$p_i' = -\frac{\partial H}{\partial q_i},$$
$$q_i' = \frac{\partial H}{\partial p_i}.$$

Write $y(x)$ as a vector variable, made up from N momenta followed by the N coordinates. That is,

$$y_i = \begin{cases} p_i, & 1 \le i \le N, \\ q_{i-N}, & N+1 \le i \le 2N. \end{cases}$$

With the understanding that H is regarded as a function of y, the differential equations can be written in the form $y' = f(y)$, where

$$f(y) = J\nabla(H), \qquad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},$$

in which I is the $N \times N$ unit matrix.

Theorem 123A *$H(y(x))$ is invariant.*

Proof. Calculate $\partial H/\partial y$ to obtain the result $\nabla(H)^{\mathsf{T}} J \nabla(H) = 0$. □

The Jacobian of this problem is equal to

$$\frac{\partial}{\partial y} f(y) = \frac{\partial}{\partial y} \left(J \nabla(H) \right) = JW(y),$$

where W is the 'Hessian' matrix defined as the $2N \times 2N$ matrix with (i, j) element equal to $\partial^2 H / \partial y_i \partial y_j$.

If the initial value $y_0 = y(x_0)$ is perturbed by a small number ϵ multiplied by a fixed vector v_0, then, to within $\mathcal{O}(\epsilon^2)$, the solution is modified by $\epsilon v + \mathcal{O}(\epsilon^2)$ where

$$v'(x) = \frac{\partial f}{\partial y} v(x) = JW(y)v(x).$$

For two such perturbations u and v, it is interesting to consider the value of the scalar $u^{\mathsf{T}} J v$.

This satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x} u^{\mathsf{T}} J v = u^{\mathsf{T}} J J W v + (JWu)^{\mathsf{T}} J v = -u^{\mathsf{T}} W v + u^{\mathsf{T}} W v = 0.$$

Hence we have:

Theorem 123B *$u^{\mathsf{T}} J v$ is invariant with time.*

In the special case of a two-dimensional Hamiltonian problem, the value of $(\epsilon u)^{\mathsf{T}} J (\epsilon v)$ can be interpreted as the area of the infinitesimal parallelogram with sides in the directions u and v. As the solution evolves, u and v might change, but the area $u^{\mathsf{T}} J v$ remains invariant. This is illustrated in Figure 123(i) for the two problems $H(p, q) = p^2/2 + q^2/2$ and $H(p, q) = p^2/2 - \cos(q)$ respectively.

124 Further differential-algebraic problems

Consider the initial value problem

$$y' = y + z, \tag{124a}$$

$$0 = z + z^3 - y, \tag{124b}$$

$$y(0) = 2, \quad z(0) = 1.$$

This is an index 1 problem, because a single differentiation of (124b) and a substitution from (124a) converts this to a differential equation system consisting of (124b) together with $z' = (y + z)/(1 + 3z^2)$. However, this reduction does not do justice to the original formulation in the sense that a solution with slightly perturbed initial values has little to do with the original index 1 problem. This emphasizes the fact that initial conditions for the differential-algebraic equation formulation must be consistent with the algebraic constraint for it to be well-posed. A more appropriate reduction is to replace (124a) by $y' = y + \phi(y)$, where $\phi(y)$ is the real value of z which satisfies (124b).

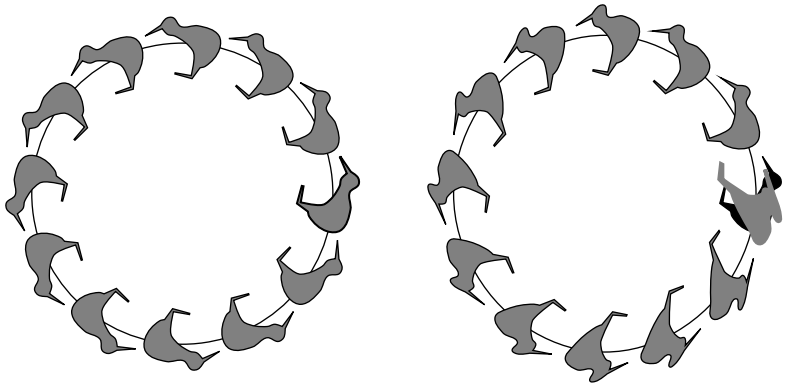

Figure 123(i) Illustration of symplectic behaviour for two problems: $H(p,q) = p^2/2 + q^2/2$ (left) and $H(p,q) = p^2/2 - \cos(q)$ (right). The underlying image depicts the North Island brown kiwi, *Apteryx mantelli*.

We next introduce an initial value problem comprising two differential equations and a single algebraic constraint:

$$y_1' = -\sin(z), \tag{124c}$$

$$y_2' = 2\cos(z) - y_1, \tag{124d}$$

$$0 = y_1^2 + y_2^2 - 1, \tag{124e}$$

$$y_1(0) = 1, \quad y_2(0) = 0, \qquad z(0) = 0.$$

An attempt to reduce this to an ordinary differential equation system by differentiating (124e) and substituting from (124c) and (124d), leads to a new algebraic constraint

$$-y_1 \sin(z) + y_2(2\cos(z) - y_1) = 0, \tag{124f}$$

and it is clear that this will be satisfied by the solution to the original problem. However, this so-called 'hidden constraint' introduces a new complexity into this type of problem. That is, for initial values to be consistent, (124f) must be satisfied at the initial time. If, for example, the initial values $y_1(0) = 1$ and $y_2(0) = 0$ are retained, but the initial value $z(0)$ is perturbed slightly, (124f) will not be satisfied and no genuine solution exists. But the hidden constraint, as the problem has actually been posed, is satisfied, and we can take the reduction towards an ordinary differential equation system to completion. Differentiate (124f) and substitute from (124c) and (124d) and we finally arrive at

$$z'(\cos^2(z) + 2\sin^2(z)) = \sin^2(z) + y_2 \sin(z) + (2\cos(z) - y_1)^2. \tag{124g}$$

Because two differentiation steps were required to reach this equation, the original system is referred to as an index 2 problem. In summary, the original index 2 problem, comprising (124c), (124d), (124e) has been reduced, first to an index 1 formulation (124c), (124d), (124f), and then to an ordinary differential equation system (124c), (124d), (124g).

Exercises 12

12.1 Show that a problem of the form

$$u' = -\alpha'(v)\gamma(u,v),$$
$$v' = \beta'(u)\gamma(u,v),$$

satisfies the assumptions of (122a) with a suitable choice of $H(u,v)$.

12.2 Write the Lotka–Volterra equations (106a), (106b) in the form given in Exercise **12.1**.

13 Difference Equation Problems

130 Introduction to difference equations

While differential equations deal with functions of a continuous variable, difference equations deal with functions of a discrete variable. Instead of a formula for the derivative of a function written in terms of the function itself, we have to consider sequences for which each member is related in some specific way to its immediate predecessor or several of its most recent predecessors. Thus we may write

$$x_n = \phi_n(x_{n-1}, x_{n-2}, \ldots, x_{n-k}),$$

where k is the 'order' of this difference equation. This equation, in which x_n depends on k previous values, can be recast in a vector setting in which members of the sequence lie not in $\mathbb{R}$ but in $\mathbb{R}^k$, and depend only on *one* previous value. Thus if

$$X_n = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-k+1} \end{bmatrix},$$

then

$$X_n = \Phi_n(X_{n-1}) = \begin{bmatrix} \phi_n(x_{n-1}, x_{n-2}, \ldots, x_{n-k}) \\ x_{n-1} \\ x_{n-2} \\ \vdots \\ x_{n-k+1} \end{bmatrix}.$$

Just as for differential equations, we can use either formulation as we please.

131 A linear problem

Consider the difference equation

$$y_n = 3y_{n-1} - 2y_{n-2} + C\theta^n, \tag{131a}$$

where C and θ are constants. We do not specify an initial value, but aim instead to find the family of all solutions. As a first step, we look at the special case in which $C = 0$. In this case, the equation becomes linear in the sense that known solutions can be combined by linear combinations. The simplified equation in matrix–vector form is

$$\begin{bmatrix} y_n \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{n-1} \\ y_{n-2} \end{bmatrix},$$

which can be rewritten as

$$\begin{bmatrix} y_n - y_{n-1} \\ -y_n + 2y_{n-1} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{n-1} - y_{n-2} \\ -y_{n-1} + 2y_{n-2} \end{bmatrix},$$

with solution defined by

$$\begin{aligned} y_n - y_{n-1} &= A2^{n-1}, \\ -y_n + 2y_{n-1} &= B, \end{aligned}$$

for constants A and B. By eliminating y_{n-1}, we find

$$y_n = A2^n + B$$

for the general solution. The fact that this combines powers of 2 and 1, the eigenvalues of the matrix

$$\begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}, \tag{131b}$$

suggests that we can look for solutions for the original formulation in the form λ^n without transforming to the matrix–vector formulation. Substitute this trial solution into (131a), with $C = 0$, and we find, apart from a factor λ^{n-2}, that the condition on λ is

$$\lambda^2 - 3\lambda + 2 = 0.$$

This is the characteristic polynomial of the matrix (131b), but it can be read off immediately from the coefficients in (131a).

To find the general solution to (131a), if $C \neq 0$, it is easy to see that we only need to find one special solution to which we can add the terms $A2^n + B$ to obtain all possible solutions. A special solution is easily found, if $\theta \neq 1$ and $\theta \neq 2$, in the form

$$y_n = \frac{C\theta^{n+2}}{(\theta - 1)(\theta - 2)}.$$

This type of special solution is not available if θ equals either 1 or 2. In these cases a special solution can be found as a multiple of n or $n2^n$, respectively. Combining these cases, we write the general solution as

$$y_n = \begin{cases} A2^n + B - Cn, & \theta = 1, \\ A2^n + B + 2Cn2^n, & \theta = 2, \\ A2^n + B + \dfrac{C\theta^2}{(\theta-1)(\theta-2)}\theta^n, & \theta \neq 1, \theta \neq 2. \end{cases}$$

132 The Fibonacci difference equation

The initial value difference equation

$$y_n = y_{n-1} + y_{n-2}, \qquad y_0 = 0, \quad y_1 = 1, \tag{132a}$$

is famous because of the mathematical, biological and even numerological significance attached to the solution values

$$1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots.$$

To find the general solution, solve the polynomial equation

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

to find the terms λ_1^n and λ_2^n, where

$$\lambda_1 = \frac{1+\sqrt{5}}{2}, \quad \lambda_2 = \frac{1-\sqrt{5}}{2} = -\lambda_1^{-1}.$$

To find the coefficients A and B in the general solution

$$y_n = A\Big(\frac{1+\sqrt{5}}{2}\Big)^n + B\Big(-\frac{1+\sqrt{5}}{2}\Big)^{-n},$$

substitute $n = 0$ and $n = 1$, to find $A = -B = 5^{-1/2}$, and therefore the specific solution to the initial value problem (132a),

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

133 Three quadratic problems

We consider the solutions to the problems

$$y_n = y_{n-1}^2, \tag{133a}$$
$$y_n = y_{n-1}^2 - 2, \tag{133b}$$
$$y_n = y_{n-1}y_{n-2}. \tag{133c}$$

If $z_n = \ln(y_n)$ in (133a), then $z_n = 2z_{n-1}$ with solution $z_n = 2^n z_0$. Hence, the general solution to (133a) is

$$y_n = y_0^{2^n}.$$

To solve (133b), substitute $y_n = z_n + z_n^{-1}$, so that

$$z_n + \frac{1}{z_n} = z_{n-1}^2 + \frac{1}{z_{n-1}^2},$$

and this is satisfied by any solution to $z_n = z_{n-1}^2$. Hence, using the known solution of (133a), we find

$$y_n = z_0^{2^n} + z_0^{-2^n},$$

where z_0 is one of the solutions to the equation

$$z_0 + \frac{1}{z_0} = y_0.$$

Finally, to solve (133c), substitute $z_n = \ln(y_n)$, and we find that

$$z_n = z_{n-1} + z_{n-2}.$$

The general solution to this is found from the Fibonacci equation, so that substituting back in terms of y_n, we find

$$y_n = A^{\left(\frac{1}{2}(1+\sqrt{5})\right)^n} \cdot B^{\left(\frac{1}{2}(1-\sqrt{5})\right)^n},$$

with A and B determined from the initial values.

134 Iterative solutions of a polynomial equation

We discuss the possible solution of the polynomial equation

$$x^2 - 2 = 0.$$

Of course this is only an example, and a similar discussion would be possible with other polynomial equations. Consider the difference equations

$$y_n = y_{n-1} - \tfrac{1}{2}y_{n-1}^2 + 1, \qquad y_0 = 0, \tag{134a}$$

$$y_n = y_{n-1} - \tfrac{1}{2}y_{n-1}^2 + 1, \qquad y_0 = 4, \tag{134b}$$

$$y_n = y_{n-1} - y_{n-1}^2 + 2, \qquad y_0 = \frac{3}{2}, \tag{134c}$$

$$y_n = \frac{y_{n-1}}{2} + \frac{1}{y_{n-1}}, \qquad y_0 = 100, \tag{134d}$$

$$y_n = \frac{y_{n-1}y_{n-2} + 2}{y_{n-1} + y_{n-2}}, \qquad y_0 = 0, \quad y_1 = 1. \tag{134e}$$

Note that each of these difference equations has $\sqrt{2}$ as a stationary point. That is, each of them is satisfied by $y_n = \sqrt{2}$, for every n. Before commenting further, it is interesting to see what happens if a few values are evaluated numerically for each sequence. These are shown in Table 134(I).

Note that (134a) seems to be converging to $\sqrt{2}$, whereas (134b) seems to have no hope of ever doing so. Of course the starting value, y_0, is the distinguishing feature, and we can perhaps investigate which values converge and which ones do not. It can be shown that the fate of the iterates for various starting values can be summarized as follows:

Table 134(I) The first few terms in the solutions of some difference equations

	Equation (134a)	Equation (134b)	Equation (134c)	Equation (134d)	Equation (134e)
y_0	0.0000000000	4.0000000000	1.5000000000	1.000000×10^{2}	0.0000000000
y_1	1.0000000000	−3.0000000000	1.2500000000	5.001000×10	1.0000000000
y_2	1.5000000000	−6.5000000000	1.6875000000	2.502500×10	2.0000000000
y_3	1.3750000000	-2.662500×10	0.8398437500	1.255246×10	1.3333333333
y_4	1.4296875000	-3.800703×10^{2}	2.1345062256	6.3558946949	1.4000000000
y_5	1.4076843262	-7.260579×10^{4}	−0.4216106015	3.3352816093	1.4146341463
y_6	1.4168967451	-2.635873×10^{9}	1.4006338992	1.9674655622	1.4142114385

$$
\begin{aligned}
y_0 \in \{-\sqrt{2}, 2+\sqrt{2}\}&: \quad \text{Convergence to } x = -\sqrt{2} \\
y_0 \in (-\sqrt{2}, 2+\sqrt{2})&: \quad \text{Convergence to } x = \sqrt{2} \\
y_0 \notin [-\sqrt{2}, 2+\sqrt{2}]&: \quad \text{Divergence}
\end{aligned}
$$

Note that the starting value $y_0 = -\sqrt{2}$, while it is a fixed point of the mapping given by (134a), is unstable; that is, any small perturbation from this initial value will send the sequence either into instability or convergence to $+\sqrt{2}$. A similar remark applies to $y_0 = 2+\sqrt{2}$, which maps immediately to $y_1 = -\sqrt{2}$.

The difference equation (134c) converges to $\pm\sqrt{2}$ in a finite number of steps for y_0 in a certain countable set; otherwise the sequence formed from this equation diverges.

Equation (134d) is the Newton method and converges quadratically to $\sqrt{2}$ for any positive y_0. By quadratic convergence, we mean that $|y_n - \sqrt{2}|$ divided by $|y_{n-1} - \sqrt{2}|^2$ is bounded. In fact, in the limit as $n \to \infty$,

$$\frac{y_n - \sqrt{2}}{(y_{n-1} - \sqrt{2})^2} \to \frac{\sqrt{2}}{4}.$$

The iteration scheme given by (134e) is based on the secant method for solving non-linear equations. To solve $\phi(y) = 0$, y_n is found by fitting a straight line through the two points $(y_{n-2}, \phi(y_{n-2}))$ and $(y_{n-1}, \phi(y_{n-1}))$ and defining y_n as the point where this line crosses the horizontal axis. In the case $\phi(y) = y^2 - 2$, this results in (134e).

It is interesting to ask if there exists an 'order' k for this sequence. In other words, assuming that convergence is actually achieved, does $k \geq 1$ exist such that

$$\frac{|y_n - \sqrt{2}|}{|y_{n-1} - \sqrt{2}|^k}$$

has a limiting value as $n \to \infty$? For the secant method k does exist, and has the value $k = \frac{1}{2}(\sqrt{5}+1)$.

135 The arithmetic-geometric mean

Let a_0 and b_0 be real numbers chosen so that $0 < b_0 < a_0$, and define the sequence of (a_n, b_n) pairs by the formulae

$$\begin{aligned} a_n &= \tfrac{1}{2}(a_{n-1} + b_{n-1}), \\ b_n &= \sqrt{a_{n-1}b_{n-1}}, \end{aligned} \qquad n = 1, 2, \ldots. \tag{135a}$$

We can verify (i) that $b_{n-1} < b_n < a_n < a_{n-1}$ for all $n \geq 1$ and (ii) that the sequence $a_0 - b_0$, $a_1 - b_1$, $a_2 - b_2$, $\ldots$ converges to zero. The truth of (i) follows from elementary properties of arithmetic and geometric means. Furthermore, (ii) can be proved from the identity

$$a_n - b_n = \frac{(a_{n-1} - b_{n-1})^2}{2\big(\sqrt{a_{n-1}} + \sqrt{b_{n-1}}\big)^2}.$$

The common limit of the a_n and b_n sequences is known as the 'arithmetic-geometric mean' of a_0 and b_0. We present a single application.

The quantities

$$\begin{aligned} F(a,b) &= \int_0^{\pi/2} \big(a^2\cos^2(\theta) + b^2\sin^2(\theta)\big)^{-1/2}\, d\theta, \\ E(a,b) &= \int_0^{\pi/2} \big(a^2\cos^2(\theta) + b^2\sin^2(\theta)\big)^{1/2}\, d\theta, \end{aligned}$$

are known as 'complete elliptic integrals' of the first and second kind, respectively. The value of $4E(a,b)$ is the length of the circumference of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Use $a_0 = a$ and $b_0 = b$ as starting values for the computation of the sequences defined by (135a), and denote by a_∞ the arithmetic-geometric mean of a_0 and b_0. Then it can be shown that

$$F(a_0, b_0) = F(a_1, b_1),$$

and therefore that

$$F(a_0, b_0) = F(a_\infty, a_\infty) = \frac{\pi}{2a_\infty}.$$

The value of $E(a_0, b_0)$ can also be found from the sequences that lead to the arithmetic-geometric mean. In fact

$$E(a_0, b_0) = \frac{\pi}{2a_\infty}\left(a_0^2 - 2a_1(a_0 - a_1) - 4a_2(a_1 - a_2) - 8a_3(a_2 - a_3) - \cdots\right).$$

Exercises 13

13.1 Write the difference equation given by (134e) in the form

$$z_n = \phi(z_{n-1}),$$

with z_0 a given initial value.

13.2 Write the difference equation system

$$\begin{aligned} u_n &= u_{n-1} + v_{n-1}, & u_0 &= 2, \\ v_n &= 2u_{n-1} + v_{n-1}^2, & v_0 &= 1, \end{aligned}$$

in the form $y_n = \phi(y_{n-1}, y_{n-2})$, with y_0 and y_1 given initial values.

13.3 Use the formula for the error in linear interpolation together with the solution to (133c) to verify the order of convergence of (134e).

13.4 Calculate $\sqrt{2}$ by applying the Newton method to the equation

$$2x^{-2} - 1 = 0.$$

13.5 Calculate the value of $\sqrt{3}$ by applying the secant method to

$$x^2 - 3 = 0.$$

13.6 Calculate the circumference of the ellipse

$$\frac{x^2}{9} + \frac{y^2}{4} = 1,$$

using the arithmetic-geometric mean.

14 Difference Equation Theory

140 Linear difference equations

The standard form for linear difference equation systems is

$$X_n = A_n X_{n-1} + \phi_n, \tag{140a}$$

which becomes an initial value problem if the value of the initial vector X_0 is specified. The corresponding system in which ϕ_n is omitted is the 'homogeneous part'.

Many linear difference equations are more naturally formulated as

$$y_n = \alpha_{n1} y_{n-1} + \alpha_{n2} y_{n-2} + \cdots + \alpha_{nk} y_{n-k} + \psi_n,$$

but these are easily recast in the form (140a) by writing

$$X_n = \begin{bmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_{n-k+1} \end{bmatrix}, \quad A_n = \begin{bmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nk} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \phi_n = \begin{bmatrix} \psi_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

To solve (140a) as an initial value problem, we need to use products of the form

$$\prod_{i=m}^{n} A_i = A_n A_{n-1} \cdots A_{m+1} A_m.$$

We have:

Theorem 140A *The problem (140a), with initial value X_0 given, has the unique solution*

$$y_n = \Big(\prod_{i=1}^{n} A_i\Big) X_0 + \Big(\prod_{i=2}^{n} A_i\Big)\phi_1 + \Big(\prod_{i=3}^{n} A_i\Big)\phi_2 + \cdots + A_n\phi_{n-1} + \phi_n.$$

Proof. The result holds for $n = 0$, and the general case follows by induction. □

141 Constant coefficients

We consider the solution of a linear difference equation with constant coefficients:

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} + \psi_n. \tag{141a}$$

The solution is found in terms of the solution to the canonical problem in which the initial information is given in the form

$$\begin{bmatrix} y_0 \\ y_{-1} \\ \vdots \\ y_{-k+2} \\ y_{-k+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

Denote the solution to this problem at step m by

$$y_m = \theta_m, \qquad m = 0, 1, 2, \ldots, n,$$

with $\theta_m = 0$ for $m < 0$. Given the difference equation (141a) with initial values $y_0, y_1, \ldots, y_{k-1}$, define linear combinations of this data by

$$\begin{bmatrix} \widetilde{y}_{k-1} \\ \widetilde{y}_{k-2} \\ \widetilde{y}_{k-3} \\ \vdots \\ \widetilde{y}_1 \\ \widetilde{y}_0 \end{bmatrix} = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_{k-2} & \theta_{k-1} \\ 0 & 1 & \theta_1 & \cdots & \theta_{k-3} & \theta_{k-2} \\ 0 & 0 & 1 & \cdots & \theta_{k-4} & \theta_{k-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \theta_1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ y_{k-3} \\ \vdots \\ y_1 \\ y_0 \end{bmatrix}. \tag{141b}$$

We are now in a position to write down the solution to (141a).

Theorem 141A *Using the notation introduced in this subsection, the solution to (141a) with given initial values $y_0, y_1, \ldots, y_{k-1}$ is given by*

$$y_n = \sum_{i=0}^{k-1} \theta_{n-i} \widetilde{y}_i + \sum_{i=k}^{n} \theta_{n-i} \psi_i. \tag{141c}$$

Proof. Substitute $n = m$, for $m = 0, 1, 2, \ldots, k-1$, into (141c), and we obtain the value

$$y_m = \widetilde{y}_m + \theta_1 \widetilde{y}_{m-1} + \cdots + \theta_m \widetilde{y}_0, \quad m = 0, 1, 2, \ldots, k-1.$$

This is equal to y_m if (141b) holds. Add the contribution to the solution from each of $m = k, k+1, \ldots, n$ and the result follows. □

142 Powers of matrices

We are interested in powers of a matrix A in terms of two questions: when is the sequence of powers bounded, and when does the sequence converge to the zero matrix? There are various equivalent formulations of the criteria for these properties of A, and we state the most widely accessible of these.

Definition 142A *A square matrix A is 'stable' if there exists a constant C such that for all $n = 0, 1, 2, \ldots$, $\|A^n\| \leq C$.*

This property is often referred to as 'power-boundedness'.

Definition 142B *A square matrix A is 'convergent' if $\lim_{n\to\infty} \|A^n\| = 0$.*

Theorem 142C *Let A denote an $m \times m$ matrix. The following statements are equivalent:*

(i) *A is stable.*

(ii) *The minimal polynomial of A has all its zeros in the closed unit disc and all its multiple zeros in the open unit disc.*

(iii) *The Jordan canonical form of A has all its eigenvalues in the closed unit disc with all eigenvalues of magnitude 1 lying in 1×1 blocks.*

(iv) *There exists a non-singular matrix S such that $\|S^{-1}AS\|_\infty \leq 1$.*

Proof. We prove that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. If A is stable but (ii) is not true, then either there exist λ and $v \neq 0$ such that $|\lambda| > 1$ and $Av = \lambda v$, or there exist λ, $u \neq 0$ and v such that $|\lambda| = 1$ and $Av = \lambda v + u$, with $Au = \lambda u$. In the first case, $A^n v = \lambda^n v$ and therefore $\|A^n\| \geq |\lambda|^n$ which is not bounded. In the second case, $A^n v = \lambda^n v + n\lambda^{n-1}u$ and therefore $\|A^n\| \geq n\|u\|/\|v\| - 1$, which also is not bounded. Given (ii), it is not possible that the conditions of (iii) are not satisfied, because the minimal polynomial of any of the Jordan blocks, and therefore of A itself, would have factors that contradict (ii). If (iii) is true, then S can be chosen to form J, the Jordan canonical form of A, with the off-diagonal elements chosen sufficiently small so that $\|J\|_\infty \leq 1$. Finally, if (iv) is true then $A^n = S(S^{-1}AS)^n S^{-1}$ so that

$$\|A^n\| \leq \|S\| \cdot \|S^{-1}AS\|^n \cdot \|S^{-1}\| \leq \|S\| \cdot \|S^{-1}\|. \qquad \square$$

A related property of the difference equation

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k}, \tag{142a}$$

that is (141a) with the inhomogeneous term omitted, is also given the same name:

Definition 142D *The difference equation* (142a) *is stable if it has a bounded solution for any initial values.*

Theorem 142E *The difference equation* (142a) *is stable if and only if it satisfies the 'root condition'; namely that the polynomial*

$$\rho(z) = z^k - \alpha_1 z^{k-1} - \alpha_2 z^{k-2} - \cdots - \alpha_k$$

has all its zeros in the closed unit disc and all zeros on the boundary are simple.

Proof. The boundedness of all solutions to the difference equation (142a) is equivalent to the stability of the companion matrix

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{k-1} & \alpha_k \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and, for A, the minimal and characteristic polynomials are identical. $\square$

Theorem 142F *Let A denote an $m \times m$ matrix. The following statements are equivalent*

(i) *A is convergent.*

(ii) *The minimal polynomial of A has all its zeros in the open unit disc.*

(iii) *The Jordan canonical form of A has all its diagonal elements in the open unit disc.*

(iv) *There exists a non-singular matrix S such that $\|S^{-1}AS\|_\infty < 1$.*

Proof. We again prove that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. If A is convergent but (ii) is not true, then there exist λ and $u \neq 0$ such that $\lambda \geq 1$ and $Au = \lambda u$. Hence, $A^n u = \lambda^n u$ and therefore $\|A^n\| \geq |\lambda|^n$, which does not converge to zero. Given (ii), it is not possible that the conditions of (iii) are not satisfied, because the minimal polynomial of any of the Jordan blocks, and therefore of A itself, would have factors that contradict (ii). If (iii) is true, then S can be chosen to form J, the Jordan canonical form of A, with the off-diagonal elements chosen sufficiently small so that $\|J\|_\infty < 1$. Finally, if (iv) is true then $A^n = S(S^{-1}AS)^n S^{-1}$ so that $\|A^n\| \leq \|S\| \cdot \|S^{-1}\| \cdot \|S^{-1}AS\|^n \to 0$. □

While the two results we have presented here are related to the convergence of difference equation solutions, the next is introduced only because of its application in later chapters.

Theorem 142G *If A is a stable $m \times m$ matrix and B an arbitrary $m \times m$ matrix, then there exists a real C such that*

$$\left\| \left(A + \frac{1}{n}B\right)^n \right\| \leq C,$$

for $n = 1, 2, \ldots$.

Proof. Without loss of generality, assume that $\|\cdot\|$ denotes the norm $\|\cdot\|_\infty$. Because S exists so that $\|S^{-1}AS\| \leq 1$, we have

$$\begin{aligned}
\left\| \left(A + \frac{1}{n}B\right)^n \right\| &\leq \|S\| \cdot \|S^{-1}\| \cdot \left\| \left(S^{-1}AS + \frac{1}{n}S^{-1}BS\right)^n \right\| \\
&\leq \|S\| \cdot \|S^{-1}\| \cdot \left(1 + \frac{1}{n}\|S^{-1}BS\|\right)^n \\
&\leq \|S\| \cdot \|S^{-1}\| \exp(\|S^{-1}BS\|).
\end{aligned}$$

□

In applying this result to sequences of vectors, the term represented by the matrix B can be replaced by a non-linear function which satisfies suitable conditions. To widen the applicability of the result a non-homogeneous term is included.

Theorem 142H *Let A be a stable $m \times m$ matrix and $\phi : \mathbb{R}^m \to \mathbb{R}^m$ be such that $\|\phi(x)\| \leq L\|x\|$, for L a positive constant and $x \in \mathbb{R}^m$. If $w = (w_1, w_2, \ldots, w_n)$*

and $v = (v_0, v_1, \ldots, v_n)$ *are sequences related by*

$$v_i = Av_{i-1} + \frac{1}{n}\phi(v_{i-1}) + w_i, \quad i = 1, 2, \ldots, n, \tag{142b}$$

then

$$\|v_n\| \le C\big(\|v_0\| + \sum_{i=1}^{n} \|w_i\|\big),$$

where C *is independent of* n.

Proof. Let S be the matrix introduced in the proof of Theorem 142C. From (142b), it follows that

$$(S^{-1}v_i) = (S^{-1}AS)(S^{-1}v_{i-1}) + \frac{1}{n}(S^{-1}\phi(v_{i-1})) + (S^{-1}w_i)$$

and hence

$$\|S^{-1}v_i\| \le \|S^{-1}AS\| \cdot \|S^{-1}v_{i-1}\| + \frac{1}{n}\|S^{-1}\phi(v_{i-1})\| + \|S^{-1}w_i\|,$$

leading to the bound

$$\|v_n\| \le \|S\| \cdot \|S^{-1}\| \exp\big(L\|S\| \cdot \|S^{-1}\|\big) \Big(\|v_0\| + \sum_{i=1}^{n} \|w_i\|\Big). \qquad \square$$

Exercises 14

14.1 Find a constant C such that $\|A^n\|_\infty \le C$, for all $n = 0, 1, \ldots$, where

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

14.2 For what values of the complex number θ is the matrix A stable, where

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

14.3 For what values of the complex number θ is the matrix A convergent, where

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

15 Location of Polynomial Zeros

150 Introduction

The questions discussed in this section concern a polynomial

$$P(z) = a_0 z^n + a_1 z_{n-1} + \cdots + a_n \tag{150a}$$

and the location of its zeros in the complex plane.

We will consider four main questions

1. Are all the zeros in the open left half-plane?
2. Are all the zeros in the open right half-plane?
3. Are all the zeros in the open unit disc?
4. Are all the zeros in the complement of the closed unit disc?

These four questions are in reality only two because 1 and 2 can be interchanged using the mapping $z \mapsto -z$ and 3 and 4 can be interchanged using reversal of coefficients $a_i \mapsto a_{n-i}$. But even the two questions 1 and 3 are interconnected because z in the open left half-plane is equivalent to $(1+z)/(1-z)$ in the open unit disc. Related questions, in which the open sets referred to are replaced by their closures, are also of importance although we will focus our attention on questions 1 and 3.

Even though location relative to the imaginary axis, and location relative to the unit circle, are related through a transformation in the complex plane, different algorithms are typically used for the two questions and some of these will be explored.

Applications of the half-plane questions are to the stability and convergence to zero of linear differential equations of the form

$$a_0 y^{(n)} + a_1 y^{(n-1)} + \cdots + y = 0,$$

while the unit disc questions are concerned with the stability and convergence to zero of linear difference equations of the form

$$a_0 y_k + a_1 y_{k-1} + \cdots + y_{k-n} = 0, \qquad k = n, n+1, \ldots.$$

An additional application of the disc question is to the solution of polynomial equations by the method of Lehmer (1961). A basic reference (Miller 1974) surveys known results on these 'polynomial type' questions.

151 Left half-plane results

We will consider only the case that the coefficients in (150a) are all real and, without loss of generality, that $a_0 > 0$. For convenience we will denote the property that all zeros are in the left half-plane as property S and write $S(P)$ to represent the statement that P possesses this property. A necessary condition for $S(P)$ to hold will be presented in Theorem 151B, following a preliminary result.

Lemma 151A *If $S(P)$ then $S(P')$.*

Proof. Use the identity

$$\frac{P'(z)}{P(z)} = \sum_{i=1}^{n} \frac{1}{z - z_i} \tag{151a}$$

and calculate the change in the value of the two sides of (151a) when z moves along a path formed by combining the line $[-R, R]\mathrm{i}$ with the semicircle $R\exp(\mathrm{i}\,\theta)$, where $\theta \in [\pi/2, 3\pi/2]$. It is assumed that R is large enough for the zeros of both P and P' to be included within the semicircular region. The result of doing this calculation gives

$$(m - n)2\pi\mathrm{i} = -2\pi\mathrm{i}\,,$$

where m is the number of zeros of P' in the left half-plane. It follows that $m = n - 1$ and that $S(P')$. □

Theorem 151B *If $a_0 > 0$ and $S(P)$ holds, then $a_i > 0$, for $i = 1, 2, \ldots, n$.*

Proof. If for some i, $a_i \le 0$, apply Lemma 151A $n - i$ times and it is found that $a_i = \prod_{j=1}^{n-i}(-\widetilde{z}_j)$, where $\widetilde{z}_i$ $(j = 1, 2, \ldots, n - i)$ are in the left half-plane. This product is positive giving a contradiction. □

For each P of degree n with all coefficients positive, define

$$Q_t(z) = P(z) - \tfrac{1}{2}tz\big(P(z) - (-1)^n P(-z)\big), \tag{151b}$$

where $t \in [0, a_0/a_1]$. Note that Q_t has degree n except for $t = a_0/a_1$, when the degree becomes $n - 1$. Furthermore, the coefficient of z^n in $Q_y(z)$ is positive for $t \in [0, a_0/a_1)$.

Theorem 151C *If $a_i > 0$, $i = 0, 1, \ldots, n$, then for $t \in [0, a_0/a_1]$, $S(Q_t)$ if and only if $S(P)$.*

Proof. From (151b) it follows that $Q_t(0) = P(0) \neq 0$. Furthermore it is not possible that for some $y > 0$, $Q_t(\mathrm{i}\,y_0) = 0$ because it would follow from (151b) and the conjugate equation that

$$P(\mathrm{i}\,y)(1 - \tfrac{1}{2}ty\mathrm{i}\,) = -\tfrac{1}{2}ty\mathrm{i}\,(-1)^n P(-\mathrm{i}\,y),$$
$$P(-\mathrm{i}\,y)(1 + \tfrac{1}{2}ty\mathrm{i}\,) = \tfrac{1}{2}ty\mathrm{i}\,(-1)^n P(\mathrm{i}\,y),$$

from which it follows that $|P(\mathrm{i}\,y)|^2|1 - \frac{1}{2}ty\mathrm{i}\,|^2 = |P(\mathrm{i}\,y)|^2| - \frac{1}{2}ty\mathrm{i}\,|^2$ implying $|P(\mathrm{i}\,y)|^2 = 0$, which contradicts $S(P)$. As t increases from 0 to a value in $[0, a_0/a_1)$, it is not possible for a zero of Q_t to move to the right half-plane because it would have to have crossed the imaginary axis. This conclusion also follows for the limiting case $t = a_0/a_1$ where the coefficient of z^n vanishes. Hence, $S(Q_t)$ holds. To prove the converse result, note that from (151b), it can be shown that $Q_t(z) - (-1)^n Q_t(-z) = P(z) - (-1)^n P(-z)$ and hence that

$$P_z = Q_t(z) + \tfrac{1}{2}tz\big(P(z) - (-1)^n P(-z)\big).$$

Hence $S(Q_t)$ implies $S(P)$ using an identical argument. □

This result becomes a test for $S(P)$ by calculating a sequence of polynomials each formed from the preceding one in the same way Q_{a_0/a_1} is formed from P. Denote the sequence by $P_n := P$, P_{n-1}, ..., P_1. These are defined as follows, where $\rho(P)$ denotes the highest degree coefficient divided by the second to highest degree coefficient:

$$P_i(z) = P_{i+1}(z) - \tfrac{1}{2}\rho(P_{i+1})z\big(P_{i+1}(z) + (-1)^i P_{i+1}(-z)\big),$$

for $i = n-1, n-2, \ldots, 1$. The criterion for $S(P)$ is that

$$\rho(P_i) > 0, \qquad i = n, n-1, n-2, \ldots, 1,$$

with early termination, showing a negative result, if any non-positive coefficients arise in any of the P_i.

152 Unit disc results

Let T denote the property that a polynomial has all its zeros in the open unit disc and for a polynomial (150a), in which the coefficients are not necessarily real, write $T(P)$ to mean that P possesses this property. An obvious necessary condition is given by

Lemma 152A *If $T(P)$ then $|a_n| < |a_0|$.*

Proof. If the zeros are z_i with $|z_i| < 1$, $(i = 1, 2, \ldots, n)$, then the product is also less than 1. □

Our aim will be to find a polynomial of degree $n-1$ which has property T, given that P satisfies the requirement of Lemma 152A. Let $\widetilde{P}$ be the polynomial given by

$$\widetilde{P}(z) = \overline{a}_n z^n + \overline{a}_{n-1} z^{n-1} + \cdots + \overline{a}_0.$$

And define

$$Q(z) = \widetilde{P}(0)P(z) - P(0)\widetilde{P}(z).$$

Theorem 152B *If $T(P)$ then $T(Q)$.*

Proof. Use the theorem of Rouché (see for example Ahlfors (1978)), to show that $P - a_0/\overline{a}_n \widetilde{P}$ has the same number of zeros in the open unit disc as P. We need to check that for $|z| = 1$, $|a_0/\overline{a}_n \widetilde{P}(z)| \le |P(z)|$. This follows from

$$|\widetilde{P}(z)| = |z^n \overline{P(z)}| = |P(z)|. \qquad \square$$

The coefficient of z^0 in $Q(z)$ is found to be

$$\overline{a}_n a_0 - a_0 \overline{a}_n = 0.$$

Hence $Q(z)/z$ is a suitable polynomial of degree $n-1$ which satisfies T if $|a_n| < |a_0|$.

Define the sequence by $P_n := P, P_{n-1}, \ldots, P_1$, by

$$P_i(z) = \left(\widetilde{P}_{i+1}(0)P_{i+1}(z) - P_{i+1}(0)\widetilde{P}_{i+1}(z)\right)/z, \qquad i = n-1, n-2, \ldots, 0,$$

and the test for $T(P)$ becomes

$$|P_i(0)| < |\widetilde{P}_i(0))|, \quad i = n, n-1, n-2, \ldots, 1.$$

This test is usually referred to as the Schur criterion.

Exercises 15

15.1 Show that the polynomial $P(z) = z^4 + 2z^3 + 4z^2 + +3z + 1$ satisfies the S condition.

15.2 For what values of a does the polynomial $P(z) = 2z^3 + z^2 + az + 2$ satisfy the S condition?

15.3 Show that $40z^3 + 62z^2 + (34 - 15\mathrm{i})z + 8 - 12\mathrm{i}$ has all its zeros in the open unit disc.

15.4 Show that $z^3 + 17z^2 + 20z + 12 + \mathrm{i}$ has no zeros in the closed unit disc.

Concluding remarks

Differential equations and difference equations belong together as a unified theory and as related areas of applicable mathematics. Furthermore, each is used to approximate the other. As we move on to numerical methods in subsequent chapters we will be, most of the time, thinking about smooth functions on an interval and values of a function at a sequence of discrete points. The smooth functions are the solution of a mathematical problem – an initial value differential equation problem – and the sequence of point values are the result of a computational process. The purpose of this book is to study, in a systematic way, the construction of sequences which approximate initial value problem solutions. The link between the smooth and the discrete is not only the numerical approximation process but, in the reverse direction, it is an interpolation process, aimed at finding values of the smooth function from the discrete values.

Many properties of differential equation solutions have discrete counterparts, and the link between them is of great importance. Which numerical methods generate stable sequences when applied to problems whose mathematical solutions may or may not be stable? When is conservative behaviour for the mathematical problem matched by related behaviour in the numerical approximations? These are appropriate questions to ask; we are typically studying physical problems and we want to know the consequences of modelling these problems using specific numerical approximations.