LECTURE NOTES ON MATHEMATICAL METHODS

Mihir Sen
Joseph M. Powers

Department of Aerospace and Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana 46556-5637
USA

updated
29 July 2012, 2:31pm

Contents

Preface

These are lecture notes for AME 60611 Mathematical Methods I, the first of a pair of courses on applied mathematics taught in the Department of Aerospace and Mechanical Engineering of the University of Notre Dame. Most of the students in this course are beginning graduate students in engineering coming from a variety of backgrounds. The course objective is to survey topics in applied mathematics, including multidimensional calculus, ordinary differential equations, perturbation methods, vectors and tensors, linear analysis, linear algebra, and non-linear dynamic systems. In short, the course fully explores linear systems and considers effects of non-linearity, especially those types that can be treated analytically. The companion course, AME 60612, covers complex variables, integral transforms, and partial differential equations.

These notes emphasize method and technique over rigor and completeness; the student should call on textbooks and other reference materials. It should also be remembered that practice is essential to learning; the student would do well to apply the techniques presented by working as many problems as possible. The notes, along with much information on the course, can be found at `http://www.nd.edu/~powers/ame.60611`. At this stage, anyone is free to use the notes under the auspices of the Creative Commons license below.

These notes have appeared in various forms over the past years. An especially general tightening of notation and language, improvement of figures, and addition of numerous small topics was implemented in 2011. Fall 2011 students were also especially diligent in identifying additional areas for improvement. We would be happy to hear further suggestions from you.

Mihir Sen
`Mihir.Sen.1@nd.edu`
`http://www.nd.edu/~msen`

Joseph M. Powers
`powers@nd.edu`
`http://www.nd.edu/~powers`

Notre Dame, Indiana; USA
29 July 2012

Chapter 1

Multi-variable calculus

see Kaplan, Chapter 2: 2.1-2.22, Chapter 3: 3.9,

Here we consider many fundamental notions from the calculus of many variables.

1.1 Implicit functions

The *implicit function theorem* is as follows:

Theorem

For a given $f(x, y)$ with $f = 0$ and $\partial f/\partial y \neq 0$ at the point (x_o, y_o), there corresponds a unique function $y(x)$ in the neighborhood of (x_o, y_o).

More generally, we can think of a relation such as

$$f(x_1, x_2, \ldots, x_N, y) = 0, \tag{1.1}$$

also written as

$$f(x_n, y) = 0, \qquad n = 1, 2, \ldots, N, \tag{1.2}$$

in some region as an implicit function of y with respect to the other variables. We cannot have $\partial f/\partial y = 0$, because then f would not depend on y in this region. In principle, we can write

$$y = y(x_1, x_2, \ldots, x_N), \qquad \text{or} \qquad y = y(x_n), \quad n = 1, \ldots, N, \tag{1.3}$$

if $\partial f/\partial y \neq 0$.

The derivative $\partial y/\partial x_n$ can be determined from $f = 0$ without explicitly solving for y. First, from the definition of the total derivative, we have

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \ldots + \frac{\partial f}{\partial x_n} dx_n + \ldots + \frac{\partial f}{\partial x_N} dx_N + \frac{\partial f}{\partial y} dy = 0. \tag{1.4}$$

Differentiating with respect to x_n while holding all the other $x_m, m \neq n$, constant, we get

$$\frac{\partial f}{\partial x_n} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial x_n} = 0, \tag{1.5}$$

so that

$$\frac{\partial y}{\partial x_n} = -\frac{\frac{\partial f}{\partial x_n}}{\frac{\partial f}{\partial y}}, \tag{1.6}$$

which can be found if $\partial f/\partial y \neq 0$. That is to say, y can be considered a function of x_n if $\partial f/\partial y \neq 0$.

Let us now consider the equations

$$\begin{aligned} f(x,y,u,v) &= 0, & (1.7)\\ g(x,y,u,v) &= 0. & (1.8) \end{aligned}$$

Under certain circumstances, we can unravel Eqs. (1.7-1.8), either algebraically or numerically, to form $u = u(x,y)$, $v = v(x,y)$. The conditions for the existence of such a functional dependency can be found by differentiation of the original equations; for example, differentiating Eq. (1.7) gives

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial u}du + \frac{\partial f}{\partial v}dv = 0. \tag{1.9}$$

Holding y constant and dividing by dx, we get

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} = 0. \tag{1.10}$$

Operating on Eq. (1.8) in the same manner, we get

$$\frac{\partial g}{\partial x} + \frac{\partial g}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial x} = 0. \tag{1.11}$$

Similarly, holding x constant and dividing by dy, we get

$$\begin{aligned} \frac{\partial f}{\partial y} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} &= 0, & (1.12)\\ \frac{\partial g}{\partial y} + \frac{\partial g}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial y} &= 0. & (1.13) \end{aligned}$$

Equations (1.10,1.11) can be solved for $\partial u/\partial x$ and $\partial v/\partial x$, and Eqs. (1.12,1.13) can be solved for $\partial u/\partial y$ and $\partial v/\partial y$ by using the well known Cramer's[1] rule; see Eq. (8.93). To solve for $\partial u/\partial x$ and $\partial v/\partial x$, we first write Eqs. (1.10,1.11) in matrix form:

$$\begin{pmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{pmatrix} \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \end{pmatrix} = \begin{pmatrix} -\frac{\partial f}{\partial x} \\ -\frac{\partial g}{\partial x} \end{pmatrix}. \tag{1.14}$$

[1] Gabriel Cramer, 1704-1752, well-traveled Swiss-born mathematician who did enunciate his well known rule, but was not the first to do so.

Thus, from Cramer's rule we have

$$\frac{\partial u}{\partial x} = \frac{\begin{vmatrix} -\frac{\partial f}{\partial x} & \frac{\partial f}{\partial v} \\ -\frac{\partial g}{\partial x} & \frac{\partial g}{\partial v} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}} \equiv -\frac{\frac{\partial(f,g)}{\partial(x,v)}}{\frac{\partial(f,g)}{\partial(u,v)}}, \qquad \frac{\partial v}{\partial x} = \frac{\begin{vmatrix} \frac{\partial f}{\partial u} & -\frac{\partial f}{\partial x} \\ \frac{\partial g}{\partial u} & -\frac{\partial g}{\partial x} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}} \equiv -\frac{\frac{\partial(f,g)}{\partial(u,x)}}{\frac{\partial(f,g)}{\partial(u,v)}}. \tag{1.15}$$

In a similar fashion, we can form expressions for $\partial u/\partial y$ and $\partial v/\partial y$:

$$\frac{\partial u}{\partial y} = \frac{\begin{vmatrix} -\frac{\partial f}{\partial y} & \frac{\partial f}{\partial v} \\ -\frac{\partial g}{\partial y} & \frac{\partial g}{\partial v} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}} \equiv -\frac{\frac{\partial(f,g)}{\partial(y,v)}}{\frac{\partial(f,g)}{\partial(u,v)}}, \qquad \frac{\partial v}{\partial y} = \frac{\begin{vmatrix} \frac{\partial f}{\partial u} & -\frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial u} & -\frac{\partial g}{\partial y} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}} \equiv -\frac{\frac{\partial(f,g)}{\partial(u,y)}}{\frac{\partial(f,g)}{\partial(u,v)}}. \tag{1.16}$$

Here we take the *Jacobian*[2] *matrix* $\mathbf{J}$ of the transformation to be defined as

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{pmatrix}. \tag{1.17}$$

This is distinguished from the *Jacobian determinant*, J, defined as

$$J = \det \mathbf{J} = \frac{\partial(f,g)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}. \tag{1.18}$$

If $J \neq 0$, the derivatives exist, and we indeed can form $u(x,y)$ and $v(x,y)$. This is the condition for existence of implicit to explicit function conversion.

Example 1.1

If

$$x + y + u^6 + u + v = 0, \tag{1.19}$$
$$xy + uv = 1, \tag{1.20}$$

find $\partial u/\partial x$.

Note that we have four unknowns in two equations. In principle we could solve for $u(x,y)$ and $v(x,y)$ and then determine all partial derivatives, such as the one desired. In practice this is not always possible; for example, there is no general solution to sixth order polynomial equations such as we have here.

Equations (1.19,1.20) are rewritten as

$$f(x,y,u,v) \quad x + y + u^6 + u + v = 0, \tag{1.21}$$
$$g(x,y,u,v) = xy + uv - 1 = 0. \tag{1.22}$$

[2]Carl Gustav Jacob Jacobi, 1804-1851, German/Prussian mathematician who used these quantities, which were first studied by Cauchy, in his work on partial differential equations.

Using the formula from Eq. (1.15) to solve for the desired derivative, we get

$$\frac{\partial u}{\partial x} = \frac{\begin{vmatrix} -\frac{\partial f}{\partial x} & \frac{\partial f}{\partial v} \\ -\frac{\partial g}{\partial x} & \frac{\partial g}{\partial v} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix}}. \tag{1.23}$$

Substituting, we get

$$\frac{\partial u}{\partial x} = \frac{\begin{vmatrix} -1 & 1 \\ -y & u \end{vmatrix}}{\begin{vmatrix} 6u^5+1 & 1 \\ v & u \end{vmatrix}} = \frac{y-u}{u(6u^5+1)-v}. \tag{1.24}$$

Note when

$$v = 6u^6 + u, \tag{1.25}$$

that the relevant Jacobian determinant is zero; at such points we can determine neither $\partial u/\partial x$ nor $\partial u/\partial y$; thus, for such points we cannot form $u(x,y)$.

At points where the relevant Jacobian determinant $\partial(f,g)/\partial(u,v) \neq 0$ (which includes nearly all of the (x,y) plane), given a local value of (x,y), we can use algebra to find a corresponding u and v, which may be multivalued, and use the formula developed to find the local value of the partial derivative.

1.2 Functional dependence

Let $u = u(x,y)$ and $v = v(x,y)$. If we can write $u = g(v)$ or $v = h(u)$, then u and v are said to be *functionally dependent*. If functional dependence between u and v exists, then we can consider $f(u,v) = 0$. So,

$$\frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} = 0, \tag{1.26}$$

$$\frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} = 0. \tag{1.27}$$

In matrix form, this is

$$\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.28}$$

Since the right hand side is zero, and we desire a non-trivial solution, the determinant of the coefficient matrix must be zero for functional dependency, i.e.

$$\begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{vmatrix} = 0. \tag{1.29}$$

Note, since $\det \mathbf{J} = \det \mathbf{J}^T$, that this is equivalent to

$$J = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} = \frac{\partial(u,v)}{\partial(x,y)} = 0. \tag{1.30}$$

That is, the Jacobian determinant J must be zero for functional dependence.

Example 1.2

Determine if

$$\begin{aligned} u &= y + z, & (1.31)\\ v &= x + 2z^2, & (1.32)\\ w &= x - 4yz - 2y^2, & (1.33) \end{aligned}$$

are functionally dependent.

The determinant of the resulting coefficient matrix, by extension to three functions of three variables, is

$$\begin{aligned} \frac{\partial(u,v,w)}{\partial(x,y,z)} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} & \frac{\partial w}{\partial z} \end{vmatrix} &= \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} & \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} & \frac{\partial v}{\partial z} & \frac{\partial w}{\partial z} \end{vmatrix}, & (1.34)\\ &= \begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & -4(y+z) \\ 1 & 4z & -4y \end{vmatrix}, & (1.35)\\ &= (-1)(-4y - (-4)(y+z)) + (1)(4z), & (1.36)\\ &= 4y - 4y - 4z + 4z, & (1.37)\\ &= 0. & (1.38) \end{aligned}$$

So, u, v, w are functionally dependent. In fact $w = v - 2u^2$.

Example 1.3

Let

$$\begin{aligned} x + y + z &= 0, & (1.39)\\ x^2 + y^2 + z^2 + 2xz &= 1. & (1.40) \end{aligned}$$

Can x and y be considered as functions of z?

If $x = x(z)$ and $y = y(z)$, then dx/dz and dy/dz must exist. If we take

$$\begin{aligned} f(x,y,z) &= x + y + z = 0, & (1.41)\\ g(x,y,z) &= x^2 + y^2 + z^2 + 2xz - 1 = 0, & (1.42)\\ df = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy &= 0, & (1.43) \end{aligned}$$

$$dg = \frac{\partial g}{\partial z}dz + \frac{\partial g}{\partial x}dx + \frac{\partial g}{\partial y}dy = 0, \tag{1.44}$$

$$\frac{\partial f}{\partial z} + \frac{\partial f}{\partial x}\frac{dx}{dz} + \frac{\partial f}{\partial y}\frac{dy}{dz} = 0, \tag{1.45}$$

$$\frac{\partial g}{\partial z} + \frac{\partial g}{\partial x}\frac{dx}{dz} + \frac{\partial g}{\partial y}\frac{dy}{dz} = 0, \tag{1.46}$$

$$\begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{dx}{dz} \\ \frac{dy}{dz} \end{pmatrix} = \begin{pmatrix} -\frac{\partial f}{\partial z} \\ -\frac{\partial g}{\partial z} \end{pmatrix}, \tag{1.47}$$

then the solution matrix $(dx/dz, dy/dz)^T$ can be obtained by Cramer's rule:

$$\frac{dx}{dz} = \frac{\begin{vmatrix} -\frac{\partial f}{\partial z} & \frac{\partial f}{\partial y} \\ -\frac{\partial g}{\partial z} & \frac{\partial g}{\partial y} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}} = \frac{\begin{vmatrix} -1 & 1 \\ -(2z+2x) & 2y \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 2x+2z & 2y \end{vmatrix}} = \frac{-2y+2z+2x}{2y-2x-2z} = -1, \tag{1.48}$$

$$\frac{dy}{dz} = \frac{\begin{vmatrix} \frac{\partial f}{\partial x} & -\frac{\partial f}{\partial z} \\ \frac{\partial g}{\partial x} & -\frac{\partial g}{\partial z} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}} = \frac{\begin{vmatrix} 1 & -1 \\ 2x+2z & -(2z+2x) \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 2x+2z & 2y \end{vmatrix}} = \frac{0}{2y-2x-2z}. \tag{1.49}$$

Note here that in the expression for dx/dz that the numerator and denominator cancel; there is no special condition defined by the Jacobian determinant of the denominator being zero. In the second, $dy/dz = 0$ if $y - x - z \neq 0$, in which case this formula cannot give us the derivative.

Now, in fact, it is easily shown by algebraic manipulations (which for more general functions are not possible) that

$$x(z) = -z \pm \frac{\sqrt{2}}{2}, \tag{1.50}$$

$$y(z) = \mp\frac{\sqrt{2}}{2}. \tag{1.51}$$

This forms two distinct lines in x, y, z space. Note that on the lines of intersection of the two surfaces that $J = 2y - 2x - 2z = \mp 2\sqrt{2}$, which is never indeterminate.

The two original functions and their loci of intersection are plotted in Fig. 1.1. It is seen that the surface represented by the linear function, Eq. (1.39), is a plane, and that represented by the quadratic function, Eq. (1.40), is an open cylindrical tube. Note that planes and cylinders may or may not intersect. If they intersect, it is most likely that the intersection will be a closed arc. However, when the plane is aligned with the axis of the cylinder, the intersection will be two non-intersecting lines; such is the case in this example.

Let us see how slightly altering the equation for the plane removes the degeneracy. Take now

$$5x + y + z = 0, \tag{1.52}$$

$$x^2 + y^2 + z^2 + 2xz = 1. \tag{1.53}$$

Can x and y be considered as functions of z? If $x = x(z)$ and $y = y(z)$, then dx/dz and dy/dz must exist. If we take

$$f(x, y, z) = 5x + y + z = 0, \tag{1.54}$$

$$g(x, y, z) = x^2 + y^2 + z^2 + 2xz - 1 = 0, \tag{1.55}$$

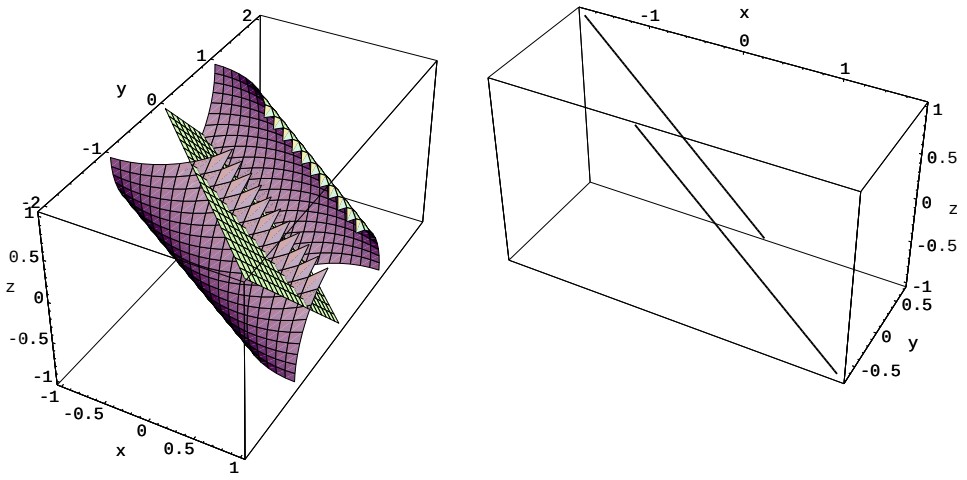

Figure 1.1: Surfaces of $x+y+z=0$ and $x^2+y^2+z^2+2xz=1$, and their loci of intersection.

then the solution matrix $(dx/dz, dy/dz)^T$ is found as before:

$$\frac{dx}{dz} = \frac{\begin{vmatrix} -\frac{\partial f}{\partial z} & \frac{\partial f}{\partial y} \\ -\frac{\partial g}{\partial z} & \frac{\partial g}{\partial y} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}} = \frac{\begin{vmatrix} -1 & 1 \\ -(2z+2x) & 2y \end{vmatrix}}{\begin{vmatrix} 5 & 1 \\ 2x+2z & 2y \end{vmatrix}} = \frac{-2y+2z+2x}{10y-2x-2z}, \tag{1.56}$$

$$\frac{dy}{dz} = \frac{\begin{vmatrix} \frac{\partial f}{\partial x} & -\frac{\partial f}{\partial z} \\ \frac{\partial g}{\partial x} & -\frac{\partial g}{\partial z} \end{vmatrix}}{\begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}} = \frac{\begin{vmatrix} 5 & -1 \\ 2x+2z & -(2z+2x) \end{vmatrix}}{\begin{vmatrix} 5 & 1 \\ 2x+2z & 2y \end{vmatrix}} = \frac{-8x-8z}{10y-2x-2z}. \tag{1.57}$$

The two original functions and their loci of intersection are plotted in Fig. 1.2.

Straightforward algebra in this case shows that an explicit dependency exists:

$$x(z) = \frac{-6z \pm \sqrt{2}\sqrt{13-8z^2}}{26}, \tag{1.58}$$

$$y(z) = \frac{-4z \mp 5\sqrt{2}\sqrt{13-8z^2}}{26}. \tag{1.59}$$

These curves represent the projection of the curve of intersection on the x, z and y, z planes, respectively. In both cases, the projections are ellipses.

1.3 Coordinate transformations

Many problems are formulated in three-dimensional Cartesian[3] space. However, many of these problems, especially those involving curved geometrical bodies, are more efficiently

[3]René Descartes, 1596-1650, French mathematician and philosopher.

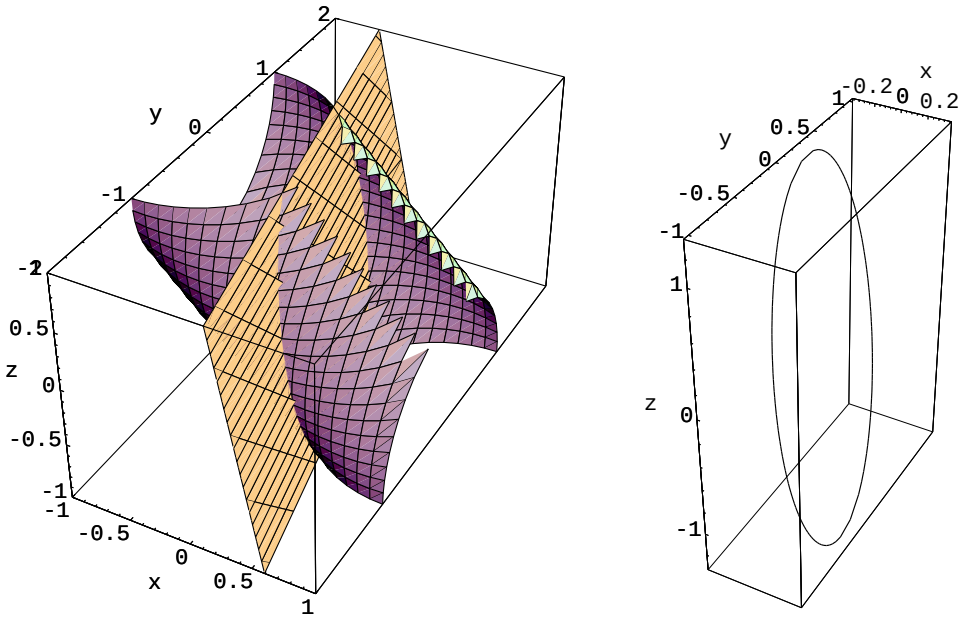

Figure 1.2: Surfaces of $5x+y+z=0$ and $x^2+y^2+z^2+2xz=1$, and their loci of intersection.

posed in a non-Cartesian, curvilinear coordinate system. To facilitate analysis involving such geometries, one needs techniques to transform from one coordinate system to another.

For this section, we will utilize an index notation, introduced by Einstein.[4] We will take untransformed Cartesian coordinates to be represented by (ξ^1, ξ^2, ξ^3). Here the superscript is an index and does not represent a power of ξ. We will denote this point by ξ^i, where $i = 1, 2, 3$. Because the space is Cartesian, we have the usual Euclidean[5] distance from Pythagoras'[6] theorem for a differential arc length ds:

$$(ds)^2 = \left(d\xi^1\right)^2 + \left(d\xi^2\right)^2 + \left(d\xi^3\right)^2, \tag{1.60}$$

$$(ds)^2 = \sum_{i=1}^{3} d\xi^i d\xi^i \equiv d\xi^i d\xi^i. \tag{1.61}$$

Here we have adopted Einstein's summation convention that when an index appears twice, a summation from 1 to 3 is understood. Though it makes little difference here, to strictly adhere to the conventions of the Einstein notation, which require a balance of sub- and superscripts, we should more formally take

$$(ds)^2 = d\xi^j \delta_{ji} d\xi^i = d\xi_i d\xi^i, \tag{1.62}$$

[4]Albert Einstein, 1879-1955, German/American physicist and mathematician.

[5]Euclid of Alexandria, $\sim$ 325 B.C.-$\sim$ 265 B.C., Greek geometer.

[6]Pythagoras of Samos, c. 570-c. 490 BC, Ionian Greek mathematician, philosopher, and mystic to whom this theorem is traditionally credited.

where δ_{ji} is the *Kronecker*[7] *delta*,

$$\delta_{ji} = \delta^{ji} = \delta^i_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1.63}$$

In matrix form, the Kronecker delta is simply the identity matrix $\mathbf{I}$, e.g.

$$\delta_{ji} = \delta^{ji} = \delta^i_j = \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.64}$$

Now let us consider a point P whose representation in Cartesian coordinates is (ξ^1, ξ^2, ξ^3) and map those coordinates so that it is now represented in a more convenient (x^1, x^2, x^3) space. This mapping is achieved by defining the following functional dependencies:

$$x^1 = x^1(\xi^1, \xi^2, \xi^3), \tag{1.65}$$

$$x^2 = x^2(\xi^1, \xi^2, \xi^3), \tag{1.66}$$

$$x^3 = x^3(\xi^1, \xi^2, \xi^3). \tag{1.67}$$

We note that in this example we make the common presumption that the entity P is invariant and that it has different representations in different coordinate systems. Thus, the coordinate axes change, but the location of P does not. This is known as an *alias* transformation. This contrasts another common approach in which a point is represented in an original space, and after application of a transformation, it is again represented in the original space in an altered state. This is known as an *alibi* transformation. The alias approach transforms the axes; the alibi approach transforms the elements of the space.

Taking derivatives can tell us whether the inverse exists.

$$dx^1 = \frac{\partial x^1}{\partial \xi^1} d\xi^1 + \frac{\partial x^1}{\partial \xi^2} d\xi^2 + \frac{\partial x^1}{\partial \xi^3} d\xi^3 = \frac{\partial x^1}{\partial \xi^j} d\xi^j, \tag{1.68}$$

$$dx^2 = \frac{\partial x^2}{\partial \xi^1} d\xi^1 + \frac{\partial x^2}{\partial \xi^2} d\xi^2 + \frac{\partial x^2}{\partial \xi^3} d\xi^3 = \frac{\partial x^2}{\partial \xi^j} d\xi^j, \tag{1.69}$$

$$dx^3 = \frac{\partial x^3}{\partial \xi^1} d\xi^1 + \frac{\partial x^3}{\partial \xi^2} d\xi^2 + \frac{\partial x^3}{\partial \xi^3} d\xi^3 = \frac{\partial x^3}{\partial \xi^j} d\xi^j, \tag{1.70}$$

$$\begin{pmatrix} dx^1 \\ dx^2 \\ dx^3 \end{pmatrix} = \begin{pmatrix} \frac{\partial x^1}{\partial \xi^1} & \frac{\partial x^1}{\partial \xi^2} & \frac{\partial x^1}{\partial \xi^3} \\ \frac{\partial x^2}{\partial \xi^1} & \frac{\partial x^2}{\partial \xi^2} & \frac{\partial x^2}{\partial \xi^3} \\ \frac{\partial x^3}{\partial \xi^1} & \frac{\partial x^3}{\partial \xi^2} & \frac{\partial x^3}{\partial \xi^3} \end{pmatrix} \begin{pmatrix} d\xi^1 \\ d\xi^2 \\ d\xi^3 \end{pmatrix}, \tag{1.71}$$

$$dx^i = \frac{\partial x^i}{\partial \xi^j} d\xi^j. \tag{1.72}$$

In order for the inverse to exist we must have a non-zero Jacobian determinant for the transformation, i.e.

$$\frac{\partial(x^1, x^2, x^3)}{\partial(\xi^1, \xi^2, \xi^3)} \neq 0. \tag{1.73}$$

[7]Leopold Kronecker, 1823-1891, German/Prussian mathematician.

As long as Eq. (1.73) is satisfied, the inverse transformation exists:

$$\xi^1 = \xi^1(x^1, x^2, x^3), \tag{1.74}$$
$$\xi^2 = \xi^2(x^1, x^2, x^3), \tag{1.75}$$
$$\xi^3 = \xi^3(x^1, x^2, x^3). \tag{1.76}$$

Likewise then,

$$d\xi^i = \frac{\partial \xi^i}{\partial x^j} dx^j. \tag{1.77}$$

1.3.1 Jacobian matrices and metric tensors

Defining the Jacobian matrix[8] $\mathbf{J}$ to be associated with the inverse transformation, Eq. (1.77), we take

$$\mathbf{J} = \frac{\partial \xi^i}{\partial x^j} = \begin{pmatrix} \frac{\partial \xi^1}{\partial x^1} & \frac{\partial \xi^1}{\partial x^2} & \frac{\partial \xi^1}{\partial x^3} \\ \frac{\partial \xi^2}{\partial x^1} & \frac{\partial \xi^2}{\partial x^2} & \frac{\partial \xi^2}{\partial x^3} \\ \frac{\partial \xi^3}{\partial x^1} & \frac{\partial \xi^3}{\partial x^2} & \frac{\partial \xi^3}{\partial x^3} \end{pmatrix}. \tag{1.78}$$

We can then rewrite $d\xi^i$ from Eq. (1.77) in Gibbs'[9] vector notation as

$$d\boldsymbol{\xi} = \mathbf{J} \cdot d\mathbf{x}. \tag{1.79}$$

Now for Euclidean spaces, distance must be independent of coordinate systems, so we require

$$(ds)^2 = d\xi^i d\xi^i = \left(\frac{\partial \xi^i}{\partial x^k} dx^k\right)\left(\frac{\partial \xi^i}{\partial x^l} dx^l\right) = dx^k \underbrace{\frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^i}{\partial x^l}}_{g_{kl}} dx^l. \tag{1.80}$$

In Gibbs' vector notation Eq. (1.80) becomes[10]

$$(ds)^2 = d\boldsymbol{\xi}^T \cdot d\boldsymbol{\xi}, \tag{1.81}$$
$$= (\mathbf{J} \cdot d\mathbf{x})^T \cdot (\mathbf{J} \cdot d\mathbf{x}). \tag{1.82}$$

[8]The definition we adopt influences the form of many of our formulæ given throughout the remainder of these notes. There are three obvious alternates: i) An argument can be made that a better definition of $\mathbf{J}$ would be the transpose of our Jacobian matrix: $\mathbf{J} \to \mathbf{J}^T$. This is because when one considers that the differential operator acts *first*, the Jacobian matrix is really $\frac{\partial}{\partial x^j}\xi^i$, and the alternative definition is more consistent with traditional matrix notation, which would have the first row as $(\frac{\partial}{\partial x^1}\xi^1, \frac{\partial}{\partial x^1}\xi^2, \frac{\partial}{\partial x^1}\xi^3)$, ii) Many others, e.g. Kay, adopt as $\mathbf{J}$ the inverse of our Jacobian matrix: $\mathbf{J} \to \mathbf{J}^{-1}$. This Jacobian matrix is thus defined in terms of the forward transformation, $\partial x^i/\partial \xi^j$, or iii) One could adopt $\mathbf{J} \to (\mathbf{J}^T)^{-1}$. As long as one realizes the implications of the notation, however, the convention adopted ultimately does not matter.

[9]Josiah Willard Gibbs, 1839-1903, prolific American mechanical engineer and mathematician with a lifetime affiliation with Yale University as well as the recipient of the first American doctorate in engineering.

[10]Common alternate formulations of vector mechanics of non-Cartesian spaces view the Jacobian as an *intrinsic* part of the dot product and would say instead that by definition $(ds)^2 = d\mathbf{x} \cdot d\mathbf{x}$. Such formulations have no need for the transpose operation, especially since they do not carry forward simply to non-Cartesian systems. The formulation used here has the advantage of *explicitly recognizing* the linear algebra operations necessary to form the scalar ds. These same alternate notations reserve the dot product for that between a vector and a vector and would hold instead that $d\boldsymbol{\xi} = \mathbf{J}d\mathbf{x}$. However, this could be confused with raising the dimension of the quantity of interest; whereas we use the dot to lower the dimension.

Now, it can be shown that $(\mathbf{J} \cdot d\mathbf{x})^T = d\mathbf{x}^T \cdot \mathbf{J}^T$ (see also Sec. 8.2.3.5), so

$$(ds)^2 = d\mathbf{x}^T \cdot \underbrace{\mathbf{J}^T \cdot \mathbf{J}}_{\mathbf{G}} \cdot d\mathbf{x}. \tag{1.83}$$

If we define the *metric tensor*, g_{kl} or $\mathbf{G}$, as follows:

$$g_{kl} = \frac{\partial \xi^i}{\partial x^k} \frac{\partial \xi^i}{\partial x^l}, \tag{1.84}$$

$$\mathbf{G} = \mathbf{J}^T \cdot \mathbf{J}, \tag{1.85}$$

then we have, equivalently in both Einstein and Gibbs notations,

$$(ds)^2 = dx^k g_{kl} dx^l, \tag{1.86}$$

$$(ds)^2 = d\mathbf{x}^T \cdot \mathbf{G} \cdot d\mathbf{x}. \tag{1.87}$$

Note that in Einstein notation, one can loosely imagine super-scripted terms in a denominator as being sub-scripted terms in a corresponding numerator. Now g_{kl} can be represented as a matrix. If we define

$$g = \det g_{kl}, \tag{1.88}$$

it can be shown that the ratio of volumes of differential elements in one space to that of the other is given by

$$d\xi^1 \; d\xi^2 \; d\xi^3 = \sqrt{g} \; dx^1 \; dx^2 \; dx^3. \tag{1.89}$$

Thus, transformations for which $g = 1$ are volume-preserving. Volume-preserving transformations also have $J = \det \mathbf{J} = \pm 1$. It can also be shown that if $J = \det \mathbf{J} > 0$, the transformation is locally orientation-preserving. If $J = \det \mathbf{J} < 0$, the transformation is orientation-reversing, and thus involves a reflection. So, if $J = \det \mathbf{J} = 1$, the transformation is volume- and orientation-preserving.

We also require dependent variables and all derivatives to take on the same values at corresponding points in each space, e.g. if ϕ ($\phi = f(\xi^1, \xi^2, \xi^3) = h(x^1, x^2, x^3)$) is a dependent variable defined at $(\hat{\xi}^1, \hat{\xi}^2, \hat{\xi}^3)$, and $(\hat{\xi}^1, \hat{\xi}^2, \hat{\xi}^3)$ maps into $(\hat{x}^1, \hat{x}^2, \hat{x}^3)$, we require $f(\hat{\xi}^1, \hat{\xi}^2, \hat{\xi}^3) = h(\hat{x}^1, \hat{x}^2, \hat{x}^3)$. The chain rule lets us transform derivatives to other spaces:

$$\begin{pmatrix} \frac{\partial \phi}{\partial x^1} & \frac{\partial \phi}{\partial x^2} & \frac{\partial \phi}{\partial x^3} \end{pmatrix} = \begin{pmatrix} \frac{\partial \phi}{\partial \xi^1} & \frac{\partial \phi}{\partial \xi^2} & \frac{\partial \phi}{\partial \xi^3} \end{pmatrix} \underbrace{\begin{pmatrix} \frac{\partial \xi^1}{\partial x^1} & \frac{\partial \xi^1}{\partial x^2} & \frac{\partial \xi^1}{\partial x^3} \\ \frac{\partial \xi^2}{\partial x^1} & \frac{\partial \xi^2}{\partial x^2} & \frac{\partial \xi^2}{\partial x^3} \\ \frac{\partial \xi^3}{\partial x^1} & \frac{\partial \xi^3}{\partial x^2} & \frac{\partial \xi^3}{\partial x^3} \end{pmatrix}}_{\mathbf{J}}, \tag{1.90}$$

$$\frac{\partial \phi}{\partial x^i} = \frac{\partial \phi}{\partial \xi^j} \frac{\partial \xi^j}{\partial x^i}. \tag{1.91}$$

Equation (1.91) can also be inverted, given that $g \neq 0$, to find $(\partial\phi/\partial\xi^1, \partial\phi/\partial\xi^2, \partial\phi/\partial\xi^3)$.

Employing Gibbs notation[11] we can write Eq. (1.91) as

$$\nabla_{\mathbf{x}}^T \phi = \nabla_{\boldsymbol{\xi}}^T \phi \cdot \mathbf{J}. \tag{1.92}$$

The fact that the gradient operator required the use of row vectors in conjunction with the Jacobian matrix, while the transformation of distance, earlier in this section, Eq. (1.79), required the use of column vectors is of fundamental importance, and will be soon examined further in Sec. 1.3.2 where we distinguish between what are known as *covariant* and *contravariant* vectors.

Transposing both sides of Eq. (1.92), we could also say

$$\nabla_{\mathbf{x}} \phi = \mathbf{J}^T \cdot \nabla_{\boldsymbol{\xi}} \phi. \tag{1.93}$$

Inverting, we then have

$$\nabla_{\boldsymbol{\xi}} \phi = (\mathbf{J}^T)^{-1} \cdot \nabla_{\mathbf{x}} \phi. \tag{1.94}$$

Thus, in general, we could say for the gradient operator

$$\nabla_{\boldsymbol{\xi}} = (\mathbf{J}^T)^{-1} \cdot \nabla_{\mathbf{x}}. \tag{1.95}$$

Contrasting Eq. (1.95) with Eq. (1.79), $d\boldsymbol{\xi} = \mathbf{J} \cdot d\mathbf{x}$, we see the gradient operation transforms in a fundamentally different way than the differential operation d, unless we restrict attention to an unusual $\mathbf{J}$, one whose transpose is equal to its inverse. We will sometimes make this restriction, and sometimes not. When we choose such a special $\mathbf{J}$, there will be many additional simplifications in the analysis; these are realized because it will be seen for many such transformations that nearly all of the original Cartesian character will be retained, albeit in a rotated, but otherwise undeformed, coordinate system. We shall later identify a matrix whose transpose is equal to its inverse as an orthogonal matrix, $\mathbf{Q}$: $\mathbf{Q}^T = \mathbf{Q}^{-1}$ and study it in detail in Secs. 6.2.1, 8.6.

One can also show the relation between $\partial \xi^i / \partial x^j$ and $\partial x^i / \partial \xi^j$ to be

$$\frac{\partial \xi^i}{\partial x^j} = \left(\left(\frac{\partial x^i}{\partial \xi^j} \right)^T \right)^{-1} = \left(\frac{\partial x^j}{\partial \xi^i} \right)^{-1}, \tag{1.96}$$

$$\begin{pmatrix} \frac{\partial \xi^1}{\partial x^1} & \frac{\partial \xi^1}{\partial x^2} & \frac{\partial \xi^1}{\partial x^3} \\ \frac{\partial \xi^2}{\partial x^1} & \frac{\partial \xi^2}{\partial x^2} & \frac{\partial \xi^2}{\partial x^3} \\ \frac{\partial \xi^3}{\partial x^1} & \frac{\partial \xi^3}{\partial x^2} & \frac{\partial \xi^3}{\partial x^3} \end{pmatrix} = \begin{pmatrix} \frac{\partial x^1}{\partial \xi^1} & \frac{\partial x^1}{\partial \xi^2} & \frac{\partial x^1}{\partial \xi^3} \\ \frac{\partial x^2}{\partial \xi^1} & \frac{\partial x^2}{\partial \xi^2} & \frac{\partial x^2}{\partial \xi^3} \\ \frac{\partial x^3}{\partial \xi^1} & \frac{\partial x^3}{\partial \xi^2} & \frac{\partial x^3}{\partial \xi^3} \end{pmatrix}^{-1}. \tag{1.97}$$

[11] In Cartesian coordinates, we take $\nabla_{\boldsymbol{\xi}} \equiv \begin{pmatrix} \frac{\partial}{\partial \xi^1} \\ \frac{\partial}{\partial \xi^2} \\ \frac{\partial}{\partial \xi^3} \end{pmatrix}$. This gives rise to the natural, albeit unconventional, notation $\nabla_{\boldsymbol{\xi}}^T = \begin{pmatrix} \frac{\partial}{\partial \xi^1} & \frac{\partial}{\partial \xi^2} & \frac{\partial}{\partial \xi^3} \end{pmatrix}$. This notion does not extend easily to non-Cartesian systems, for which index notation is preferred. Here, for convenience, we will take $\nabla_{\mathbf{x}}^T \equiv \begin{pmatrix} \frac{\partial}{\partial x^1} & \frac{\partial}{\partial x^2} & \frac{\partial}{\partial x^3} \end{pmatrix}$, and a similar column version for $\nabla_{\mathbf{x}}$.

Thus, the Jacobian matrix $\mathbf{J}$ of the transformation is simply the inverse of the Jacobian matrix of the inverse transformation. Note that in the very special case for which the transpose is the inverse, that we can replace the inverse by the transpose. Note that the transpose of the transpose is the original matrix and determines that $\partial\xi^i/\partial x^j = \partial x^i/\partial\xi^j$. This allows the i to remain "upstairs" and the j to remain "downstairs." Such a transformation will be seen to be a pure rotation or reflection.

Example 1.4

Transform the Cartesian equation

$$\frac{\partial\phi}{\partial\xi^1} + \frac{\partial\phi}{\partial\xi^2} = \left(\xi^1\right)^2 + \left(\xi^2\right)^2. \tag{1.98}$$

under the following:

1. Cartesian to linearly homogeneous affine coordinates.

Consider the following linear non-orthogonal transformation:

$$x^1 = \frac{2}{3}\xi^1 + \frac{2}{3}\xi^2, \tag{1.99}$$

$$x^2 = -\frac{2}{3}\xi^1 + \frac{1}{3}\xi^2, \tag{1.100}$$

$$x^3 = \xi^3. \tag{1.101}$$

This transformation is of the class of *affine* transformations, which are of the form

$$x^i = A^i_j\xi^j + b^i, \tag{1.102}$$

where A^i_j and b^i are constants. Affine transformations for which $b^i = 0$ are further distinguished as *linear homogeneous* transformations. The transformation of this example is both affine and linear homogeneous.

Equations (1.99-1.101) form a linear system of three equations in three unknowns; using standard techniques of linear algebra allows us to solve for ξ^1, ξ^2, ξ^3 in terms of x^1, x^2, x^3; that is, we find the inverse transformation, which is

$$\xi^1 = \frac{1}{2}x^1 - x^2, \tag{1.103}$$

$$\xi^2 = x^1 + x^2, \tag{1.104}$$

$$\xi^3 = x^3. \tag{1.105}$$

Lines of constant x^1 and x^2 in the ξ^1, ξ^2 plane as well as lines of constant ξ^1 and ξ^2 in the x^1, x^2 plane are plotted in Fig. 1.3. Also shown is a unit square in the Cartesian ξ^1, ξ^2 plane, with vertices A, B, C, D. The image of this rectangle is plotted as a parallelogram in the x^1, x^2 plane. It is seen the orientation has been preserved in what amounts to a clockwise rotation accompanied by stretching; moreover, the area (and thus the volume in three dimensions) has been decreased.

The appropriate Jacobian matrix for the inverse transformation is

$$\mathbf{J} = \frac{\partial\xi^i}{\partial x^j} = \begin{pmatrix} \frac{\partial\xi^1}{\partial x^1} & \frac{\partial\xi^1}{\partial x^2} & \frac{\partial\xi^1}{\partial x^3} \\ \frac{\partial\xi^2}{\partial x^1} & \frac{\partial\xi^2}{\partial x^2} & \frac{\partial\xi^2}{\partial x^3} \\ \frac{\partial\xi^3}{\partial x^1} & \frac{\partial\xi^3}{\partial x^2} & \frac{\partial\xi^3}{\partial x^3} \end{pmatrix}, \tag{1.106}$$

$$\mathbf{J} = \begin{pmatrix} \frac{1}{2} & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.107}$$

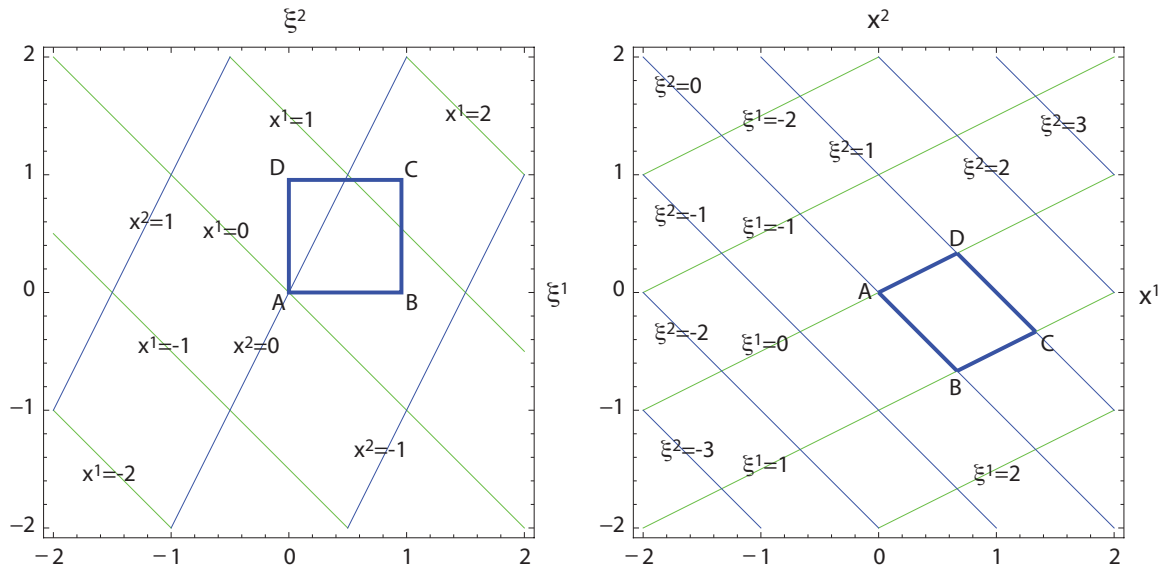

Figure 1.3: Lines of constant x^1 and x^2 in the ξ^1, ξ^2 plane and lines of constant ξ^1 and ξ^2 in the x^1, x^2 plane for the homogeneous affine transformation of example problem.

The Jacobian determinant is

$$J = \det \mathbf{J} = (1)\left(\left(\frac{1}{2}\right)(1) - (-1)(1)\right) = \frac{3}{2}. \tag{1.108}$$

So a unique transformation, $\boldsymbol{\xi} = \mathbf{J} \cdot \mathbf{x}$, always exists, since the Jacobian determinant is never zero. Inversion gives $\mathbf{x} = \mathbf{J}^{-1} \cdot \boldsymbol{\xi}$. Since $J > 0$, the transformation preserves the orientation of geometric entities. Since $J > 1$, a unit volume element in $\boldsymbol{\xi}$ space is larger than its image in $\mathbf{x}$ space.

The metric tensor is

$$g_{kl} = \frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^i}{\partial x^l} = \frac{\partial \xi^1}{\partial x^k}\frac{\partial \xi^1}{\partial x^l} + \frac{\partial \xi^2}{\partial x^k}\frac{\partial \xi^2}{\partial x^l} + \frac{\partial \xi^3}{\partial x^k}\frac{\partial \xi^3}{\partial x^l}. \tag{1.109}$$

For example for $k = 1, l = 1$ we get

$$g_{11} = \frac{\partial \xi^i}{\partial x^1}\frac{\partial \xi^i}{\partial x^1} = \frac{\partial \xi^1}{\partial x^1}\frac{\partial \xi^1}{\partial x^1} + \frac{\partial \xi^2}{\partial x^1}\frac{\partial \xi^2}{\partial x^1} + \frac{\partial \xi^3}{\partial x^1}\frac{\partial \xi^3}{\partial x^1}, \tag{1.110}$$

$$g_{11} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + (1)(1) + (0)(0) = \frac{5}{4}. \tag{1.111}$$

Repeating this operation for all terms of g_{kl}, we find the complete metric tensor is

$$g_{kl} = \begin{pmatrix} \frac{5}{4} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{1.112}$$

$$g = \det g_{kl} = (1)\left(\left(\frac{5}{4}\right)(2) - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right) = \frac{9}{4}. \tag{1.113}$$

This is equivalent to the calculation in Gibbs notation:

$$\mathbf{G} = \mathbf{J}^T \cdot \mathbf{J}, \tag{1.114}$$

$$\mathbf{G} = \begin{pmatrix} \frac{1}{2} & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{1.115}$$

$$\mathbf{G} = \begin{pmatrix} \frac{5}{4} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.116}$$

Distance in the transformed system is given by

$$(ds)^2 = dx^k \; g_{kl} \; dx^l, \tag{1.117}$$

$$(ds)^2 = d\mathbf{x}^T \cdot \mathbf{G} \cdot d\mathbf{x}, \tag{1.118}$$

$$(ds)^2 = \begin{pmatrix} dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} \frac{5}{4} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} dx^1 \\ dx^2 \\ dx^3 \end{pmatrix}, \tag{1.119}$$

$$(ds)^2 = \underbrace{\begin{pmatrix} (\frac{5}{4} \; dx^1 + \frac{1}{2} \; dx^2) & (\frac{1}{2} \; dx^1 + 2 \; dx^2) & dx^3 \end{pmatrix}}_{=dx_l = dx^k g_{kl}} \underbrace{\begin{pmatrix} dx^1 \\ dx^2 \\ dx^3 \end{pmatrix}}_{=dx^l} = dx_l dx^l, \tag{1.120}$$

$$(ds)^2 = \frac{5}{4} \left(dx^1\right)^2 + 2 \left(dx^2\right)^2 + \left(dx^3\right)^2 + dx^1 \; dx^2. \tag{1.121}$$

Detailed algebraic manipulation employing the so-called method of quadratic forms, to be discussed in Sec. 8.12, reveals that the previous equation can be rewritten as follows:

$$(ds)^2 = \frac{9}{20} \left(dx^1 + 2dx^2\right)^2 + \frac{1}{5} \left(-2dx^1 + dx^2\right)^2 + \left(dx^3\right)^2. \tag{1.122}$$

Direct expansion reveals the two forms for $(ds)^2$ to be identical. Note:

- The Jacobian matrix $\mathbf{J}$ is not symmetric.
- The metric tensor $\mathbf{G} = \mathbf{J}^T \cdot \mathbf{J}$ is symmetric.
- The fact that the metric tensor has non-zero off-diagonal elements is a consequence of the transformation being non-orthogonal.
- We identify here a new representation of the differential distance vector in the transformed space: $dx_l = dx^k g_{kl}$ whose significance will soon be discussed in Sec. 1.3.2.
- The distance is guaranteed to be positive. This will be true for all affine transformations in ordinary three-dimensional Euclidean space. In the generalized space-time continuum suggested by the theory of relativity, the generalized distance may in fact be negative; this generalized distance ds for an infinitesimal change in space and time is given by $ds^2 = \left(d\xi^1\right)^2 + \left(d\xi^2\right)^2 + \left(d\xi^3\right)^2 - \left(d\xi^4\right)^2$, where the first three coordinates are the ordinary Cartesian space coordinates and the fourth is $\left(d\xi^4\right)^2 = (c \; dt)^2$, where c is the speed of light.

Also we have the volume ratio of differential elements as

$$d\xi^1 \; d\xi^2 \; d\xi^3 = \sqrt{\frac{9}{4}} \; dx^1 \; dx^2 \; dx^3, \tag{1.123}$$

$$= \frac{3}{2} \; dx^1 \; dx^2 \; dx^3. \tag{1.124}$$

Now we use Eq. (1.94) to find the appropriate derivatives of ϕ. We first note that

$$(\mathbf{J}^T)^{-1} = \begin{pmatrix} \frac{1}{2} & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{2}{3} & -\frac{2}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.125}$$

So

$$\begin{pmatrix} \frac{\partial\phi}{\partial\xi^1} \\ \frac{\partial\phi}{\partial\xi^2} \\ \frac{\partial\phi}{\partial\xi^3} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & -\frac{2}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial\phi}{\partial x^1} \\ \frac{\partial\phi}{\partial x^2} \\ \frac{\partial\phi}{\partial x^3} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial x^1}{\partial\xi^1} & \frac{\partial x^2}{\partial\xi^1} & \frac{\partial x^3}{\partial\xi^1} \\ \frac{\partial x^1}{\partial\xi^2} & \frac{\partial x^2}{\partial\xi^2} & \frac{\partial x^3}{\partial\xi^2} \\ \frac{\partial x^1}{\partial\xi^3} & \frac{\partial x^2}{\partial\xi^3} & \frac{\partial x^3}{\partial\xi^3} \end{pmatrix}}_{(\mathbf{J}^T)^{-1}} \begin{pmatrix} \frac{\partial\phi}{\partial x^1} \\ \frac{\partial\phi}{\partial x^2} \\ \frac{\partial\phi}{\partial x^3} \end{pmatrix}. \tag{1.126}$$

Thus, by inspection,

$$\frac{\partial\phi}{\partial\xi^1} = \frac{2}{3}\frac{\partial\phi}{\partial x^1} - \frac{2}{3}\frac{\partial\phi}{\partial x^2}, \tag{1.127}$$

$$\frac{\partial\phi}{\partial\xi^2} = \frac{2}{3}\frac{\partial\phi}{\partial x^1} + \frac{1}{3}\frac{\partial\phi}{\partial x^2}. \tag{1.128}$$

So the transformed version of Eq. (1.98) becomes

$$\left(\frac{2}{3}\frac{\partial\phi}{\partial x^1} - \frac{2}{3}\frac{\partial\phi}{\partial x^2}\right) + \left(\frac{2}{3}\frac{\partial\phi}{\partial x^1} + \frac{1}{3}\frac{\partial\phi}{\partial x^2}\right) = \left(\frac{1}{2}x^1 - x^2\right)^2 + \left(x^1 + x^2\right)^2, \tag{1.129}$$

$$\frac{4}{3}\frac{\partial\phi}{\partial x^1} - \frac{1}{3}\frac{\partial\phi}{\partial x^2} = \frac{5}{4}\left(x^1\right)^2 + x^1x^2 + 2\left(x^2\right)^2. \tag{1.130}$$

2. Cartesian to cylindrical coordinates.

The transformations are

$$x^1 = \pm\sqrt{\left(\xi^1\right)^2 + \left(\xi^2\right)^2}, \tag{1.131}$$

$$x^2 = \tan^{-1}\left(\frac{\xi^2}{\xi^1}\right), \tag{1.132}$$

$$x^3 = \xi^3. \tag{1.133}$$

Here we have taken the unusual step of admitting negative x^1. This is admissible mathematically, but does not make sense according to our geometric intuition as it corresponds to a negative radius. Note further that this system of equations is non-linear, and that the transformation as defined is non-unique. For such systems, we cannot always find an explicit algebraic expression for the inverse transformation. In this case, some straightforward algebraic and trigonometric manipulation reveals that we can find an explicit representation of the inverse transformation, which is

$$\xi^1 = x^1\cos x^2, \tag{1.134}$$

$$\xi^2 = x^1\sin x^2, \tag{1.135}$$

$$\xi^3 = x^3. \tag{1.136}$$

Lines of constant x^1 and x^2 in the ξ^1, ξ^2 plane and lines of constant ξ^1 and ξ^2 in the x^1, x^2 plane are plotted in Fig. 1.4. Notice that the lines of constant x^1 are orthogonal to lines of constant x^2 in the Cartesian ξ^1, ξ^2 plane; the analog holds for the x^1, x^2 plane. For general transformations, this will not be the case. Also note that a square of area $1/2 \times 1/2$ is marked in the ξ^1, ξ^2 plane. Its image in the x^1, x^2 plane is also indicated. The non-uniqueness of the mapping from one plane to the other is evident.

The appropriate Jacobian matrix for the inverse transformation is

$$\mathbf{J} = \frac{\partial\xi^i}{\partial x^j} = \begin{pmatrix} \frac{\partial\xi^1}{\partial x^1} & \frac{\partial\xi^1}{\partial x^2} & \frac{\partial\xi^1}{\partial x^3} \\ \frac{\partial\xi^2}{\partial x^1} & \frac{\partial\xi^2}{\partial x^2} & \frac{\partial\xi^2}{\partial x^3} \\ \frac{\partial\xi^3}{\partial x^1} & \frac{\partial\xi^3}{\partial x^2} & \frac{\partial\xi^3}{\partial x^3} \end{pmatrix}, \tag{1.137}$$

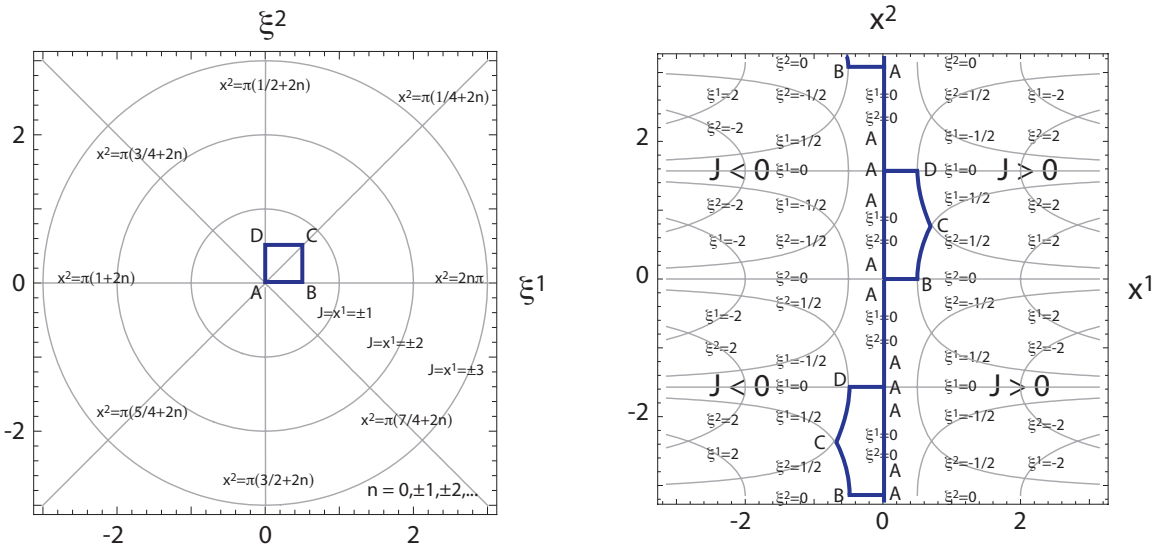

Figure 1.4: Lines of constant x^1 and x^2 in the ξ^1, ξ^2 plane and lines of constant ξ^1 and ξ^2 in the x^1, x^2 plane for cylindrical coordinates transformation of example problem.

$$\mathbf{J} = \begin{pmatrix} \cos x^2 & -x^1 \sin x^2 & 0 \\ \sin x^2 & x^1 \cos x^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.138}$$

The Jacobian determinant is

$$J = x^1 \cos^2 x^2 + x^1 \sin^2 x^2 = x^1. \tag{1.139}$$

So a unique transformation fails to exist when $x^1 = 0$. For $x^1 > 0$, the transformation is orientation-preserving. For $x^1 = 1$, the transformation is volume-preserving. For $x^1 < 0$, the transformation is orientation-reversing. This is a fundamental mathematical reason why we do not consider negative radius. It fails to preserve the orientation of a mapped element. For $x^1 \in (0, 1)$, a unit element in $\boldsymbol{\xi}$ space is smaller than a unit element in $\mathbf{x}$ space; the converse holds for $x^1 \in (1, \infty)$.

The metric tensor is

$$g_{kl} = \frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^i}{\partial x^l} = \frac{\partial \xi^1}{\partial x^k}\frac{\partial \xi^1}{\partial x^l} + \frac{\partial \xi^2}{\partial x^k}\frac{\partial \xi^2}{\partial x^l} + \frac{\partial \xi^3}{\partial x^k}\frac{\partial \xi^3}{\partial x^l}. \tag{1.140}$$

For example for $k = 1, l = 1$ we get

$$g_{11} = \frac{\partial \xi^i}{\partial x^1}\frac{\partial \xi^i}{\partial x^1} = \frac{\partial \xi^1}{\partial x^1}\frac{\partial \xi^1}{\partial x^1} + \frac{\partial \xi^2}{\partial x^1}\frac{\partial \xi^2}{\partial x^1} + \frac{\partial \xi^3}{\partial x^1}\frac{\partial \xi^3}{\partial x^1}, \tag{1.141}$$

$$g_{11} = \cos^2 x^2 + \sin^2 x^2 + 0 = 1. \tag{1.142}$$

Repeating this operation, we find the complete metric tensor is

$$g_{kl} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{1.143}$$

$$g = \det g_{kl} = \left(x^1\right)^2. \tag{1.144}$$

This is equivalent to the calculation in Gibbs notation:

$$\mathbf{G} = \mathbf{J}^T \cdot \mathbf{J}, \tag{1.145}$$

$$\mathbf{G} = \begin{pmatrix} \cos x^2 & \sin x^2 & 0 \\ -x^1 \sin x^2 & x^1 \cos x^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos x^2 & -x^1 \sin x^2 & 0 \\ \sin x^2 & x^1 \cos x^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{1.146}$$

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.147}$$

Distance in the transformed system is given by

$$(ds)^2 = dx^k \; g_{kl} \; dx^l, \tag{1.148}$$

$$(ds)^2 = d\mathbf{x}^T \cdot \mathbf{G} \cdot d\mathbf{x}, \tag{1.149}$$

$$(ds)^2 = \begin{pmatrix} dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} dx^1 \\ dx^2 \\ dx^3 \end{pmatrix}, \tag{1.150}$$

$$(ds)^2 = \underbrace{\begin{pmatrix} dx^1 & (x^1)^2 dx^2 & dx^3 \end{pmatrix}}_{dx_l = dx^k g_{kl}} \underbrace{\begin{pmatrix} dx^1 \\ dx^2 \\ dx^3 \end{pmatrix}}_{=dx^l} = dx_l dx^l, \tag{1.151}$$

$$(ds)^2 = \left(dx^1\right)^2 + \left(x^1 dx^2\right)^2 + \left(dx^3\right)^2. \tag{1.152}$$

Note:

- The fact that the metric tensor is diagonal can be attributed to the transformation being orthogonal.
- Since the product of any matrix with its transpose is guaranteed to yield a symmetric matrix, the metric tensor is always symmetric.

Also we have the volume ratio of differential elements as

$$d\xi^1 \; d\xi^2 \; d\xi^3 = x^1 \; dx^1 \; dx^2 \; dx^3. \tag{1.153}$$

Now we use Eq. (1.94) to find the appropriate derivatives of ϕ. We first note that

$$(\mathbf{J}^T)^{-1} = \begin{pmatrix} \cos x^2 & \sin x^2 & 0 \\ -x^1 \sin x^2 & x^1 \cos x^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \cos x^2 & -\frac{\sin x^2}{x^1} & 0 \\ \sin x^2 & \frac{\cos x^2}{x^1} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.154}$$

So

$$\begin{pmatrix} \frac{\partial \phi}{\partial \xi^1} \\ \frac{\partial \phi}{\partial \xi^2} \\ \frac{\partial \phi}{\partial \xi^3} \end{pmatrix} = \begin{pmatrix} \cos x^2 & -\frac{\sin x^2}{x^1} & 0 \\ \sin x^2 & \frac{\cos x^2}{x^1} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial \phi}{\partial x^1} \\ \frac{\partial \phi}{\partial x^2} \\ \frac{\partial \phi}{\partial x^3} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial x^1}{\partial \xi^1} & \frac{\partial x^2}{\partial \xi^1} & \frac{\partial x^3}{\partial \xi^1} \\ \frac{\partial x^1}{\partial \xi^2} & \frac{\partial x^2}{\partial \xi^2} & \frac{\partial x^3}{\partial \xi^2} \\ \frac{\partial x^1}{\partial \xi^3} & \frac{\partial x^2}{\partial \xi^3} & \frac{\partial x^3}{\partial \xi^3} \end{pmatrix}}_{(\mathbf{J}^T)^{-1}} \begin{pmatrix} \frac{\partial \phi}{\partial x^1} \\ \frac{\partial \phi}{\partial x^2} \\ \frac{\partial \phi}{\partial x^3} \end{pmatrix}. \tag{1.155}$$

Thus, by inspection,

$$\frac{\partial \phi}{\partial \xi^1} = \cos x^2 \frac{\partial \phi}{\partial x^1} - \frac{\sin x^2}{x^1} \frac{\partial \phi}{\partial x^2}, \tag{1.156}$$

$$\frac{\partial \phi}{\partial \xi^2} = \sin x^2 \frac{\partial \phi}{\partial x^1} + \frac{\cos x^2}{x^1} \frac{\partial \phi}{\partial x^2}. \tag{1.157}$$

So the transformed version of Eq. (1.98) becomes

$$\left(\cos x^2 \frac{\partial \phi}{\partial x^1} - \frac{\sin x^2}{x^1}\frac{\partial \phi}{\partial x^2}\right) + \left(\sin x^2 \frac{\partial \phi}{\partial x^1} + \frac{\cos x^2}{x^1}\frac{\partial \phi}{\partial x^2}\right) = \left(x^1\right)^2, \tag{1.158}$$

$$\left(\cos x^2 + \sin x^2\right)\frac{\partial \phi}{\partial x^1} + \left(\frac{\cos x^2 - \sin x^2}{x^1}\right)\frac{\partial \phi}{\partial x^2} = \left(x^1\right)^2. \tag{1.159}$$

1.3.2 Covariance and contravariance

Quantities known as *contravariant vectors* transform locally according to

$$\bar{u}^i = \frac{\partial \bar{x}^i}{\partial x^j} u^j. \tag{1.160}$$

We note that "local" refers to the fact that the transformation is locally linear. Eq. (1.160) is not a general recipe for a global transformation rule. Quantities known as *covariant vectors* transform locally according to

$$\bar{u}_i = \frac{\partial x^j}{\partial \bar{x}^i} u_j. \tag{1.161}$$

Here we have considered general transformations from one non-Cartesian coordinate system (x^1, x^2, x^3) to another $(\bar{x}^1, \bar{x}^2, \bar{x}^3)$. Note that indices associated with contravariant quantities appear as superscripts, and those associated with covariant quantities appear as subscripts.

In the special case where the barred coordinate system is Cartesian, we take U to denote the Cartesian vector and say

$$U^i = \frac{\partial \xi^i}{\partial x^j} u^j, \qquad U_i = \frac{\partial x^j}{\partial \xi^i} u_j. \tag{1.162}$$

Example 1.5

Let's say (x, y, z) is a normal Cartesian system and define the transformation

$$\bar{x} = \lambda x, \qquad \bar{y} = \lambda y, \qquad \bar{z} = \lambda z. \tag{1.163}$$

Now we can assign velocities in both the unbarred and barred systems:

$$u^x = \frac{dx}{dt}, \qquad u^y = \frac{dy}{dt}, \qquad u^z = \frac{dz}{dt}, \tag{1.164}$$

$$\bar{u}^{\bar{x}} = \frac{d\bar{x}}{dt}, \qquad \bar{u}^{\bar{y}} = \frac{d\bar{y}}{dt}, \qquad \bar{u}^{\bar{z}} = \frac{d\bar{z}}{dt}, \tag{1.165}$$

$$\bar{u}^{\bar{x}} = \frac{\partial \bar{x}}{\partial x}\frac{dx}{dt}, \qquad \bar{u}^{\bar{y}} = \frac{\partial \bar{y}}{\partial y}\frac{dy}{dt}, \qquad \bar{u}^{\bar{z}} = \frac{\partial \bar{z}}{\partial z}\frac{dz}{dt}, \tag{1.166}$$

$$\bar{u}^{\bar{x}} = \lambda u^x, \qquad \bar{u}^{\bar{y}} = \lambda u^y, \qquad \bar{u}^{\bar{z}} = \lambda u^z, \tag{1.167}$$

$$\bar{u}^{\bar{x}} = \frac{\partial \bar{x}}{\partial x} u^x, \qquad \bar{u}^{\bar{y}} = \frac{\partial \bar{y}}{\partial y} u^y, \qquad \bar{u}^{\bar{z}} = \frac{\partial \bar{z}}{\partial z} u^z. \tag{1.168}$$

This suggests the velocity vector is contravariant.

Now consider a vector which is the gradient of a function $f(x,y,z)$. For example, let

$$f(x,y,z) = x + y^2 + z^3, \tag{1.169}$$

$$u_x = \frac{\partial f}{\partial x}, \qquad u_y = \frac{\partial f}{\partial y}, \qquad u_z = \frac{\partial f}{\partial z}, \tag{1.170}$$

$$u_x = 1, \qquad u_y = 2y, \qquad u_z = 3z^2. \tag{1.171}$$

In the new coordinates

$$f\left(\frac{\bar{x}}{\lambda}, \frac{\bar{y}}{\lambda}, \frac{\bar{z}}{\lambda}\right) = \frac{\bar{x}}{\lambda} + \frac{\bar{y}^2}{\lambda^2} + \frac{\bar{z}^3}{\lambda^3}, \tag{1.172}$$

so

$$\bar{f}\left(\bar{x},\bar{y},\bar{z}\right) = \frac{\bar{x}}{\lambda} + \frac{\bar{y}^2}{\lambda^2} + \frac{\bar{z}^3}{\lambda^3}. \tag{1.173}$$

Now

$$\bar{u}_{\bar{x}} = \frac{\partial \bar{f}}{\partial \bar{x}}, \qquad \bar{u}_{\bar{y}} = \frac{\partial \bar{f}}{\partial \bar{y}}, \qquad \bar{u}_{\bar{z}} = \frac{\partial \bar{f}}{\partial \bar{z}}, \tag{1.174}$$

$$\bar{u}_{\bar{x}} = \frac{1}{\lambda}, \qquad \bar{u}_{\bar{y}} = \frac{2\bar{y}}{\lambda^2}, \qquad \bar{u}_{\bar{z}} = \frac{3\bar{z}^2}{\lambda^3}. \tag{1.175}$$

In terms of x, y, z, we have

$$\bar{u}_{\bar{x}} = \frac{1}{\lambda}, \qquad \bar{u}_{\bar{y}} = \frac{2y}{\lambda}, \qquad \bar{u}_{\bar{z}} = \frac{3z^2}{\lambda}. \tag{1.176}$$

So it is clear here that, in contrast to the velocity vector,

$$\bar{u}_{\bar{x}} = \frac{1}{\lambda} u_x, \qquad \bar{u}_{\bar{y}} = \frac{1}{\lambda} u_y, \qquad \bar{u}_{\bar{z}} = \frac{1}{\lambda} u_z. \tag{1.177}$$

More generally, we find for this case that

$$\bar{u}_{\bar{x}} = \frac{\partial x}{\partial \bar{x}} u_x, \qquad \bar{u}_{\bar{y}} = \frac{\partial y}{\partial \bar{y}} u_y, \qquad \bar{u}_{\bar{z}} = \frac{\partial z}{\partial \bar{z}} u_z, \tag{1.178}$$

which suggests the gradient vector is covariant.

Contravariant tensors transform locally according to

$$\bar{v}^{ij} = \frac{\partial \bar{x}^i}{\partial x^k} \frac{\partial \bar{x}^j}{\partial x^l} v^{kl}. \tag{1.179}$$

Covariant tensors transform locally according to

$$\bar{v}_{ij} = \frac{\partial x^k}{\partial \bar{x}^i} \frac{\partial x^l}{\partial \bar{x}^j} v_{kl}. \tag{1.180}$$

Mixed tensors transform locally according to

$$\bar{v}^i_j = \frac{\partial \bar{x}^i}{\partial x^k} \frac{\partial x^l}{\partial \bar{x}^j} v^k_l. \tag{1.181}$$

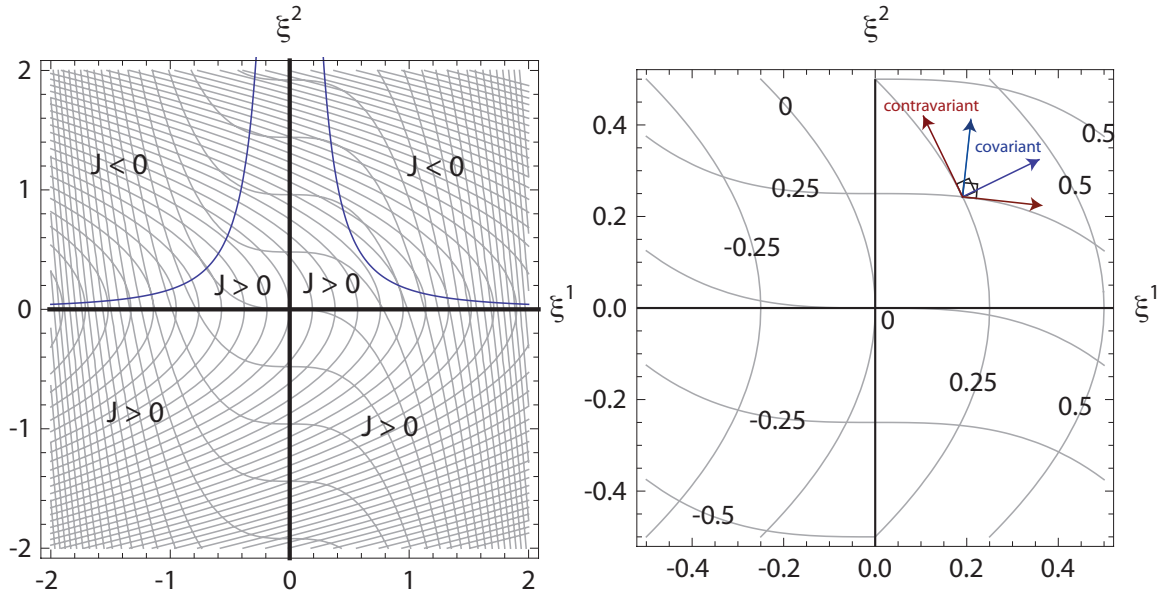

Figure 1.5: Contours for the transformation $x^1 = \xi^1 + (\xi^2)^2$, $x^2 = \xi^2 + (\xi^1)^3$ (left) and a blown-up version (right) including a pair of contravariant basis vectors, which are tangent to the contours, and covariant basis vectors, which are normal to the contours.

Recall that *variance* is another term for *gradient* and that *co-* denotes *with*. A vector which is co-variant is aligned with the variance or the gradient. Recalling next that *contra-* denotes *against*, a vector which is contra-variant is aligned against the variance or the gradient. This results in a set of contravariant basis vectors being tangent to lines of $x^i = C$, while covariant basis vectors are normal to lines of $x^i = C$. A vector in space has two natural representations, one on a contravariant basis, and the other on a covariant basis. The contravariant representation seems more natural because it is similar to the familiar $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ for Cartesian systems, though both can be used to obtain equivalent results.

For the transformation $x^1 = \xi^1 + (\xi^2)^2$, $x^2 = \xi^2 + (\xi^1)^3$, Figure 1.5 gives a plot of a set of lines of constant x^1 and x^2 in the Cartesian ξ^1, ξ^2 plane, along with a local set of contravariant and covariant basis vectors. Note the covariant basis vectors, because they are directly related to the gradient vector, point in the direction of most rapid change of x^1 and x^2 and are orthogonal to contours on which x^1 and x^2 are constant. The contravariant vectors are tangent to the contours. It can be shown that the contravariant vectors are aligned with the columns of $\mathbf{J}$, and the covariant vectors are aligned with the rows of $\mathbf{J}^{-1}$. This transformation has some special properties. Near the origin, the higher order terms become negligible, and the transformation reduces to the identity mapping $x^1 \sim \xi^1$, $x^2 \sim \xi^2$. As such, in the neighborhood of the origin, one has $\mathbf{J} = \mathbf{I}$, and there is no change in area or orientation of an element. Moreover, on each of the coordinate axes $x^1 = \xi^1$ and $x^2 = \xi^2$; additionally, on each of the coordinate axes $J = 1$, so in those special locations the transformation is area- and orientation-preserving. This non-linear transformation can be

shown to be singular where $J = 0$; this occurs when $\xi^2 = 1/(6(\xi^1)^2)$. As $J \to 0$, the contours of ξ^1 align more and more with the contours of ξ^2, and thus the contravariant basis vectors come closer to paralleling each other. When $J = 0$, the two contours of each osculate. At such points there is only one linearly independent contravariant basis vector, which is not enough to represent an arbitrary vector in a linear combination. An analog holds for the covariant basis vectors. In the first and fourth quadrants and some of the second and third, the transformation is orientation-reversing. The transformation is orientation-preserving in most of the second and third quadrants.

Example 1.6

Consider the vector fields defined in Cartesian coordinates by

$$\text{a)} \quad U^i = \begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix}, \qquad \text{b)} \quad U^i = \begin{pmatrix} \xi^1 \\ 2\xi^2 \end{pmatrix}. \tag{1.182}$$

At the point

$$P : \begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{1.183}$$

find the covariant and contravariant representations of both cases of U^i in cylindrical coordinates.

a) At P in the Cartesian system, we have the contravariant

$$U^i = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\Bigg|_{\xi_1=1,\xi_2=1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{1.184}$$

For a Cartesian coordinate system, the metric tensor $g_{ij} = \delta_{ij} = g_{ji} = \delta_{ji}$. Thus, the covariant representation in the Cartesian system is

$$U_j = g_{ji}U^i = \delta_{ji}U^i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{1.185}$$

Now consider cylindrical coordinates: $\xi^1 = x^1 \cos x^2$, $\xi^2 = x^1 \sin x^2$. For the inverse transformation, let us insist that $J > 0$, so $x^1 = \sqrt{(\xi^1)^2 + (\xi^2)^2}$, $x^2 = \tan^{-1}(\xi^2/\xi^1)$. Thus, at P we have a representation of

$$P : \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ \frac{\pi}{4} \end{pmatrix}. \tag{1.186}$$

For the transformation, we have

$$\mathbf{J} = \begin{pmatrix} \cos x^2 & -x^1 \sin x^2 \\ \sin x^2 & x^1 \cos x^2 \end{pmatrix}, \qquad \mathbf{G} = \mathbf{J}^T \cdot \mathbf{J} = \begin{pmatrix} 1 & 0 \\ 0 & (x^1)^2 \end{pmatrix}. \tag{1.187}$$

At P, we thus have

$$\mathbf{J} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -1 \\ \frac{\sqrt{2}}{2} & 1 \end{pmatrix}, \qquad \mathbf{G} = \mathbf{J}^T \cdot \mathbf{J} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{1.188}$$

Now, specializing Eq. (1.160) by considering the barred coordinate to be Cartesian, we can say

$$U^i = \frac{\partial \xi^i}{\partial x^j} u^j. \tag{1.189}$$

Locally, we can use the Gibbs notation and say $\mathbf{U} = \mathbf{J} \cdot \mathbf{u}$, and thus get $\mathbf{u} = \mathbf{J}^{-1} \cdot \mathbf{U}$, so that the contravariant representation is

$$\begin{pmatrix} u^1 \\ u^2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -1 \\ \frac{\sqrt{2}}{2} & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}. \tag{1.190}$$

In Gibbs notation, one can interpret this as $1\mathbf{i}+1\mathbf{j} = \sqrt{2}\mathbf{e}_r+0\mathbf{e}_\theta$. Note that this representation is different than the simple polar coordinates of P given by Eq. (1.186). Let us look closer at the cylindrical basis vectors $\mathbf{e}_r$ and $\mathbf{e}_\theta$. In cylindrical coordinates, the contravariant representations of the unit basis vectors must be $\mathbf{e}_r = (1,0)^T$ and $\mathbf{e}_\theta = (0,1)^T$. So in Cartesian coordinates those basis vectors are represented as

$$\mathbf{e}_r = \mathbf{J} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos x^2 & -x^1 \sin x^2 \\ \sin x^2 & x^1 \cos x^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos x^2 \\ \sin x^2 \end{pmatrix}, \tag{1.191}$$

$$\mathbf{e}_\theta = \mathbf{J} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \cos x^2 & -x^1 \sin x^2 \\ \sin x^2 & x^1 \cos x^2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -x^1 \sin x^2 \\ x^1 \cos x^2 \end{pmatrix}. \tag{1.192}$$

In general a unit vector in the transformed space is not a unit vector in the Cartesian space. Note that $\mathbf{e}_\theta$ is a unit vector in Cartesian space only when $x^1 = 1$; this is also the condition for $J = 1$. Lastly, we see the covariant representation is given by $u_j = u^i g_{ij}$. Since g_{ij} is symmetric, we can transpose this to get $u_j = g_{ji} u^i$:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mathbf{G} \cdot \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}. \tag{1.193}$$

This simple vector field has an identical contravariant and covariant representation. The appropriate invariant quantities are independent of the representation:

$$U_i U^i = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2, \tag{1.194}$$

$$u_i u^i = \begin{pmatrix} \sqrt{2} & 0 \end{pmatrix} \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = 2. \tag{1.195}$$

Thought tempting, we note that there is no clear way to form the representation $x_i x^i$ to demonstrate any additional invariance.

b) At P in the Cartesian system, we have the contravariant

$$U^i = \left. \begin{pmatrix} \xi_1 \\ 2\xi_2 \end{pmatrix} \right|_{\xi_1=1, \xi_2=1} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{1.196}$$

In the same fashion as demonstrated in part a), we find the contravariant representation of U^i in cylindrical coordinates at P is

$$\begin{pmatrix} u^1 \\ u^2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -1 \\ \frac{\sqrt{2}}{2} & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{2}} \\ \frac{1}{2} \end{pmatrix}. \tag{1.197}$$

In Gibbs notation, we could interpret this as $1\mathbf{i} + 2\mathbf{j} = (3/\sqrt{2})\mathbf{e}_r + (1/2)\mathbf{e}_\theta$.

The covariant representation is given once again by $u_j = g_{ji} u^i$:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mathbf{G} \cdot \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} \frac{3}{\sqrt{2}} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{2}} \\ 1 \end{pmatrix}. \tag{1.198}$$

This less simple vector field has distinct contravariant and covariant representations. However, the appropriate invariant quantities are independent of the representation:

$$U_i U^i = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 5, \tag{1.199}$$

$$u_i u^i = \begin{pmatrix} \frac{3}{\sqrt{2}} & 1 \end{pmatrix} \begin{pmatrix} \frac{3}{\sqrt{2}} \\ \frac{1}{2} \end{pmatrix} = 5. \tag{1.200}$$

The idea of covariant and contravariant derivatives play an important role in mathematical physics, namely in that the equations should be formulated such that they are invariant under coordinate transformations. This is not particularly difficult for Cartesian systems, but for non-orthogonal systems, one cannot use differentiation in the ordinary sense but must instead use the notion of covariant and contravariant derivatives, depending on the problem. The role of these terms was especially important in the development of the theory of relativity.

Consider a contravariant vector u^i defined in x^i which has corresponding components U^i in the Cartesian ξ^i. Take w^i_j and W^i_j to represent the covariant spatial derivative of u^i and U^i, respectively. Let's use the chain rule and definitions of tensorial quantities to arrive at a formula for covariant differentiation. From the definition of contravariance, Eq. (1.160),

$$U^i = \frac{\partial \xi^i}{\partial x^l} u^l. \tag{1.201}$$

Take the derivative in Cartesian space and then use the chain rule:

$$W^i_j = \frac{\partial U^i}{\partial \xi^j} = \frac{\partial U^i}{\partial x^k} \frac{\partial x^k}{\partial \xi^j}, \tag{1.202}$$

$$= \left(\frac{\partial}{\partial x^k} \underbrace{\left(\frac{\partial \xi^i}{\partial x^l} \, u^l \right)}_{=U^i} \right) \frac{\partial x^k}{\partial \xi^j}, \tag{1.203}$$

$$= \left(\frac{\partial^2 \xi^i}{\partial x^k \partial x^l} u^l + \frac{\partial \xi^i}{\partial x^l} \frac{\partial u^l}{\partial x^k} \right) \frac{\partial x^k}{\partial \xi^j}, \tag{1.204}$$

$$W^p_q = \left(\frac{\partial^2 \xi^p}{\partial x^k \partial x^l} u^l + \frac{\partial \xi^p}{\partial x^l} \frac{\partial u^l}{\partial x^k} \right) \frac{\partial x^k}{\partial \xi^q}. \tag{1.205}$$

From the definition of a mixed tensor, Eq. (1.181),

$$w^i_j = W^p_q \frac{\partial x^i}{\partial \xi^p} \frac{\partial \xi^q}{\partial x^j}, \tag{1.206}$$

$$= \underbrace{\left(\frac{\partial^2 \xi^p}{\partial x^k \partial x^l} u^l + \frac{\partial \xi^p}{\partial x^l}\frac{\partial u^l}{\partial x^k}\right)\frac{\partial x^k}{\partial \xi^q}}_{=W_q^p}\frac{\partial x^i}{\partial \xi^p}\frac{\partial \xi^q}{\partial x^j}, \tag{1.207}$$

$$= \frac{\partial^2 \xi^p}{\partial x^k \partial x^l}\frac{\partial x^k}{\partial \xi^q}\frac{\partial x^i}{\partial \xi^p}\frac{\partial \xi^q}{\partial x^j}u^l + \frac{\partial \xi^p}{\partial x^l}\frac{\partial x^k}{\partial \xi^q}\frac{\partial x^i}{\partial \xi^p}\frac{\partial \xi^q}{\partial x^j}\frac{\partial u^l}{\partial x^k}, \tag{1.208}$$

$$= \frac{\partial^2 \xi^p}{\partial x^k \partial x^l}\underbrace{\frac{\partial x^k}{\partial x^j}}_{\delta_j^k}\frac{\partial x^i}{\partial \xi^p}u^l + \underbrace{\frac{\partial x^i}{\partial x^l}}_{\delta_l^i}\underbrace{\frac{\partial x^k}{\partial x^j}}_{\delta_j^k}\frac{\partial u^l}{\partial x^k}, \tag{1.209}$$

$$= \frac{\partial^2 \xi^p}{\partial x^k \partial x^l}\delta_j^k\frac{\partial x^i}{\partial \xi^p}u^l + \delta_l^i\delta_j^k\frac{\partial u^l}{\partial x^k}, \tag{1.210}$$

$$= \frac{\partial^2 \xi^p}{\partial x^j \partial x^l}\frac{\partial x^i}{\partial \xi^p}u^l + \frac{\partial u^i}{\partial x^j}. \tag{1.211}$$

Here, we have used the identity that

$$\frac{\partial x^i}{\partial x^j} = \delta_j^i, \tag{1.212}$$

where δ_j^i is another form of the Kronecker delta. We define the *Christoffel*[12] *symbols* Γ_{jl}^i as follows:

$$\Gamma_{jl}^i = \frac{\partial^2 \xi^p}{\partial x^j \partial x^l}\frac{\partial x^i}{\partial \xi^p}, \tag{1.213}$$

and use the term Δ_j to represent the covariant derivative. Thus, the covariant derivative of a contravariant vector u^i is as follows:

$$\Delta_j u^i = w_j^i = \frac{\partial u^i}{\partial x^j} + \Gamma_{jl}^i u^l. \tag{1.214}$$

Example 1.7

Find $\nabla^T \cdot \mathbf{u}$ in cylindrical coordinates. The transformations are

$$x^1 = +\sqrt{(\xi^1)^2 + (\xi^2)^2}, \tag{1.215}$$

$$x^2 = \tan^{-1}\left(\frac{\xi^2}{\xi^1}\right), \tag{1.216}$$

$$x^3 = \xi^3. \tag{1.217}$$

The inverse transformation is

$$\xi^1 = x^1 \cos x^2, \tag{1.218}$$

$$\xi^2 = x^1 \sin x^2, \tag{1.219}$$

$$\xi^3 = x^3. \tag{1.220}$$

[12]Elwin Bruno Christoffel, 1829-1900, German mathematician.

This corresponds to finding

$$\Delta_i u^i = w_i^i = \frac{\partial u^i}{\partial x^i} + \Gamma^i_{il} u^l. \tag{1.221}$$

Now for $i = j$

$$\Gamma^i_{il} u^l = \frac{\partial^2 \xi^p}{\partial x^i \partial x^l} \frac{\partial x^i}{\partial \xi^p} u^l, \tag{1.222}$$

$$= \frac{\partial^2 \xi^1}{\partial x^i \partial x^l} \frac{\partial x^i}{\partial \xi^1} u^l + \frac{\partial^2 \xi^2}{\partial x^i \partial x^l} \frac{\partial x^i}{\partial \xi^2} u^l + \underbrace{\frac{\partial^2 \xi^3}{\partial x^i \partial x^l}}_{=0} \frac{\partial x^i}{\partial \xi^3} u^l. \tag{1.223}$$

Noting that all second partials of ξ^3 are zero,

$$\Gamma^i_{il} u^l = \frac{\partial^2 \xi^1}{\partial x^i \partial x^l} \frac{\partial x^i}{\partial \xi^1} u^l + \frac{\partial^2 \xi^2}{\partial x^i \partial x^l} \frac{\partial x^i}{\partial \xi^2} u^l. \tag{1.224}$$

Expanding the i summation,

$$\begin{aligned} \Gamma^i_{il} u^l = & \frac{\partial^2 \xi^1}{\partial x^1 \partial x^l} \frac{\partial x^1}{\partial \xi^1} u^l + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^l} \frac{\partial x^2}{\partial \xi^1} u^l + \frac{\partial^2 \xi^1}{\partial x^3 \partial x^l} \underbrace{\frac{\partial x^3}{\partial \xi^1}}_{=0} u^l \\ & + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^l} \frac{\partial x^1}{\partial \xi^2} u^l + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^l} \frac{\partial x^2}{\partial \xi^2} u^l + \frac{\partial^2 \xi^2}{\partial x^3 \partial x^l} \underbrace{\frac{\partial x^3}{\partial \xi^2}}_{=0} u^l. \end{aligned} \tag{1.225}$$

Noting that partials of x^3 with respect to ξ^1 and ξ^2 are zero,

$$\Gamma^i_{il} u^l = \frac{\partial^2 \xi^1}{\partial x^1 \partial x^l} \frac{\partial x^1}{\partial \xi^1} u^l + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^l} \frac{\partial x^2}{\partial \xi^1} u^l + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^l} \frac{\partial x^1}{\partial \xi^2} u^l + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^l} \frac{\partial x^2}{\partial \xi^2} u^l. \tag{1.226}$$

Expanding the l summation, we get

$$\begin{aligned} \Gamma^i_{il} u^l = & \frac{\partial^2 \xi^1}{\partial x^1 \partial x^1} \frac{\partial x^1}{\partial \xi^1} u^1 + \frac{\partial^2 \xi^1}{\partial x^1 \partial x^2} \frac{\partial x^1}{\partial \xi^1} u^2 + \underbrace{\frac{\partial^2 \xi^1}{\partial x^1 \partial x^3}}_{=0} \frac{\partial x^1}{\partial \xi^1} u^3 \\ & + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^1} \frac{\partial x^2}{\partial \xi^1} u^1 + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^2} \frac{\partial x^2}{\partial \xi^1} u^2 + \underbrace{\frac{\partial^2 \xi^1}{\partial x^2 \partial x^3}}_{=0} \frac{\partial x^2}{\partial \xi^1} u^3 \\ & + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^1} \frac{\partial x^1}{\partial \xi^2} u^1 + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^2} \frac{\partial x^1}{\partial \xi^2} u^2 + \underbrace{\frac{\partial^2 \xi^2}{\partial x^1 \partial x^3}}_{=0} \frac{\partial x^1}{\partial \xi^2} u^3 \\ & + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^1} \frac{\partial x^2}{\partial \xi^2} u^1 + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^2} \frac{\partial x^2}{\partial \xi^2} u^2 + \underbrace{\frac{\partial^2 \xi^2}{\partial x^2 \partial x^3}}_{=0} \frac{\partial x^2}{\partial \xi^2} u^3. \end{aligned} \tag{1.227}$$

Again removing the x^3 variation, we get

$$\begin{aligned} \Gamma^i_{il} u^l = & \frac{\partial^2 \xi^1}{\partial x^1 \partial x^1} \frac{\partial x^1}{\partial \xi^1} u^1 + \frac{\partial^2 \xi^1}{\partial x^1 \partial x^2} \frac{\partial x^1}{\partial \xi^1} u^2 + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^1} \frac{\partial x^2}{\partial \xi^1} u^1 + \frac{\partial^2 \xi^1}{\partial x^2 \partial x^2} \frac{\partial x^2}{\partial \xi^1} u^2 \\ & + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^1} \frac{\partial x^1}{\partial \xi^2} u^1 + \frac{\partial^2 \xi^2}{\partial x^1 \partial x^2} \frac{\partial x^1}{\partial \xi^2} u^2 + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^1} \frac{\partial x^2}{\partial \xi^2} u^1 + \frac{\partial^2 \xi^2}{\partial x^2 \partial x^2} \frac{\partial x^2}{\partial \xi^2} u^2. \end{aligned} \tag{1.228}$$

Substituting for the partial derivatives, we find

$$\begin{aligned}\Gamma^i_{il}u^l &= 0u^1 - \sin x^2 \cos x^2 u^2 - \sin x^2 \left(\frac{-\sin x^2}{x^1}\right) u^1 - x^1 \cos x^2 \left(\frac{-\sin x^2}{x^1}\right) u^2 \\ &\quad +0u^1 + \cos x^2 \sin x^2 u^2 + \cos x^2 \left(\frac{\cos x^2}{x^1}\right) u^1 - x^1 \sin x^2 \left(\frac{\cos x^2}{x^1}\right) u^2, \qquad (1.229)\\ &= \frac{u^1}{x^1}. \qquad (1.230)\end{aligned}$$

So, in cylindrical coordinates

$$\nabla^T \cdot \mathbf{u} = \frac{\partial u^1}{\partial x^1} + \frac{\partial u^2}{\partial x^2} + \frac{\partial u^3}{\partial x^3} + \frac{u^1}{x^1}. \qquad (1.231)$$

Note: In standard cylindrical notation, $x^1 = r, x^2 = \theta, x^3 = z$. Considering u to be a velocity vector, we get

$$\begin{aligned}\nabla^T \cdot \mathbf{u} &= \frac{\partial}{\partial r}\left(\frac{dr}{dt}\right) + \frac{\partial}{\partial \theta}\left(\frac{d\theta}{dt}\right) + \frac{\partial}{\partial z}\left(\frac{dz}{dt}\right) + \frac{1}{r}\left(\frac{dr}{dt}\right), \qquad (1.232)\\ \nabla^T \cdot \mathbf{u} &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{dr}{dt}\right) + \frac{1}{r}\frac{\partial}{\partial \theta}\left(r\frac{d\theta}{dt}\right) + \frac{\partial}{\partial z}\left(\frac{dz}{dt}\right), \qquad (1.233)\\ \nabla^T \cdot \mathbf{u} &= \frac{1}{r}\frac{\partial}{\partial r}\left(ru_r\right) + \frac{1}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{\partial u_z}{\partial z}. \qquad (1.234)\end{aligned}$$

Here we have also used the more traditional $u_\theta = r(d\theta/dt) = x^1 u^2$, along with $u_r = u^1, u_z = u^3$. For practical purposes, this insures that u_r, u_θ, u_z all have the same dimensions.

Example 1.8

Calculate the acceleration vector $d\mathbf{u}/dt$ in cylindrical coordinates.

Start by expanding the total derivative as

$$\frac{d\mathbf{u}}{dt} = \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}^T \cdot \nabla \mathbf{u}.$$

Now, we take $\mathbf{u}$ to be a contravariant velocity vector and the gradient operation to be a covariant derivative. Employ index notation to get

$$\begin{aligned}\frac{d\mathbf{u}}{dt} &= \frac{\partial u^i}{\partial t} + u^j \Delta_j u^i, \qquad (1.235)\\ &= \frac{\partial u^i}{\partial t} + u^j \left(\frac{\partial u^i}{\partial x^j} + \Gamma^i_{jl} u^l\right). \qquad (1.236)\end{aligned}$$

After an extended calculation similar to the previous example, one finds after expanding all terms that

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} \frac{\partial u^1}{\partial t} \\ \frac{\partial u^2}{\partial t} \\ \frac{\partial u^3}{\partial t} \end{pmatrix} + \begin{pmatrix} u^1 \frac{\partial u^1}{\partial x^1} + u^2 \frac{\partial u^1}{\partial x^2} + u^3 \frac{\partial u^1}{\partial x^3} \\ u^1 \frac{\partial u^2}{\partial x^1} + u^2 \frac{\partial u^2}{\partial x^2} + u^3 \frac{\partial u^2}{\partial x^3} \\ u^1 \frac{\partial u^3}{\partial x^1} + u^2 \frac{\partial u^3}{\partial x^2} + u^3 \frac{\partial u^3}{\partial x^3} \end{pmatrix} + \begin{pmatrix} -x^1 \left(u^2\right)^2 \\ 2\frac{u^1 u^2}{x^1} \\ 0 \end{pmatrix}. \qquad (1.237)$$

The last term is related to the well known Coriolis[13] and centripetal acceleration terms. However, these are not in the standard form to which most are accustomed. To arrive at that standard form, one must return to a so-called physical representation. Here again take $x^1 = r$, $x^2 = \theta$, and $x^3 = z$. Also take $u_r = dr/dt = u^1$, $u_\theta = r(d\theta/dt) = x^1 u^2$, $u_z = dz/dt = u^3$. Then the r acceleration equation becomes

$$\frac{du_r}{dt} = \frac{\partial u_r}{\partial t} + u_r \frac{\partial u_r}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_r}{\partial \theta} + u_z \frac{\partial u_r}{\partial z} - \underbrace{\frac{u_\theta^2}{r}}_{\text{centripetal}} . \tag{1.238}$$

Here the final term is the traditional centripetal acceleration. The θ acceleration is slightly more complicated. First one writes

$$\frac{d}{dt}\left(\frac{d\theta}{dt}\right) = \frac{\partial}{\partial t}\left(\frac{d\theta}{dt}\right) + \frac{dr}{dt}\frac{\partial}{\partial r}\left(\frac{d\theta}{dt}\right) + \frac{d\theta}{dt}\frac{\partial}{\partial \theta}\left(\frac{d\theta}{dt}\right) + \frac{dz}{dt}\frac{\partial}{\partial z}\left(\frac{d\theta}{dt}\right) + 2\frac{\frac{dr}{dt}\frac{d\theta}{dt}}{r}. \tag{1.239}$$

Now, here one is actually interested in du_θ/dt, so both sides are multiplied by r and then one operates to get

$$\begin{aligned}
\frac{du_\theta}{dt} &= r\frac{\partial}{\partial t}\left(\frac{d\theta}{dt}\right) + r\frac{dr}{dt}\frac{\partial}{\partial r}\left(\frac{d\theta}{dt}\right) + r\frac{d\theta}{dt}\frac{\partial}{\partial \theta}\left(\frac{d\theta}{dt}\right) + r\frac{dz}{dt}\frac{\partial}{\partial z}\left(\frac{d\theta}{dt}\right) + 2\frac{dr}{dt}\frac{d\theta}{dt}, && (1.240)\\
&= \frac{\partial}{\partial t}\left(r\frac{d\theta}{dt}\right) + \frac{dr}{dt}\left(\frac{\partial}{\partial r}\left(r\frac{d\theta}{dt}\right) - \frac{d\theta}{dt}\right) + \frac{r\frac{d\theta}{dt}}{r}\frac{\partial}{\partial \theta}\left(r\frac{d\theta}{dt}\right) + \frac{dz}{dt}\frac{\partial}{\partial z}\left(r\frac{d\theta}{dt}\right) + 2\frac{dr}{dt}\frac{\left(r\frac{d\theta}{dt}\right)}{r}, && (1.241)\\
&= \frac{\partial u_\theta}{\partial t} + u_r\frac{\partial u_\theta}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_\theta}{\partial \theta} + u_z\frac{\partial u_\theta}{\partial z} + \underbrace{\frac{u_r u_\theta}{r}}_{\text{Coriolis}} . && (1.242)
\end{aligned}$$

The final term here is the Coriolis acceleration. The z acceleration then is easily seen to be

$$\frac{du_z}{dt} = \frac{\partial u_z}{\partial t} + u_r\frac{\partial u_z}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_z}{\partial \theta} + u_z\frac{\partial u_z}{\partial z}. \tag{1.243}$$

We summarize some useful identities, all of which can be proved, as well as some other common notation, as follows

$$\begin{aligned}
g_{kl} &= \frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^i}{\partial x^l}, && (1.244)\\
g &= \det g_{ij}, && (1.245)\\
g_{ik}g^{kj} &= g_i^j = g_j^i = \delta_i^j = \delta_j^i = \delta_{ij} = \delta^{ij}, && (1.246)\\
u_j &= u^i g_{ij}, && (1.247)\\
u^i &= g^{ij}u_j, && (1.248)\\
\mathbf{u}^T \cdot \mathbf{v} &= u_i v^i = u^i v_i = u^i g_{ij} v^j = u_i g^{ij} v_j, && (1.249)\\
\mathbf{u} \times \mathbf{v} &= \epsilon^{ijk} g_{jm} g_{kn} u^m v^n = \epsilon^{ijk} u_j v_k, && (1.250)
\end{aligned}$$

[13] Gaspard-Gustave Coriolis, 1792-1843, French mechanician.

$$\Gamma^i_{jk} = \frac{\partial^2 \xi^p}{\partial x^j \partial x^k}\frac{\partial x^i}{\partial \xi^p} = \frac{1}{2}g^{ip}\left(\frac{\partial g_{pj}}{\partial x^k} + \frac{\partial g_{pk}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^p}\right), \tag{1.251}$$

$$\nabla \mathbf{u} = \Delta_j u^i = u^i_{,j} = \frac{\partial u^i}{\partial x^j} + \Gamma^i_{jl}u^l, \tag{1.252}$$

$$\text{div } \mathbf{u} = \nabla^T \cdot \mathbf{u} = \Delta_i u^i = u^i_{,i} = \frac{\partial u^i}{\partial x^i} + \Gamma^i_{il}u^l = \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^i}\left(\sqrt{g}\ u^i\right), \tag{1.253}$$

$$\text{curl } \mathbf{u} = \nabla \times \mathbf{u} = \epsilon^{ijk}u_{k,j} = \epsilon^{ijk}g_{kp}u^p_{,j} = \epsilon^{ijk}g_{kp}\left(\frac{\partial u^p}{\partial x^j} + \Gamma^p_{jl}u^l\right), \tag{1.254}$$

$$\frac{d\mathbf{u}}{dt} = \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}^T \cdot \nabla \mathbf{u} = \frac{\partial u^i}{\partial t} + u^j\frac{\partial u^i}{\partial x^j} + \Gamma^i_{jl}u^l u^j, \tag{1.255}$$

$$\text{grad } \phi = \nabla\phi = \phi_{,i} = \frac{\partial \phi}{\partial x^i}, \tag{1.256}$$

$$\text{div grad } \phi = \nabla^2\phi = \nabla^T \cdot \nabla\phi = g^{ij}\phi_{,ij} = \frac{\partial}{\partial x^j}\left(g^{ij}\frac{\partial \phi}{\partial x^i}\right) + \Gamma^j_{jk}g^{ik}\frac{\partial \phi}{\partial x^i}, \tag{1.257}$$

$$= \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}\left(\sqrt{g}\ g^{ij}\frac{\partial \phi}{\partial x^i}\right), \tag{1.258}$$

$$\nabla \mathsf{T} = T^{ij}_{,k} = \frac{\partial T^{ij}}{\partial x^k} + \Gamma^i_{lk}T^{lj} + \Gamma^j_{lk}T^{il}, \tag{1.259}$$

$$\text{div } \mathsf{T} = \nabla^T \cdot \mathsf{T} = T^{ij}_{,j} = \frac{\partial T^{ij}}{\partial x^j} + \Gamma^i_{lj}T^{lj} + \Gamma^j_{lj}T^{il}, \tag{1.260}$$

$$= \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}\left(\sqrt{g}\ T^{ij}\right) + \Gamma^i_{jk}T^{jk} = \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}\left(\sqrt{g}\ T^{kj}\frac{\partial \xi^i}{\partial x^k}\right). \tag{1.261}$$

1.3.3 Orthogonal curvilinear coordinates

In this section we specialize our discussion to widely used orthogonal curvilinear coordinate transformations. Such transformations admit non-constant diagonal metric tensors. Because of the diagonal nature of the metric tensor, many simplifications arise. For such systems, subscripts alone suffice. Here, we simply summarize the results.

For an orthogonal curvilinear coordinate system (q_1, q_2, q_3), we have

$$ds^2 = (h_1 dq_1)^2 + (h_2 dq_2)^2 + (h_3 dq_3)^2, \tag{1.262}$$

where

$$h_i = \sqrt{\left(\frac{\partial x_1}{\partial q_i}\right)^2 + \left(\frac{\partial x_2}{\partial q_i}\right)^2 + \left(\frac{\partial x_3}{\partial q_i}\right)^2}. \tag{1.263}$$

We can show that

$$\text{grad } \phi = \nabla\phi = \frac{1}{h_1}\frac{\partial \phi}{\partial q_1}\mathbf{e}_1 + \frac{1}{h_2}\frac{\partial \phi}{\partial q_2}\mathbf{e}_2 + \frac{1}{h_3}\frac{\partial \phi}{\partial q_3}\mathbf{e}_3, \tag{1.264}$$

$$\text{div } \mathbf{u} = \nabla^T \cdot \mathbf{u} = \frac{1}{h_1 h_2 h_3}\left(\frac{\partial}{\partial q_1}(u_1 h_2 h_3) + \frac{\partial}{\partial q_2}(u_2 h_3 h_1) + \frac{\partial}{\partial q_3}(u_3 h_1 h_2)\right), \tag{1.265}$$

$$\text{curl } \mathbf{u} = \nabla \times \mathbf{u} = \frac{1}{h_1 h_2 h_3} \begin{vmatrix} h_1 \mathbf{e}_1 & h_2 \mathbf{e}_2 & h_3 \mathbf{e}_3, \\ \frac{\partial}{\partial q_1} & \frac{\partial}{\partial q_2} & \frac{\partial}{\partial q_3} \\ u_1 h_1 & u_2 h_2 & u_3 h_3 \end{vmatrix}, \tag{1.266}$$

$$\text{div grad } \phi = \nabla^2 \phi = \frac{1}{h_1 h_2 h_3} \left(\frac{\partial}{\partial q_1} \left(\frac{h_2 h_3}{h_1} \frac{\partial \phi}{\partial q_1} \right) + \frac{\partial}{\partial q_2} \left(\frac{h_3 h_1}{h_2} \frac{\partial \phi}{\partial q_2} \right) + \frac{\partial}{\partial q_3} \left(\frac{h_1 h_2}{h_3} \frac{\partial \phi}{\partial q_3} \right) \right). \tag{1.267}$$

Example 1.9

Find expressions for the gradient, divergence, and curl in cylindrical coordinates (r, θ, z) where

$$x_1 = r \cos\theta, \tag{1.268}$$
$$x_2 = r \sin\theta, \tag{1.269}$$
$$x_3 = z. \tag{1.270}$$

The 1,2 and 3 directions are associated with r, θ, and z, respectively. From Eq. (1.263), the scale factors are

$$h_r = \sqrt{\left(\frac{\partial x_1}{\partial r}\right)^2 + \left(\frac{\partial x_2}{\partial r}\right)^2 + \left(\frac{\partial x_3}{\partial r}\right)^2}, \tag{1.271}$$
$$= \sqrt{\cos^2\theta + \sin^2\theta}, \tag{1.272}$$
$$= 1, \tag{1.273}$$
$$h_\theta = \sqrt{\left(\frac{\partial x_1}{\partial \theta}\right)^2 + \left(\frac{\partial x_2}{\partial \theta}\right)^2 + \left(\frac{\partial x_3}{\partial \theta}\right)^2}, \tag{1.274}$$
$$= \sqrt{r^2 \sin^2\theta + r^2 \cos^2\theta}, \tag{1.275}$$
$$= r, \tag{1.276}$$
$$h_z = \sqrt{\left(\frac{\partial x_1}{\partial z}\right)^2 + \left(\frac{\partial x_2}{\partial z}\right)^2 + \left(\frac{\partial x_3}{\partial z}\right)^2}, \tag{1.277}$$
$$= 1, \tag{1.278}$$

so that

$$\text{grad } \phi = \frac{\partial \phi}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial \phi}{\partial \theta}\mathbf{e}_\theta + \frac{\partial \phi}{\partial z}\mathbf{e}_z, \tag{1.279}$$
$$\text{div } \mathbf{u} = \frac{1}{r}\left(\frac{\partial}{\partial r}(u_r r) + \frac{\partial}{\partial \theta}(u_\theta) + \frac{\partial}{\partial z}(u_z r)\right) = \frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{1}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{\partial u_z}{\partial z}, \tag{1.280}$$
$$\text{curl } \mathbf{u} = \frac{1}{r} \begin{vmatrix} \mathbf{e}_r & r\mathbf{e}_\theta & \mathbf{e}_z \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial z} \\ u_r & u_\theta r & u_z \end{vmatrix}. \tag{1.281}$$

1.4 Maxima and minima

Consider the real function $f(x)$, where $x \in [a,b]$. Extrema are at $x = x_m$, where $f'(x_m) = 0$, if $x_m \in [a,b]$. It is a local minimum, a local maximum, or an inflection point according to whether $f''(x_m)$ is positive, negative or zero, respectively.

Now consider a function of two variables $f(x,y)$, with $x \in [a,b]$, $y \in [c,d]$. A necessary condition for an extremum is

$$\frac{\partial f}{\partial x}(x_m, y_m) = \frac{\partial f}{\partial y}(x_m, y_m) = 0. \tag{1.282}$$

where $x_m \in [a,b]$, $y_m \in [c,d]$. Next, we find the Hessian[14] matrix:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}. \tag{1.283}$$

We use $\mathbf{H}$ and its elements to determine the character of the local extremum:

- f is a maximum if $\partial^2 f/\partial x^2 < 0$, $\partial^2 f/\partial y^2 < 0$, and $\partial^2 f/\partial x \partial y < \sqrt{(\partial^2 f/\partial x^2)(\partial^2 f/\partial y^2)}$,
- f is a minimum if $\partial^2 f/\partial x^2 > 0$, $\partial^2 f/\partial y^2 > 0$, and $\partial^2 f/\partial x \partial y < \sqrt{(\partial^2 f/\partial x^2)(\partial^2 f/\partial y^2)}$,
- f is a saddle otherwise, as long as $\det \mathbf{H} \neq 0$, and
- if $\det \mathbf{H} = 0$, higher order terms need to be considered.

Note that the first two conditions for maximum and minimum require that terms on the diagonal of $\mathbf{H}$ must dominate those on the off-diagonal with diagonal terms further required to be of the same sign. For higher dimensional systems, one can show that if all the eigenvalues of $\mathbf{H}$ are negative, f is maximized, and if all the eigenvalues of $\mathbf{H}$ are positive, f is minimized.

One can begin to understand this by considering a Taylor[15] series expansion of $f(x,y)$. Taking $\mathbf{x} = (x,y)^T$ and $d\mathbf{x} = (dx, dy)^T$, multi-variable Taylor series expansion gives

$$f(\mathbf{x} + d\mathbf{x}) = f(\mathbf{x}) + d\mathbf{x}^T \cdot \underbrace{\nabla f}_{=0} + d\mathbf{x}^T \cdot \mathbf{H} \cdot d\mathbf{x} + \ldots. \tag{1.284}$$

At an extremum, $\nabla f = 0$, so

$$f(\mathbf{x} + d\mathbf{x}) = f(\mathbf{x}) + d\mathbf{x}^T \cdot \mathbf{H} \cdot d\mathbf{x} + \ldots. \tag{1.285}$$

Later (see p. 276 and Sec. 8.2.3.8), we shall see that, by virtue of the definition of the term "positive definite," if the Hessian $\mathbf{H}$ is positive definite, then for all $d\mathbf{x}$, $d\mathbf{x}^T \cdot \mathbf{H} \cdot d\mathbf{x} > 0$, which corresponds to a minimum. For negative definite $\mathbf{H}$, we have a maximum.

[14] Ludwig Otto Hesse, 1811-1874, German mathematician, studied under Jacobi.
[15] Brook Taylor, 1685-1731, English mathematician, musician, and painter.

Example 1.10

Consider extrema of

$$f = x^2 - y^2. \tag{1.286}$$

Equating partial derivatives with respect to x and to y to zero, we get

$$\frac{\partial f}{\partial x} = 2x = 0, \tag{1.287}$$

$$\frac{\partial f}{\partial y} = -2y = 0. \tag{1.288}$$

This gives $x = 0, y = 0$. For these values we find that

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}, \tag{1.289}$$

$$= \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}. \tag{1.290}$$

Since $\det \mathbf{H} = -4 \neq 0$, and $\partial^2 f/\partial x^2$ and $\partial^2 f/\partial y^2$ have different signs, the equilibrium is a saddle point.

1.4.1 Derivatives of integral expressions

Often functions are expressed in terms of integrals. For example

$$y(x) = \int_{a(x)}^{b(x)} f(x,t)\ dt. \tag{1.291}$$

Here t is a dummy variable of integration. *Leibniz's*[16] *rule* tells us how to take derivatives of functions in integral form:

$$y(x) = \int_{a(x)}^{b(x)} f(x,t)\ dt, \tag{1.292}$$

$$\frac{dy(x)}{dx} = f(x,b(x))\frac{db(x)}{dx} - f(x,a(x))\frac{da(x)}{dx} + \int_{a(x)}^{b(x)} \frac{\partial f(x,t)}{\partial x}\ dt. \tag{1.293}$$

Inverting this arrangement in a special case, we note if

$$y(x) = y(x_o) + \int_{x_0}^{x} f(t)\ dt, \tag{1.294}$$

then

[16] Gottfried Wilhelm von Leibniz, 1646-1716, German mathematician and philosopher of great influence; co-inventor with Sir Isaac Newton, 1643-1727, of the calculus.

$$\frac{dy(x)}{dx} = f(x)\frac{dx}{dx} - f(x_0)\frac{dx_o}{dx} + \int_{x_0}^{x} \frac{\partial f(t)}{\partial x}\, dt, \tag{1.295}$$

$$\frac{dy(x)}{dx} = f(x). \tag{1.296}$$

Note that the integral expression naturally includes the initial condition that when $x = x_0$, $y = y(x_0)$. This needs to be expressed separately for the differential version of the equation.

Example 1.11

Find dy/dx if

$$y(x) = \int_{x}^{x^2} (x+1)t^2\, dt. \tag{1.297}$$

Using Leibniz's rule we get

$$\frac{dy(x)}{dx} = ((x+1)x^4)(2x) - ((x+1)x^2)(1) + \int_{x}^{x^2} t^2\, dt, \tag{1.298}$$

$$= 2x^6 + 2x^5 - x^3 - x^2 + \left(\frac{t^3}{3}\right)\Bigg|_{x}^{x^2}, \tag{1.299}$$

$$= 2x^6 + 2x^5 - x^3 - x^2 + \frac{x^6}{3} - \frac{x^3}{3}, \tag{1.300}$$

$$= \frac{7x^6}{3} + 2x^5 - \frac{4x^3}{3} - x^2. \tag{1.301}$$

$$\tag{1.302}$$

In this case it is possible to integrate explicitly to achieve the same result:

$$y(x) = (x+1)\int_{x}^{x^2} t^2\, dt, \tag{1.303}$$

$$= (x+1)\left(\frac{t^3}{3}\right)\Bigg|_{x}^{x^2}, \tag{1.304}$$

$$= (x+1)\left(\frac{x^6}{3} - \frac{x^3}{3}\right), \tag{1.305}$$

$$y(x) = \frac{x^7}{3} + \frac{x^6}{3} - \frac{x^4}{3} - \frac{x^3}{3}, \tag{1.306}$$

$$\frac{dy(x)}{dx} = \frac{7x^6}{3} + 2x^5 - \frac{4x^3}{3} - x^2. \tag{1.307}$$

So the two methods give identical results.

1.4.2 Calculus of variations

The problem is to find the function $y(x)$, with $x \in [x_1, x_2]$, and boundary conditions $y(x_1) = y_1, y(x_2) = y_2$, such that

$$I = \int_{x_1}^{x_2} f(x, y, y')\ dx, \tag{1.308}$$

is an extremum. Here, we find an operation of mapping a function $y(x)$ into a scalar I, which can be expressed as $I = \mathcal{F}(y)$. The operator $\mathcal{F}$ which performs this task is known as a *functional*.

If $y(x)$ is the desired solution, let $Y(x) = y(x) + \epsilon h(x)$, where $h(x_1) = h(x_2) = 0$. Thus, $Y(x)$ also satisfies the boundary conditions; also $Y'(x) = y'(x) + \epsilon h'(x)$. We can write

$$I(\epsilon) = \int_{x_1}^{x_2} f(x, Y, Y')\ dx. \tag{1.309}$$

Taking $dI/d\epsilon$, utilizing Leibniz's rule, Eq. (1.293), we get

$$\frac{dI}{d\epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial x} \underbrace{\frac{\partial x}{\partial \epsilon}}_{0} + \frac{\partial f}{\partial Y} \underbrace{\frac{\partial Y}{\partial \epsilon}}_{h(x)} + \frac{\partial f}{\partial Y'} \underbrace{\frac{\partial Y'}{\partial \epsilon}}_{h'(x)} \right) dx. \tag{1.310}$$

Evaluating, we find

$$\frac{dI}{d\epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial x} 0 + \frac{\partial f}{\partial Y} h(x) + \frac{\partial f}{\partial Y'} h'(x) \right) dx. \tag{1.311}$$

Since I is an extremum at $\epsilon = 0$, we have $dI/d\epsilon = 0$ for $\epsilon = 0$. This gives

$$0 = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial Y} h(x) + \frac{\partial f}{\partial Y'} h'(x) \right) \Bigg|_{\epsilon=0} dx. \tag{1.312}$$

Also when $\epsilon = 0$, we have $Y = y, Y' = y'$, so

$$0 = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} h(x) + \frac{\partial f}{\partial y'} h'(x) \right) dx. \tag{1.313}$$

Look at the second term in this integral. Since from integration by parts we get

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'} h'(x)\ dx = \int_{x_1}^{x_2} \frac{\partial f}{\partial y'} \frac{dh}{dx}\ dx = \int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\ dh, \tag{1.314}$$

$$= \underbrace{\frac{\partial f}{\partial y'} h(x) \Big|_{x_1}^{x_2}}_{=0} - \int_{x_1}^{x_2} \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right)\ h(x)\ dx, \tag{1.315}$$

$$= -\int_{x_1}^{x_2} \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right)\ h(x)\ dx. \tag{1.316}$$

The first term in Eq. (1.315) is zero because of our conditions on $h(x_1)$ and $h(x_2)$. Thus, substituting Eq. (1.316) into the original equation, Eq. (1.313), we find

$$\int_{x_1}^{x_2} \underbrace{\left(\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right)}_{0} h(x)\, dx = 0. \tag{1.317}$$

The equality holds for all $h(x)$, so that we must have

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0. \tag{1.318}$$

This is called the *Euler*[17]*-Lagrange*[18] *equation*; sometimes it is simply called Euler's equation.

While this is, in general, the preferred form of the Euler-Lagrange equation, its explicit dependency on the two end conditions is better displayed by considering a slightly different form. By expanding the total derivative term, that is

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}(x, y, y')\right) = \frac{\partial^2 f}{\partial y' \partial x}\underbrace{\frac{dx}{dx}}_{=1} + \frac{\partial^2 f}{\partial y' \partial y}\underbrace{\frac{dy}{dx}}_{y'} + \frac{\partial^2 f}{\partial y' \partial y'}\underbrace{\frac{dy'}{dx}}_{y''}, \tag{1.319}$$

$$= \frac{\partial^2 f}{\partial y' \partial x} + \frac{\partial^2 f}{\partial y' \partial y} y' + \frac{\partial^2 f}{\partial y' \partial y'} y'', \tag{1.320}$$

the Euler-Lagrange equation, Eq. (1.318), after slight rearrangement becomes

$$\frac{\partial^2 f}{\partial y' \partial y'} y'' + \frac{\partial^2 f}{\partial y' \partial y} y' + \frac{\partial^2 f}{\partial y' \partial x} - \frac{\partial f}{\partial y} = 0, \tag{1.321}$$

$$f_{y'y'} \frac{d^2 y}{dx^2} + f_{y'y} \frac{dy}{dx} + (f_{y'x} - f_y) = 0. \tag{1.322}$$

This is clearly a second order differential equation for $f_{y'y'} \neq 0$, and in general, non-linear. If $f_{y'y'}$ is always non-zero, the problem is said to be *regular*. If $f_{y'y'} = 0$ at any point, the equation is no longer second order, and the problem is said to be *singular* at such points. Note that satisfaction of two boundary conditions becomes problematic for equations less than second order.

There are several special cases of the function f.

- $f = f(x, y)$:

  The Euler-Lagrange equation is

  $$\frac{\partial f}{\partial y} = 0, \tag{1.323}$$

  which is easily solved:

  $$f(x, y) = A(x), \tag{1.324}$$

  which, knowing f, is then solved for $y(x)$.

[17] Leonhard Euler, 1707-1783, prolific Swiss mathematician, born in Basel, died in St. Petersburg.
[18] Joseph-Louis Lagrange, 1736-1813, Italian-born French mathematician.

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

  The Euler-Lagrange equation is

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0, \tag{1.325}$$

  which yields

$$\frac{\partial f}{\partial y'} = A, \tag{1.326}$$

$$f(x, y') = Ay' + B(x). \tag{1.327}$$

  Again, knowing f, the equation is solved for y' and then integrated to find $y(x)$.

- $f = f(y, y')$:

  The Euler-Lagrange equation is

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}(y, y')\right) = 0, \tag{1.328}$$

$$\frac{\partial f}{\partial y} - \left(\frac{\partial^2 f}{\partial y \partial y'}\frac{dy}{dx} + \frac{\partial^2 f}{\partial y' \partial y'}\frac{dy'}{dx}\right) = 0, \tag{1.329}$$

$$\frac{\partial f}{\partial y} - \frac{\partial^2 f}{\partial y \partial y'}\frac{dy}{dx} - \frac{\partial^2 f}{\partial y' \partial y'}\frac{d^2 y}{dx^2} = 0. \tag{1.330}$$

  Multiply by y' to get

$$y'\left(\frac{\partial f}{\partial y} - \frac{\partial^2 f}{\partial y \partial y'}\frac{dy}{dx} - \frac{\partial^2 f}{\partial y' \partial y'}\frac{d^2 y}{dx^2}\right) = 0. \tag{1.331}$$

  Add and subtract $(\partial f/\partial y')y''$ to get

$$y'\left(\frac{\partial f}{\partial y} - \frac{\partial^2 f}{\partial y \partial y'}\frac{dy}{dx} - \frac{\partial^2 f}{\partial y' \partial y'}\frac{d^2 y}{dx^2}\right) + \frac{\partial f}{\partial y'}y'' - \frac{\partial f}{\partial y'}y'' = 0. \tag{1.332}$$

  Regroup to get

$$\underbrace{\frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y''}_{=df/dx} - \underbrace{\left(y'\left(\frac{\partial^2 f}{\partial y \partial y'}\frac{dy}{dx} + \frac{\partial^2 f}{\partial y' \partial y'}\frac{d^2 y}{dx^2}\right) + \frac{\partial f}{\partial y'}y''\right)}_{=d/dx(y'\partial f/\partial y')} = 0. \tag{1.333}$$

  Regroup again to get

$$\frac{d}{dx}\left(f - y'\frac{\partial f}{\partial y'}\right) = 0, \tag{1.334}$$

which can be integrated. Thus,

$$f(y, y') - y' \frac{\partial f}{\partial y'} = K, \tag{1.335}$$

where K is an arbitrary constant. What remains is a first order ordinary differential equation which can be solved. Another integration constant arises. This second constant, along with K, are determined by the two end point conditions.

Example 1.12

Find the curve of minimum length between the points (x_1, y_1) and (x_2, y_2).

If $y(x)$ is the curve, then $y(x_1) = y_1$ and $y(x_2) = y_2$. The length of the curve is

$$L = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\, dx. \tag{1.336}$$

So our f reduces to $f(y') = \sqrt{1 + (y')^2}$. The Euler-Lagrange equation is

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{1 + (y')^2}} \right) = 0, \tag{1.337}$$

which can be integrated to give

$$\frac{y'}{\sqrt{1 + (y')^2}} = K. \tag{1.338}$$

Solving for y' we get

$$y' = \sqrt{\frac{K^2}{1 - K^2}} \equiv A, \tag{1.339}$$

from which

$$y = Ax + B. \tag{1.340}$$

The constants A and B are obtained from the boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. *The shortest distance between two points is a straight line.*

Example 1.13

Find the curve through the points (x_1, y_1) and (x_2, y_2), such that the surface area of the body of revolution by rotating the curve around the x-axis is a minimum.

We wish to minimize

$$I = \int_{x_1}^{x_2} y\sqrt{1 + (y')^2}\, dx. \tag{1.341}$$

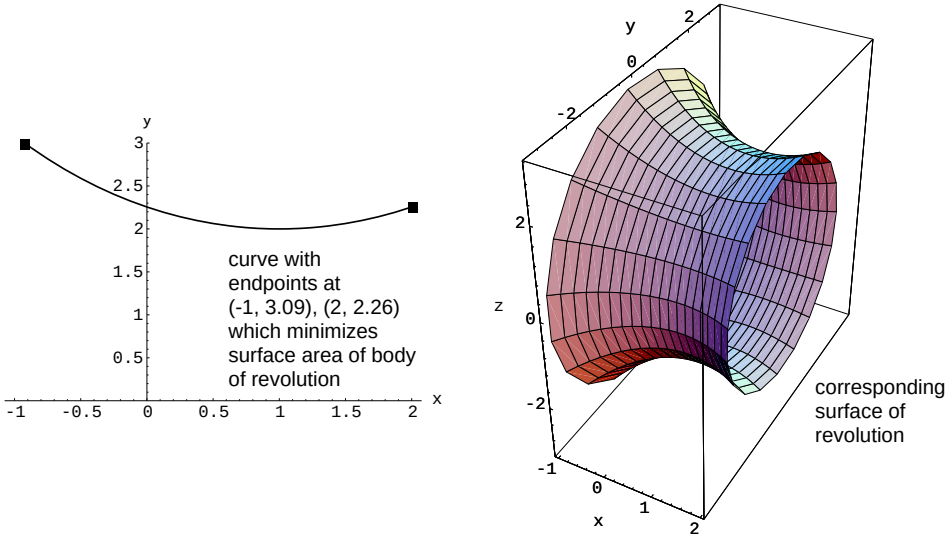

Figure 1.6: Body of revolution of minimum surface area for $(x_1, y_1) = (-1, 3.08616)$ and $(x_2, y_2) = (2, 2.25525)$.

Here f reduces to $f(y, y') = y\sqrt{1 + (y')^2}$. So the Euler-Lagrange equation reduces to

$$\begin{aligned} f(y, y') - y' \frac{\partial f}{\partial y'} &= A, & (1.342)\\ y\sqrt{1 + y'^2} - y' y \frac{y'}{\sqrt{1 + y'^2}} &= A, & (1.343)\\ y(1 + y'^2) - y y'^2 &= A\sqrt{1 + y'^2}, & (1.344)\\ y &= A\sqrt{1 + y'^2}, & (1.345)\\ y' &= \sqrt{\left(\frac{y}{A}\right)^2 - 1}, & (1.346)\\ y(x) &= A \cosh \frac{x - B}{A}. & (1.347) \end{aligned}$$

This is a catenary. The constants A and B are determined from the boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. In general this requires a trial and error solution of simultaneous algebraic equations. If $(x_1, y_1) = (-1, 3.08616)$ and $(x_2, y_2) = (2, 2.25525)$, one finds solution of the resulting algebraic equations gives $A = 2, B = 1$.

For these conditions, the curve $y(x)$ along with the resulting body of revolution of minimum surface area are plotted in Fig. 1.6.

1.5 Lagrange multipliers

Suppose we have to determine the extremum of $f(x_1, x_2, \ldots, x_M)$ subject to the n constraints

$$g_n(x_1, x_2, \ldots, x_M) = 0, \qquad n = 1, 2, \ldots, N. \tag{1.348}$$

Define

$$f^* = f - \lambda_1 g_1 - \lambda_2 g_2 - \ldots - \lambda_N g_N, \tag{1.349}$$

where the λ_n $(n = 1, 2, \cdots, N)$ are unknown constants called *Lagrange multipliers*. To get the extremum of f^*, we equate to zero its derivative with respect to $x_1, x_2, \ldots, x_M$. Thus, we have

$$\frac{\partial f^*}{\partial x_m} = 0, \ m = 1, \ldots, M, \tag{1.350}$$

$$g_n = 0, \ n = 1, \ldots, N. \tag{1.351}$$

which are $(M + N)$ equations that can be solved for x_m $(m = 1, 2, \ldots, M)$ and λ_n $(n = 1, 2, \ldots, N)$.

Example 1.14

Extremize $f = x^2 + y^2$ subject to the constraint $g = 5x^2 - 6xy + 5y^2 - 8 = 0$.

Let

$$f^* = x^2 + y^2 - \lambda(5x^2 - 6xy + 5y^2 - 8), \tag{1.352}$$

from which

$$\frac{\partial f^*}{\partial x} = 2x - 10\lambda x + 6\lambda y = 0, \tag{1.353}$$

$$\frac{\partial f^*}{\partial y} = 2y + 6\lambda x - 10\lambda y = 0, \tag{1.354}$$

$$g = 5x^2 - 6xy + 5y^2 - 8 = 0. \tag{1.355}$$

From Eq. (1.353),

$$\lambda = \frac{2x}{10x - 6y}, \tag{1.356}$$

which, when substituted into Eq. (1.354), gives

$$x = \pm y. \tag{1.357}$$

Equation (1.357), when solved in conjunction with Eq. (1.355), gives the extrema to be at $(x, y) = (\sqrt{2}, \sqrt{2}), (-\sqrt{2}, -\sqrt{2}), (1/\sqrt{2}, -1/\sqrt{2}), (-1/\sqrt{2}, 1/\sqrt{2})$. The first two sets give $f = 4$ (maximum) and the last two $f = 1$ (minimum). The function to be maximized along with the constraint function and its image are plotted in Fig. 1.7.

A similar technique can be used for the extremization of a functional with constraint. We wish to find the function $y(x)$, with $x \in [x_1, x_2]$, and $y(x_1) = y_1, y(x_2) = y_2$, such that the integral

$$I = \int_{x_1}^{x_2} f(x, y, y') \, dx, \tag{1.358}$$

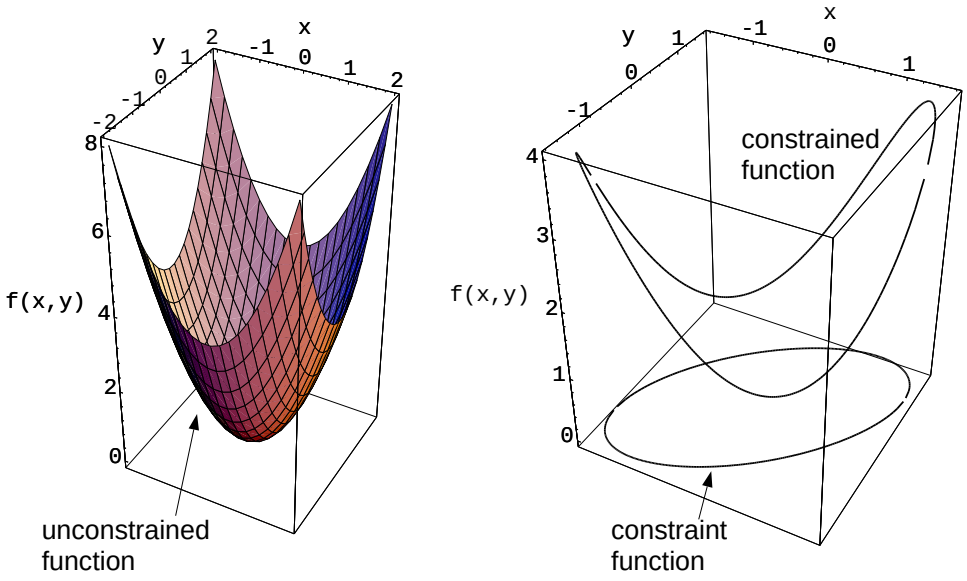

Figure 1.7: Unconstrained function $f(x,y)$ along with constrained function and constraint function (image of constrained function.)

is an extremum, and satisfies the constraint

$$g = 0. \tag{1.359}$$

Define

$$I^* = I - \lambda g, \tag{1.360}$$

and continue as before.

Example 1.15

Extremize I, where

$$I = \int_0^a y\sqrt{1+(y')^2}\,dx, \tag{1.361}$$

with $y(0) = y(a) = 0$, and subject to the constraint

$$\int_0^a \sqrt{1+(y')^2}\,dx = \ell. \tag{1.362}$$

That is, find the maximum surface area of a body of revolution which has a constant length.

Let

$$g = \int_0^a \sqrt{1+(y')^2}\,dx - \ell = 0. \tag{1.363}$$

Then let

$$I^* = I - \lambda g = \int_0^a y\sqrt{1+(y')^2}\,dx - \lambda\int_0^a \sqrt{1+(y')^2}\,dx + \lambda\ell, \tag{1.364}$$

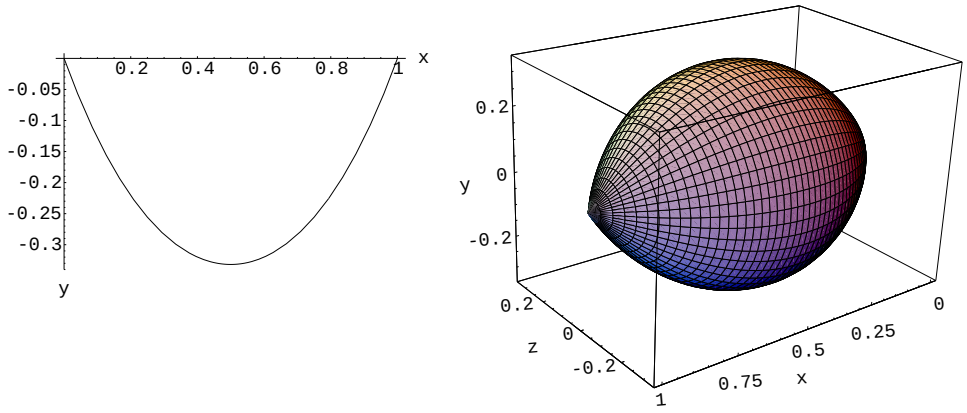

Figure 1.8: Curve of length $\ell = 5/4$ with $y(0) = y(1) = 0$ whose surface area of corresponding body of revolution (also shown) is maximum.

$$= \int_0^a (y-\lambda)\sqrt{1+(y')^2}\,dx + \lambda\ell, \tag{1.365}$$

$$= \int_0^a \left((y-\lambda)\sqrt{1+(y')^2} + \frac{\lambda\ell}{a}\right) dx. \tag{1.366}$$

With $f^* = (y-\lambda)\sqrt{1+(y')^2} + \lambda\ell/a$, we have the Euler-Lagrange equation

$$\frac{\partial f^*}{\partial y} - \frac{d}{dx}\left(\frac{\partial f^*}{\partial y'}\right) = 0. \tag{1.367}$$

Integrating from an earlier developed relationship, Eq. (1.335), when $f = f(y, y')$, and absorbing $\lambda\ell/a$ into a constant A, we have

$$(y-\lambda)\sqrt{1+(y')^2} - y'(y-\lambda)\frac{y'}{\sqrt{1+(y')^2}} = A, \tag{1.368}$$

from which

$$(y-\lambda)(1+(y')^2) - (y')^2(y-\lambda) = A\sqrt{1+(y')^2}, \tag{1.369}$$

$$(y-\lambda)\left(1+(y')^2-(y')^2\right) = A\sqrt{1+(y')^2}, \tag{1.370}$$

$$y-\lambda = A\sqrt{1+(y')^2}, \tag{1.371}$$

$$y' = \sqrt{\left(\frac{y-\lambda}{A}\right)^2 - 1}, \tag{1.372}$$

$$y = \lambda + A\cosh\frac{x-B}{A}. \tag{1.373}$$

Here A, B, λ have to be numerically determined from the three conditions $y(0) = y(a) = 0$, $g = 0$. If we take the case where $a = 1, \ell = 5/4$, we find that $A = 0.422752, B = 1/2, \lambda = -0.754549$. For these values, the curve of interest, along with the surface of revolution, is plotted in Fig. 1.8.

Problems

1. If
$$z^3 + zx + x^4 y = 2y^3,$$
   (a) find a general expression for
$$\left.\frac{\partial z}{\partial x}\right|_y, \left.\frac{\partial z}{\partial y}\right|_x,$$
   (b) evaluate
$$\left.\frac{\partial z}{\partial x}\right|_y, \left.\frac{\partial z}{\partial y}\right|_x,$$
   at $(x, y) = (1, 2)$, considering only real values of x, y, z, i.e. $x, y, z \in \mathbb{R}^1$.
   (c) Give a computer generated plot of the surface $z(x, y)$ for $x \in [-2, 2]$, $y \in [-2, 2]$, $z \in [-2, 2]$.

2. Determine the general curve $y(x)$, with $x \in [x_1, x_2]$, of total length L with endpoints $y(x_1) = y_1$ and $y(x_2) = y_2$ fixed, for which the area under the curve, $\int_{x_1}^{x_2} y \, dx$, is a maximum. Show that if $(x_1, y_1) = (0, 0); (x_2, y_2) = (1, 1); L = 3/2$, that the curve which maximizes the area and satisfies all constraints is the circle, $(y + 0.254272)^2 + (x - 1.2453)^2 = (1.26920)^2$. Plot this curve. What is the area? Verify that each constraint is satisfied. What function $y(x)$ minimizes the area and satisfies all constraints? Plot this curve. What is the area? Verify that each constraint is satisfied.

3. Show that if a ray of light is reflected from a mirror, the shortest distance of travel is when the angle of incidence on the mirror is equal to the angle of reflection.

4. The speed of light in different media separated by a planar interface is c_1 and c_2. Show that if the time taken for light to go from a fixed point in one medium to another in the second is a minimum, the angle of incidence, α_i, and the angle of refraction, α_r, are related by
$$\frac{\sin \alpha_i}{\sin \alpha_r} = \frac{c_1}{c_2}.$$

5. $\mathcal{F}$ is a quadrilateral with perimeter P. Find the form of $\mathcal{F}$ such that its area is a maximum. What is this area?

6. A body slides due to gravity from point A to point B along the curve $y = f(x)$. There is no friction and the initial velocity is zero. If points A and B are fixed, find $f(x)$ for which the time taken will be the least. What is this time? If $A : (x, y) = (1, 2), B : (x, y) = (0, 0)$, where distances are in meters, plot the minimum time curve, and find the minimum time if the gravitational acceleration is $\mathbf{g} = -9.81 \ m/s^2 \mathbf{j}$.

7. Consider the integral $I = \int_0^1 (y' - y + e^x)^2 \, dx$. What kind of extremum does this integral have (maximum or minimum)? What should $y(x)$ be for this extremum? What does the solution of the Euler-Lagrange equation give, if $y(0) = 0$ and $y(1) = -e$? Find the value of the extremum. Plot $y(x)$ for the extremum. If $y_0(x)$ is the solution of the Euler-Lagrange equation, compute I for $y_1(x) = y_0(x) + h(x)$, where you can take any $h(x)$ you like, but with $h(0) = h(1) = 0$.

8. Find the length of the shortest curve between two points with cylindrical coordinates $(r, \theta, z) = (a, 0, 0)$ and $(r, \theta, z) = (a, \Theta, Z)$ along the surface of the cylinder $r = a$.

9. Determine the shape of a parallelogram with a given area which has the least perimeter.

10. Find the extremum of the functional

$$\int_0^1 (x^2 y'^2 + 40x^4 y)\ dx,$$

with $y(0) = 0$ and $y(1) = 1$. Plot $y(x)$ which renders the integral at an extreme point.

11. Find the point on the plane $ax + by + cz = d$ which is nearest to the origin.

12. Extremize the integral

$$\int_0^1 y'^2\ dx,$$

subject to the end conditions $y(0) = 0$, $y(1) = 0$, and also the constraint

$$\int_0^1 y\ dx = 1.$$

Plot the function $y(x)$ which extremizes the integral and satisfies all constraints.

13. Show that the functions

$$\begin{aligned} u &= \frac{x+y}{x-y}, \\ v &= \frac{xy}{(x-y)^2}, \end{aligned}$$

are functionally dependent.

14. Find the point on the curve of intersection of $z - xy = 10$ and $x + y + z = 1$, that is closest to the origin.

15. Find a function $y(x)$ with $y(0) = 1$, $y(1) = 0$ that extremizes the integral

$$I = \int_0^1 \frac{\sqrt{1 + \left(\frac{dy}{dx}\right)^2}}{y}\ dx.$$

Plot $y(x)$ for this function.

16. For elliptic cylindrical coordinates

$$\begin{aligned} \xi^1 &= \cosh x^1 \cos x^2, \\ \xi^2 &= \sinh x^1 \sin x^2, \\ \xi^3 &= x^3. \end{aligned}$$

Find the Jacobian matrix $\mathbf{J}$ and the metric tensor $\mathbf{G}$. Find the transformation $x^i = x^i(\xi^j)$. Plot lines of constant x^1 and x^2 in the ξ^1 and ξ^2 plane.

17. For the elliptic coordinate system of the previous problem, find $\nabla^T \cdot \mathbf{u}$ where $\mathbf{u}$ is an arbitrary vector.

18. For parabolic coordinates

$$\begin{aligned} \xi^1 &= x^1 x^2 \cos x^3, \\ \xi^2 &= x^1 x^2 \sin x^3, \\ \xi^3 &= \frac{1}{2}\left((x^2)^2 - (x^1)^2\right). \end{aligned}$$

Find the Jacobian matrix $\mathbf{J}$ and the metric tensor $\mathbf{G}$. Find the transformation $x^i = x^i(\xi^j)$. Plot lines of constant x^1 and x^2 in the ξ^1 and ξ^2 plane.

19. For the parabolic coordinate system of the previous problem, find $\nabla^T \cdot \mathbf{u}$ where $\mathbf{u}$ is an arbitrary vector.

20. Find the covariant derivative of the contravariant velocity vector in cylindrical coordinates.

21. Prove Eq. (1.293) using the chain rule.

Chapter 2

First-order ordinary differential equations

see Kaplan, 9.1-9.3,
see Lopez, Chapters 1-3,
see Riley, Hobson, and Bence, Chapter 12,
see Bender and Orszag, 1.6.

We consider here the solution of so-called *first-order ordinary differential equations*. Such equations are of the form

$$F(x, y, y') = 0, \tag{2.1}$$

where $y' = dy/dx$. Note this is fully non-linear. A first order equation typically requires the solution to be specified at one point, though for non-linear equations, this does not guarantee uniqueness. An example, which we will not try to solve analytically, is

$$\left(xy^2 \left(\frac{dy}{dx} \right)^3 + 2\frac{dy}{dx} + \ln\left(\sin xy\right) \right)^2 - 1 = 0, \qquad y(1) = 1. \tag{2.2}$$

Fortunately, many first order equations, even non-linear ones, can be solved by techniques presented in this chapter.

2.1 Separation of variables

Equation (2.1) is separable if it can be written in the form

$$P(x)dx = Q(y)dy, \tag{2.3}$$

which can then be integrated.

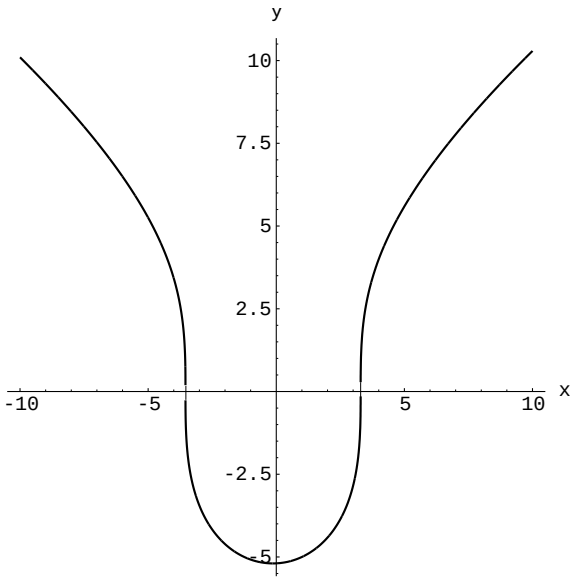

Figure 2.1: $y(x)$ which solves $yy' = (8x+1)/y$ with $y(1) = -5$.

Example 2.1

Solve

$$yy' = \frac{8x+1}{y}, \text{ with } y(1) = -5. \tag{2.4}$$

Separating variables

$$y^2 dy = 8x dx + dx. \tag{2.5}$$

Integrating, we have

$$\frac{y^3}{3} = 4x^2 + x + C. \tag{2.6}$$

The initial condition gives $C = -140/3$, so that the solution is

$$y^3 = 12x^2 + 3x - 140. \tag{2.7}$$

The solution is plotted in Fig. 2.1.

2.2 Homogeneous equations

A first order differential equation is defined by many[1] as *homogeneous* if it can be written in the form

$$y' = f\left(\frac{y}{x}\right). \tag{2.8}$$

Defining

$$u = \frac{y}{x}, \tag{2.9}$$

we get

$$y = ux, \tag{2.10}$$

from which

$$y' = u + xu'. \tag{2.11}$$

Substituting in Eq. (2.8) and separating variables, we have

$$\begin{aligned} u + xu' &= f(u), & (2.12)\\ u + x\frac{du}{dx} &= f(u), & (2.13)\\ x\frac{du}{dx} &= f(u) - u, & (2.14)\\ \frac{du}{f(u) - u} &= \frac{dx}{x}, & (2.15) \end{aligned}$$

which can be integrated.

Equations of the form

$$y' = f\left(\frac{a_1 x + a_2 y + a_3}{a_4 x + a_5 y + a_6}\right), \tag{2.16}$$

can be similarly integrated.

Example 2.2

Solve

$$xy' = 3y + \frac{y^2}{x}, \text{ with } y(1) = 4. \tag{2.17}$$

This can be written as

$$y' = 3\left(\frac{y}{x}\right) + \left(\frac{y}{x}\right)^2. \tag{2.18}$$

Let $u = y/x$. Then

$$f(u) = 3u + u^2. \tag{2.19}$$

[1]The word "homogeneous" has two distinct interpretations in differential equations. In the present section, the word actually refers to the function f, which is better considered as a so-called homogeneous function of degree zero, which implies $f(tx, ty) = f(x, y)$. Obviously $f(y/x)$ satisfies this criteria. A more common interpretation is that an equation of the form $\mathbf{L}(y) = f$ is homogeneous iff $f = 0$.

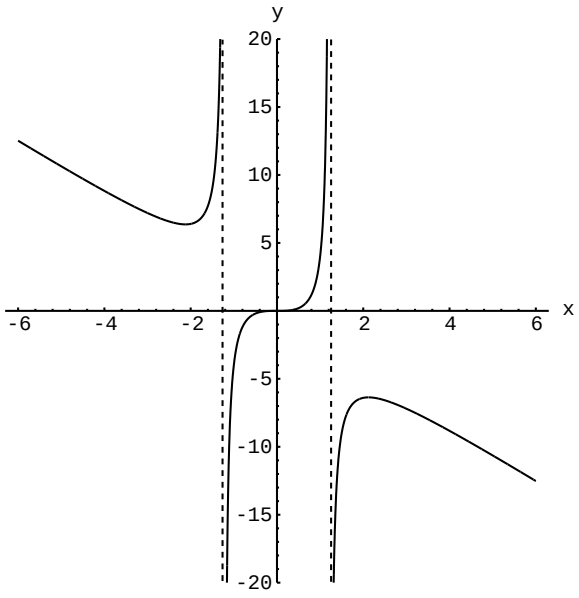

Figure 2.2: $y(x)$ which solves $xy' = 3y + y^2/x$ with $y(1) = 4$.

Using our developed formula, Eq. (2.15), we get

$$\frac{du}{2u+u^2} = \frac{dx}{x}. \tag{2.20}$$

Since by partial fraction expansion we have

$$\frac{1}{2u+u^2} = \frac{1}{2u} - \frac{1}{4+2u}, \tag{2.21}$$

Eq. (2.20) can be rewritten as

$$\frac{du}{2u} - \frac{du}{4+2u} = \frac{dx}{x}. \tag{2.22}$$

Both sides can be integrated to give

$$\frac{1}{2}\left(\ln|u| - \ln|2+u|\right) = \ln|x| + C. \tag{2.23}$$

The initial condition gives $C = (1/2)\ln(2/3)$, so that the solution can be reduced to

$$\left|\frac{y}{2x+y}\right| = \frac{2}{3}x^2.$$

This can be solved explicitly for $y(x)$ for each case of the absolute value. The first case

$$y(x) = \frac{\frac{4}{3}x^3}{1-\frac{2}{3}x^2}, \tag{2.24}$$

is seen to satisfy the condition at $x = 1$. The second case is discarded as it does not satisfy the condition at $x = 1$. The solution is plotted in Fig. 2.2.

2.3 Exact equations

A differential equation is exact if it can be written in the form

$$dF(x,y) = 0, \tag{2.25}$$

where $F(x,y) = 0$ is a solution to the differential equation. The chain rule is used to expand the derivative of $F(x,y)$ as

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = 0. \tag{2.26}$$

So, for an equation of the form

$$P(x,y)dx + Q(x,y)dy = 0, \tag{2.27}$$

we have an exact differential if

$$\frac{\partial F}{\partial x} = P(x,y), \qquad \frac{\partial F}{\partial y} = Q(x,y), \tag{2.28}$$

$$\frac{\partial^2 F}{\partial x \partial y} = \frac{\partial P}{\partial y}, \qquad \frac{\partial^2 F}{\partial y \partial x} = \frac{\partial Q}{\partial x}. \tag{2.29}$$

As long as $F(x,y)$ is continuous and differentiable, the mixed second partials are equal, thus,

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}. \tag{2.30}$$

must hold if $F(x,y)$ is to exist and render the original differential equation to be exact.

Example 2.3

Solve

$$\frac{dy}{dx} = \frac{e^{x-y}}{e^{x-y}-1}, \tag{2.31}$$

$$\underbrace{\left(e^{x-y}\right)}_{=P}dx + \underbrace{\left(1-e^{x-y}\right)}_{=Q}dy = 0, \tag{2.32}$$

$$\frac{\partial P}{\partial y} = -e^{x-y}, \tag{2.33}$$

$$\frac{\partial Q}{\partial x} = -e^{x-y}. \tag{2.34}$$

Since $\partial P/\partial y = \partial Q/\partial x$, the equation is exact. Thus,

$$\frac{\partial F}{\partial x} = P(x,y), \tag{2.35}$$

$$\frac{\partial F}{\partial x} = e^{x-y}, \tag{2.36}$$

$$F(x,y) = e^{x-y} + A(y), \tag{2.37}$$

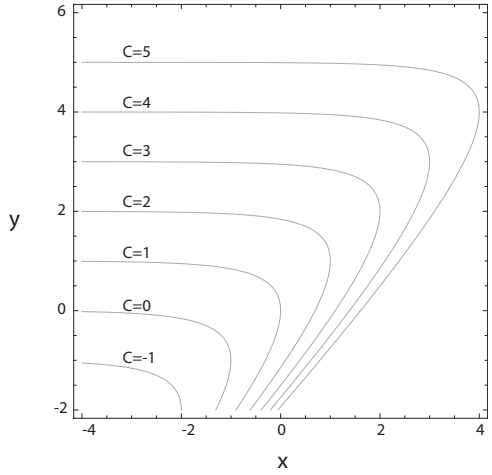

Figure 2.3: $y(x)$ which solves $y' = \exp(x-y)/(\exp(x-y)-1)$.

$$\frac{\partial F}{\partial y} = -e^{x-y} + \frac{dA}{dy} = Q(x,y) = 1 - e^{x-y}, \tag{2.38}$$

$$\frac{dA}{dy} = 1, \tag{2.39}$$

$$A(y) = y - C, \tag{2.40}$$

$$F(x,y) = e^{x-y} + y - C = 0, \tag{2.41}$$

$$e^{x-y} + y = C. \tag{2.42}$$

The solution for various values of C is plotted in Fig. 2.3.

2.4 Integrating factors

Sometimes, an equation of the form of Eq. (2.27) is not exact, but can be made so by multiplication by a function $u(x,y)$, where u is called the *integrating factor*. It is not always obvious that integrating factors exist; sometimes they do not. When one exists, it may not be unique.

Example 2.4

Solve

$$\frac{dy}{dx} = \frac{2xy}{x^2 - y^2}. \tag{2.43}$$

Separating variables, we get

$$(x^2 - y^2)\, dy = 2xy\, dx. \tag{2.44}$$

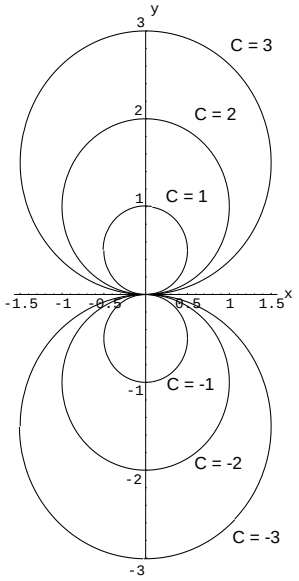

Figure 2.4: $y(x)$ which solves $y'(x) = 2xy/(x^2 - y^2)$.

This is not exact according to criterion (2.30). It turns out that the integrating factor is y^{-2}, so that on multiplication, we get

$$\frac{2x}{y}\,dx - \left(\frac{x^2}{y^2} - 1\right)\,dy = 0. \tag{2.45}$$

This can be written as

$$d\,\left(\frac{x^2}{y} + y\right) = 0, \tag{2.46}$$

which gives

$$\frac{x^2}{y} + y = C, \tag{2.47}$$

$$x^2 + y^2 = Cy. \tag{2.48}$$

The solution for various values of C is plotted in Fig. 2.4.

The general first-order linear equation

$$\frac{dy(x)}{dx} + P(x)\,y(x) = Q(x), \tag{2.49}$$

with

$$y(x_o) = y_o, \tag{2.50}$$

can be solved using the integrating factor

$$e^{\int_a^x P(s)ds} = e^{(F(x)-F(a))}. \tag{2.51}$$

We choose a such that

$$F(a) = 0. \tag{2.52}$$

Multiply by the integrating factor and proceed:

$$\begin{aligned}
\left(e^{\int_a^x P(s)ds}\right)\frac{dy(x)}{dx} + \left(e^{\int_a^x P(s)ds}\right)P(x)\ y(x) &= \left(e^{\int_a^x P(s)ds}\right)Q(x), && (2.53)\\
\text{product rule:}\qquad \frac{d}{dx}\left(e^{\int_a^x P(s)ds}y(x)\right) &= \left(e^{\int_a^x P(s)ds}\right)Q(x), && (2.54)\\
\text{replace } x \text{ by } t\text{:}\qquad \frac{d}{dt}\left(e^{\int_a^t P(s)ds}y(t)\right) &= \left(e^{\int_a^t P(s)ds}\right)Q(t), && (2.55)\\
\text{integrate:}\ \int_{x_o}^{x}\frac{d}{dt}\left(e^{\int_a^t P(s)ds}y(t)\right)dt &= \int_{x_o}^{x}\left(e^{\int_a^t P(s)ds}\right)Q(t)dt, && (2.56)\\
e^{\int_a^x P(s)ds}y(x) - e^{\int_a^{x_o} P(s)ds}y(x_o) &= \int_{x_o}^{x}\left(e^{\int_a^t P(s)ds}\right)Q(t)\ dt, && (2.57)
\end{aligned}$$

which yields

$$y(x) = e^{-\int_a^x P(s)ds}\left(e^{\int_a^{x_o} P(s)ds}y_o + \int_{x_o}^{x}\left(e^{\int_a^t P(s)ds}\right)Q(t)dt\right). \tag{2.58}$$

Example 2.5

Solve

$$y' - y = e^{2x}; \qquad y(0) = y_o. \tag{2.59}$$

Here

$$P(x) = -1, \tag{2.60}$$

or

$$\begin{aligned}
P(s) &= -1, && (2.61)\\
\int_a^x P(s)ds &= \int_a^x (-1)ds, && (2.62)\\
&= -s|_a^x, && (2.63)\\
&= a - x. && (2.64)
\end{aligned}$$

So

$$F(\tau) = -\tau. \tag{2.65}$$

For $F(a) = 0$, take $a = 0$. So the integrating factor is

$$e^{\int_a^x P(s)ds} = e^{a-x} = e^{0-x} = e^{-x}. \tag{2.66}$$

Multiplying and rearranging, we get

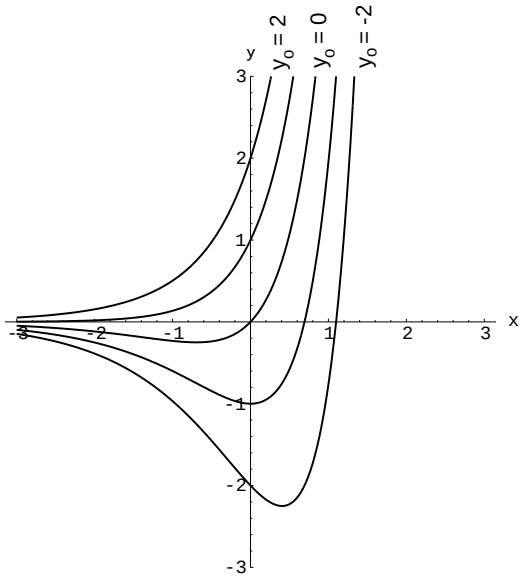

Figure 2.5: $y(x)$ which solves $y' - y = e^{2x}$ with $y(0) = y_o$.

$$e^{-x}\frac{dy(x)}{dx} - e^{-x}y(x) = e^x, \tag{2.67}$$

$$\frac{d}{dx}\left(e^{-x}y(x)\right) = e^x, \tag{2.68}$$

$$\frac{d}{dt}\left(e^{-t}y(t)\right) = e^t, \tag{2.69}$$

$$\int_{x_o=0}^{x} \frac{d}{dt}\left(e^{-t}y(t)\right)\, dt = \int_{x_o=0}^{x} e^t dt, \tag{2.70}$$

$$e^{-x}y(x) - e^{-0}y(0) = e^x - e^0, \tag{2.71}$$

$$e^{-x}y(x) - y_o = e^x - 1, \tag{2.72}$$

$$y(x) = e^x\left(y_o + e^x - 1\right), \tag{2.73}$$

$$y(x) = e^{2x} + \left(y_o - 1\right)e^x. \tag{2.74}$$

The solution for various values of y_o is plotted in Fig. 2.5.

2.5 Bernoulli equation

Some first-order non-linear equations also have analytical solutions. An example is the *Bernoulli*[2] equation

$$y' + P(x)y = Q(x)y^n. \tag{2.75}$$

[2]Jacob Bernoulli, 1654-1705, Swiss-born member of a prolific mathematical family.

where $n \neq 1$. Let

$$u = y^{1-n}, \tag{2.76}$$

so that

$$y = u^{\frac{1}{1-n}}. \tag{2.77}$$

The derivative is

$$y' = \frac{1}{1-n}\left(u^{\frac{n}{1-n}}\right)u'. \tag{2.78}$$

Substituting in Eq. (2.75), we get

$$\frac{1}{1-n}\left(u^{\frac{n}{1-n}}\right)u' + P(x)u^{\frac{1}{1-n}} = Q(x)u^{\frac{n}{1-n}}. \tag{2.79}$$

This can be written as

$$u' + (1-n)P(x)u = (1-n)Q(x), \tag{2.80}$$

which is a first-order linear equation of the form of Eq. (2.49) and can be solved.

2.6 Riccati equation

A *Riccati*[3] equation is of the form

$$\frac{dy}{dx} = P(x)y^2 + Q(x)y + R(x). \tag{2.81}$$

Studied by several Bernoullis and two Riccatis, it was solved by Euler. If we know a specific solution $y = S(x)$ of this equation, the general solution can then be found. Let

$$y = S(x) + \frac{1}{z(x)}. \tag{2.82}$$

thus

$$\frac{dy}{dx} = \frac{dS}{dx} - \frac{1}{z^2}\frac{dz}{dx}. \tag{2.83}$$

Substituting into Eq. (2.81), we get

$$\frac{dS}{dx} - \frac{1}{z^2}\frac{dz}{dx} = P\left(S + \frac{1}{z}\right)^2 + Q\left(S + \frac{1}{z}\right) + R, \tag{2.84}$$

$$\frac{dS}{dx} - \frac{1}{z^2}\frac{dz}{dx} = P\left(S^2 + \frac{2S}{z} + \frac{1}{z^2}\right) + Q\left(S + \frac{1}{z}\right) + R, \tag{2.85}$$

$$\underbrace{\frac{dS}{dx} - \left(PS^2 + QS + R\right)}_{=0} - \frac{1}{z^2}\frac{dz}{dx} = P\left(\frac{2S}{z} + \frac{1}{z^2}\right) + Q\left(\frac{1}{z}\right), \tag{2.86}$$

$$-\frac{dz}{dx} = P\left(2Sz + 1\right) + Qz, \tag{2.87}$$

$$\frac{dz}{dx} + \left(2P(x)S(x) + Q(x)\right)z = -P(x). \tag{2.88}$$

[3]Jacopo Riccati, 1676-1754, Venetian mathematician.

Again this is a first order linear equation in z and x of the form of Eq. (2.49) and can be solved.

Example 2.6

Solve

$$y' = \frac{e^{-3x}}{x} y^2 - \frac{1}{x} y + 3e^{3x}. \tag{2.89}$$

One solution is

$$y = S(x) = e^{3x}. \tag{2.90}$$

Verify:

$$\begin{aligned} 3e^{3x} &= \frac{e^{-3x}}{x} e^{6x} - \frac{1}{x} e^{3x} + 3e^{3x}, & (2.91)\\ 3e^{3x} &= \frac{e^{3x}}{x} - \frac{e^{3x}}{x} + 3e^{3x}, & (2.92)\\ 3e^{3x} &= 3e^{3x}, & (2.93) \end{aligned}$$

so let

$$y = e^{3x} + \frac{1}{z}. \tag{2.94}$$

Also we have

$$\begin{aligned} P(x) &= \frac{e^{-3x}}{x}, & (2.95)\\ Q(x) &= -\frac{1}{x}, & (2.96)\\ R(x) &= 3e^{3x}. & (2.97) \end{aligned}$$

Substituting into Eq. (2.88), we get

$$\begin{aligned} \frac{dz}{dx} + \left(2\frac{e^{-3x}}{x} e^{3x} - \frac{1}{x} \right) z &= -\frac{e^{-3x}}{x}, & (2.98)\\ \frac{dz}{dx} + \frac{z}{x} &= -\frac{e^{-3x}}{x}. & (2.99) \end{aligned}$$

The integrating factor here is

$$e^{\int \frac{dx}{x}} = e^{\ln x} = x \tag{2.100}$$

Multiplying by the integrating factor x

$$\begin{aligned} x\frac{dz}{dx} + z &= -e^{-3x}, & (2.101)\\ \frac{d(xz)}{dx} &= -e^{-3x}, & (2.102) \end{aligned}$$

which can be integrated as

$$z = \frac{e^{-3x}}{3x} + \frac{C}{x} = \frac{e^{-3x} + 3C}{3x}. \tag{2.103}$$

Since $y = S(x) + 1/z$, the solution is thus

$$y = e^{3x} + \frac{3x}{e^{-3x} + 3C}. \tag{2.104}$$

The solution for various values of C is plotted in Fig. 2.6.

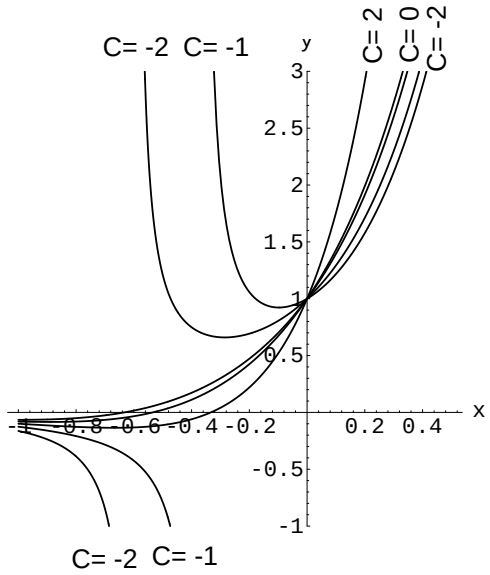

Figure 2.6: $y(x)$ which solves $y' = \exp(-3x)/x - y/x + 3\exp(3x)$.

2.7 Reduction of order

There are higher order equations that can be reduced to first-order equations and then solved.

2.7.1 y absent

If

$$f(x, y', y'') = 0, \tag{2.105}$$

then let $u(x) = y'$. Thus, $u'(x) = y''$, and the equation reduces to

$$f\left(x, u, \frac{du}{dx}\right) = 0, \tag{2.106}$$

which is an equation of first order.

Example 2.7

Solve

$$xy'' + 2y' = 4x^3. \tag{2.107}$$

Let $u = y'$, so that

$$x\frac{du}{dx} + 2u = 4x^3. \tag{2.108}$$

Multiplying by x

$$x^2\frac{du}{dx} + 2xu = 4x^4, \tag{2.109}$$

$$\frac{d}{dx}(x^2u) = 4x^4. \tag{2.110}$$

This can be integrated to give

$$u = \frac{4}{5}x^3 + \frac{C_1}{x^2}, \tag{2.111}$$

from which

$$y = \frac{1}{5}x^4 - \frac{C_1}{x} + C_2, \tag{2.112}$$

for $x \neq 0$.

2.7.2 x absent

If

$$f(y, y', y'') = 0, \tag{2.113}$$

let $u(x) = y'$, so that

$$y'' = \frac{dy'}{dx} = \frac{dy'}{dy}\frac{dy}{dx} = \frac{du}{dy}u, \tag{2.114}$$

Equation (2.113) becomes

$$f\left(y, u, u\frac{du}{dy}\right) = 0, \tag{2.115}$$

which is also an equation of first order. Note however that the independent variable is now y while the dependent variable is u.

Example 2.8

Solve

$$y'' - 2yy' = 0; \qquad y(0) = y_o, \qquad y'(0) = y'_o. \tag{2.116}$$

Let $u = y'$, so that $y'' = du/dx = (dy/dx)(du/dy) = u(du/dy)$. The equation becomes

$$u\frac{du}{dy} - 2yu = 0. \tag{2.117}$$

Now

$$u = 0, \tag{2.118}$$

satisfies Eq. (2.117). Thus,

$$\frac{dy}{dx} = 0, \tag{2.119}$$

$$y = C, \tag{2.120}$$

applying one initial condition: $$y = y_o \tag{2.121}$$

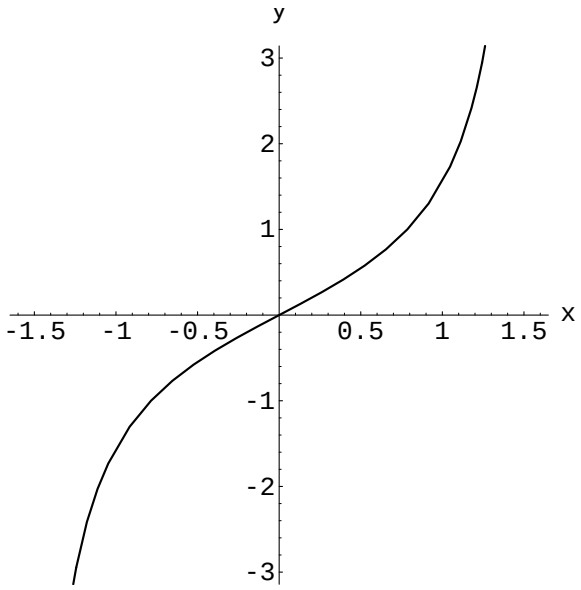

Figure 2.7: $y(x)$ which solves $y'' - 2yy' = 0$ with $y(0) = 0, y'(0) = 1$.

This satisfies the initial conditions only under special circumstances, i.e. $y'_o = 0$. For $u \neq 0$,

$$\frac{du}{dy} = 2y, \tag{2.122}$$

$$u = y^2 + C_1, \tag{2.123}$$

$$\text{apply I.C.'s:} \quad y'_o = y_o^2 + C_1, \tag{2.124}$$

$$C_1 = y'_o - y_o^2, \tag{2.125}$$

$$\frac{dy}{dx} = y^2 + y'_o - y_o^2, \tag{2.126}$$

$$\frac{dy}{y^2 + y'_o - y_o^2} = dx, \tag{2.127}$$

from which for $y'_o - y_o^2 > 0$

$$\frac{1}{\sqrt{y'_o - y_o^2}} \tan^{-1}\left(\frac{y}{\sqrt{y'_o - y_o^2}}\right) = x + C_2, \tag{2.128}$$

$$\frac{1}{\sqrt{y'_o - y_o^2}} \tan^{-1}\left(\frac{y_o}{\sqrt{y'_o - y_o^2}}\right) = C_2, \tag{2.129}$$

$$y(x) = \sqrt{y'_o - y_o^2} \tan\left(x\sqrt{y'_o - y_o^2} + \tan^{-1}\left(\frac{y_o}{\sqrt{y'_o - y_o^2}}\right)\right). \tag{2.130}$$

The solution for $y_o = 0, y'_o = 1$ is plotted in Fig. 2.7.

For $y'_o - y_o^2 = 0$,

$$\frac{dy}{dx} = y^2, \tag{2.131}$$

$$\frac{dy}{y^2} = dx, \tag{2.132}$$

$$-\frac{1}{y} = x + C_2, \tag{2.133}$$
$$-\frac{1}{y_o} = C_2, \tag{2.134}$$
$$-\frac{1}{y} = x - \frac{1}{y_o} \tag{2.135}$$
$$y = \frac{1}{\frac{1}{y_o} - x}. \tag{2.136}$$

For $y_o' - y_o^2 < 0$, one would obtain solutions in terms of hyperbolic trigonometric functions; see Sec. 10.3.

2.8 Uniqueness and singular solutions

Not all differential equations have solutions, as can be seen by considering

$$y' = \frac{y}{x} \ln y, \qquad y(0) = 2. \tag{2.137}$$

The general solution of the differential equation is $y = e^{Cx}$, but no finite value of C allows the initial condition to be satisfied. Let's check this by direct substitution:

$$y = e^{Cx}, \tag{2.138}$$
$$y' = Ce^{Cx}, \tag{2.139}$$
$$\frac{y}{x} \ln y = \frac{e^{Cx}}{x} \ln e^{Cx}, \tag{2.140}$$
$$= \frac{e^{Cx}}{x} Cx, \tag{2.141}$$
$$= Ce^{Cx}, \tag{2.142}$$
$$= y'. \tag{2.143}$$

So the differential equation is satisfied for all values of C. Now to satisfy the initial condition, we must have

$$2 = e^{C(0)}, \tag{2.144}$$
$$2 = 1? \tag{2.145}$$

There is no finite value of C that allows satisfaction of the initial condition. The original differential equation can be written as $xy' = y \ln y$. The point $x = 0$ is *singular* since at that point, the highest derivative is multiplied by 0 leaving only $0 = y \ln y$ at $x = 0$. For the very special initial condition $y(0) = 1$, the solution $y = e^{Cx}$ is valid for *all* values of C. Thus, for

this singular equation, for most initial conditions, no solution exists. For one special initial condition, a solution exists, but it is not unique.

Theorem

Let $f(x,y)$ be continuous and satisfy $|f(x,y)| \leq m$ and the *Lipschitz*[4] *condition* $|f(x,y) - f(x,y_0)| \leq k|y-y_0|$ in a bounded region $\mathcal{R}$. Then the equation $y' = f(x,y)$ has one and only one solution containing the point (x_0, y_0).

A stronger condition is that if $f(x,y)$ and $\partial f/\partial y$ are finite and continuous at (x_0,y_0), then a solution of $y' = f(x,y)$ exists and is unique in the neighborhood of this point.

Example 2.9

Analyze the uniqueness of the solution of

$$\frac{dy}{dt} = -K\sqrt{y}, \qquad y(T) = 0. \tag{2.146}$$

Here, t is the independent variable instead of x. Taking,

$$f(t,y) = -K\sqrt{y}, \tag{2.147}$$

we have

$$\frac{\partial f}{\partial y} = -\frac{K}{2\sqrt{y}}, \tag{2.148}$$

which is not finite at $y = 0$. So the solution cannot be guaranteed to be unique. In fact, one solution is

$$y(t) = \frac{1}{4}K^2(t-T)^2. \tag{2.149}$$

Another solution which satisfies the initial condition and differential equation is

$$y(t) = 0. \tag{2.150}$$

Obviously the solution is not unique.

Example 2.10

Consider the differential equation and initial condition

$$\frac{dy}{dx} = 3y^{2/3}, \qquad y(2) = 0. \tag{2.151}$$

On separating variables and integrating, we get

$$3y^{1/3} = 3x + 3C, \tag{2.152}$$

[4]Rudolf Otto Sigismund Lipschitz, 1832-1903, German mathematician.

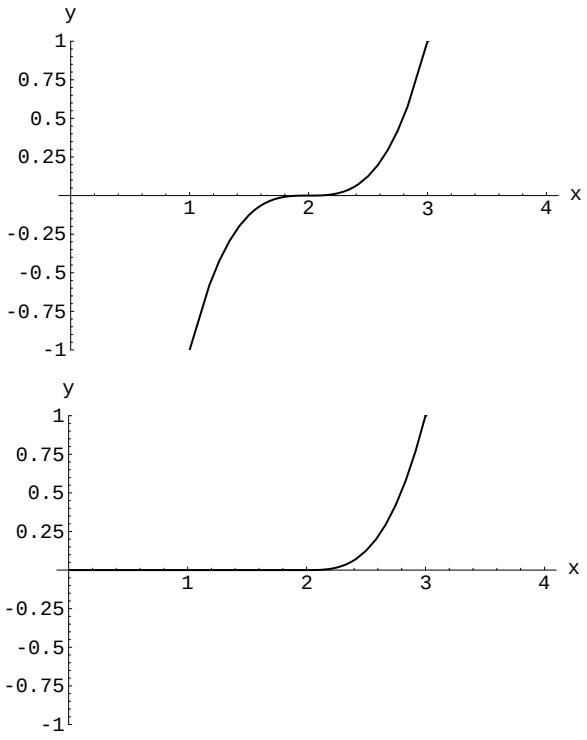

Figure 2.8: Two solutions $y(x)$ which satisfy $y' = 3y^{2/3}$ with $y(2) = 0$.

so that the *general* solution is

$$y = (x + C)^3. \tag{2.153}$$

Applying the initial condition, we find

$$y = (x - 2)^3. \tag{2.154}$$

However,

$$y = 0, \tag{2.155}$$

and

$$y = \begin{cases} (x-2)^3 & \text{if } x \geq 2, \\ 0 & \text{if } x < 2. \end{cases} \tag{2.156}$$

are also solutions. These *singular* solutions cannot be obtained from the general solution. However, values of y' and y are the same at intersections. Both satisfy the differential equation. The two solutions are plotted in Fig. 2.8.

2.9 Clairaut equation

The solution of a Clairaut[5] equation

$$y = xy' + f(y'), \tag{2.157}$$

[5]Alexis Claude Clairaut, 1713-1765, Parisian/French mathematician.

can be obtained by letting $y' = u(x)$, so that

$$y = xu + f(u). \tag{2.158}$$

Differentiating with respect to x, we get

$$\begin{aligned} y' &= xu' + u + \frac{df}{du}u', & (2.159)\\ u &= xu' + u + \frac{df}{du}u', & (2.160)\\ \left(x + \frac{df}{du}\right)u' &= 0. & (2.161) \end{aligned}$$

There are two possible solutions to this, $u' = 0$ or $x + df/du = 0$. If we consider the first and take

$$u' = \frac{du}{dx} = 0, \tag{2.162}$$

we can integrate to get

$$u = C, \tag{2.163}$$

where C is a constant. Then, from Eq. (2.158), we get the general solution

$$y = Cx + f(C). \tag{2.164}$$

Applying an initial condition $y(x_o) = y_o$ gives what we will call the *regular* solution.

But if we take the second

$$x + \frac{df}{du} = 0, \tag{2.165}$$

and rearrange to get

$$x = -\frac{df}{du}, \tag{2.166}$$

then Eq. (2.166) along with the rearranged Eq. (2.158)

$$y = -u\frac{df}{du} + f(u), \tag{2.167}$$

form a set of parametric equations for what we call the *singular* solution. It is singular because the coefficient on the highest derivative in Eq. (2.161) is itself 0.

Example 2.11

Solve

$$y = xy' + (y')^3, \qquad y(0) = y_o. \tag{2.168}$$

Take

$$u = y'. \tag{2.169}$$

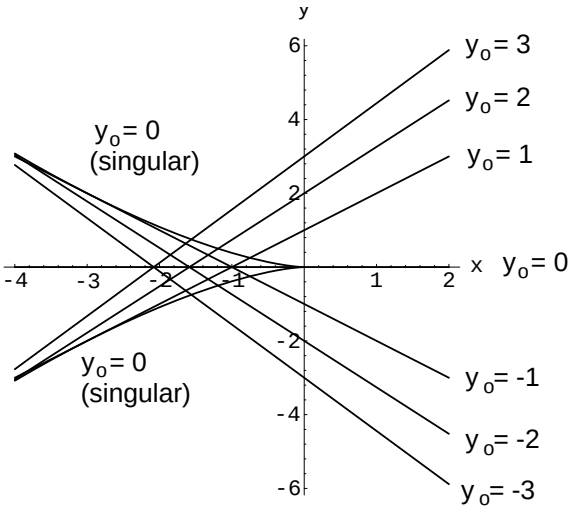

Figure 2.9: Two solutions $y(x)$ which satisfy $y = xy' + (y')^3$ with $y(0) = y_o$.

Then

$$f(u) = u^3, \tag{2.170}$$
$$\frac{df}{du} = 3u^2, \tag{2.171}$$

so specializing Eq. (2.164) gives

$$y = Cx + C^3$$

as the general solution. Use the initial condition to evaluate C and get the regular solution:

$$y_o = C(0) + C^3, \tag{2.172}$$
$$C = y_o^{1/3}, \tag{2.173}$$
$$y = y_o^{1/3}x + y_o. \tag{2.174}$$

Note if $y_o \in \mathbb{R}^1$, there are actually three roots for C: $C = y_o^{1/3}, (-1/2 \pm i\sqrt{3}/2)y_o^{1/3}$. So the solution is non-unique. However, if we confine our attention to real valued solutions, there is a unique real solution, with $C = y_o^{1/3}$.

The parametric form of the singular solution is

$$y = -2u^3, \tag{2.175}$$
$$x = -3u^2. \tag{2.176}$$

Eliminating the parameter u, we obtain

$$y = \pm 2\left(-\frac{x}{3}\right)^{3/2}, \tag{2.177}$$

as the explicit form of the singular solution.

The regular solutions and singular solution are plotted in Fig. 2.9. Note

- In contrast to solutions for equations linear in y', the trajectories $y(x; y_o)$ cross at numerous locations in the $x - y$ plane. This is a consequence of the differential equation's *non-linearity*

- While the singular solution satisfies the differential equation, it satisfies this initial condition only when $y_o = 0$
- For real valued x and y, the singular solution is only valid for $x \leq 0$.
- Because of non-linearity, addition of the regular and singular solutions does not yield a solution to the differential equation.

Problems

1. Find the general solution of the differential equation

$$y' + x^2 y(1+y) = 1 + x^3(1+x).$$

   Plot solutions for $y(0) = -2, 0, 2$.

2. Solve

$$\dot{x} = 2tx + te^{-t^2}x^2.$$

   Plot a solution for $x(0) = 1$.

3. Solve

$$3x^2y^2\ dx + 2x^3y\ dy = 0.$$

4. Solve

$$\frac{dy}{dx} = \frac{x-y}{x+y}.$$

5. Solve the non-linear equation $(y' - x)y'' + 2y' = 2x$.
6. Solve $xy'' + 2y' = x$. Plot a solution for $y(1) = 1, y'(1) = 1$.
7. Solve $y'' - 2yy' = 0$. Plot a solution for $y(0) = 0, y'(0) = 3$.
8. Given that $y_1 = x^{-1}$ is one solution of $y'' + (3/x)y' + (1/x^2)y = 0$, find the other solution.
9. Solve
   (a) $y' \tan y + 2 \sin x \sin(\frac{\pi}{2} + x) + \ln x = 0$
   (b) $xy' - 2y - x^4 - y^2 = 0$
   (c) $y' \cos y \cos x + \sin y \sin x = 0$
   (d) $y' + y \cot x = e^x$
   (e) $x^5y' + y + e^{x^2}(x^6 - 1)y^3 = 0$, with $y(1) = e^{-1/2}$
   (f) $y' + y^2 - xy - 1 = 0$
   (g) $y'(x + y^2) - y = 0$
   (h) $y' = \frac{x+2y-5}{-2x-y+4}$
   (i) $y' + xy = y$

   Plot solutions, when possible, for $y(0) = -1, 0, 1$.

10. Find all solutions of
$$(x+1)(y')^2 + (x-y)y' - y = 0$$

11. Find an a for which a *unique* real solution of
$$(y')^4 + 8(y')^3 + (3a+16)(y')^2 + 12ay' + 2a^2 = 0, \text{ with } y(1) = -2$$
exists. Find the solution.

12. Solve
$$y' - \frac{1}{x^2}y^2 + \frac{1}{x}y = 1$$

13. Find the most general solution to
$$(y'-1)(y'+1) = 0$$

14. Solve
$$(D-1)(D-2)y = x$$

Chapter 3

Linear ordinary differential equations

see Kaplan, 9.1-9.4,
see Lopez, Chapter 5,
see Bender and Orszag, 1.1-1.5,
see Riley, Hobson, and Bence, Chapter 13, Chapter 15.6,
see Friedman, Chapter 3.

We consider in this chapter *linear ordinary differential equations*. We will mainly be concerned with equations which are of second order or higher in a single dependent variable.

3.1 Linearity and linear independence

An ordinary differential equation can be written in the form

$$\mathbf{L}(y) = f(x), \tag{3.1}$$

where $y(x)$ is an unknown function. The equation is said to be *homogeneous* if $f(x) = 0$, giving then

$$\mathbf{L}(y) = 0. \tag{3.2}$$

This is the most common usage for the term "homogeneous." The operator $\mathbf{L}$ is composed of a combination of derivatives $d/dx, d^2/dx^2$, etc. The operator $\mathbf{L}$ is linear if

$$\mathbf{L}(y_1 + y_2) = \mathbf{L}(y_1) + \mathbf{L}(y_2), \tag{3.3}$$

and

$$\mathbf{L}(\alpha y) = \alpha \mathbf{L}(y), \tag{3.4}$$

where α is a scalar. We can contrast this definition of linearity with the definition of more general term "affine" given by Eq. (1.102), which, while similar, admits a constant inhomogeneity.

For the remainder of this chapter, we will take $\mathbf{L}$ to be a linear differential operator. The general form of $\mathbf{L}$ is

$$\mathbf{L} = P_N(x)\frac{d^N}{dx^N} + P_{N-1}(x)\frac{d^{N-1}}{dx^{N-1}} + \ldots + P_1(x)\frac{d}{dx} + P_0(x). \tag{3.5}$$

The ordinary differential equation, Eq. (3.1), is then linear when $\mathbf{L}$ has the form of Eq. (3.5).

Definition: The functions $y_1(x), y_2(x), \ldots, y_N(x)$ are said to be *linearly independent* when $C_1y_1(x) + C_2y_2(x) + \ldots + C_Ny_N(x) = 0$ is true only when $C_1 = C_2 = \ldots = C_N = 0$.

A homogeneous equation of order N can be shown to have N linearly independent solutions. These are called *complementary functions*. If y_n $(n = 1, \ldots, N)$ are the complementary functions of Eq. (3.2), then

$$y(x) = \sum_{n=1}^{N} C_n y_n(x), \tag{3.6}$$

is the general solution of the homogeneous Eq. (3.2). In language to be defined in a future chapter, Sec. 7.3, we can say the complementary functions are linearly independent and span the space of solutions of the homogeneous equation; they are the bases of the null space of the differential operator $\mathbf{L}$. If $y_p(x)$ is any *particular solution* of Eq. (3.1), the general solution to Eq. (3.2) is then

$$y(x) = y_p(x) + \sum_{n=1}^{N} C_n y_n(x). \tag{3.7}$$

Now we would like to show that any solution $\phi(x)$ to the homogeneous equation $\mathbf{L}(y) = 0$ can be written as a linear combination of the N complementary functions $y_n(x)$:

$$C_1y_1(x) + C_2y_2(x) + \ldots + C_Ny_N(x) = \phi(x). \tag{3.8}$$

We can form additional equations by taking a series of derivatives up to $N - 1$:

$$C_1y_1'(x) + C_2y_2'(x) + \ldots + C_Ny_N'(x) = \phi'(x), \tag{3.9}$$

$$\vdots$$

$$C_1y_1^{(N-1)}(x) + C_2y_2^{(N-1)}(x) + \ldots + C_Ny_N^{(N-1)}(x) = \phi^{(N-1)}(x). \tag{3.10}$$

This is a linear system of algebraic equations:

$$\begin{pmatrix} y_1 & y_2 & \cdots & y_N \\ y_1' & y_2' & \cdots & y_N' \\ \vdots & \vdots & \cdots & \vdots \\ y_1^{(N-1)} & y_2^{(N-1)} & \cdots & y_N^{(N-1)} \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_N \end{pmatrix} = \begin{pmatrix} \phi(x) \\ \phi'(x) \\ \vdots \\ \phi^{(N-1)}(x) \end{pmatrix}. \tag{3.11}$$

We could solve Eq. (3.11) by Cramer's rule, which requires the use of determinants. For a unique solution, we need the determinant of the coefficient matrix of Eq. (3.11) to be non-zero. This particular determinant is known as the *Wronskian*[1] W of $y_1(x), y_2(x), \ldots, y_N(x)$ and is defined as

$$W = \begin{vmatrix} y_1 & y_2 & \cdots & y_N \\ y_1' & y_2' & \cdots & y_N' \\ \vdots & \vdots & \cdots & \vdots \\ y_1^{(N-1)} & y_2^{(N-1)} & \cdots & y_N^{(N-1)} \end{vmatrix}. \tag{3.12}$$

The condition $W \neq 0$ indicates linear independence of the functions $y_1(x), y_2(x), \ldots, y_N(x)$, since if $\phi(x) \equiv 0$, the only solution is $C_n = 0, n = 1, \ldots, N$. Unfortunately, the converse is not always true; that is, if $W = 0$, the complementary functions may or may not be linearly dependent, though in most cases $W = 0$ indeed implies linear dependence.

Example 3.1

Determine the linear independence of (a) $y_1 = x$ and $y_2 = 2x$, (b) $y_1 = x$ and $y_2 = x^2$, and (c) $y_1 = x^2$ and $y_2 = x|x|$ for $x \in (-1, 1)$.

(a) $W = \begin{vmatrix} x & 2x \\ 1 & 2 \end{vmatrix} = 0$, linearly dependent.

(b) $W = \begin{vmatrix} x & x^2 \\ 1 & 2x \end{vmatrix} = x^2 \neq 0$, linearly independent, except at $x = 0$.

(c) We can restate y_2 as

$$y_2(x) = -x^2 \qquad x \in (-1, 0], \tag{3.13}$$

$$y_2(x) = x^2 \qquad x \in (0, 1), \tag{3.14}$$

so that

$$W = \begin{vmatrix} x^2 & -x^2 \\ 2x & -2x \end{vmatrix} = -2x^3 + 2x^3 = 0, \qquad x \in (-1, 0], \tag{3.15}$$

$$W = \begin{vmatrix} x^2 & x^2 \\ 2x & 2x \end{vmatrix} = 2x^3 - 2x^3 = 0, \qquad x \in (0, 1). \tag{3.16}$$

Thus, $W = 0$ for $x \in (-1, 1)$, which suggests the functions may be linearly dependent. However, when we seek C_1 and C_2 such that $C_1 y_1 + C_2 y_2 = 0$, we find the only solution is $C_1 = 0, C_2 = 0$; therefore, the functions are in fact linearly independent, despite the fact that $W = 0$! Let's check this. For $x \in (-1, 0]$,

$$C_1 x^2 + C_2(-x^2) = 0, \tag{3.17}$$

so we will need $C_1 = C_2$ at a minimum. For $x \in (0, 1)$,

$$C_1 x^2 + C_2 x^2 = 0, \tag{3.18}$$

which gives the requirement that $C_1 = -C_2$. Substituting the first condition into the second gives $C_2 = -C_2$, which is only satisfied if $C_2 = 0$, thus requiring that $C_1 = 0$; hence, the functions are indeed linearly independent.

[1] Józef Maria Hoene-Wroński, 1778-1853, Polish-born French mathematician.

Example 3.2

Determine the linear independence of the set of polynomials,

$$y_n(x) = \left\{1, x, \frac{x^2}{2}, \frac{x^3}{6}, \ldots, \frac{x^{N-1}}{(N-1)!}\right\}. \tag{3.19}$$

The Wronskian is

$$W = \begin{vmatrix} 1 & x & \frac{1}{2}x^2 & \frac{1}{6}x^3 & \ldots & \frac{1}{(N-1)!}x^{N-1} \\ 0 & 1 & x & \frac{1}{2}x^2 & \ldots & \frac{1}{(N-2)!}x^{N-2} \\ 0 & 0 & 1 & x & \ldots & \frac{1}{(N-3)!}x^{N-3} \\ 0 & 0 & 0 & 1 & \ldots & \frac{1}{(N-4)!}x^{N-4} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{vmatrix} = 1. \tag{3.20}$$

The determinant is unity, $\forall N$. As such, the polynomials are linearly independent.

3.2 Complementary functions

This section will consider solutions to the homogeneous part of the differential equation.

3.2.1 Equations with constant coefficients

First consider equations with constant coefficients.

3.2.1.1 Arbitrary order

Consider the homogeneous equation with constant coefficients

$$A_N y^{(N)} + A_{N-1} y^{(N-1)} + \ldots + A_1 y' + A_0 y = 0, \tag{3.21}$$

where A_n, $(n = 0, \ldots, N)$ are constants. To find the solution of Eq. (3.21), we let $y = e^{rx}$. Substituting we get

$$A_N r^N e^{rx} + A_{N-1} r^{(N-1)} e^{rx} + \ldots + A_1 r^1 e^{rx} + A_0 e^{rx} = 0. \tag{3.22}$$

Eliminating the non-zero common factor e^{rx}, we get

$$A_N r^N + A_{N-1} r^{(N-1)} + \ldots + A_1 r^1 + A_0 r^0 = 0, \tag{3.23}$$

$$\sum_{n=0}^{N} A_n r^n = 0. \tag{3.24}$$

This is called the *characteristic* equation. It is an n^{th} order polynomial which has N roots (some of which could be repeated, some of which could be complex), r_n $(n = 1, \ldots, N)$ from which N linearly independent complementary functions $y_n(x)$ $(n = 1, \ldots, N)$ have to be obtained. The general solution is then given by Eq. (3.6).

If all roots are real and distinct, then the complementary functions are simply $e^{r_n x}$, $(n = 1, \ldots, N)$. If, however, k of these roots are repeated, i.e. $r_1 = r_2 = \ldots = r_k = r$, then the linearly independent complementary functions are obtained by multiplying e^{rx} by $1, x, x^2, \ldots, x^{k-1}$. For a pair of complex conjugate roots $p \pm qi$, one can use de Moivre's formula (see Appendix, Eq. (10.91)) to show that the complementary functions are $e^{px}\cos qx$ and $e^{px}\sin qx$.

Example 3.3

Solve

$$\frac{d^4y}{dx^4} - 2\frac{d^3y}{dx^3} + \frac{d^2y}{dx^2} + 2\frac{dy}{dx} - 2y = 0. \tag{3.25}$$

Substituting $y = e^{rx}$, we get a characteristic equation

$$r^4 - 2r^3 + r^2 + 2r - 2 = 0, \tag{3.26}$$

which can be factored as

$$(r + 1)(r - 1)(r^2 - 2r + 2) = 0, \tag{3.27}$$

from which

$$r_1 = -1, \qquad r_2 = 1 \qquad r_3 = 1 + i \qquad r_4 = 1 - i. \tag{3.28}$$

The general solution is

$$\begin{aligned}
y(x) &= C_1e^{-x} + C_2e^{x} + C_3'e^{(1+i)x} + C_4'e^{(1-i)x}, && (3.29)\\
&= C_1e^{-x} + C_2e^{x} + C_3'e^{x}e^{ix} + C_4'e^{x}e^{-ix}, && (3.30)\\
&= C_1e^{-x} + C_2e^{x} + e^{x}\left(C_3'e^{ix} + C_4'e^{-ix}\right), && (3.31)\\
&= C_1e^{-x} + C_2e^{x} + e^{x}\left(C_3'\left(\cos x + i\sin x\right) + C_4'\left(\cos(-x) + i\sin(-x)\right)\right), && (3.32)\\
&= C_1e^{-x} + C_2e^{x} + e^{x}\left((C_3' + C_4')\cos x + i(C_3' - C_4')\sin x\right), && (3.33)\\
y(x) &= C_1e^{-x} + C_2e^{x} + e^{x}(C_3\cos x + C_4\sin x), && (3.34)
\end{aligned}$$

where $C_3 = C_3' + C_4'$ and $C_4 = i(C_3' - C_4')$.

3.2.1.2 First order

The characteristic polynomial of the first order equation

$$ay' + by = 0, \tag{3.35}$$

is

$$ar + b = 0. \tag{3.36}$$

So

$$r = -\frac{b}{a}, \tag{3.37}$$

thus, the complementary function for Eq. (3.35) is simply

$$y = Ce^{-\frac{b}{a}x}. \tag{3.38}$$

3.2.1.3 Second order

The characteristic polynomial of the second order equation

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = 0, \tag{3.39}$$

is

$$ar^2 + br + c = 0. \tag{3.40}$$

Depending on the coefficients of this quadratic equation, there are three cases to be considered.

- $b^2 - 4ac > 0$: two distinct real roots r_1 and r_2. The complementary functions are $y_1 = e^{r_1 x}$ and $y_2 = e^{r_2 x}$,
- $b^2 - 4ac = 0$: one real root. The complementary functions are $y_1 = e^{rx}$ and $y_2 = xe^{rx}$, or
- $b^2 - 4ac < 0$: two complex conjugate roots $p \pm qi$. The complementary functions are $y_1 = e^{px}\cos qx$ and $y_2 = e^{px}\sin qx$.

Example 3.4

Solve

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2y = 0. \tag{3.41}$$

The characteristic equation is

$$r^2 - 3r + 2 = 0, \tag{3.42}$$

with solutions

$$r_1 = 1, \qquad r_2 = 2. \tag{3.43}$$

The general solution is then

$$y = C_1 e^x + C_2 e^{2x}. \tag{3.44}$$

Example 3.5

Solve

$$\frac{d^2y}{dx^2} - 2\frac{dy}{dx} + y = 0. \tag{3.45}$$

The characteristic equation is

$$r^2 - 2r + 1 = 0, \tag{3.46}$$

with repeated roots

$$r_1 = 1, \qquad r_2 = 1. \tag{3.47}$$

The general solution is then

$$y = C_1 e^x + C_2 x e^x. \tag{3.48}$$

Example 3.6

Solve

$$\frac{d^2y}{dx^2} - 2\frac{dy}{dx} + 10y = 0. \tag{3.49}$$

The characteristic equation is

$$r^2 - 2r + 10 = 0, \tag{3.50}$$

with solutions

$$r_1 = 1 + 3i, \qquad r_2 = 1 - 3i. \tag{3.51}$$

The general solution is then

$$y = e^x(C_1 \cos 3x + C_2 \sin 3x). \tag{3.52}$$

3.2.2 Equations with variable coefficients

3.2.2.1 One solution to find another

If $y_1(x)$ is a known solution of

$$y'' + P(x)y' + Q(x)y = 0, \tag{3.53}$$

let the other solution be $y_2(x) = u(x)y_1(x)$. We then form derivatives of y_2 and substitute into the original differential equation. First compute the derivatives:

$$\begin{aligned} y_2' &= uy_1' + u'y_1, & (3.54)\\ y_2'' &= uy_1'' + u'y_1' + u'y_1' + u''y_1, & (3.55)\\ y_2'' &= uy_1'' + 2u'y_1' + u''y_1. & (3.56) \end{aligned}$$

Substituting into Eq. (3.53), we get

$$\underbrace{(uy_1'' + 2u'y_1' + u''y_1)}_{y_2''} + P(x)\underbrace{(uy_1' + u'y_1)}_{y_2'} + Q(x)\underbrace{uy_1}_{y_2} = 0, \tag{3.57}$$

$$u''y_1 + u'(2y_1' + P(x)y_1) + u\underbrace{(y_1'' + P(x)y_1' + Q(x)y_1)}_{=0} = 0, \tag{3.58}$$

$$\text{cancel coefficient on } u: \qquad u''y_1 + u'(2y_1' + P(x)y_1) = 0. \tag{3.59}$$

This can be written as a first-order equation in v, where $v = u'$:

$$v'y_1 + v(2y_1' + P(x)y_1) = 0, \tag{3.60}$$

which is solved for $v(x)$ using known methods for first order equations.

3.2.2.2 Euler equation

An equation of the type

$$x^2\frac{d^2y}{dx^2} + Ax\frac{dy}{dx} + By = 0, \tag{3.61}$$

where A and B are constants, can be solved by a change of independent variables. Let

$$z = \ln x, \tag{3.62}$$

so that

$$x = e^z. \tag{3.63}$$

Then

$$\frac{dz}{dx} = \frac{1}{x} = e^{-z}, \tag{3.64}$$

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = e^{-z}\frac{dy}{dz}, \qquad \text{so} \qquad \frac{d}{dx} = e^{-z}\frac{d}{dz}, \tag{3.65}$$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right), \tag{3.66}$$

$$= e^{-z}\frac{d}{dz}\left(e^{-z}\frac{dy}{dz}\right), \tag{3.67}$$

$$= e^{-2z}\left(\frac{d^2y}{dz^2} - \frac{dy}{dz}\right). \tag{3.68}$$

Substituting into Eq. (3.61), we get

$$\frac{d^2y}{dz^2} + (A-1)\frac{dy}{dz} + By = 0, \tag{3.69}$$

which is an equation with constant coefficients.

In what amounts to the same approach, one can alternatively assume a solution of the form $y = Cx^r$. This leads to a characteristic polynomial for r of

$$r(r-1) + Ar + B = 0. \tag{3.70}$$

The two roots for r induce two linearly independent complementary functions.

Example 3.7

Solve

$$x^2y'' - 2xy' + 2y = 0, \text{ for } x > 0. \tag{3.71}$$

Here $A = -2$ and $B = 2$ in Eq. (3.61). Using this, along with $x = e^z$, we get Eq. (3.69) to reduce to

$$\frac{d^2y}{dz^2} - 3\frac{dy}{dz} + 2y = 0. \tag{3.72}$$

The solution is

$$y = C_1e^z + C_2e^{2z} = C_1x + C_2x^2. \tag{3.73}$$

Note that this equation can also be solved by letting $y = Cx^r$. Substituting into the equation, we get $r^2 - 3r + 2 = 0$, so that $r_1 = 1$ and $r_2 = 2$. The solution is then obtained as a linear combination of x^{r_1} and x^{r_2}.

Example 3.8

Solve

$$x^2\frac{d^2y}{dx^2} + 3x\frac{dy}{dx} + 15y = 0. \tag{3.74}$$

Let us assume here that $y = Cx^r$. Substituting this assumption into Eq. (3.74) yields

$$x^2Cr(r-1)x^{r-2} + 3xCrx^{r-1} + 15Cx^r = 0. \tag{3.75}$$

For $x \neq 0$, $C \neq 0$, we divide by Cx^r to get

$$\begin{aligned} r(r-1) + 3r + 15 &= 0, & (3.76)\\ r^2 + 2r + 15 &= 0. & (3.77) \end{aligned}$$

Solving gives

$$r = -1 \pm i\sqrt{14}. \tag{3.78}$$

Thus, we see there are two linearly independent complementary functions:

$$y(x) = C_1x^{-1+i\sqrt{14}} + C_2x^{-1-i\sqrt{14}}. \tag{3.79}$$

Factoring gives

$$y(x) = \frac{1}{x}\left(C_1x^{i\sqrt{14}} + C_2x^{-i\sqrt{14}}\right). \tag{3.80}$$

Expanding in terms of exponentials and logarithms gives

$$\begin{aligned} y(x) &= \frac{1}{x}\left(C_1(\exp(\ln x))^{i\sqrt{14}} + C_2(\exp(\ln x))^{-i\sqrt{14}}\right), && (3.81)\\ &= \frac{1}{x}\left(C_1\exp(i\sqrt{14}\ln x) + C_2\exp(i\sqrt{14}\ln x)\right), && (3.82)\\ &= \frac{1}{x}\left(\hat{C}_1\cos(\sqrt{14}\ln x) + \hat{C}_2\sin(\sqrt{14}\ln x)\right). && (3.83)\end{aligned}$$

3.3 Particular solutions

We will now consider particular solutions of the inhomogeneous Eq. (3.1).

3.3.1 Method of undetermined coefficients

Guess a solution with unknown coefficients, and then substitute in the equation to determine these coefficients. The number of undetermined coefficients has no relation to the order of the differential equation.

Example 3.9

Consider

$$y'' + 4y' + 4y = 169\sin 3x. \tag{3.84}$$

Thus

$$\begin{aligned} r^2 + 4r + 4 &= 0, && (3.85)\\ (r+2)(r+2) &= 0, && (3.86)\\ r_1 = -2, \qquad r_2 &= -2. && (3.87)\end{aligned}$$

Since the roots are repeated, the complementary functions are

$$y_1 = e^{-2x}, \qquad y_2 = xe^{-2x}. \tag{3.88}$$

For the particular function, guess

$$y_p = a\sin 3x + b\cos 3x, \tag{3.89}$$

so

$$\begin{aligned} y_p' &= 3a\cos 3x - 3b\sin 3x, && (3.90)\\ y_p'' &= -9a\sin 3x - 9b\cos 3x. && (3.91)\end{aligned}$$

Substituting into Eq. (3.84), we get

$$\underbrace{(-9a\sin 3x - 9b\cos 3x)}_{y_p''} + 4\underbrace{(3a\cos 3x - 3b\sin 3x)}_{y_p'} + 4\underbrace{(a\sin 3x + b\cos 3x)}_{y_p} = 169\sin 3x, \tag{3.92}$$

$$(-5a - 12b)\sin 3x + (12a - 5b)\cos 3x = 169\sin 3x, \tag{3.93}$$

$$\underbrace{(-5a - 12b - 169)}_{=0}\sin 3x + \underbrace{(12a - 5b)}_{=0}\cos 3x = 0. \tag{3.94}$$

Now sine and cosine can be shown to be linearly independent. Because of this, since the right hand side of Eq. (3.94) is zero, the constants on the sine and cosine functions must also be zero. This yields the simple system of linear algebraic equations

$$\begin{pmatrix} -5 & -12 \\ 12 & -5 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 169 \\ 0 \end{pmatrix}, \tag{3.95}$$

we find that $a = -5$ and $b = -12$. The solution is then

$$y(x) = (C_1 + C_2 x)e^{-2x} - 5\sin 3x - 12\cos 3x. \tag{3.96}$$

Example 3.10

Solve

$$y'''' - 2y''' + y'' + 2y' - 2y = x^2 + x + 1. \tag{3.97}$$

Let the particular integral be of the form $y_p = ax^2 + bx + c$. Substituting and reducing, we get

$$\underbrace{-(2a+1)}_{=0}x^2 + \underbrace{(4a - 2b - 1)}_{=0}x + \underbrace{(2a + 2b - 2c - 1)}_{=0} = 0. \tag{3.98}$$

Since x^2, x^1 and x^0 are linearly independent, their coefficients in Eq. (3.98) must be zero, from which $a = -1/2$, $b = -3/2$, and $c = -5/2$. Thus,

$$y_p = -\frac{1}{2}(x^2 + 3x + 5). \tag{3.99}$$

The solution of the homogeneous equation was found in a previous example, see Eq. (3.34), so that the general solution is

$$y = C_1 e^{-x} + C_2 e^x + e^x(C_3\cos x + C_4\sin x) - \frac{1}{2}(x^2 + 3x + 5). \tag{3.100}$$

A variant must be attempted if any term of $f(x)$ is a complementary function.

Example 3.11

Solve

$$y'' + 4y = 6\sin 2x. \tag{3.101}$$

Since $\sin 2x$ is a complementary function, we will try

$$y_p = x(a\sin 2x + b\cos 2x), \tag{3.102}$$

from which

$$\begin{aligned} y_p' &= 2x(a\cos 2x - b\sin 2x) + (a\sin 2x + b\cos 2x), & (3.103)\\ y_p'' &= -4x(a\sin 2x + b\cos 2x) + 4(a\cos 2x - b\sin 2x). & (3.104) \end{aligned}$$

Substituting into Eq. (3.101), we compare coefficients and get $a = 0$, $b = -3/2$. The general solution is then

$$y = C_1\sin 2x + C_2\cos 2x - \frac{3}{2}x\cos 2x. \tag{3.105}$$

Example 3.12

Solve

$$y'' + 2y' + y = xe^{-x}. \tag{3.106}$$

The complementary functions are e^{-x} and xe^{-x}. To get the particular solution we have to choose a function of the kind $y_p = ax^3e^{-x}$. On substitution we find that $a = 1/6$. Thus, the general solution is

$$y = C_1e^{-x} + C_2xe^{-x} + \frac{1}{6}x^3e^{-x}. \tag{3.107}$$

3.3.2 Variation of parameters

For an equation of the class

$$P_N(x)y^{(N)} + P_{N-1}(x)y^{(N-1)} + \ldots + P_1(x)y' + P_0(x)y = f(x), \tag{3.108}$$

we propose

$$y_p = \sum_{n=1}^{N} u_n(x)y_n(x), \tag{3.109}$$

where $y_n(x)$, $(n = 1, \ldots, N)$ are complementary functions of the equation, and $u_n(x)$, $(n = 1, \ldots, N)$ are N unknown functions. Differentiating Eq. (3.109), we find

$$y_p' = \underbrace{\sum_{n=1}^{N} u_n' y_n}_{\text{choose to be } 0} + \sum_{n=1}^{N} u_n y_n'. \tag{3.110}$$

We set $\sum_{n=1}^{N} u_n' y_n$ to zero as a first condition. Differentiating the rest of Eq. (3.110), we obtain

$$y_p'' = \underbrace{\sum_{n=1}^{N} u_n' y_n'}_{\text{choose to be } 0} + \sum_{n=1}^{N} u_n y_n''. \tag{3.111}$$

Again we set the first term on the right side of Eq. (3.111) to zero as a second condition. Following this procedure repeatedly we arrive at

$$y_p^{(N-1)} = \underbrace{\sum_{n=1}^{N} u_n' y_n^{(N-2)}}_{\text{choose to be } 0} + \sum_{n=1}^{N} u_n y_n^{(N-1)}. \tag{3.112}$$

The vanishing of the first term on the right gives us the $(N-1)$'th condition. Substituting these into Eq. (3.108), the last condition

$$P_N(x) \sum_{n=1}^{N} u_n' y_n^{(N-1)} + \sum_{n=1}^{N} u_n \underbrace{\left(P_N y_n^{(N)} + P_{N-1} y_n^{(N-1)} + \ldots + P_1 y_n' + P_0 y_n\right)}_{=0} = f(x), \tag{3.113}$$

is obtained. Since each of the functions y_n is a complementary function, the term within brackets is zero.

To summarize, we have the following N equations in the N unknowns u_n', $(n = 1, \ldots, N)$ that we have obtained:

$$\begin{aligned}
\sum_{n=1}^{N} u_n' y_n &= 0, \\
\sum_{n=1}^{N} u_n' y_n' &= 0, \\
&\vdots \\
\sum_{n=1}^{N} u_n' y_n^{(N-2)} &= 0, \\
P_N(x) \sum_{n=1}^{N} u_n' y_n^{(N-1)} &= f(x).
\end{aligned} \tag{3.114}$$

These can be solved for u_n', and then integrated to give the u_n's.

Example 3.13

Solve

$$y'' + y = \tan x. \tag{3.115}$$

The complementary functions are

$$y_1 = \cos x, \qquad y_2 = \sin x. \tag{3.116}$$

The equations for $u_1(x)$ and $u_2(x)$ are

$$\begin{aligned} u_1' y_1 + u_2' y_2 &= 0, & (3.117)\\ u_1' y_1' + u_2' y_2' &= \tan x. & (3.118) \end{aligned}$$

Solving this system, which is linear in u_1' and u_2', we get

$$\begin{aligned} u_1' &= -\sin x \tan x, & (3.119)\\ u_2' &= \cos x \tan x. & (3.120) \end{aligned}$$

Integrating, we get

$$\begin{aligned} u_1 &= \int -\sin x \tan x \; dx = \sin x - \ln|\sec x + \tan x|, & (3.121)\\ u_2 &= \int \cos x \tan x \; dx = -\cos x. & (3.122) \end{aligned}$$

The particular solution is

$$\begin{aligned} y_p &= u_1 y_1 + u_2 y_2, & (3.123)\\ &= (\sin x - \ln|\sec x + \tan x|)\cos x - \cos x \sin x, & (3.124)\\ &= -\cos x \ln|\sec x + \tan x|. & (3.125) \end{aligned}$$

The complete solution, obtained by adding the complementary and particular, is

$$y = C_1 \cos x + C_2 \sin x - \cos x \ln|\sec x + \tan x|. \tag{3.126}$$

3.3.3 Green's functions

A similar goal can be achieved for boundary value problems involving a more general linear operator $\mathbf{L}$, where $\mathbf{L}$ is given by Eq. (3.5). If on the closed interval $a \leq x \leq b$ we have a two point boundary problem for a general linear differential equation of the form:

$$\mathbf{L}y = f(x), \tag{3.127}$$

where the highest derivative in $\mathbf{L}$ is order N and with general homogeneous boundary conditions at $x = a$ and $x = b$ on linear combinations of y and $N - 1$ of its derivatives:

$$\mathbf{A}\left(y(a), y'(a), \ldots, y^{(N-1)}(a)\right)^T + \mathbf{B}\left(y(b), y'(b), \ldots, y^{(N-1)}(b)\right)^T = 0, \tag{3.128}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ constant coefficient matrices. Then, knowing $\mathbf{L}$, $\mathbf{A}$ and $\mathbf{B}$, we can form a solution of the form:

$$y(x) = \int_a^b f(s)g(x,s)ds. \tag{3.129}$$

This is desirable as

- once $g(x, s)$ is known, the solution is defined for *all* f including
  - forms of f for which no simple explicit integrals can be written, and
  - piecewise continuous forms of f,
- numerical solution of the quadrature problem is more robust than direct numerical solution of the original differential equation,
- the solution will automatically satisfy all boundary conditions, and
- the solution is useful in experiments in which the system dynamics are well characterized (e.g. mass-spring-damper) but the forcing may be erratic (perhaps digitally specified).

If the boundary conditions are inhomogeneous, a simple transformation of the dependent variables can be effected to render the boundary conditions to be homogeneous.

We now define the *Green's*[2] *function:* $g(x, s)$ and proceed to show that with this definition, we are guaranteed to achieve the solution to the differential equation in the desired form as shown at the beginning of the section. We take $g(x, s)$ to be the Green's function for the linear differential operator $\mathbf{L}$, as defined by Eq. (3.5), if it satisfies the following conditions:

- $\mathbf{L}g(x, s) = \delta(x - s)$,
- $g(x, s)$ satisfies all boundary conditions given on x,
- $g(x, s)$ is a solution of $\mathbf{L}g = 0$ on $a \leq x < s$ and on $s < x \leq b$,
- $g(x, s), g'(x, s), \ldots, g^{(N-2)}(x, s)$ are continuous for $x \in [a, b]$,
- $g^{(N-1)}(x, s)$ is continuous for $[a, b]$ except at $x = s$ where it has a jump of $1/P_N(s)$; the jump is defined from left to right.

[2]George Green, 1793-1841, English corn-miller and mathematician of humble origin and uncertain education, though he generated modern mathematics of the first rank.

Also *for purposes of these conditions*, s is thought of as a constant parameter. In the actual Green's function representation of the solution, s is a dummy variable. The *Dirac delta function* $\delta(x-s)$ is discussed in the Appendix, Sec. 10.7.10, and in Sec. 7.20 in Kaplan.

These conditions are not all independent; nor is the dependence obvious. Consider for example,

$$\mathbf{L} = P_2(x)\frac{d^2}{dx^2} + P_1(x)\frac{d}{dx} + P_o(x). \tag{3.130}$$

Then we have

$$P_2(x)\frac{d^2g}{dx^2} + P_1(x)\frac{dg}{dx} + P_o(x)g = \delta(x-s), \tag{3.131}$$

$$\frac{d^2g}{dx^2} + \frac{P_1(x)}{P_2(x)}\frac{dg}{dx} + \frac{P_o(x)}{P_2(x)}g = \frac{\delta(x-s)}{P_2(x)}. \tag{3.132}$$

Now integrate both sides with respect to x in a small neighborhood enveloping $x = s$:

$$\int_{s-\epsilon}^{s+\epsilon} \frac{d^2g}{dx^2}\,dx + \int_{s-\epsilon}^{s+\epsilon} \frac{P_1(x)}{P_2(x)}\frac{dg}{dx}\,dx + \int_{s-\epsilon}^{s+\epsilon} \frac{P_o(x)}{P_2(x)}g\,dx = \int_{s-\epsilon}^{s+\epsilon} \frac{\delta(x-s)}{P_2(x)}\,dx. \tag{3.133}$$

Since $P's$ are continuous, as we let $\epsilon \to 0$ we get

$$\int_{s-\epsilon}^{s+\epsilon} \frac{d^2g}{dx^2}\,dx + \frac{P_1(s)}{P_2(s)}\int_{s-\epsilon}^{s+\epsilon} \frac{dg}{dx}\,dx + \frac{P_o(s)}{P_2(s)}\int_{s-\epsilon}^{s+\epsilon} g\,dx = \frac{1}{P_2(s)}\int_{s-\epsilon}^{s+\epsilon} \delta(x-s)\,dx. \tag{3.134}$$

Integrating, we find

$$\left.\frac{dg}{dx}\right|_{s+\epsilon} - \left.\frac{dg}{dx}\right|_{s-\epsilon} + \frac{P_1(s)}{P_2(s)}\underbrace{\left(g|_{s+\epsilon} - g|_{s-\epsilon}\right)}_{\to 0} + \frac{P_o(s)}{P_2(s)}\underbrace{\int_{s-\epsilon}^{s+\epsilon} g\,dx}_{\to 0} = \frac{1}{P_2(s)}\underbrace{H(x-s)|_{s-\epsilon}^{s+\epsilon}}_{\to 1}. \tag{3.135}$$

Since g is continuous, this reduces to

$$\left.\frac{dg}{dx}\right|_{s+\epsilon} - \left.\frac{dg}{dx}\right|_{s-\epsilon} = \frac{1}{P_2(s)}. \tag{3.136}$$

This is consistent with the final point, that the second highest derivative of g suffers a jump at $x = s$.

Next, we show that applying this definition of $g(x,s)$ to our desired result lets us recover the original differential equation, rendering $g(x,s)$ to be appropriately defined. This can be easily shown by direct substitution:

$$y(x) = \int_a^b f(s)g(x,s)ds, \tag{3.137}$$

$$\mathbf{L}y = \mathbf{L}\int_a^b f(s)g(x,s)ds. \tag{3.138}$$

Now $\mathbf{L}$ behaves as $\partial^N/\partial x^N$, via Leibniz's rule, Eq. (1.293)

$$\mathbf{L}y = \int_a^b f(s) \underbrace{\mathbf{L}g(x,s)}_{\delta(x-s)} ds, \tag{3.139}$$

$$= \int_a^b f(s)\delta(x-s)ds, \tag{3.140}$$

$$= f(x). \tag{3.141}$$

Example 3.14

Find the Green's function and the corresponding solution integral of the differential equation

$$\frac{d^2y}{dx^2} = f(x), \tag{3.142}$$

subject to boundary conditions

$$y(0) = 0, \qquad y(1) = 0. \tag{3.143}$$

Verify the solution integral if $f(x) = 6x$.

Here

$$\mathbf{L} = \frac{d^2}{dx^2}. \tag{3.144}$$

Now 1) break the problem up into two domains: a) $x < s$, b) $x > s$, 2) Solve $\mathbf{L}g = 0$ in both domains; four constants arise, 3) Use boundary conditions for two constants, 4) use conditions at $x = s$: continuity of g and a jump of dg/dx, for the other two constants.

a) $x < s$

$$\frac{d^2g}{dx^2} = 0, \tag{3.145}$$

$$\frac{dg}{dx} = C_1, \tag{3.146}$$

$$g = C_1x + C_2, \tag{3.147}$$

$$g(0) = 0 = C_1(0) + C_2, \tag{3.148}$$

$$C_2 = 0, \tag{3.149}$$

$$g(x,s) = C_1x, \qquad x < s. \tag{3.150}$$

b) $x > s$

$$\frac{d^2g}{dx^2} = 0, \tag{3.151}$$

$$\frac{dg}{dx} = C_3, \tag{3.152}$$

$$g = C_3x + C_4, \tag{3.153}$$

$$g(1) = 0 = C_3(1) + C_4, \tag{3.154}$$

$$C_4 = -C_3, \tag{3.155}$$

$$g(x,s) = C_3\,(x-1)\,, \qquad x > s \tag{3.156}$$

Continuity of $g(x,s)$ when $x = s$:

$$C_1 s = C_3 (s-1), \tag{3.157}$$

$$C_1 = C_3 \frac{s-1}{s}, \tag{3.158}$$

$$g(x,s) = C_3 \frac{s-1}{s} x, \qquad x < s, \tag{3.159}$$

$$g(x,s) = C_3 (x-1), \qquad x > s. \tag{3.160}$$

Jump in dg/dx at $x = s$ (note $P_2(x) = 1$):

$$\left.\frac{dg}{dx}\right|_{s+\epsilon} - \left.\frac{dg}{dx}\right|_{s-\epsilon} = 1, \tag{3.161}$$

$$C_3 - C_3 \frac{s-1}{s} = 1, \tag{3.162}$$

$$C_3 = s, \tag{3.163}$$

$$g(x,s) = x(s-1), \qquad x < s, \tag{3.164}$$

$$g(x,s) = s(x-1), \qquad x > s. \tag{3.165}$$

Note some properties of $g(x,s)$ which are common in such problems:

- it is broken into two domains,
- it is continuous in and through both domains,
- its $N-1$ (here $N = 2$, so first) derivative is discontinuous at $x = s$,
- it is symmetric in s and x across the two domains, and
- it is seen by inspection to satisfy both boundary conditions.

The general solution in integral form can be written by breaking the integral into two pieces as

$$y(x) = \int_0^x f(s)\ s(x-1)\ ds + \int_x^1 f(s)\ x(s-1)\ ds, \tag{3.166}$$

$$= (x-1) \int_0^x f(s)\ s\ ds + x \int_x^1 f(s)\ (s-1)\ ds. \tag{3.167}$$

Now evaluate the integral if $f(x) = 6x$ (thus $f(s) = 6s$).

$$y(x) = (x-1) \int_0^x (6s)\ s\ ds + x \int_x^1 (6s)\ (s-1)\ ds, \tag{3.168}$$

$$= (x-1) \int_0^x 6s^2\ ds + x \int_x^1 \left(6s^2 - 6s\right)\ ds, \tag{3.169}$$

$$= (x-1)\ \left(2s^3\right)\Big|_0^x + x\ \left(2s^3 - 3s^2\right)\Big|_x^1, \tag{3.170}$$

$$= (x-1)(2x^3 - 0) + x((2-3) - (2x^3 - 3x^2)), \tag{3.171}$$

$$= 2x^4 - 2x^3 - x - 2x^4 + 3x^3, \tag{3.172}$$

$$y(x) = x^3 - x. \tag{3.173}$$

Note the original differential equation and both boundary conditions are automatically satisfied by the solution. The solution is plotted in Fig. 3.1.

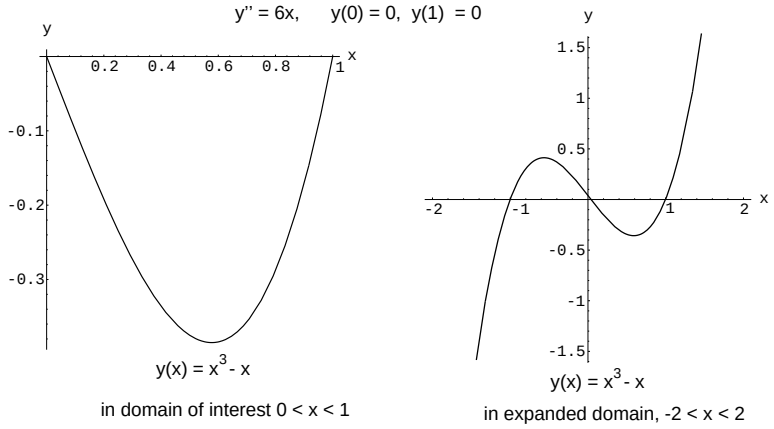

Figure 3.1: Sketch of problem solution, $y'' = 6x, y(0) = y(1) = 0$.

3.3.4 Operator D

The linear operator $\mathbf{D}$ is defined by

$$\mathbf{D}(y) = \frac{dy}{dx}, \tag{3.174}$$

or, in terms of the operator alone,

$$\mathbf{D} = \frac{d}{dx}. \tag{3.175}$$

The operator can be repeatedly applied, so that

$$\mathbf{D}^n(y) = \frac{d^n y}{dx^n}. \tag{3.176}$$

Another example of its use is

$$\begin{aligned} (\mathbf{D} - a)(\mathbf{D} - b)f(x) &= (\mathbf{D} - a)((\mathbf{D} - b)f(x)), & (3.177)\\ &= (\mathbf{D} - a)\left(\frac{df}{dx} - bf\right), & (3.178)\\ &= \frac{d^2 f}{dx^2} - (a+b)\frac{df}{dx} + abf. & (3.179) \end{aligned}$$

Negative powers of $\mathbf{D}$ are related to integrals. This comes from

$$\frac{dy(x)}{dx} = f(x) \qquad y(x_o) = y_o, \tag{3.180}$$

$$y(x) = y_o + \int_{x_o}^{x} f(s)\, ds, \tag{3.181}$$

then

$$\begin{aligned}
\text{substituting:} \qquad \mathbf{D}(y(x)) &= f(x), && (3.182)\\
\text{apply inverse:} \qquad \mathbf{D}^{-1}(\mathbf{D}(y(x))) &= \mathbf{D}^{-1}(f(x)), && (3.183)\\
y(x) &= \mathbf{D}^{-1}(f(x)), && (3.184)\\
&= y_o + \int_{x_o}^{x} f(s)\, ds, && (3.185)\\
\text{so} \qquad \mathbf{D}^{-1} &= y_o + \int_{x_o}^{x} (\ldots)\, ds. && (3.186)
\end{aligned}$$

We can evaluate $h(x)$ where

$$h(x) = \frac{1}{\mathbf{D}-a} f(x), \tag{3.187}$$

in the following way

$$\begin{aligned}
(\mathbf{D}-a)\, h(x) &= (\mathbf{D}-a)\left(\frac{1}{\mathbf{D}-a} f(x)\right), && (3.188)\\
(\mathbf{D}-a)\, h(x) &= f(x), && (3.189)\\
\frac{dh(x)}{dx} - ah(x) &= f(x), && (3.190)\\
e^{-ax}\frac{dh(x)}{dx} - ae^{-ax}h(x) &= f(x)e^{-ax}, && (3.191)\\
\frac{d}{dx}\left(e^{-ax}h(x)\right) &= f(x)e^{-ax}, && (3.192)\\
\frac{d}{ds}\left(e^{-as}h(s)\right) &= f(s)e^{-as}, && (3.193)\\
\int_{x_o}^{x} \frac{d}{ds}\left(e^{-as}h(s)\right)\, ds &= \int_{x_o}^{x} f(s)e^{-as}\, ds, && (3.194)\\
e^{-ax}h(x) - e^{-ax_o}h(x_o) &= \int_{x_o}^{x} f(s)e^{-as}\, ds, && (3.195)\\
h(x) &= e^{a(x-x_o)}h(x_o) + e^{ax}\int_{x_o}^{x} f(s)e^{-as}\, ds, && (3.196)\\
\frac{1}{\mathbf{D}-a} f(x) &= e^{a(x-x_o)}h(x_o) + e^{ax}\int_{x_o}^{x} f(s)e^{-as}\, ds. && (3.197)
\end{aligned}$$

This gives us $h(x)$ explicitly in terms of the known function f such that h satisfies $\mathbf{D}(h) - ah = f$.

We can find the solution to higher order equations such as

$$(\mathbf{D}-a)(\mathbf{D}-b)y(x) = f(x), \qquad y(x_o) = y_o,\ y'(x_o) = y'_o, \tag{3.198}$$

$$(\mathbf{D}-b)y(x) = \frac{1}{\mathbf{D}-a}f(x), \tag{3.199}$$

$$(\mathbf{D}-b)y(x) = h(x), \tag{3.200}$$

$$y(x) = y_o e^{b(x-x_o)} + e^{bx}\int_{x_o}^{x} h(s)e^{-bs}\,ds. \tag{3.201}$$

Note that

$$\frac{dy}{dx} = y_o b e^{b(x-x_o)} + h(x) + be^{bx}\int_{x_o}^{x} h(s)e^{-bs}\,ds, \tag{3.202}$$

$$\frac{dy}{dx}(x_o) = y_o' = y_o b + h(x_o), \tag{3.203}$$

which can be rewritten as

$$(\mathbf{D}-b)(y(x_o)) = h(x_o), \tag{3.204}$$

which is what one would expect.

Returning to the problem at hand, we take our expression for $h(x)$, evaluate it at $x = s$ and substitute into the expression for $y(x)$ to get

$$y(x) = y_o e^{b(x-x_o)} + e^{bx}\int_{x_o}^{x}\left(h(x_o)e^{a(s-x_o)} + e^{as}\int_{x_o}^{s} f(t)e^{-at}\,dt\right)e^{-bs}\,ds, \tag{3.205}$$

$$= y_o e^{b(x-x_o)} + e^{bx}\int_{x_o}^{x}\left((y_o' - y_o b)\,e^{a(s-x_o)} + e^{as}\int_{x_o}^{s} f(t)e^{-at}dt\right)e^{-bs}ds, \tag{3.206}$$

$$= y_o e^{b(x-x_o)} + e^{bx}\int_{x_o}^{x}\left((y_o' - y_o b)\,e^{(a-b)s-ax_o} + e^{(a-b)s}\int_{x_o}^{s} f(t)e^{-at}dt\right)ds, \tag{3.207}$$

$$= y_o e^{b(x-x_o)} + e^{bx}\,(y_o' - y_o b)\int_{x_o}^{x} e^{(a-b)s-ax_o}ds + e^{bx}\int_{x_o}^{x} e^{(a-b)s}\left(\int_{x_o}^{s} f(t)e^{-at}dt\right)ds, \tag{3.208}$$

$$= y_o e^{b(x-x_o)} + e^{bx}\,(y_o' - y_o b)\,\frac{e^{a(x-x_o)-xb} - e^{-bx_o}}{a-b} + e^{bx}\int_{x_o}^{x} e^{(a-b)s}\left(\int_{x_o}^{s} f(t)e^{-at}dt\right)ds, \tag{3.209}$$

$$= y_o e^{b(x-x_o)} + (y_o' - y_o b)\,\frac{e^{a(x-x_o)} - e^{b(x-x_o)}}{a-b} + e^{bx}\int_{x_o}^{x} e^{(a-b)s}\left(\int_{x_o}^{s} f(t)e^{-at}dt\right)ds, \tag{3.210}$$

$$= y_o e^{b(x-x_o)} + (y_o' - y_o b)\,\frac{e^{a(x-x_o)} - e^{b(x-x_o)}}{a-b} + e^{bx}\int_{x_o}^{x}\int_{x_o}^{s} e^{(a-b)s}f(t)e^{-at}dt\,ds. \tag{3.211}$$

Changing the order of integration and integrating on s, we get

$$y(x) = y_o e^{b(x-x_o)} + (y_o' - y_o b)\,\frac{e^{a(x-x_o)} - e^{b(x-x_o)}}{a-b} + e^{bx}\int_{x_o}^{x}\int_{t}^{x} e^{(a-b)s}f(t)e^{-at}ds\,dt,$$

$$= y_o e^{b(x-x_o)} + (y_o' - y_o b)\,\frac{e^{a(x-x_o)} - e^{b(x-x_o)}}{a-b} + e^{bx}\int_{x_o}^{x} f(t)e^{-at}\left(\int_t^x e^{(a-b)s}ds\right)\,dt, \tag{3.212}$$

$$= y_o e^{b(x-x_o)} + (y_o' - y_o b)\,\frac{e^{a(x-x_o)} - e^{b(x-x_o)}}{a-b} + \int_{x_o}^{x} \frac{f(t)}{a-b}\left(e^{a(x-t)} - e^{b(x-t)}\right)\,dt. \tag{3.213}$$

$$\tag{3.214}$$

Thus, we have a solution to the second order linear differential equation with constant coefficients and arbitrary forcing expressed in integral form. A similar alternate expression can be developed when $a = b$.

Problems

1. Find the general solution of the differential equation
$$y' + x^2 y(1+y) = 1 + x^3(1+x).$$

2. Show that the functions $y_1 = \sin x$, $y_2 = x\cos x$, and $y_3 = x$ are linearly independent. Find the lowest order differential equation of which they are the complementary functions.

3. Solve the following initial value problem for (a) $C = 6$, (b) $C = 4$, and (c) $C = 3$ with $y(0) = 1$ and $y'(0) = -3$.
$$\frac{d^2y}{dt^2} + C\frac{dy}{dt} + 4y = 0.$$
Plot your results.

4. Solve
   (a) $\frac{d^3y}{dx^3} - 3\frac{d^2y}{dx^2} + 4y = 0$,
   (b) $\frac{d^4y}{dx^4} - 5\frac{d^3y}{dx^3} + 11\frac{d^2y}{dx^2} - 7\frac{dy}{dx} = 12$,
   (c) $y'' + 2y = 6e^x + \cos 2x$,
   (d) $x^2y'' - 3xy' - 5y = x^2\log x$,
   (e) $\frac{d^2y}{dx^2} + y = 2e^x\cos x + (e^x - 2)\sin x$.

5. Find a particular solution to the following ODE using (a) variation of parameters and (b) undetermined coefficients.
$$\frac{d^2y}{dx^2} - 4y = \cosh 2x.$$

6. Solve the boundary value problem
$$\frac{d^2y}{dx^2} + y\frac{dy}{dx} = 0,$$
with boundary conditions $y(0) = 0$ and $y(\pi/2) = -1$ Plot your result.

7. Solve
$$2x^2\frac{d^3y}{dx^3} + 2x\frac{d^2y}{dx^2} - 8\frac{dy}{dx} = 1,$$
with $y(1) = 4$, $y'(1) = 8$, $y(2) = 11$. Plot your result.

8. Solve
$$x^2y'' + xy' - 4y = 6x.$$

9. Find the general solution of
$$y'' + 2y' + y = xe^{-x}.$$

10. Find the Green's function solution of
$$y'' + y' - 2y = f(x),$$
with $y(0) = 0$, $y'(1) = 0$. Determine $y(x)$ if $f(x) = 3\sin x$. Plot your result.

11. Find the Green's function solution of
$$y'' + 4y = f(x),$$
with $y(0) = y(1)$, $y'(0) = 0$. Verify this is the correct solution when $f(x) = x^2$. Plot your result.

12. Solve $y''' - 2y'' - y' + 2y = \sin^2 x$.

13. Solve $y''' + 6y'' + 12y' + 8y = e^x - 3\sin x - 8e^{-2x}$.

14. Solve $x^4y'''' + 7x^3y''' + 8x^2y'' = 4x^{-3}$.

15. Show that x^{-1} and x^5 are solutions of the equation
$$x^2y'' - 3xy' - 5y = 0.$$
Thus, find the general solution of
$$x^2y'' - 3xy' - 5y = x^2.$$

16. Solve the equation
$$2y'' - 4y' + 2y = \frac{e^x}{x},$$
where $x > 0$.

Chapter 4

Series solution methods

see Kaplan, Chapter 6,
see Hinch, Chapters 1, 2, 5, 6, 7,
see Bender and Orszag,
see Kervorkian and Cole,
see Van Dyke,
see Murdock,
see Holmes,
see Lopez, Chapters 7-11, 14,
see Riley, Hobson, and Bence, Chapter 14.

This chapter will deal with series solution methods. Such methods are useful in solving both algebraic and differential equations. The first method is formally exact in that an infinite number of terms can often be shown to have absolute and uniform convergence properties. The second method, asymptotic series solutions, is less rigorous in that convergence is not always guaranteed; in fact convergence is rarely examined because the problems tend to be intractable. Still asymptotic methods will be seen to be quite useful in interpreting the results of highly non-linear equations in local domains.

4.1 Power series

Solutions to many differential equations cannot be found in a closed form solution expressed for instance in terms of polynomials and transcendental functions such as sin and cos. Often, instead, the solutions can be expressed as an infinite series of polynomials. It is desirable to get a complete expression for the n^{th} term of the series so that one can make statements regarding absolute and uniform convergence of the series. Such solutions are approximate in that if one uses a finite number of terms to represent the solution, there is a truncation error. Formally though, for series which converge, an infinite number of terms gives a true representation of the actual solution, and hence the method is exact.

A function $f(x)$ is said to be *analytic* if it is an infinitely differentiable function such that

the Taylor series, $\sum_{n=0}^{\infty} f^{(n)}(x_o)(x-x_o)^n/n!$, at any point $x = x_o$ in its domain converges to $f(x)$ in a neighborhood of $x = x_o$.

4.1.1 First-order equation

An equation of the form

$$\frac{dy}{dx} + P(x)y = Q(x), \tag{4.1}$$

where $P(x)$ and $Q(x)$ are analytic at $x = a$, has a power series solution

$$y(x) = \sum_{n=0}^{\infty} a_n (x-a)^n, \tag{4.2}$$

around this point.

Example 4.1

Find the power series solution of

$$\frac{dy}{dx} = y \qquad y(0) = y_o, \tag{4.3}$$

around $x = 0$.

Let

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots, \tag{4.4}$$

so that

$$\frac{dy}{dx} = a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + \cdots. \tag{4.5}$$

Substituting into Eq. (4.3), we have

$$\underbrace{a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + \cdots}_{dy/dx} = \underbrace{a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots}_{y}, \tag{4.6}$$

$$\underbrace{(a_1 - a_0)}_{=0} + \underbrace{(2a_2 - a_1)}_{=0} x + \underbrace{(3a_3 - a_2)}_{=0} x^2 + \underbrace{(4a_4 - a_3)}_{=0} x^3 + \cdots = 0 \tag{4.7}$$

Because the polynomials x^0, x^1, x^2, ... are linearly independent, the coefficients must be all zero. Thus,

$$a_1 = a_0, \tag{4.8}$$

$$a_2 = \frac{1}{2} a_1 = \frac{1}{2} a_0, \tag{4.9}$$

$$a_3 = \frac{1}{3} a_2 = \frac{1}{3!} a_0, \tag{4.10}$$

$$a_4 = \frac{1}{4} a_3 = \frac{1}{4!} a_0, \tag{4.11}$$

$$\vdots$$

so that

$$y(x) = a_0 \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \right). \tag{4.12}$$

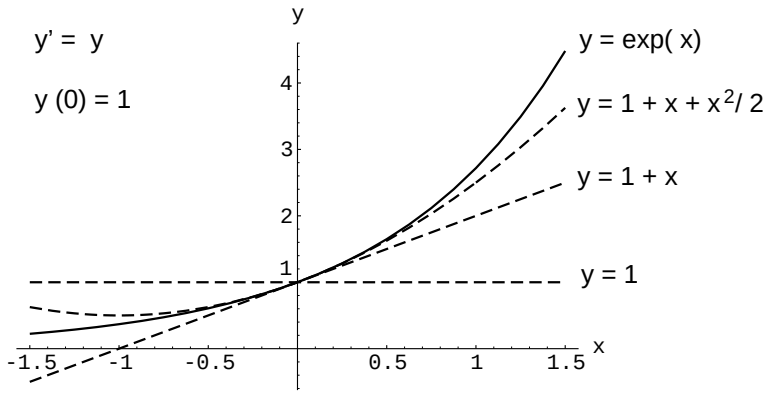

Figure 4.1: Comparison of truncated series and exact solutions.

Applying the initial condition at $x = 0$ gives $a_0 = y_o$ so

$$y(x) = y_o \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots\right). \tag{4.13}$$

Of course this power series is the Taylor series expansion, see Sec. 10.1, of the closed form solution $y = y_o e^x$ about $x = 0$. The power series solution about a different point will give a different solution.

For $y_o = 1$ the exact solution and three approximations to the exact solution are shown in Figure 4.1. Alternatively, one can use a compact summation notation. Thus,

$$\begin{aligned}
y &= \sum_{n=0}^{\infty} a_n x^n, && (4.14)\\
\frac{dy}{dx} &= \sum_{n=0}^{\infty} n a_n x^{n-1}, && (4.15)\\
&= \sum_{n=1}^{\infty} n a_n x^{n-1}, && (4.16)\\
m = n - 1 \qquad &= \sum_{m=0}^{\infty} (m+1) a_{m+1} x^m, && (4.17)\\
&= \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n. && (4.18)
\end{aligned}$$

Thus, the differential equation becomes

$$\begin{aligned}
\underbrace{\sum_{n=0}^{\infty} (n+1) a_{n+1} x^n}_{dy/dx} &= \underbrace{\sum_{n=0}^{\infty} a_n x^n}_{y}, && (4.19)\\
\sum_{n=0}^{\infty} \underbrace{\left((n+1)a_{n+1} - a_n\right)}_{=0} x^n &= 0, && (4.20)\\
(n+1)a_{n+1} &= a_n, && (4.21)\\
a_{n+1} &= \frac{1}{n+1} a_n, && (4.22)
\end{aligned}$$

$$a_n = \frac{a_0}{n!}, \tag{4.23}$$

$$y = a_0 \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{4.24}$$

$$y = y_o \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{4.25}$$

The ratio test tells us that

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \frac{1}{n+1} \to 0, \tag{4.26}$$

so the series converges absolutely.

If a series is *uniformly* convergent in a domain, it converges at the same rate for all x in that domain. We can use the Weierstrass[1] M-test for uniform convergence. That is for a series

$$\sum_{n=0}^{\infty} u_n(x), \tag{4.27}$$

to be convergent, we need a convergent series of constants M_n to exist

$$\sum_{n=0}^{\infty} M_n, \tag{4.28}$$

such that

$$|u_n(x)| \leq M_n, \tag{4.29}$$

for all x in the domain. For our problem, we take $x \in [-A, A]$, where $A > 0$.

So for *uniform* convergence we must have

$$\left| \frac{x^n}{n!} \right| \leq M_n. \tag{4.30}$$

So take

$$M_n = \frac{A^n}{n!}. \tag{4.31}$$

(Note M_n is thus strictly positive). So

$$\sum_{n=0}^{\infty} M_n = \sum_{n=0}^{\infty} \frac{A^n}{n!}. \tag{4.32}$$

By the ratio test, this is convergent if

$$\lim_{n\to\infty} \left| \frac{\frac{A^{n+1}}{(n+1)!}}{\frac{A^n}{(n)!}} \right| \leq 1, \tag{4.33}$$

$$\lim_{n\to\infty} \left| \frac{A}{n+1} \right| \leq 1. \tag{4.34}$$

This holds for all A, so for $x \in (-\infty, \infty)$ the series converges absolutely and uniformly.

[1] Karl Theodor Wilhelm Weierstrass, 1815-1897, Westphalia-born German mathematician.

4.1.2 Second-order equation

We consider series solutions of

$$P(x)\frac{d^2y}{dx^2} + Q(x)\frac{dy}{dx} + R(x)y = 0, \tag{4.35}$$

around $x = a$. There are three different cases, depending of the behavior of $P(a)$, $Q(a)$ and $R(a)$, in which $x = a$ is classified as an ordinary point, a regular singular point, or an irregular singular point. These are described next.

4.1.2.1 Ordinary point

If $P(a) \neq 0$ and Q/P, R/P are analytic at $x = a$, this point is called an *ordinary* point. The general solution is $y = C_1y_1(x) + C_2y_2(x)$ where y_1 and y_2 are of the form $\sum_{n=0}^{\infty} a_n(x-a)^n$. The radius of convergence of the series is the distance to the nearest complex singularity, i.e. the distance between $x = a$ and the closest point on the complex plane at which Q/P or R/P is not analytic.

Example 4.2

Find the series solution of

$$y'' + xy' + y = 0, \qquad y(0) = y_o, \qquad y'(0) = y'_o, \tag{4.36}$$

around $x = 0$.

The point $x = 0$ is an ordinary point, so that we have

$$y = \sum_{n=0}^{\infty} a_n x^n, \tag{4.37}$$

$$y' = \sum_{n=1}^{\infty} n a_n x^{n-1}, \tag{4.38}$$

$$xy' = \sum_{n=1}^{\infty} n a_n x^n, \tag{4.39}$$

$$xy' = \sum_{n=0}^{\infty} n a_n x^n, \tag{4.40}$$

$$y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2}, \tag{4.41}$$

$$m = n-2, \qquad y'' = \sum_{m=0}^{\infty} (m+1)(m+2) a_{m+2} x^m, \tag{4.42}$$

$$= \sum_{n=0}^{\infty} (n+1)(n+2) a_{n+2} x^n. \tag{4.43}$$

Substituting into Eq. (4.36), we get

$$\sum_{n=0}^{\infty} \underbrace{((n+1)(n+2)a_{n+2} + na_n + a_n)}_{=0} x^n = 0. \tag{4.44}$$

Equating the coefficients to zero, we get

$$a_{n+2} = -\frac{1}{n+2} a_n, \tag{4.45}$$

so that

$$y = a_0 \left(1 - \frac{x^2}{2} + \frac{x^4}{4\cdot 2} - \frac{x^6}{6\cdot 4\cdot 2} + \cdots\right) + a_1 \left(x - \frac{x^3}{3} + \frac{x^5}{5\cdot 3} - \frac{x^7}{7\cdot 5\cdot 3} + \cdots\right), \tag{4.46}$$

$$y = y_o \left(1 - \frac{x^2}{2} + \frac{x^4}{4\cdot 2} - \frac{x^6}{6\cdot 4\cdot 2} + \cdots\right) + y_o' \left(x - \frac{x^3}{3} + \frac{x^5}{5\cdot 3} - \frac{x^7}{7\cdot 5\cdot 3} + \cdots\right), \tag{4.47}$$

$$y = y_o \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} x^{2n} + y_o' \sum_{n=1}^{\infty} \frac{(-1)^{n-1} 2^n n!}{(2n)!} x^{2n-1}, \tag{4.48}$$

$$y = y_o \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{-x^2}{2}\right)^n - \frac{y_o'}{x} \sum_{n=1}^{\infty} \frac{n!}{(2n)!} \left(-2x^2\right)^n. \tag{4.49}$$

The series converges for all x. For $y_o = 1, y_o' = 0$ the exact solution, which can be shown to be

$$y = \exp\left(-\frac{x^2}{2}\right), \tag{4.50}$$

and two approximations to the exact solution are shown in Fig. 4.2. For arbitrary y_o and y_o', the solution can be shown to be

$$y = \exp\left(-\frac{x^2}{2}\right) \left(y_o + \sqrt{\frac{\pi}{2}} y_o' \text{erfi}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{4.51}$$

Here "erfi" is the so-called *imaginary error function*; see Sec. 10.7.4 of the Appendix.

4.1.2.2 Regular singular point

If $P(a) = 0$, then $x = a$ is a *singular* point. Furthermore, if $(x-a)Q/P$ and $(x-a)^2 R/P$ are both analytic at $x = a$, this point is called a *regular* singular point. Then there exists at least *one* solution of the form

$$y(x) = (x-a)^r \sum_{n=0}^{\infty} a_n (x-a)^n = \sum_{n=0}^{\infty} a_n (x-a)^{n+r}. \tag{4.52}$$

This is known as the *Frobenius*[2] method. The radius of convergence of the series is again the distance to the nearest complex singularity.

An equation for r is called the *indicial* equation. The following are the different kinds of solutions of the indicial equation possible:

[2]Ferdinand Georg Frobenius, 1849-1917, Prussian/German mathematician.

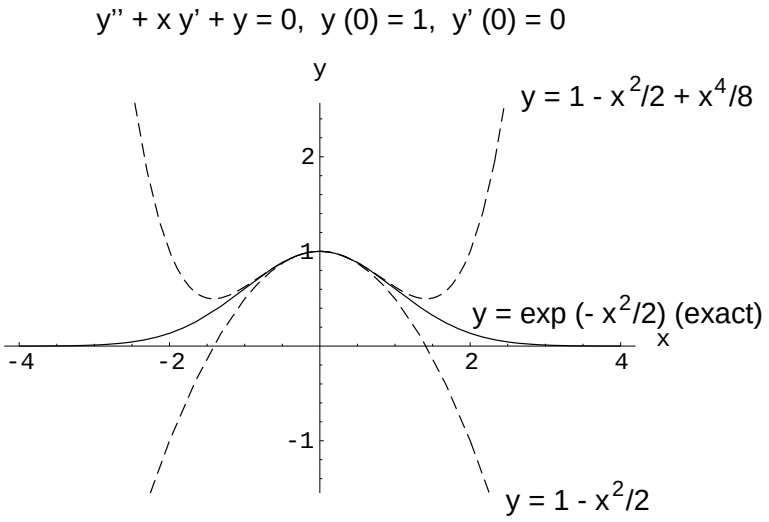

Figure 4.2: Comparison of truncated series and exact solutions.

- $r_1 \neq r_2$, and $r_1 - r_2$ not an integer. Then

$$y_1 = (x-a)^{r_1} \sum_{n=0}^{\infty} a_n (x-a)^n = \sum_{n=0}^{\infty} a_n (x-a)^{n+r_1}, \tag{4.53}$$

$$y_2 = (x-a)^{r_2} \sum_{n=0}^{\infty} b_n (x-a)^n = \sum_{n=0}^{\infty} a_n (x-a)^{n+r_2}. \tag{4.54}$$

- $r_1 = r_2 = r$. Then

$$y_1 = (x-a)^{r} \sum_{n=0}^{\infty} a_n (x-a)^n = \sum_{n=0}^{\infty} a_n (x-a)^{n+r}, \tag{4.55}$$

$$y_2 = y_1 \ln x + (x-a)^{r} \sum_{n=0}^{\infty} b_n (x-a)^n = y_1 \ln x + \sum_{n=0}^{\infty} b_n (x-a)^{n+r}. \tag{4.56}$$

- $r_1 \neq r_2$, and $r_1 - r_2$ is a positive integer.

$$y_1 = (x-a)^{r_1} \sum_{n=0}^{\infty} a_n (x-a)^n = \sum_{n=0}^{\infty} a_n (x-a)^{n+r_1}, \tag{4.57}$$

$$y_2 = k y_1 \ln x + (x-a)^{r_2} \sum_{n=0}^{\infty} b_n (x-a)^n = k y_1 \ln x + \sum_{n=0}^{\infty} b_n (x-a)^{n+r_2}. \tag{4.58}$$

The constants a_n and k are determined by the differential equation. The general solution is

$$y(x) = C_1 y_1(x) + C_2 y_2(x). \tag{4.59}$$

Example 4.3

Find the series solution of

$$4xy'' + 2y' + y = 0, \tag{4.60}$$

around $x = 0$.

The point $x = 0$ is a regular singular point. So we have $a = 0$ and take

$$y = x^r \sum_{n=0}^{\infty} a_n x^n, \tag{4.61}$$

$$y = \sum_{n=0}^{\infty} a_n x^{n+r}, \tag{4.62}$$

$$y' = \sum_{n=0}^{\infty} a_n (n+r) x^{n+r-1}, \tag{4.63}$$

$$y'' = \sum_{n=0}^{\infty} a_n (n+r)(n+r-1) x^{n+r-2}, \tag{4.64}$$

$$\underbrace{4 \sum_{n=0}^{\infty} a_n (n+r)(n+r-1) x^{n+r-1}}_{=4xy''} + \underbrace{2 \sum_{n=0}^{\infty} a_n (n+r) x^{n+r-1}}_{=2y'} + \underbrace{\sum_{n=0}^{\infty} a_n x^{n+r}}_{=y} = 0, \tag{4.65}$$

$$2 \sum_{n=0}^{\infty} a_n (n+r)(2n+2r-1) x^{n+r-1} + \sum_{n=0}^{\infty} a_n x^{n+r} = 0, \tag{4.66}$$

$m = n - 1$

$$2 \sum_{m=-1}^{\infty} a_{m+1} (m+1+r)(2(m+1)+2r-1) x^{m+r} + \sum_{n=0}^{\infty} a_n x^{n+r} = 0, \tag{4.67}$$

$$2 \sum_{n=-1}^{\infty} a_{n+1} (n+1+r)(2(n+1)+2r-1) x^{n+r} + \sum_{n=0}^{\infty} a_n x^{n+r} = 0, \tag{4.68}$$

$$2a_0 r(2r-1) x^{-1+r} + 2 \sum_{n=0}^{\infty} a_{n+1} (n+1+r)(2(n+1)+2r-1) x^{n+r} + \sum_{n=0}^{\infty} a_n x^{n+r} = 0. \tag{4.69}$$

The first term ($n = -1$) gives the indicial equation:

$$r(2r-1) = 0, \tag{4.70}$$

from which $r = 0, 1/2$. We then have

$$2 \sum_{n=0}^{\infty} a_{n+1} (n+r+1)(2n+2r+1) x^{n+r} + \sum_{n=0}^{\infty} a_n x^{n+r} = 0, \tag{4.71}$$

$$\sum_{n=0}^{\infty} \underbrace{\left(2a_{n+1}(n+r+1)(2n+2r+1) + a_n\right)}_{=0} x^{n+r} = 0. \tag{4.72}$$

For $r = 0$

$$a_{n+1} = -a_n \frac{1}{(2n+2)(2n+1)}, \tag{4.73}$$

$$y_1 = a_0 \left(1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \cdots\right). \tag{4.74}$$

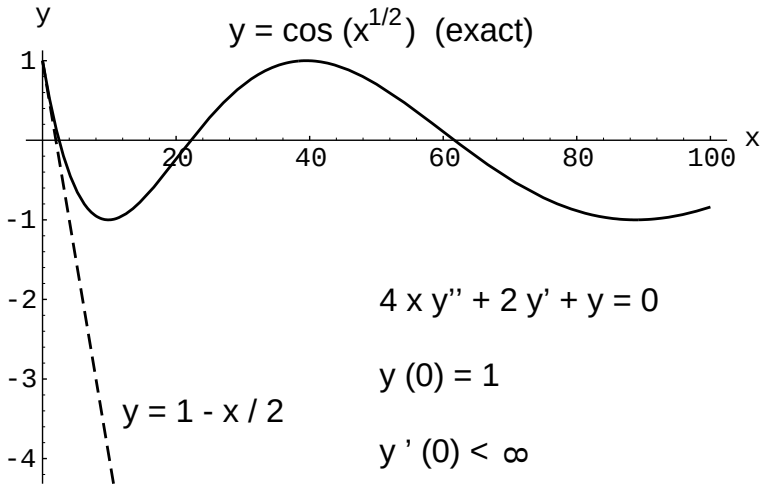

Figure 4.3: Comparison of truncated series and exact solutions.

For $r = 1/2$

$$a_{n+1} = -a_n \frac{1}{2(2n+3)(n+1)}, \tag{4.75}$$

$$y_2 = a_0 x^{1/2} \left(1 - \frac{x}{3!} + \frac{x^2}{5!} - \frac{x^3}{7!} + \cdots\right). \tag{4.76}$$

The series converges for all x to $y_1 = \cos\sqrt{x}$ and $y_2 = \sin\sqrt{x}$. The general solution is

$$y = C_1 y_1 + C_2 y_2, \tag{4.77}$$

or

$$y(x) = C_1 \cos\sqrt{x} + C_2 \sin\sqrt{x}. \tag{4.78}$$

Note that $y(x)$ is real and non-singular for $x \in [0, \infty)$. However, the first derivative

$$y'(x) = -C_1 \frac{\sin\sqrt{x}}{2\sqrt{x}} + C_2 \frac{\cos\sqrt{x}}{2\sqrt{x},} \tag{4.79}$$

is singular at $x = 0$. The nature of the singularity is seen from a Taylor series expansion of $y'(x)$ about $x = 0$, which gives

$$y'(x) \sim C_1 \left(-\frac{1}{2} + \frac{x}{12} + \ldots\right) + C_2 \left(\frac{1}{2\sqrt{x}} - \frac{\sqrt{x}}{4} + \ldots\right). \tag{4.80}$$

So there is a weak $1/\sqrt{x}$ singularity in $y'(x)$ at $x = 0$.

For $y(0) = 1$, $y'(0) < \infty$, the exact solution and the linear approximation to the exact solution are shown in Fig. 4.3. For this case, one has $C_1 = 1$ to satisfy the condition on $y(0)$, and one must have $C_2 = 0$ to satisfy the non-singular condition on $y'(0)$.

Example 4.4

Find the series solution of

$$xy'' - y = 0, \tag{4.81}$$

around $x = 0$.

Let $y = \sum_{n=0}^{\infty} a_n x^{n+r}$. Then, from Eq. (4.81)

$$r(r-1)a_0 x^{r-1} + \sum_{n=1}^{\infty} \left((n+r)(n+r-1)a_n - a_{n-1}\right) x^{n+r-1} = 0. \tag{4.82}$$

The indicial equation is $r(r-1) = 0$, from which $r = 0, 1$.

Consider the larger of the two, i.e. $r = 1$. For this we get

$$\begin{aligned} a_n &= \frac{1}{n(n+1)} a_{n-1}, & (4.83)\\ &= \frac{1}{n!(n+1)!} a_0. & (4.84) \end{aligned}$$

Thus,

$$y_1(x) = x + \frac{1}{2}x^2 + \frac{1}{12}x^3 + \frac{1}{144}x^4 + \ldots. \tag{4.85}$$

From Eq. (4.58), the second solution is

$$y_2(x) = k y_1(x) \ln x + \sum_{n=0}^{\infty} b_n x^n. \tag{4.86}$$

It has derivatives

$$\begin{aligned} y_2'(x) &= k\frac{y_1(x)}{x} + k y_1'(x) \ln x + \sum_{n=0}^{\infty} n b_n x^{n-1}, & (4.87)\\ y_2''(x) &= -k\frac{y_1(x)}{x^2} + 2k\frac{y_1'(x)}{x} + k y_1''(x) \ln x + \sum_{n=0}^{\infty} n(n-1) b_n x^{n-2}. & (4.88) \end{aligned}$$

To take advantage of Eq. (4.81), let us multiply the second derivative by x.

$$xy_2''(x) = -k\frac{y_1(x)}{x} + 2k y_1'(x) + k \underbrace{x y_1''(x)}_{=y_1(x)} \ln x + \sum_{n=0}^{\infty} n(n-1) b_n x^{n-1}. \tag{4.89}$$

Now since y_1 is a solution of Eq. (4.81), we have $xy_1'' = y_1$; thus,

$$xy_2''(x) = -k\frac{y_1(x)}{x} + 2k y_1'(x) + k y_1(x) \ln x + \sum_{n=0}^{\infty} n(n-1) b_n x^{n-1}. \tag{4.90}$$

Now subtract Eq. (4.86) from both sides and then enforce Eq. (4.81) to get

$$\begin{aligned} 0 = xy_2''(x) - y_2(x) &= -k\frac{y_1(x)}{x} + 2k y_1'(x) + k y_1(x) \ln x + \sum_{n=0}^{\infty} n(n-1) b_n x^{n-1}\\ &\quad - \left(k y_1(x) \ln x + \sum_{n=1}^{\infty} b_n x^n \right). \end{aligned} \tag{4.91}$$

Simplifying and rearranging, we get

$$-\frac{ky_1(x)}{x} + 2ky_1'(x) + \sum_{n=0}^{\infty} n(n-1)b_n x^{n-1} - \sum_{n=0}^{\infty} b_n x^n = 0. \tag{4.92}$$

Substituting the solution $y_1(x)$ already obtained, we get

$$\begin{aligned} 0 &= -k\left(1 + \frac{1}{2}x + \frac{1}{12}x^2 + \ldots\right) + 2k\left(1 + x + \frac{1}{2}x^2 + \ldots\right) \\ &\quad + \left(2b_2x + 6b_3x^2 + \ldots\right) - \left(b_0 + b_1x + b_2x^2 + \ldots\right). \end{aligned} \tag{4.93}$$

Collecting terms, we have

$$k = b_0, \tag{4.94}$$

$$b_{n+1} = \frac{1}{n(n+1)}\left(b_n - \frac{k(2n+1)}{n!(n+1)!}\right) \text{ for } n = 1, 2, \ldots. \tag{4.95}$$

Thus,

$$\begin{aligned} y_2(x) &= b_0 y_1 \ln x + b_0\left(1 - \frac{3}{4}x^2 - \frac{7}{36}x^3 - \frac{35}{1728}x^4 - \ldots\right) \\ &\quad + b_1 \underbrace{\left(x + \frac{1}{2}x^2 + \frac{1}{12}x^3 + \frac{1}{144}x^4 + \ldots\right)}_{=y_1}. \end{aligned} \tag{4.96}$$

Since the last part of the series, shown in an under-braced term, is actually $y_1(x)$, and we already have C_1y_1 as part of the solution, we choose $b_1 = 0$. Because we also allow for a C_2, we can then set $b_0 = 1$. Thus, we take

$$y_2(x) = y_1 \ln x + \left(1 - \frac{3}{4}x^2 - \frac{7}{36}x^3 - \frac{35}{1728}x^4 - \ldots\right). \tag{4.97}$$

The general solution, $y = C_1y_1 + C_2y_2$, is

$$\begin{aligned} y(x) &= C_1 \underbrace{\left(x + \frac{1}{2}x^2 + \frac{1}{12}x^3 + \frac{1}{144}x^4 + \ldots\right)}_{y_1} \\ &\quad + C_2 \underbrace{\left(\left(x + \frac{1}{2}x^2 + \frac{1}{12}x^3 + \frac{1}{144}x^4 + \ldots\right)\ln x + \left(1 - \frac{3}{4}x^2 - \frac{7}{36}x^3 - \frac{35}{1728}x^4 - \ldots\right)\right)}_{y_2}. \end{aligned} \tag{4.98}$$

It can also be shown that the solution can be represented compactly as

$$y(x) = \sqrt{x}\left(C_1I_1(2\sqrt{x}) + C_2K_1(2\sqrt{x})\right), \tag{4.99}$$

where I_1 and K_1 are what is known as *modified Bessel functions of the first and second kinds, respectively, both of order 1.* The function $I_1(s)$ is non-singular, while $K_1(s)$ is singular at $s = 0$.

4.1.2.3 Irregular singular point

If $P(a) = 0$ and in addition either $(x-a)Q/P$ or $(x-a)^2R/P$ is not analytic at $x = a$, this point is an *irregular* singular point. In this case a series solution cannot be guaranteed.

4.1.3 Higher order equations

Similar techniques can sometimes be used for equations of higher order.

Example 4.5

Solve

$$y''' - xy = 0, \tag{4.100}$$

around $x = 0$.

Let

$$y = \sum_{n=0}^{\infty} a_n x^n, \tag{4.101}$$

from which

$$xy = \sum_{n=1}^{\infty} a_{n-1} x^n, \tag{4.102}$$

$$y''' = 6a_3 + \sum_{n=1}^{\infty} (n+1)(n+2)(n+3) a_{n+3} x^n. \tag{4.103}$$

Substituting into Eq. (4.100), we find that

$$a_3 = 0, \tag{4.104}$$

$$a_{n+3} = \frac{1}{(n+1)(n+2)(n+3)} a_{n-1}, \tag{4.105}$$

which gives the general solution

$$\begin{aligned} y(x) = & \; a_0 \left(1 + \frac{1}{24}x^4 + \frac{1}{8064}x^8 + \ldots\right) \\ & + a_1 x \left(1 + \frac{1}{60}x^4 + \frac{1}{30240}x^8 + \ldots\right) \\ & + a_2 x^2 \left(1 + \frac{1}{120}x^4 + \frac{1}{86400}x^8 + \ldots\right). \end{aligned} \tag{4.106}$$

For $y(0) = 1, y'(0) = 0, y''(0) = 0$, we get $a_0 = 1$, $a_1 = 0$, and $a_2 = 0$. The exact solution and the linear approximation to the exact solution, $y \sim 1 + x^4/24$, are shown in Fig. 4.4. The exact solution is expressed in terms of one of the hypergeometric functions, see Sec. 10.7.8 of the Appendix, and is

$$y = {}_0F_2\left(\{\}\,; \left\{\frac{1}{2}, \frac{3}{4}\right\}; \frac{x^4}{64}\right). \tag{4.107}$$

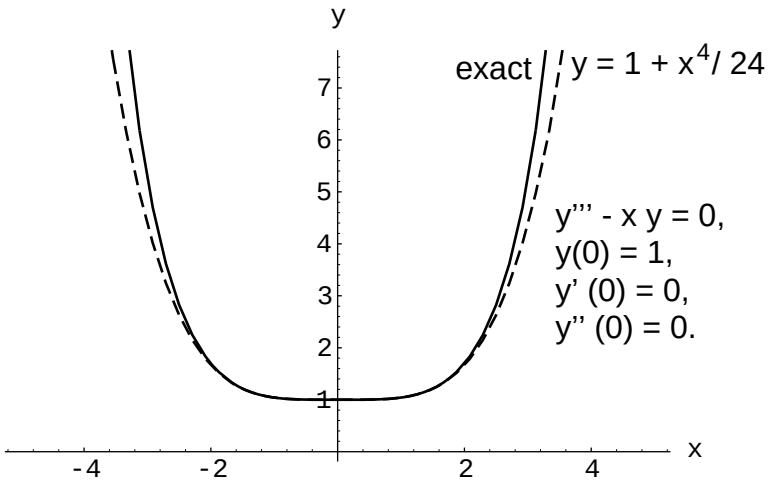

Figure 4.4: Comparison of truncated series and exact solutions.

4.2 Perturbation methods

Perturbation methods, also known as linearization or asymptotic techniques, are not as rigorous as infinite series methods in that usually it is impossible to make a statement regarding convergence. Nevertheless, the methods have proven to be powerful in many regimes of applied mathematics, science, and engineering.

The method hinges on the identification of a small parameter ϵ, $0 < \epsilon \ll 1$. Typically there is an easily obtained solution when $\epsilon = 0$. One then uses this solution as a seed to construct a linear theory about it. The resulting set of linear equations are then solved giving a solution which is valid in a regime near $\epsilon = 0$.

4.2.1 Algebraic and transcendental equations

To illustrate the method of solution, we begin with quadratic algebraic equations for which exact solutions are available. We can then easily see the advantages and limitations of the method.

Example 4.6

For $0 < \epsilon \ll 1$ solve

$$x^2 + \epsilon x - 1 = 0. \tag{4.108}$$

Let

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots. \tag{4.109}$$

Substituting into Eq. (4.108),

$$\underbrace{\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots\right)^2}_{=x^2} + \epsilon \underbrace{\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots\right)}_{=x} - 1 = 0, \tag{4.110}$$

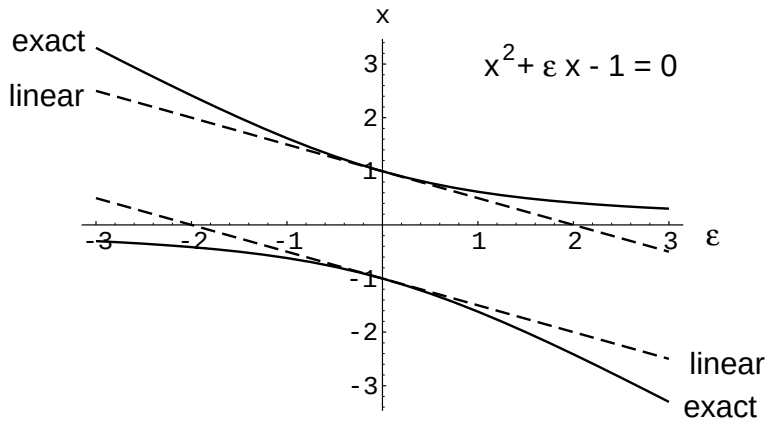

Figure 4.5: Comparison of asymptotic and exact solutions.

expanding the square by polynomial multiplication,

$$\left(x_0^2 + 2x_1x_0\epsilon + \left(x_1^2 + 2x_2x_0\right)\epsilon^2 + \ldots\right) + \left(x_0\epsilon + x_1\epsilon^2 + \ldots\right) - 1 = 0. \tag{4.111}$$

Regrouping, we get

$$\underbrace{\left(x_0^2 - 1\right)}_{=0}\epsilon^0 + \underbrace{\left(2x_1x_0 + x_0\right)}_{=0}\epsilon^1 + \underbrace{\left(x_1^2 + 2x_0x_2 + x_1\right)}_{=0}\epsilon^2 + \ldots = 0. \tag{4.112}$$

Because $\epsilon^0, \epsilon^1, \epsilon^2, \ldots$, are linearly independent, the coefficients in Eq. (4.112) must each equal zero. Thus, we get

$$\begin{array}{lrcll} O(\epsilon^0): & x_0^2 - 1 & = & 0 \Rightarrow \quad x_0 = 1, & -1, \\ O(\epsilon^1): & 2x_0x_1 + x_0 & = & 0 \Rightarrow \quad x_1 = -\frac{1}{2}, & -\frac{1}{2}, \\ O(\epsilon^2): & x_1^2 + 2x_0x_2 + x_1 & = & 0 \Rightarrow \quad x_2 = \frac{1}{8}, & -\frac{1}{8}, \\ \vdots & & & & \end{array} \tag{4.113}$$

The solutions are

$$x = 1 - \frac{\epsilon}{2} + \frac{\epsilon^2}{8} + \cdots, \tag{4.114}$$

and

$$x = -1 - \frac{\epsilon}{2} - \frac{\epsilon^2}{8} + \cdots. \tag{4.115}$$

The exact solutions can also be expanded

$$x = \frac{1}{2}\left(-\epsilon \pm \sqrt{\epsilon^2 + 4}\,\right), \tag{4.116}$$

$$= \pm 1 - \frac{\epsilon}{2} \pm \frac{\epsilon^2}{8} + \ldots, \tag{4.117}$$

to give the same results. The exact solution and the linear approximation are shown in Fig. 4.5.

Example 4.7

For $0 < \epsilon \ll 1$ solve

$$\epsilon x^2 + x - 1 = 0. \tag{4.118}$$

Note as $\epsilon \to 0$, the equation becomes singular. Let

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots. \tag{4.119}$$

Substituting into Eq. (4.118), we get

$$\epsilon \underbrace{\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots\right)^2}_{x^2} + \underbrace{\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots\right)}_{x} = 0. \tag{4.120}$$

Expanding the quadratic term gives

$$\epsilon \left(x_0^2 + 2\epsilon x_0 x_1 + \cdots\right) + \left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots\right) - 1 = 0, \tag{4.121}$$

$$\underbrace{(x_0 - 1)}_{=0}\epsilon^0 + \underbrace{(x_0^2 + x_1)}_{=0}\epsilon^1 + \underbrace{(2x_0 x_1 + x_2)}_{=0}\epsilon^2 + \cdots = 0. \tag{4.122}$$

Because of linear independence of $\epsilon^0, \epsilon^1, \epsilon^2, \ldots$, their coefficients must be zero. Thus, collecting different powers of ϵ, we get

$$\begin{array}{lrcll} O(\epsilon^0): & x_0 - 1 & = & 0 \Rightarrow & x_0 = 1, \\ O(\epsilon^1): & x_0^2 + x_1 & = & 0 \Rightarrow & x_1 = -1, \\ O(\epsilon^2): & 2x_0 x_1 + x_2 & = & 0 \Rightarrow & x_2 = 2, \\ \vdots & & & & \end{array} \tag{4.123}$$

This gives one solution

$$x = 1 - \epsilon + 2\epsilon^2 + \cdots. \tag{4.124}$$

To get the other solution, let

$$X = \frac{x}{\epsilon^\alpha}. \tag{4.125}$$

Equation (4.118) becomes

$$\epsilon^{2\alpha+1} X^2 + \epsilon^\alpha X - 1 = 0. \tag{4.126}$$

The first two terms are of the same order if $2\alpha + 1 = \alpha$. This demands $\alpha = -1$. With this,

$$X = x\epsilon, \qquad \epsilon^{-1} X^2 + \epsilon^{-1} X - 1 = 0. \tag{4.127}$$

This gives

$$X^2 + X - \epsilon = 0. \tag{4.128}$$

We expand

$$X = X_0 + \epsilon X_1 + \epsilon^2 X_2 + \cdots, \tag{4.129}$$

so

$$\underbrace{\left(X_0 + \epsilon X_1 + \epsilon^2 X_2 + \cdots\right)^2}_{X^2} + \underbrace{\left(X_0 + \epsilon X_1 + \epsilon^2 X_2 + \cdots\right)}_{X} - \epsilon = 0, \tag{4.130}$$

$$\left(X_0^2 + 2\epsilon X_0 X_1 + \epsilon^2 (X_1^2 + 2X_0 X_2) + \cdots\right) + \left(X_0 + \epsilon X_1 + \epsilon^2 X_2 + \cdots\right) - \epsilon = 0. \tag{4.131}$$

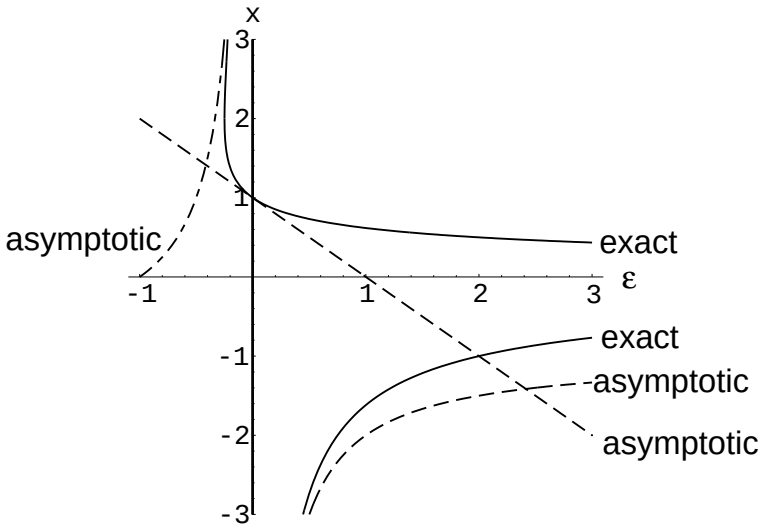

Figure 4.6: Comparison of asymptotic and exact solutions.

Collecting terms of the same order

$$\begin{array}{lrcll} O(\epsilon^0): & X_0^2 + X_0 & = & 0 \Rightarrow & X_0 = -1, \quad 0, \\ O(\epsilon^1): & 2X_0X_1 + X_1 & = & 1 \Rightarrow & X_1 = -1, \quad 1, \\ O(\epsilon^2): & X_1^2 + 2X_0X_2 + X_2 & = & 0 \Rightarrow & X_2 = 1, \quad -1, \\ \vdots & & & & \end{array} \tag{4.132}$$

gives the two solutions

$$X = -1 - \epsilon + \epsilon^2 + \cdots, \tag{4.133}$$

$$X = \epsilon - \epsilon^2 + \cdots, \tag{4.134}$$

or, with $X = x\epsilon$,

$$x = \frac{1}{\epsilon}\left(-1 - \epsilon + \epsilon^2 + \cdots\right), \tag{4.135}$$

$$x = 1 - \epsilon + \cdots. \tag{4.136}$$

Expansion of the exact solutions

$$x = \frac{1}{2\epsilon}\left(-1 \pm \sqrt{1 + 4\epsilon}\right), \tag{4.137}$$

$$= \frac{1}{2\epsilon}\left(-1 \pm \left(1 + 2\epsilon - 2\epsilon^2 + 4\epsilon^4 + \cdots\right)\right), \tag{4.138}$$

gives the same results. The exact solution and the linear approximation are shown in Fig. 4.6.

Example 4.8

Solve

$$\cos x = \epsilon \sin(x + \epsilon), \tag{4.139}$$

for x near $\pi/2$.

Fig. 4.7 shows a plot of $\cos x$ and $\epsilon \sin(x+\epsilon)$ for $\epsilon = 0.1$. It is seen that there are multiple intersections near $x = \left(n + \frac{1}{2}\pi\right)$, where $n = 0, \pm 1, \pm 2, \ldots$. We seek only one of these. When we

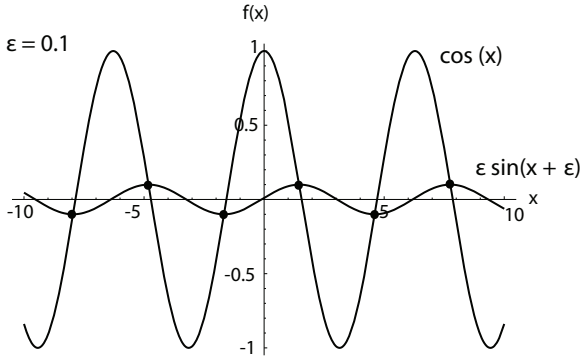

Figure 4.7: Location of roots.

substitute

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots, \tag{4.140}$$

into Eq. (4.139), we find

$$\cos(\underbrace{x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots}_{x}) = \epsilon \sin(\underbrace{x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots}_{x} + \epsilon). \tag{4.141}$$

Now we expand both the left and right hand sides in a Taylor series in ϵ about $\epsilon = 0$. We note that a general function $f(\epsilon)$ has such a Taylor series of $f(\epsilon) \sim f(0) + \epsilon f'(0) + (\epsilon^2/2) f''(0) + \ldots$ Expanding the left hand side, we get

$$\begin{aligned} \underbrace{\cos(x_0 + \epsilon x_1 + \ldots)}_{=\cos x} &= \underbrace{\cos(x_0 + \epsilon x_1 + \ldots)|_{\epsilon=0}}_{=\cos x|_{\epsilon=0}} \\ &\quad + \epsilon \overbrace{\underbrace{(-\sin(x_0 + \epsilon x_1 + \ldots))}_{=d/dx(\cos x)|_{\epsilon=0}} \underbrace{(x_1 + 2\epsilon x_2 + \ldots)}_{=dx/d\epsilon|_{\epsilon=0}}\Big|_{\epsilon=0}}^{=d/d\epsilon(\cos x)|_{\epsilon=0}} + \ldots, \end{aligned} \tag{4.142}$$

$$\cos(x_0 + \epsilon x_1 + \ldots) = \cos x_0 - \epsilon x_1 \sin x_0 + \ldots. \tag{4.143}$$

The right hand side is similar. We then arrive at Eq. (4.139) being expressed as

$$\cos x_0 - \epsilon x_1 \sin x_0 + \ldots = \epsilon(\sin x_0 + \ldots). \tag{4.144}$$

Collecting terms

$$\begin{aligned} &O(\epsilon^0): && \cos x_0 &&= 0 \Rightarrow && x_0 = \tfrac{\pi}{2}, \\ &O(\epsilon^1): && -x_1 \sin x_0 - \sin x_0 &&= 0 \Rightarrow && x_1 = -1, \\ &\vdots \end{aligned} \tag{4.145}$$

The solution is

$$x = \frac{\pi}{2} - \epsilon + \cdots. \tag{4.146}$$

4.2.2 Regular perturbations

Differential equations can also be solved using perturbation techniques.

Example 4.9

For $0 < \epsilon \ll 1$ solve

$$y'' + \epsilon y^2 = 0, \tag{4.147}$$

$$y(0) = 1, \qquad y'(0) = 0. \tag{4.148}$$

Let

$$\begin{aligned} y(x) &= y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2(x) + \cdots, & (4.149)\\ y'(x) &= y_0'(x) + \epsilon y_1'(x) + \epsilon^2 y_2'(x) + \cdots, & (4.150)\\ y''(x) &= y_0''(x) + \epsilon y_1''(x) + \epsilon^2 y_2''(x) + \cdots. & (4.151) \end{aligned}$$

Substituting into Eq. (4.147),

$$\underbrace{\left(y_0''(x) + \epsilon y_1''(x) + \epsilon^2 y_2''(x) + \cdots\right)}_{y''} + \epsilon \underbrace{\left(y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2(x) + \cdots\right)^2}_{y^2} = 0, \tag{4.152}$$

$$\left(y_0''(x) + \epsilon y_1''(x) + \epsilon^2 y_2''(x) + \cdots\right) + \epsilon\left(y_0^2(x) + 2\epsilon y_1(x) y_o(x) + \cdots\right) = 0. \tag{4.153}$$

Substituting into the boundary conditions, Eq. (4.148):

$$y_0(0) + \epsilon y_1(0) + \epsilon^2 y_2(0) + \cdots = 1, \tag{4.154}$$

$$y_0'(0) + \epsilon y_1'(0) + \epsilon^2 y_2'(0) + \cdots = 0. \tag{4.155}$$

Collecting terms

$$\begin{array}{lllll} O(\epsilon^0): & y_0'' = 0, & y_0(0) = 1, & y_0'(0) = 0 \Rightarrow & y_0 = 1, \\ O(\epsilon^1): & y_1'' = -y_0^2, & y_1(0) = 0, & y_1'(0) = 0 \Rightarrow & y_1 = -\frac{x^2}{2}, \\ O(\epsilon^2): & y_2'' = -2y_0 y_1, & y_2(0) = 0, & y_2'(0) = 0 \Rightarrow & y_2 = \frac{x^4}{12}, \\ \vdots & & & & \end{array} \tag{4.156}$$

The solution is

$$y = 1 - \epsilon\frac{x^2}{2} + \epsilon^2\frac{x^4}{12} + \cdots. \tag{4.157}$$

For validity of the asymptotic solution, we must have

$$1 \gg \epsilon\frac{x^2}{2}. \tag{4.158}$$

This solution becomes invalid when the first term is as large or larger than the second:

$$1 \leq \epsilon\frac{x^2}{2}, \tag{4.159}$$

$$|x| \geq \sqrt{\frac{2}{\epsilon}}. \tag{4.160}$$

Using the techniques of the previous chapter it is seen that Eqs. (4.147, 4.148) possess an exact solution. With

$$u = \frac{dy}{dx}, \qquad \frac{d^2y}{dx^2} = \frac{dy'}{dy}\frac{dy}{dx} = \frac{du}{dy}u, \tag{4.161}$$

Eq. (4.147) becomes

$$\begin{aligned} u\frac{du}{dy} + \epsilon y^2 &= 0, && (4.162)\\ udu &= -\epsilon y^2 dy, && (4.163)\\ \frac{u^2}{2} &= -\frac{\epsilon}{3}y^3 + C_1, && (4.164)\\ u = 0 \qquad \text{when} \qquad y &= 1 \qquad \text{so} \qquad C = \frac{\epsilon}{3}, && (4.165)\\ u &= \pm\sqrt{\frac{2\epsilon}{3}(1-y^3)}, && (4.166)\\ \frac{dy}{dx} &= \pm\sqrt{\frac{2\epsilon}{3}(1-y^3)}, && (4.167)\\ dx &= \pm\frac{dy}{\sqrt{\frac{2\epsilon}{3}(1-y^3)}}, && (4.168)\\ x &= \pm\int_1^y \frac{ds}{\sqrt{\frac{2\epsilon}{3}(1-s^3)}}. && (4.169) \end{aligned}$$

It can be shown that this integral can be represented in terms of a) the Gamma function, Γ, (see Sec. 10.7.1 of the Appendix), and b) Gauss's[3] hypergeometric function, ${}_2F_1(a,b,c,z)$, (see Sec. 10.7.8 of the Appendix), as follows:

$$x = \mp\sqrt{\frac{\pi}{6\epsilon}}\frac{\Gamma\left(\frac{1}{3}\right)}{\Gamma\left(\frac{5}{6}\right)} \pm \sqrt{\frac{3}{2\epsilon}}\, y\, \left({}_2F_1\left(\frac{1}{3},\frac{1}{2},\frac{4}{3},y^3\right)\right). \tag{4.170}$$

It is likely difficult to invert either Eq. (4.169) or (4.170) to get $y(x)$ explicitly. For small ϵ, the essence of the solution is better conveyed by the asymptotic solution. A portion of the asymptotic and exact solutions for $\epsilon = 0.1$ are shown in Fig. 4.8. For this value, the asymptotic solution is expected to be invalid for $|x| \geq \sqrt{2/\epsilon} = 4.47$.

Example 4.10

Solve

$$y'' + \epsilon y^2 = 0, \qquad y(0) = 1, \qquad y'(0) = \epsilon. \tag{4.171}$$

Let

$$y(x) = y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2(x) + \cdots. \tag{4.172}$$

[3]Johann Carl Friedrich Gauss, 1777-1855, Brunswick-born German mathematician of tremendous influence.

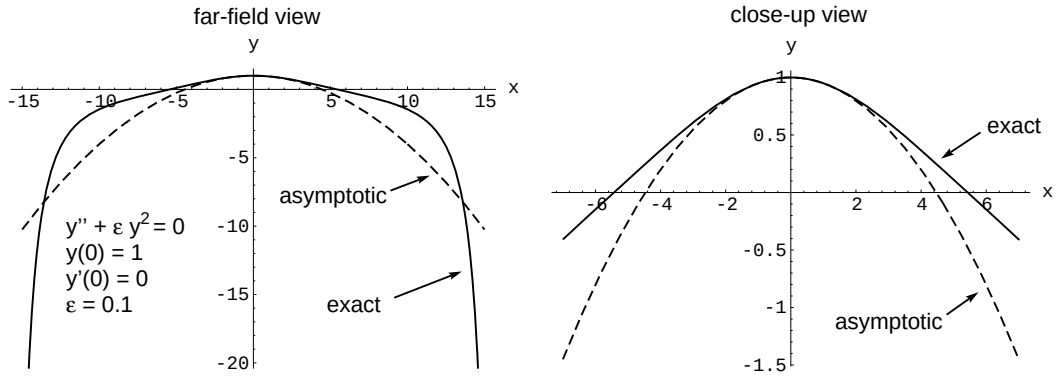

Figure 4.8: Comparison of asymptotic and exact solutions.

Substituting into Eq. (4.171) and collecting terms

$$\begin{array}{lrcll} O(\epsilon^0): & y_0'' & = & 0, & y_0(0)=1, \quad y_0'(0)=0 \Rightarrow \quad y_0 = 1, \\ O(\epsilon^1): & y_1'' & = & -y_0^2, & y_1(0)=0, \quad y_1'(0)=1 \Rightarrow \quad y_1 = -\frac{x^2}{2} + x, \\ O(\epsilon^2): & y_2'' & = & -2y_0y_1, & y_2(0)=0, \quad y_2'(0)=0 \Rightarrow \quad y_2 = \frac{x^4}{12} - \frac{x^3}{3}, \\ \vdots & & & & & \end{array} \tag{4.173}$$

The solution is

$$y = 1 - \epsilon\left(\frac{x^2}{2} - x\right) + \epsilon^2\left(\frac{x^4}{12} - \frac{x^3}{3}\right) + \cdots. \tag{4.174}$$

A portion of the asymptotic and exact solutions for $\epsilon = 0.1$ are shown in Fig. 4.9. Compared to the

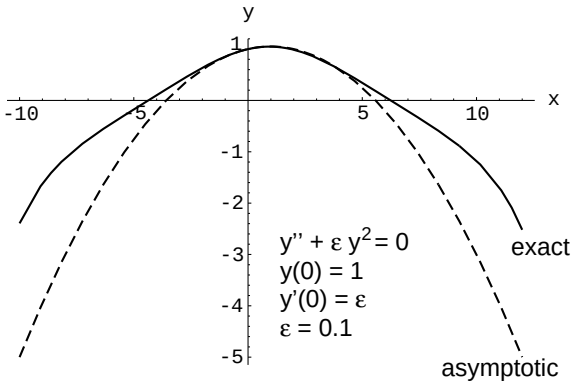

Figure 4.9: Comparison of asymptotic and exact solutions.

previous example, there is a slight offset from the y axis.

4.2.3 Strained coordinates

The regular perturbation expansion may not be valid over the complete domain of interest. The method of *strained coordinates*, also known as the Poincaré[4]-Lindstedt[5] method, is designed to address this. In a slightly different context this method is known as Lighthill's[6] method.

Example 4.11

Find an approximate solution of the Duffing equation:

$$\ddot{x} + x + \epsilon x^3 = 0, \qquad x(0) = 1, \qquad \dot{x}(0) = 0. \tag{4.175}$$

First let's give some physical motivation, as also outlined in Section 10.2 of Kaplan. One problem in which Duffing's equation arises is the undamped motion of a mass subject to a non-linear spring force. Consider a body of mass m moving in the horizontal x plane. Initially the body is given a small positive displacement $x(0) = x_o$. The body has zero initial velocity $dx/dt(0) = 0$. The body is subjected to a non-linear spring force F_s oriented such that it will pull the body towards $x = 0$:

$$F_s = (k_0 + k_1 x^2)x. \tag{4.176}$$

Here k_0 and k_1 are dimensional constants with SI units N/m and N/m^3 respectively. Newton's second law gives us

$$m\frac{d^2x}{dt^2} = -(k_0 + k_1 x^2)x, \tag{4.177}$$

$$m\frac{d^2x}{dt^2} + (k_0 + k_1 x^2)x = 0, \qquad x(0) = x_o, \frac{dx}{dt}(0) = 0. \tag{4.178}$$

Choose an as yet arbitrary length scale L and an as yet arbitrary time scale T with which to scale the problem and take:

$$\tilde{x} = \frac{x}{L}, \qquad \tilde{t} = \frac{t}{T}. \tag{4.179}$$

Substitute

$$\frac{mL}{T^2}\frac{d^2\tilde{x}}{d\tilde{t}^2} + k_0 L\tilde{x} + k_1 L^3\tilde{x}^3 = 0, \qquad L\tilde{x}(0) = x_o, \qquad \frac{L}{T}\frac{d\tilde{x}}{d\tilde{t}}(0) = 0. \tag{4.180}$$

Rearrange to make all terms dimensionless:

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \frac{k_0 T^2}{m}\tilde{x} + \frac{k_1 L^2 T^2}{m}\tilde{x}^3 = 0, \qquad \tilde{x}(0) = \frac{x_o}{L}, \qquad \frac{d\tilde{x}}{d\tilde{t}}(0) = 0. \tag{4.181}$$

Now we want to examine the effect of small non-linearities. Choose the length and time scales such that the leading order motion has an amplitude which is $O(1)$ and a frequency which is $O(1)$. So take

$$T \equiv \sqrt{\frac{m}{k_0}}, \qquad L \equiv x_o. \tag{4.182}$$

So

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \tilde{x} + \frac{k_1 x_o^2 \frac{m}{k_0}}{m}\tilde{x}^3 = 0, \qquad \tilde{x}(0) = 1, \qquad \frac{d\tilde{x}}{d\tilde{t}}(0) = 0. \tag{4.183}$$

[4]Henri Poincaré, 1854-1912, French polymath.

[5]Anders Lindstedt, 1854-1939, Swedish mathematician, astronomer, and actuarial scientist.

[6]Sir Michael James Lighthill, 1924-1998, British applied mathematician and noted open-water swimmer.

Choosing

$$\epsilon \equiv \frac{k_1 x_o^2}{k_0}, \tag{4.184}$$

we get

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \tilde{x} + \epsilon\tilde{x}^3 = 0, \qquad \tilde{x}(0) = 1, \qquad \frac{d\tilde{x}}{d\tilde{t}}(0) = 0. \tag{4.185}$$

So our asymptotic theory will be valid for

$$\epsilon \ll 1, \qquad k_1 x_o^2 \ll k_0. \tag{4.186}$$

Now, let's drop the superscripts and focus on the mathematics. An accurate numerical approximation to the exact solution $x(t)$ for $\epsilon = 0.2$ and the so-called phase plane for this solution, giving dx/dt versus x are shown in Fig. 4.10.

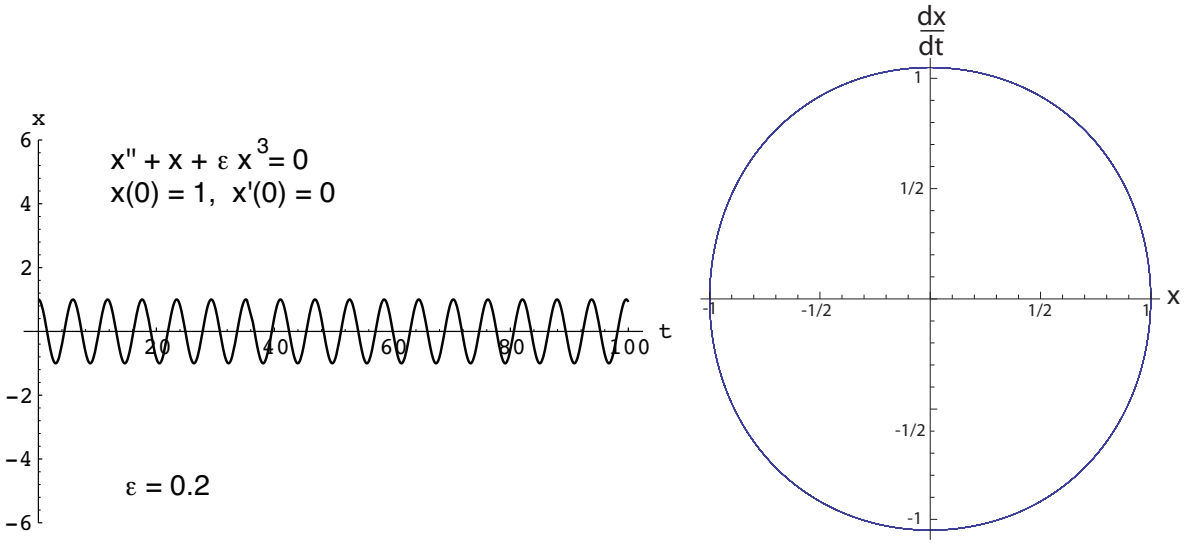

Figure 4.10: Numerical solution $x(t)$ and phase plane trajectory, dx/dt versus x for Duffing's equation, $\epsilon = 0.2$.

Note if $\epsilon = 0$, the solution is $x(t) = \cos t$, and thus $dx/dt = -\sin t$. Thus, for $\epsilon = 0$, $x^2 + (dx/dt)^2 = \cos^2 t + \sin^2 t = 1$. Thus, the $\epsilon = 0$ phase plane solution is a unit circle. The phase plane portrait of Fig. 4.10 displays a small deviation from a circle. This deviation would be more pronounced for larger ϵ.

Let's use an asymptotic method to try to capture this solution. Using the expansion

$$x(t) = x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + \cdots, \tag{4.187}$$

and collecting terms, we find

$$\begin{array}{lrcll} O(\epsilon^0): & \ddot{x}_0 + x_0 & = & 0, \quad x_0(0) = 1, \quad \dot{x}_0(0) = 0 \Rightarrow & x_0 = \cos t, \\ O(\epsilon^1): & \ddot{x}_1 + x_1 & = & -x_0^3, \quad x_1(0) = 0, \quad \dot{x}_1(0) = 0 \Rightarrow & x_1 = \frac{1}{32}(-\cos t + \cos 3t - 12t\sin t), \\ \vdots & & & & \end{array} \tag{4.188}$$

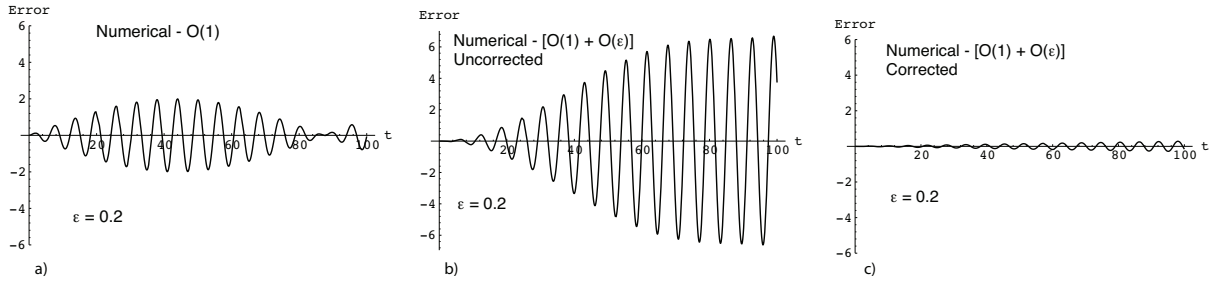

Figure 4.11: Error plots for various approximations from the method of strained coordinates to Duffing's equation with $\epsilon = 0.2$. Difference between high accuracy numerical solution and: a) leading order asymptotic solution, b) uncorrected $O(\epsilon)$ asymptotic solution, c) corrected $O(\epsilon)$ asymptotic solution.

The difference between the exact solution and the leading order solution, $x_{exact}(t) - x_0(t)$ is plotted in Fig. 4.11a. The error is the same order of magnitude as the solution itself for moderate values of t. This is undesirable.

To $O(\epsilon)$ the solution is

$$x = \cos t + \frac{\epsilon}{32}\left(-\cos t + \cos 3t - \underbrace{12t\sin t}_{\text{secular term}}\right) + \cdots. \tag{4.189}$$

This series has a so-called "secular term," $-\epsilon\frac{3}{8}t\sin t$, that grows without bound. Thus, our solution is only valid for $t \ll \epsilon^{-1}$.

Now nature may or may not admit unbounded growth depending on the problem. Let us return to the original Eq. (4.175) to consider whether or not unbounded growth is admissible. Eq. (4.175) can be integrated once via the following steps

$$\begin{aligned}
\dot{x}\left(\ddot{x} + x + \epsilon x^3\right) &= 0, && (4.190)\\
\dot{x}\ddot{x} + \dot{x}x + \epsilon\dot{x}x^3 &= 0, && (4.191)\\
\frac{d}{dt}\left(\frac{1}{2}\dot{x}^2 + \frac{1}{2}x^2 + \frac{\epsilon}{4}x^4\right) &= 0, && (4.192)\\
\frac{1}{2}\dot{x}^2 + \frac{1}{2}x^2 + \frac{\epsilon}{4}x^4 &= \left(\frac{1}{2}\dot{x}^2 + \frac{1}{2}x^2 + \frac{\epsilon}{4}x^4\right)\Big|_{t=0}, && (4.193)\\
\frac{1}{2}\dot{x}^2 + \frac{1}{2}x^2 + \frac{\epsilon}{4}x^4 &= \frac{1}{4}(2+\epsilon), && (4.194)
\end{aligned}$$

indicating that the solution is bounded. The difference between the exact solution and the leading order solution, $x_{exact}(t) - (x_0(t) + \epsilon x_1(t))$ is plotted in Fig. 4.11b. There is some improvement for early time, but the solution is actually worse for later time. This is because of the secularity.

To have a solution valid for all time, we strain the time coordinate

$$t = (1 + c_1\epsilon + c_2\epsilon^2 + \cdots)\tau, \tag{4.195}$$

where τ is the new time variable. The c_i's should be chosen to avoid secular terms.

Differentiating

$$\dot{x} = \frac{dx}{d\tau}\frac{d\tau}{dt} = \frac{dx}{d\tau}\left(\frac{dt}{d\tau}\right)^{-1}, \tag{4.196}$$

$$= \frac{dx}{d\tau}(1 + c_1\epsilon + c_2\epsilon^2 + \cdots)^{-1}, \tag{4.197}$$

$$\ddot{x} = \frac{d^2x}{d\tau^2}(1 + c_1\epsilon + c_2\epsilon^2 + \cdots)^{-2}, \tag{4.198}$$

$$= \frac{d^2x}{d\tau^2}\left(1 - c_1\epsilon + (c_1^2 - c_2)\epsilon^2 + \cdots\right)^2, \tag{4.199}$$

$$= \frac{d^2x}{d\tau^2}(1 - 2c_1\epsilon + (3c_1^2 - 2c_2)\epsilon^2 + \cdots). \tag{4.200}$$

Furthermore, we write

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \ldots \tag{4.201}$$

Substituting into Eq. (4.175), we get

$$\underbrace{\left(\frac{d^2x_0}{d\tau^2} + \epsilon\frac{d^2x_1}{d\tau^2} + \epsilon^2\frac{d^2x_2}{d\tau^2} + \cdots\right)(1 - 2c_1\epsilon + (3c_1^2 - 2c_2)\epsilon^2 + \cdots)}_{\ddot{x}}$$
$$+ \underbrace{(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots)}_{x} + \epsilon\underbrace{(x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots)^3}_{x^3} = 0. \tag{4.202}$$

Collecting terms, we get

$$\begin{array}{lrcll}
O(\epsilon^0): & \frac{d^2x_0}{d\tau^2} + x_0 &=& 0, & x_0(0) = 1, \quad \frac{dx_0}{d\tau}(0) = 0, \\
& x_0(\tau) &=& \cos\tau, & \\
O(\epsilon^1): & \frac{d^2x_1}{d\tau^2} + x_1 &=& 2c_1\frac{d^2x_0}{d\tau^2} - x_0^3, & x_1(0) = 0, \quad \frac{dx_1}{d\tau}(0) = 0, \\
& &=& -2c_1\cos\tau - \cos^3\tau, & \\
& &=& -(2c_1 + \frac{3}{4})\cos\tau - \frac{1}{4}\cos 3\tau, & \\
& x_1(\tau) &=& \frac{1}{32}(-\cos\tau + \cos 3\tau), \qquad \text{if we choose } c_1 = -\frac{3}{8}. &
\end{array} \tag{4.203}$$

Thus,

$$x(\tau) = \cos\tau + \epsilon\frac{1}{32}(-\cos\tau + \cos 3\tau) + \cdots. \tag{4.204}$$

Since

$$t = \left(1 - \epsilon\frac{3}{8} + \cdots\right)\tau, \tag{4.205}$$

$$\tau = \left(1 + \epsilon\frac{3}{8} + \cdots\right)t, \tag{4.206}$$

we get the corrected solution approximation to be

$$x(t) = \cos\left(\underbrace{\left(1 + \epsilon\frac{3}{8} + \cdots\right)}_{\text{Frequency Modulation (FM)}} t\right)$$
$$+\epsilon\frac{1}{32}\left(-\cos\left(\left(1 + \epsilon\frac{3}{8} + \cdots\right)t\right) + \cos\left(3\left(1 + \epsilon\frac{3}{8} + \cdots\right)t\right)\right) + \cdots. \tag{4.207}$$

The difference between the exact solution and the leading order solution, $x_{exact}(t) - (x_0(t) + \epsilon x_1(t))$ for the corrected solution to $O(\epsilon)$ is plotted in Fig. 4.11c. The error is much smaller relative to the previous cases; there does appear to be a slight growth in the amplitude of the error with time. This might not be expected, but in fact is a characteristic behavior of the *truncation error of the numerical method* used to generate the exact solution.

Example 4.12

Find the amplitude of the limit cycle oscillations of the van der Pol[7] equation

$$\ddot{x} - \epsilon(1 - x^2)\dot{x} + x = 0, \qquad x(0) = A, \qquad \dot{x}(0) = 0, \qquad 0 < \epsilon \ll 1. \tag{4.208}$$

Here A is the amplitude and is considered to be an adjustable parameter in this problem. If a limit cycle exists, it will be valid as $t \to \infty$. Note this could be thought of as a model for a mass-spring-damper system with a non-linear damping coefficient of $-\epsilon(1 - x^2)$. For small $|x|$, the damping coefficient is negative. From our intuition from linear mass-spring-damper systems, we recognize that this will lead to amplitude growth, at least for sufficiently small $|x|$. However, when the amplitude grows to $|x| > 1/\sqrt{\epsilon}$, the damping coefficient again becomes positive, thus decaying the amplitude. We might expect a *limit cycle* amplitude where there exists a balance between the tendency for amplitude to grow or decay.

Let

$$t = (1 + c_1\epsilon + c_2\epsilon^2 + \cdots)\tau, \tag{4.209}$$

so that Eq. (4.208) becomes

$$\frac{d^2x}{d\tau^2}(1 - 2c_1\epsilon + \ldots) - \epsilon(1 - x^2)\frac{dx}{d\tau}(1 - c_1\epsilon + \ldots) + x = 0. \tag{4.210}$$

We also use

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \ldots. \tag{4.211}$$

Thus, we get

$$x_0 = A\cos\tau, \tag{4.212}$$

to $O(\epsilon^0)$. To $O(\epsilon)$, the equation is

$$\frac{d^2x_1}{d\tau^2} + x_1 = -2c_1 A\cos\tau - A\left(1 - \frac{A^2}{4}\right)\sin\tau + \frac{A^3}{4}\sin 3\tau. \tag{4.213}$$

Choosing $c_1 = 0$ and $A = 2$ in order to suppress secular terms, we get

$$x_1 = \frac{3}{4}\sin\tau - \frac{1}{4}\sin 3\tau. \tag{4.214}$$

The amplitude, to lowest order, is

$$A = 2, \tag{4.215}$$

so to $O(\epsilon)$ the solution is

$$x(t) = 2\cos\left(t + O(\epsilon^2)\right) + \epsilon\left(\frac{3}{4}\sin\left(t + O(\epsilon^2)\right) - \frac{1}{4}\sin\left(3\left(t + O(\epsilon^2)\right)\right)\right) + O(\epsilon^2). \tag{4.216}$$

The exact solution, x_{exact}, $\dot{x}_{exact}$, calculated by high precision numerics in the x, $\dot{x}$ phase plane, $x(t)$, and the difference between the exact solution and the asymptotic leading order solution, $x_{exact}(t) - x_0(t)$, and the difference between the exact solution and the asymptotic solution corrected to $O(\epsilon)$: $x_{exact}(t) - (x_0(t) + \epsilon x_1(t))$ is plotted in Fig. 4.12. Because of the special choice of initial conditions, the solution trajectory lies for all time on the limit cycle of the phase plane. Note that the leading order solution is only marginally better than the corrected solution at this value of ϵ. For smaller values of ϵ, the relative errors between the two approximations would widen; that is, the asymptotic correction would become relatively speaking, more accurate.

[7] Balthasar van der Pol, 1889-1959, Dutch physicist.

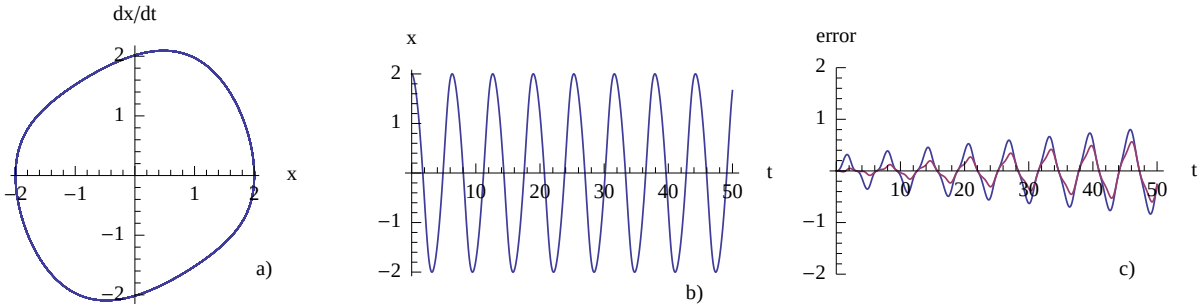

Figure 4.12: Results for van der Pol equation, $d^2x/dt^2 - \epsilon(1-x^2)dx/dt + x = 0$, $x(0) = 2, \dot{x}(0) = 0, \epsilon = 0.3$: a) high precision numerical phase plane, b) high precision numerical calculation of $x(t)$, c) difference between exact and asymptotic leading order solution (blue), and difference between exact and corrected asymptotic solution to $O(\epsilon)$ (red) from the method of strained coordinates.

4.2.4 Multiple scales

The method of multiple scales is a strategy for isolating features of a solution which may evolve on widely disparate scales.

Example 4.13

Solve

$$\frac{d^2x}{dt^2} - \epsilon(1-x^2)\frac{dx}{dt} + x = 0, \qquad x(0) = 0, \qquad \frac{dx}{dt}(0) = 1, \qquad 0 < \epsilon \ll 1. \tag{4.217}$$

Let $x = x(\tau, \tilde{\tau})$, where the fast time scale is

$$\tau = (1 + a_1\epsilon + a_2\epsilon^2 + \cdots)t, \tag{4.218}$$

and the slow time scale is

$$\tilde{\tau} = \epsilon t. \tag{4.219}$$

Since

$$x = x(\tau, \tilde{\tau}), \tag{4.220}$$

$$\frac{dx}{dt} = \frac{\partial x}{\partial \tau}\frac{d\tau}{dt} + \frac{\partial x}{\partial \tilde{\tau}}\frac{d\tilde{\tau}}{dt}. \tag{4.221}$$

The first derivative is

$$\frac{dx}{dt} = \frac{\partial x}{\partial \tau}(1 + a_1\epsilon + a_2\epsilon^2 + \cdots) + \frac{\partial x}{\partial \tilde{\tau}}\epsilon, \tag{4.222}$$

so

$$\frac{d}{dt} = (1 + a_1\epsilon + a_2\epsilon^2 + \cdots)\frac{\partial}{\partial \tau} + \epsilon\frac{\partial}{\partial \tilde{\tau}}. \tag{4.223}$$

Applying this operator to dx/dt, we get

$$\frac{d^2x}{dt^2} = (1 + a_1\epsilon + a_2\epsilon^2 + \cdots)^2\frac{\partial^2 x}{\partial\tau^2} + 2(1 + a_1\epsilon + a_2\epsilon^2 + \cdots)\epsilon\frac{\partial^2 x}{\partial\tau\partial\tilde{\tau}} + \epsilon^2\frac{\partial^2 x}{\partial\tilde{\tau}^2}. \tag{4.224}$$

Introduce

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \cdots. \tag{4.225}$$

So to $O(\epsilon)$, Eq. (4.217) becomes

$$\underbrace{(1 + 2a_1\epsilon + \cdots)\frac{\partial^2 (x_0 + \epsilon x_1 + \cdots)}{\partial\tau^2} + 2\epsilon\frac{\partial^2 (x_0 + \cdots)}{\partial\tau\partial\tilde{\tau}} + \cdots}_{\ddot{x}}$$
$$\underbrace{-\epsilon\,(1 - x_0^2 - \cdots)\frac{\partial(x_0 + \cdots)}{\partial\tau} + \cdots}_{(1-x^2)\dot{x}} + \underbrace{(x_0 + \epsilon x_1 + \cdots)}_{x} = 0. \tag{4.226}$$

Collecting terms of $O(\epsilon^0)$, we have

$$\frac{\partial^2 x_0}{\partial\tau^2} + x_0 = 0 \text{ with } x_0(0,0) = 0,\ \tfrac{\partial x_0}{\partial\tau}(0,0) = 1. \tag{4.227}$$

The solution is

$$x_0 = A(\tilde{\tau})\cos\tau + B(\tilde{\tau})\sin\tau \text{ with } A(0) = 0,\ B(0) = 1. \tag{4.228}$$

The terms of $O(\epsilon^1)$ give

$$\frac{\partial^2 x_1}{\partial\tau^2} + x_1 = -2a_1\frac{\partial^2 x_0}{\partial\tau^2} - 2\frac{\partial^2 x_0}{\partial\tau\partial\tilde{\tau}} + (1 - x_0^2)\frac{\partial x_0}{\partial\tau}, \tag{4.229}$$
$$= \left(2a_1B + 2A' - A + \frac{A}{4}(A^2 + B^2)\right)\sin\tau + \left(2a_1A - 2B' + B - \frac{B}{4}(A^2 + B^2)\right)\cos\tau + \frac{A}{4}(A^2 - 3B^2)\sin 3\tau - \frac{B}{4}(3A^2 - B^2)\cos 3\tau. \tag{4.230}$$

with

$$x_1(0,0) = 0, \tag{4.231}$$
$$\frac{\partial x_1}{\partial\tau}(0,0) = -a_1\frac{\partial x_0}{\partial\tau}(0,0) - \frac{\partial x_0}{\partial\tilde{\tau}}(0,0), \tag{4.232}$$
$$= -a_1 - \frac{\partial x_0}{\partial\tilde{\tau}}(0,0). \tag{4.233}$$

Since ϵt is already represented in $\tilde{\tau}$, choose $a_1 = 0$. Then

$$2A' - A + \frac{A}{4}(A^2 + B^2) = 0, \tag{4.234}$$
$$2B' - B + \frac{B}{4}(A^2 + B^2) = 0. \tag{4.235}$$

Since $A(0) = 0$, try $A(\tilde{\tau}) = 0$. Then

$$2B' - B + \frac{B^3}{4} = 0. \tag{4.236}$$

Multiplying by B, we get

$$2BB' - B^2 + \frac{B^4}{4} = 0, \tag{4.237}$$

$$(B^2)' - B^2 + \frac{B^4}{4} = 0. \tag{4.238}$$

Taking $F \equiv B^2$, we get

$$F' - F + \frac{F^2}{4} = 0. \tag{4.239}$$

This is a first order ODE in F, which can be easily solved. Separating variables, integrating, and transforming from F back to B, we get

$$\frac{B^2}{1 - \frac{B^2}{4}} = Ce^{\tilde{\tau}}. \tag{4.240}$$

Since $B(0) = 1$, we get $C = 4/3$. From this

$$B = \frac{2}{\sqrt{1 + 3e^{-\tilde{\tau}}}}, \tag{4.241}$$

so that

$$x(\tau, \tilde{\tau}) = \frac{2}{\sqrt{1 + 3e^{-\tilde{\tau}}}} \sin \tau + O(\epsilon), \tag{4.242}$$

$$x(t) = \underbrace{\frac{2}{\sqrt{1 + 3e^{-\epsilon t}}}}_{\text{Amplitude Modulation (AM)}} \sin\left((1 + O(\epsilon^2))t\right) + O(\epsilon). \tag{4.243}$$

The high precision numerical approximation for the solution trajectory in the $(x, \dot{x})$ phase plane, the high precision numerical solution $x_{exact}(t)$, and the difference between the exact solution and the asymptotic leading order solution, $x_{exact}(t) - x_0(t)$, and the difference between the exact solution and the asymptotic solution corrected to $O(\epsilon)$: $x_{exact}(t) - (x_0(t) + \epsilon x_1(t))$ are plotted in Fig. 4.13. Note that the amplitude, which is initially 1, grows to a value of 2, the same value which was obtained in the previous example. This is evident in the phase plane, where the initial condition does not lie on the long time limit cycle. Here, we have additionally obtained the time scale for the growth of the amplitude change. Note also that the leading order approximation is poor for $t > 1/\epsilon$, while the corrected approximation is relatively good. Also note that for $\epsilon = 0.3$, the segregation in time scales is not dramatic. The "fast" time scale is that of the oscillation and is $O(1)$. The slow time scale is $O(1/\epsilon)$, which here is around 3. For smaller ϵ, the effect would be more dramatic.

4.2.5 Boundary layers

The method of boundary layers, also known as matched asymptotic expansion, can be used in some cases. It is most appropriate for cases in which a small parameter multiplies the highest order derivative. In such cases a regular perturbation scheme will fail since we lose a boundary condition at leading order.

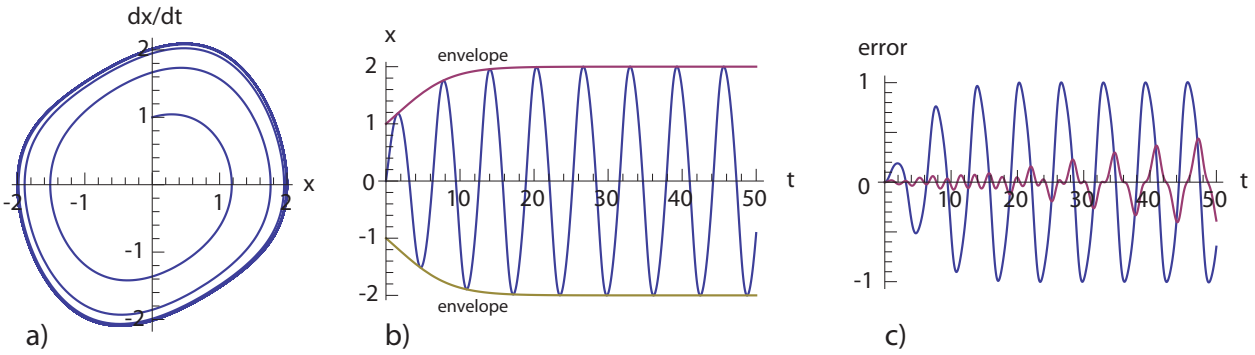

Figure 4.13: Results for van der Pol equation, $d^2x/dt^2 - \epsilon(1-x^2)dx/dt + x = 0$, $x(0) = 0, \dot{x}(0) = 1, \epsilon = 0.3$: a) high precision numerical phase plane, b) high precision numerical calculation of $x(t)$, along with the envelope $2/\sqrt{1+3e^{-\epsilon t}}$, c) difference between exact and asymptotic leading order solution (blue), and difference between exact and corrected asymptotic solution to $O(\epsilon)$ (red) from the method of multiple scales.

Example 4.14

Solve

$$\epsilon y'' + y' + y = 0, \qquad y(0) = 0, \qquad y(1) = 1. \tag{4.244}$$

An exact solution to this equation exists, namely

$$y(x) = \exp\left(\frac{1-x}{2\epsilon}\right) \frac{\sinh\left(\frac{x\sqrt{1-4\epsilon}}{2\epsilon}\right)}{\sinh\left(\frac{\sqrt{1-4\epsilon}}{2\epsilon}\right)}. \tag{4.245}$$

We could in principle simply expand this in a Taylor series in ϵ. However, for more difficult problems, exact solutions are not available. So here we will just use the exact solution to verify the validity of the method.

We begin with a regular perturbation expansion

$$y(x) = y_0 + \epsilon y_1(x) + \epsilon^2 y_2(x) + \cdots. \tag{4.246}$$

Substituting and collecting terms, we get

$$O(\epsilon^0) : y_0' + y_0 = 0, \qquad y_0(0) = 0, \qquad y_0(1) = 1, \tag{4.247}$$

the solution to which is

$$y_0 = ae^{-x}. \tag{4.248}$$

It is not possible for the solution to satisfy the two boundary conditions simultaneously since we only have one free variable, a. So, we divide the region of interest $x \in [0,1]$ into two parts, a thin *inner region* or *boundary layer* around $x = 0$, and an *outer region* elsewhere.

Equation (4.248) gives the solution in the outer region. To satisfy the boundary condition $y_0(1) = 1$, we find that $a = e$, so that

$$y = e^{1-x} + \cdots. \tag{4.249}$$

In the inner region, we choose a new independent variable X defined as $X = x/\epsilon$, so that the equation becomes

$$\frac{d^2y}{dX^2} + \frac{dy}{dX} + \epsilon y = 0. \tag{4.250}$$

Using a perturbation expansion, the lowest order equation is

$$\frac{d^2y_0}{dX^2} + \frac{dy_0}{dX} = 0, \tag{4.251}$$

with a solution

$$y_0 = A + Be^{-X}. \tag{4.252}$$

Applying the boundary condition $y_0(0) = 0$, we get

$$y_0 = A(1 - e^{-X}). \tag{4.253}$$

Matching of the inner and outer solutions is achieved by (Prandtl's[8] method)

$$y_{inner}(X \to \infty) = y_{outer}(x \to 0), \tag{4.254}$$

which gives $A = e$. The solution is

$$y(x) = e(1 - e^{-x/\epsilon}) + \cdots, \text{ in the inner region,} \tag{4.255}$$

$$\lim_{x\to\infty} y = e, \tag{4.256}$$

and

$$y(x) = e^{1-x} + \cdots, \text{ in the outer region,} \tag{4.257}$$

$$\lim_{x\to 0} y = e. \tag{4.258}$$

A composite solution can also be written by adding the two solutions. However, one must realize that this induces a double counting in the region where the inner layer solution matches onto the outer layer solution. Thus, we need to subtract one term to account for this overlap. This is known as the *common part.* Thus, the correct composite solution is the summation of the inner and outer parts, with the common part subtracted:

$$y(x) = \underbrace{\left(e(1 - e^{-x/\epsilon}) + \cdots\right)}_{\text{inner}} + \underbrace{\left(e^{1-x} + \cdots\right)}_{\text{outer}} - \underbrace{e}_{\text{common part}}, \tag{4.259}$$

$$y = e(e^{-x} - e^{-x/\epsilon}) + \cdots. \tag{4.260}$$

The exact solution, the inner layer solution, the outer layer solution, and the composite solution are plotted in Fig. 4.14.

Example 4.15

Obtain the solution of the previous problem

$$\epsilon y'' + y' + y = 0, \qquad y(0) = 0, \qquad y(1) = 1, \tag{4.261}$$

[8]Ludwig Prandtl, 1875-1953, German engineer based in Göttingen.

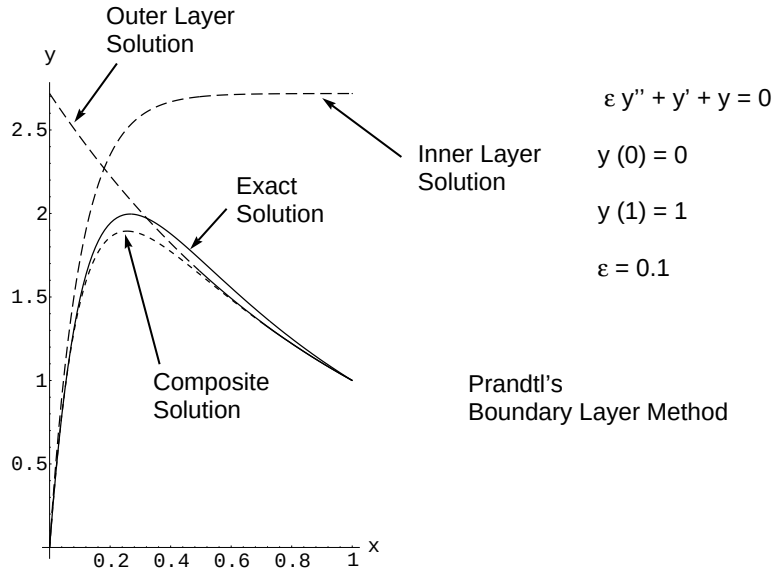

Figure 4.14: Exact, inner layer solution, outer layer solution, and composite solution for boundary layer problem.

to the next order.

Keeping terms of the next order in ϵ, we have

$$y = e^{1-x} + \epsilon((1-x)e^{1-x}) + \ldots, \tag{4.262}$$

for the outer solution, and

$$y = A(1-e^{-X}) + \epsilon\left(B - AX - (B+AX)e^{-X}\right) + \ldots, \tag{4.263}$$

for the inner solution.

Higher order matching (Van Dyke's[9] method) is obtained by expanding the outer solution in terms of the inner variable, the inner solution in terms of the outer variable, and comparing. Thus, the outer solution is, as $\epsilon \to 0$

$$\begin{aligned} y &= e^{1-\epsilon X} + \epsilon\left((1-\epsilon X)e^{1-\epsilon X}\right) + \ldots, & (4.264)\\ &= e(1-\epsilon X) + \epsilon e(1-\epsilon X)^2. & (4.265) \end{aligned}$$

Ignoring terms which are $> O(\epsilon^2)$, we get

$$\begin{aligned} y &= e(1-\epsilon X) + \epsilon e, & (4.266)\\ &= e + \epsilon e(1-X), & (4.267)\\ &= e + \epsilon e\left(1 - \frac{x}{\epsilon}\right), & (4.268)\\ &= e + \epsilon e - ex. & (4.269) \end{aligned}$$

Similarly, the inner solution as $\epsilon \to 0$ is

$$\begin{aligned} y &= A(1-e^{-x/\epsilon}) + \epsilon\left(B - A\frac{x}{\epsilon} - \left(B + A\frac{x}{\epsilon}\right)e^{-x/\epsilon}\right) + \ldots, & (4.270)\\ &= A + B\epsilon - Ax. & (4.271) \end{aligned}$$

[9]Milton Denman Van Dyke, 1922-2010, American engineer and applied mathematician.

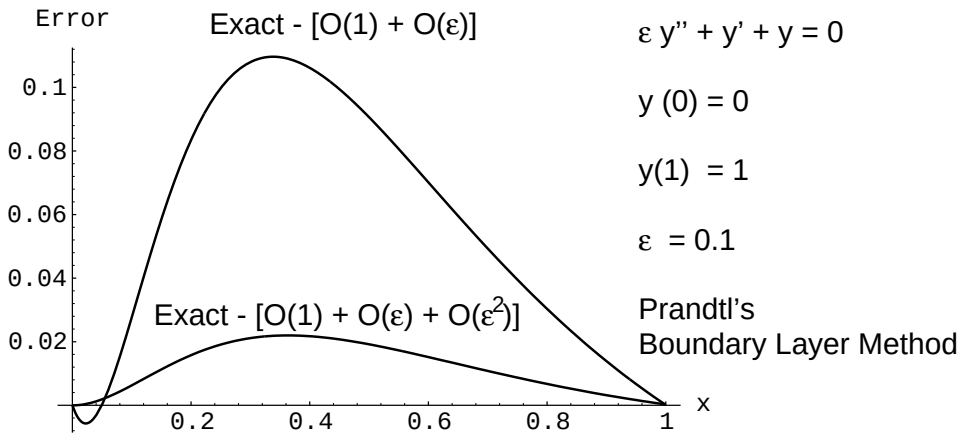

Figure 4.15: Difference between exact and asymptotic solutions for two different orders of approximation for a boundary layer problem.

Comparing, we get $A = B = e$, so that

$$y(x) = e(1 - e^{-x/\epsilon}) + e\left(\epsilon - x - (\epsilon + x)e^{-x/\epsilon}\right) + \cdots \text{ in the inner region,} \tag{4.272}$$

and

$$y(x) = e^{1-x} + \epsilon(1 - x)e^{1-x} \cdots \text{ in the outer region,} \tag{4.273}$$

The composite solution, inner plus outer minus common part, reduces to

$$y = e^{1-x} - (1 + x)e^{1-x/\epsilon} + \epsilon\left((1 - x)e^{1-x} - e^{1-x/\epsilon}\right) + \cdots. \tag{4.274}$$

The difference between the exact solution and the approximation from the previous example, and the difference between the exact solution and approximation from this example are plotted in Fig. 4.15.

Example 4.16

In the same problem, investigate the possibility of having the boundary layer at $x = 1$. The outer solution now satisfies the condition $y(0) = 0$, giving $y = 0$. Let

$$X = \frac{x - 1}{\epsilon}. \tag{4.275}$$

The lowest order inner solution satisfying $y(X = 0) = 1$ is

$$y = A + (1 - A)e^{-X}. \tag{4.276}$$

However, as $X \to -\infty$, this becomes unbounded and cannot be matched with the outer solution. Thus, a boundary layer at $x = 1$ is not possible.

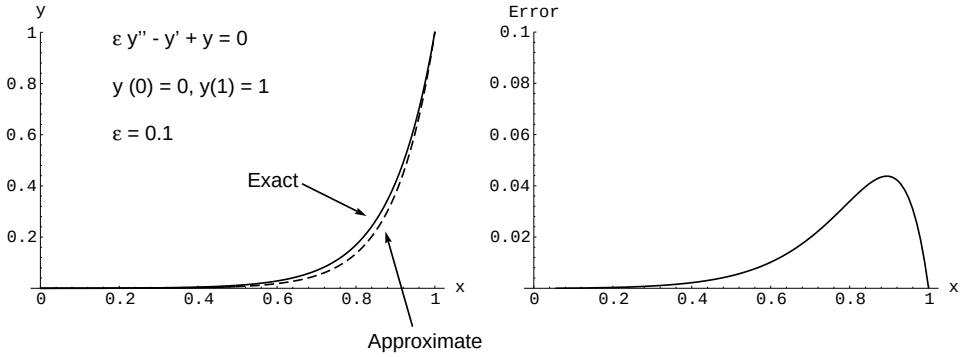

Figure 4.16: Exact, approximate, and difference in predictions for a boundary layer problem.

Example 4.17

Solve

$$\epsilon y'' - y' + y = 0, \text{ with } y(0) = 0, y(1) = 1. \tag{4.277}$$

The boundary layer is at $x = 1$. The outer solution is $y = 0$. Taking

$$X = \frac{x-1}{\epsilon} \tag{4.278}$$

the inner solution is

$$y = A + (1 - A)e^X + \ldots \tag{4.279}$$

Matching, we get

$$A = 0, \tag{4.280}$$

so that we have a composite solution

$$y(x) = e^{(x-1)/\epsilon} + \ldots. \tag{4.281}$$

The exact solution, the approximate solution to $O(\epsilon)$, and the difference between the exact solution and the approximation, are plotted in Fig. 4.16.

4.2.6 WKBJ method

Any equation of the form

$$\frac{d^2v}{dx^2} + P(x)\frac{dv}{dx} + Q(x)v = 0, \tag{4.282}$$

can be written as

$$\frac{d^2y}{dx^2} + R(x)y = 0, \tag{4.283}$$

where

$$v(x) = y(x) \exp\left(-\frac{1}{2}\int_0^x P(s)ds\right), \tag{4.284}$$

$$R(x) = Q(x) - \frac{1}{2}\frac{dP}{dx} - \frac{1}{4}\left(P(x)\right)^2. \tag{4.285}$$

So it is sufficient to study equations of the form of Eq. (4.283). The Wentzel,[10] Kramers,[11] Brillouin,[12] Jeffreys,[13] (WKBJ) method is used for equations of the kind

$$\epsilon^2 \frac{d^2y}{dx^2} = f(x)y, \tag{4.286}$$

where ϵ is a small parameter. This also includes an equation of the type

$$\epsilon^2 \frac{d^2y}{dx^2} = (\lambda^2 p(x) + q(x))y, \tag{4.287}$$

where λ is a large parameter. Alternatively, by taking $x = \epsilon t$, Eq. (4.286) becomes

$$\frac{d^2y}{dt^2} = f(\epsilon t)y. \tag{4.288}$$

We can also write Eq. (4.286) as

$$\frac{d^2y}{dx^2} = g(x)y, \tag{4.289}$$

where $g(x)$ is slowly varying in the sense that $g'/g^{3/2} \sim O(\epsilon)$.

We seek solutions to Eq. (4.286) of the form

$$y(x) = \exp\left(\frac{1}{\epsilon}\int_{x_0}^x (S_0(s) + \epsilon S_1(s) + \epsilon^2 S_2(s) + \cdots)ds\right). \tag{4.290}$$

The derivatives are

$$\frac{dy}{dx} = \frac{1}{\epsilon}\left(S_0(x) + \epsilon S_1(x) + \epsilon^2 S_2(x) + \cdots\right)\ y(x), \tag{4.291}$$

$$\frac{d^2y}{dx^2} = \frac{1}{\epsilon^2}\left(S_0(x) + \epsilon S_1(x) + \epsilon^2 S_2(x) + \cdots\right)^2\ y(x), + \frac{1}{\epsilon}\left(\frac{dS_0}{dx} + \epsilon\frac{dS_1}{dx} + \epsilon^2\frac{dS_2}{dx} + \cdots\right)\ y(x). \tag{4.292}$$

[10] Gregor Wentzel, 1898-1978, German physicist.
[11] Hendrik Anthony Kramers, 1894-1952, Dutch physicist.
[12] Léon Brillouin, 1889-1969, French physicist.
[13] Harold Jeffreys, 1891-1989, English mathematician.

Substituting into Eq. (4.286), we get

$$\underbrace{\left((S_0(x))^2 + 2\epsilon S_0(x)S_1(x) + \cdots\right)\ y(x) + \epsilon\left(\frac{dS_0}{dx} + \cdots\right)\ y(x)}_{=\epsilon^2 d^2y/dx^2} = f(x)y(x). \tag{4.293}$$

Collecting terms, at $O(\epsilon^0)$ we have

$$S_0^2(x) = f(x), \tag{4.294}$$

from which

$$S_0(x) = \pm\sqrt{f(x)}. \tag{4.295}$$

To $O(\epsilon^1)$ we have

$$2S_0(x)S_1(x) + \frac{dS_0}{dx} = 0, \tag{4.296}$$

from which

$$\begin{aligned} S_1(x) &= -\frac{\frac{dS_0}{dx}}{2S_0(x)}, && (4.297)\\ &= -\frac{\pm\frac{1}{2\sqrt{f(x)}}\frac{df}{dx}}{2\left(\pm\sqrt{f(x)}\right)}, && (4.298)\\ &= -\frac{\frac{df}{dx}}{4f(x)}. && (4.299) \end{aligned}$$

Thus, we get the general solution

$$\begin{aligned} y(x) &= C_1\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}(S_0(s) + \epsilon S_1(s) + \cdots)ds\right)\\ &\quad + C_2\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}(S_0(s) + \epsilon S_1(s) + \cdots)ds\right), && (4.300)\\ y(x) &= C_1\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}(\sqrt{f(s)} - \epsilon\frac{\frac{df}{ds}}{4f(s)} + \cdots)ds\right)\\ &\quad + C_2\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}(-\sqrt{f(s)} - \epsilon\frac{\frac{df}{ds}}{4f(s)} + \cdots)ds\right), && (4.301)\\ y(x) &= C_1\exp\left(-\int_{f(x_0)}^{f(x)}\frac{df}{4f}\right)\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}(\sqrt{f(s)} + \cdots)ds\right)\\ &\quad + C_2\exp\left(-\int_{f(x_0)}^{f(x)}\frac{df}{4f}\right)\exp\left(-\frac{1}{\epsilon}\int_{x_0}^{x}(\sqrt{f(s)} + \cdots)ds\right), && (4.302)\\ y(x) &= \frac{\hat{C}_1}{(f(x))^{1/4}}\exp\left(\frac{1}{\epsilon}\int_{x_0}^{x}\sqrt{f(s)}ds\right) + \frac{\hat{C}_2}{(f(x))^{1/4}}\exp\left(-\frac{1}{\epsilon}\int_{x_0}^{x}\sqrt{f(s)}ds\right) + \cdots. && (4.303) \end{aligned}$$

This solution is not valid near $x = a$ for which $f(a) = 0$. These are called *turning points*. At such points the solution changes from an oscillatory to an exponential character.

Example 4.18

Find an approximate solution of the Airy[14] equation

$$\epsilon^2 y'' + xy = 0, \text{ for } x > 0. \tag{4.304}$$

In this case

$$f(x) = -x. \tag{4.305}$$

Thus, $x = 0$ is a turning point. We find that

$$S_0(x) = \pm i\sqrt{x}, \tag{4.306}$$

and

$$S_1(x) = -\frac{S_0'}{2S_0} = -\frac{1}{4x}. \tag{4.307}$$

The solutions are of the form

$$y = \exp\left(\pm\frac{i}{\epsilon}\int \sqrt{x}\, dx - \int \frac{dx}{4x}\right) + \cdots, \tag{4.308}$$

$$= \frac{1}{x^{1/4}} \exp\left(\pm\frac{2x^{3/2}i}{3\epsilon}\right) + \cdots. \tag{4.309}$$

The general approximate solution is

$$y = \frac{C_1}{x^{1/4}} \sin\left(\frac{2x^{3/2}}{3\epsilon}\right) + \frac{C_2}{x^{1/4}} \cos\left(\frac{2x^{3/2}}{3\epsilon}\right) + \cdots. \tag{4.310}$$

The exact solution can be shown to be

$$y = C_1 \text{Ai}\left(-\epsilon^{-2/3}x\right) + C_2 \text{Bi}\left(-\epsilon^{-2/3}x\right). \tag{4.311}$$

Here Ai and Bi are Airy functions of the first and second kind, respectively. See Sec. 10.7.9 in the Appendix.

Example 4.19

Find a solution of $x^3 y'' = y$, for small, positive x.

Let $\epsilon^2 X = x$, so that X is of $O(1)$ when x is small. Then the equation becomes

$$\epsilon^2 \frac{d^2 y}{dX^2} = X^{-3} y. \tag{4.312}$$

[14] George Biddell Airy, 1801-1892, English applied mathematician, First Wrangler at Cambridge, holder of the Lucasian Chair (that held by Newton) at Cambridge, Astronomer Royal who had some role in delaying the identification of Neptune as predicted by John Couch Adams' perturbation theory in 1845.

The WKBJ method is applicable. We have $f = X^{-3}$. The general solution is

$$y = C_1' X^{3/4} \exp\left(-\frac{2}{\epsilon\sqrt{X}}\right) + C_2' X^{3/4} \exp\left(\frac{2}{\epsilon\sqrt{X}}\right) + \cdots. \tag{4.313}$$

In terms of the original variables

$$y = C_1 x^{3/4} \exp\left(-\frac{2}{\sqrt{x}}\right) + C_2 x^{3/4} \exp\left(\frac{2}{\sqrt{x}}\right) + \cdots. \tag{4.314}$$

The exact solution can be shown to be

$$y = \sqrt{x}\left(C_1 I_1\left(\frac{2}{\sqrt{x}}\right) + C_2 K_1\left(\frac{2}{\sqrt{x}}\right)\right). \tag{4.315}$$

Here I_1 is a modified Bessel function of the first kind of order one, and K_1 is a modified Bessel function of the second kind of order one.

4.2.7 Solutions of the type $e^{S(x)}$

Example 4.20

Solve

$$x^3 y'' = y, \tag{4.316}$$

for small, positive x.

Let $y = e^{S(x)}$, so that $y' = S' e^S$, $y'' = (S')^2 e^S + S'' e^S$, from which

$$S'' + (S')^2 = x^{-3}. \tag{4.317}$$

Assume that $S'' \ll (S')^2$ (to be checked later). Thus, $S' = \pm x^{-3/2}$, and $S = \pm 2x^{-1/2}$. Checking we get $S''/(S')^2 = x^{1/2} \to 0$ as $x \to 0$, confirming the assumption. Now we add a correction term so that $S(x) = 2x^{-1/2} + C(x)$, where we have taken the positive sign. Assume that $C \ll 2x^{-1/2}$. Substituting in the equation, we have

$$\frac{3}{2}x^{-5/2} + C'' - 2x^{-3/2}C' + (C')^2 = 0. \tag{4.318}$$

Since $C \ll 2x^{-1/2}$, we have $C' \ll x^{-3/2}$ and $C'' \ll (3/2)x^{-5/2}$. Thus

$$\frac{3}{2}x^{-5/2} - 2x^{-3/2}C' = 0, \tag{4.319}$$

from which $C' = (3/4)x^{-1}$ and $C = (3/4)\ln x$. We can now check the assumption on C.

We have $S(x) = 2x^{-1/2} + (3/4)\ln x$, so that

$$y = x^{3/4} \exp\left(-\frac{2}{\sqrt{x}}\right) + \cdots. \tag{4.320}$$

Another solution is obtained by taking $S(x) = -2x^{-1/2} + C(x)$. This procedure is similar to that of the WKBJ method, and the solution is identical. The exact solution is of course the same as the previous example.

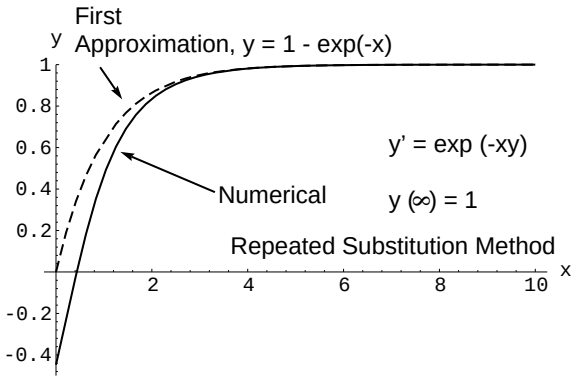

Figure 4.17: Numerical and first approximate solution for repeated substitution problem.

4.2.8 Repeated substitution

This technique sometimes works if the range of the independent variable is such that some term is small.

Example 4.21

Solve

$$y' = e^{-xy}, \qquad y(\infty) \to c, \quad c > 0, \tag{4.321}$$

for $y > 0$ and large x.

As $x \to \infty$, $y' \to 0$, so that $y \to c$. Substituting $y = c$ into Eq. (4.321), we get

$$y' = e^{-cx}, \tag{4.322}$$

which can be integrated to get, after application of the boundary condition,

$$y = c - \frac{1}{c} e^{-cx}. \tag{4.323}$$

Substituting Eq. (4.323) into the original Eq. (4.321), we find

$$y' = \exp\left(-x\left(c - \frac{1}{c}e^{-cx}\right)\right), \tag{4.324}$$

$$= e^{-cx}\left(1 + \frac{x}{c}e^{-cx} + \ldots\right). \tag{4.325}$$

which can be integrated to give

$$y = c - \frac{1}{c}e^{-cx} - \frac{1}{c^2}\left(x + \frac{1}{2c}\right)e^{-2cx} + \cdots. \tag{4.326}$$

The series converges for large x. An accurate numerical solution along with the first approximation are plotted in Fig. 4.17.

Problems

1. Solve as a series in x for $x > 0$ about the point $x = 0$:

   (a) $x^2y'' - 2xy' + (x+1)y = 0; \qquad y(1) = 1,\ y(4) = 0.$

   (b) $xy'' + y' + 2x^2y = 0; \qquad |y(0)| < \infty,\ y(1) = 1.$

   In each case find the exact solution with a symbolic computation program, and compare graphically the first four terms of your series solution with the exact solution.

2. Find two-term expansions for each of the roots of

$$(x-1)(x+3)(x-3\lambda) + 1 = 0,$$

   where λ is large.

3. Find two terms of an approximate solution of

$$y'' + \frac{\lambda}{\lambda + x}y = 0,$$

   with $y(0) = 0, y(1) = 1$, where λ is a large parameter. For $\lambda = 20$, plot $y(x)$ for the two-term expansion. Also compute the exact solution by numerical integration. Plot the difference between the asymptotic and numerical solution versus x.

4. Find the leading order solution for

$$(x - \epsilon y)\frac{dy}{dx} + xy = e^{-x},$$

   where $y(1) = 1$, and $x \in [0, 1]$, $\epsilon \ll 1$. For $\epsilon = 0.2$, plot the asymptotic solution, the exact solution and the difference versus x.

5. The motion of a pendulum is governed by the equation

$$\frac{d^2x}{dt^2} + \sin(x) = 0,$$

   with $x(0) = \epsilon$, $\frac{dx}{dt}(0) = 0$. Using strained coordinates, find the approximate solution of $x(t)$ for small ϵ through $O(\epsilon^2)$. Plot your results for both your asymptotic results and those obtained by a numerical integration of the full equation.

6. Find an approximate solution for

$$y'' - ye^{y/10} = 0,$$

   with $y(0) = 1$, $y(1) = e$.

7. Find an approximate solution for the following problem:

$$\ddot{y} - ye^{y/12} = 0, \text{ with } y(0) = 0.1,\ \dot{y}(0) = 1.2.$$

   Compare with the numerical solution for $0 \le x \le 1$.

8. Find the lowest order solution for

$$\epsilon^2 y'' + \epsilon y^2 - y + 1 = 0,$$

   with $y(0) = 1$, $y(1) = 3$, where ϵ is small. For $\epsilon = 0.2$, plot the asymptotic and exact solutions.

9. Show that for small ϵ the solution of
$$\frac{dy}{dt} - y = \epsilon e^t,$$
with $y(0) = 1$ can be approximated as an exponential on a slightly different time scale.

10. Obtain approximate general solutions of the following equations near $x = 0$.
    (a) $xy'' + y' + xy = 0$, through $O(x^6)$,
    (b) $xy'' + y = 0$, through $O(x^2)$.

11. Find all solutions through $O(\epsilon^2)$, where ϵ is a small parameter, and compare with the exact result for $\epsilon = 0.01$.
    (a) $4x^4 + 4(\epsilon + 1)x^3 + 3(2\epsilon - 5)x^2 + (2\epsilon - 16)x - 4 = 0$,
    (b) $2\epsilon x^4 + 2(2\epsilon + 1)x^3 + (7 - 2\epsilon)x^2 - 5x - 4 = 0$.

12. Find three terms of a solution of
$$x + \epsilon \cos(x + 2\epsilon) = \frac{\pi}{2},$$
where ϵ is a small parameter. For $\epsilon = 0.2$, compare the best asymptotic solution with the exact solution.

13. Find three terms of the solution of
$$\dot{x} + 2x + \epsilon x^2 = 0, \text{ with } x(0) = \cosh \epsilon,$$
where ϵ is a small parameter. Compare graphically with the exact solution for $\epsilon = 0.3$ and $0 \leq t \leq 2$.

14. Write down an approximation for
$$\int_0^{\pi/2} \sqrt{1 + \epsilon \cos^2 x}\, dx,$$
if $\epsilon = 0.1$, so that the absolute error is less than 2×10^{-4}.

15. Solve
$$y'' + y = e^{\epsilon \sin x}, \text{ with } y(0) = y(1) = 0,$$
through $O(\epsilon)$, where ϵ is a small parameter. For $\epsilon = 0.25$ graphically compare the asymptotic solution with a numerically obtained solution.

16. The solution of the matrix equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{y}$ can be written as $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{y}$. Find the perturbation solution of $(\mathbf{A} + \epsilon \mathbf{B}) \cdot \mathbf{x} = \mathbf{y}$, where ϵ is a small parameter.

17. Find all solutions of $\epsilon x^4 + x - 2 = 0$ approximately, if ϵ is small and positive. If $\epsilon = 0.001$, compare the exact solution obtained numerically with the asymptotic solution.

18. Obtain the first two terms of an approximate solution to
$$\ddot{x} + 3(1 + \epsilon)\dot{x} + 2x = 0, \text{ with } x(0) = 2(1 + \epsilon),\ \dot{x}(0) = -3(1 + 2\epsilon),$$
for small ϵ. Compare the approximate and exact solutions graphically in the range $0 \leq x \leq 1$ for (a) $\epsilon = 0.1$, (b) $\epsilon = 0.25$, and (c) $\epsilon = 0.5$.

19. Find an approximate solution to
$$\ddot{x} + (1 + \epsilon)x = 0, \text{ with } x(0) = A,\ \dot{x}(0) = B,$$
for small, positive ϵ. Compare with the exact solution. Plot both the exact solution and the approximate solution on the same graph for $A = 1, B = 0, \epsilon = 0.3$.

20. Find an approximate solution to the following problem for small ϵ

$$\epsilon^2 \ddot{y} - y = -1, \text{ with } y(0) = 0,\ y(1) = 0.$$

Compare graphically with the exact solution for $\epsilon = 0.1$.

21. Solve to leading order

$$\epsilon y'' + y y' - y = 0, \text{ with } y(0) = 0,\ y(1) = 3.$$

Compare graphically to the exact solution for $\epsilon = 0.2$.

22. If $\ddot{x} + x + \epsilon x^3 = 0$ with $x(0) = A, \dot{x}(0) = 0$ where ϵ is small, a regular expansion gives $x(t) \approx A\cos t + \epsilon(A^3/32)(-\cos t + \cos 3t - 12t \sin t)$. Explain why this is not valid for all time, and obtain a better solution by inserting $t = (1 + a_1\epsilon + \ldots)\tau$ into this solution, expanding in terms of ϵ, and choosing $a_1, a_2, \cdots$ properly (Pritulo's method).

23. Use perturbations to find an approximate solution to

$$y'' + \lambda y' = \lambda, \text{ with } y(0) = 0,\ y(1) = 0,$$

where $\lambda \gg 1$.

24. Find the complementary functions of

$$y''' - xy = 0,$$

in terms of expansions near $x = 0$. Retain only two terms for each function.

25. Find, correct to $O(\epsilon)$, the solution of

$$\ddot{x} + (1 + \epsilon\cos 2t)\, x = 0, \text{ with } x(0) = 1, \text{ and } \dot{x}(0) = 0,$$

that is bounded for all t, where $\epsilon \ll 1$.

26. Find the function f to $O(\epsilon)$ where it satisfies the integral equation

$$x = \int_0^{x + \epsilon \sin x} f(\xi)\ d\xi.$$

27. Find three terms of a perturbation solution of

$$y'' + \epsilon y^2 = 0,$$

with $y(0) = 0$, $y(1) = 1$ for $\epsilon \ll 1$. For $\epsilon = 2.5$, compare the $O(1), O(\epsilon)$, and $O(\epsilon^2)$ solutions to a numerically obtained solution in $x \in [0, 1]$.

28. Obtain a power series solution (in summation form) for $y' + ky = 0$ about $x = 0$, where k is an arbitrary, nonzero constant. Compare to a Taylor series expansion of the exact solution.

29. Obtain two terms of an approximate solution for $\epsilon e^x = \cos x$ when ϵ is small. Graphically compare to the actual values (obtained numerically) when $\epsilon = 0.2, 0.1, 0.01$.

30. Obtain three terms of a perturbation solution for the roots of the equation $(1 - \epsilon)x^2 - 2x + 1 = 0$. (Hint: The expansion $x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \ldots$ will not work.)

31. The solution of the matrix equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{y}$ can be written as $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{y}$. Find the n^{th} term of the perturbation solution of $(\mathbf{A} + \epsilon\mathbf{B}) \cdot \mathbf{x} = \mathbf{y}$, where ϵ is a small parameter. Obtain the first three terms of the solution for

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 2 & 3 \end{pmatrix},\ \mathbf{B} = \begin{pmatrix} 1/10 & 1/2 & 1/10 \\ 0 & 1/5 & 0 \\ 1/2 & 1/10 & 1/2 \end{pmatrix},\ \mathbf{y} = \begin{pmatrix} 1/2 \\ 1/5 \\ 1/10 \end{pmatrix}.$$

32. Obtain leading and first order terms for u and v, governed by the following set of coupled differential equations, for small ϵ:

$$\frac{d^2u}{dx^2} + \epsilon v \frac{du}{dx} = 1, \ u(0) = 0, \ u(1) = \frac{1}{2} + \frac{1}{120}\epsilon,$$

$$\frac{d^2v}{dx^2} + \epsilon u \frac{dv}{dx} = x, \ v(0) = 0, \ v(1) = \frac{1}{6} + \frac{1}{80}\epsilon.$$

Compare asymptotic and numerically obtained results for $\epsilon = 0.2$.

33. Obtain two terms of a perturbation solution to $\epsilon f_{xx} + f_x = -e^{-x}$ with boundary conditions $f(0) = 0$, $f(1) = 1$. Graph the solution for $\epsilon = 0.2, 0.1, 0.05, 0.025$ on $0 \leq x \leq 1$.

34. Find two uniformly valid approximate solutions of

$$\ddot{u} + \frac{\omega^2 u}{1+u^2} = 0, \text{ with } u(0) = 0,$$

up to the first order. Note that ω is not small.

35. Using a two-variable expansion, find the lowest order solution of

(a) $\ddot{x} + \epsilon\dot{x} + x = 0$ with $x(0) = 0, \ \dot{x}(0) = 1$,

(b) $\ddot{x} + \epsilon\dot{x}^3 + x = 0$ with $x(0) = 0, \ \dot{x}(0) = 1$.

where $\epsilon \ll 1$. Compare asymptotic and numerically obtained results for $\epsilon = 0.01$.

36. Obtain a three-term solution of

$$\epsilon\ddot{x} - \dot{x} = 1, \text{ with } x(0) = 0, \ x(1) = 2,$$

where $\epsilon \ll 1$.

37. Find an approximate solution to the following problem for small ϵ

$$\epsilon^2\ddot{y} - y = -1 \text{ with } y(0) = 0, \ y(1) = 0.$$

Compare graphically with the exact solution for $\epsilon = 0.1$.

38. A projectile of mass m is launched at an angle α with respect to the horizontal, and with an initial velocity V. Find the time it takes to reach its maximum height. Assume that the air resistance is small and can be written as k times the square of the velocity of the projectile. Choosing appropriate values for the parameters, compare with the numerical result.

39. For small ϵ, solve using WKBJ

$$\epsilon^2 y'' = (1+x^2)^2 y, \text{ with } y(0) = 0, \ y(1) = 1.$$

40. Obtain a general series solution of

$$y'' + k^2 y = 0,$$

about $x = 0$.

41. Find a general solution of

$$y'' + e^x y = 1,$$

near $x = 0$.

42. Solve
$$x^2 y'' + x\left(\frac{1}{2} + 2x\right) y' + \left(x - \frac{1}{2}\right) y = 0,$$
around $x = 0$.

43. Solve $y'' - \sqrt{x}y = 0$, $x > 0$ in each one of the following ways:
    (a) Substitute $x = \epsilon^{-4/5} X$, and then use WKBJ.
    (b) Substitute $x = \epsilon^{2/5} X$, and then use regular perturbation.
    (c) Find an approximate solution of the kind $y = e^{S(x)}$.

    where ϵ is small

44. Find a solution of
$$y''' - \sqrt{x}y = 0,$$
for small $x \geq 0$.

45. Find an approximate general solution of
$$(x \sin x)\ y'' + (2x \cos x + x^2 \sin x)\ y' + (x \sin x + \sin x + x^2 \cos x)\ y = 0,$$
valid near $x = 0$.

46. A bead can slide along a circular hoop in a vertical plane. The bead is initially at the lowest position, $\theta = 0$, and given an initial velocity of $2\sqrt{gR}$, where g is the acceleration due to gravity and R is the radius of the hoop. If the friction coefficient is μ, find the maximum angle θ_{max} reached by the bead. Compare perturbation and numerical results. Present results on a θ_{max} vs. μ plot, for $0 \leq \mu \leq 0.3$.

47. The initial velocity downwards of a body of mass m immersed in a very viscous fluid is V. Find the velocity of the body as a function of time. Assume that the viscous force is proportional to the velocity. Assume that the inertia of the body is small, but not negligible, relative to viscous and gravity forces. Compare perturbation and exact solutions graphically.

48. For small ϵ, solve to lowest order using the method of multiple scales
$$\ddot{x} + \epsilon \dot{x} + x = 0, \text{ with } x(0) = 0,\ \dot{x}(0) = 1.$$
    Compare exact and asymptotic results for $\epsilon = 0.3$.

49. For small ϵ, solve using WKBJ
$$\epsilon^2 y'' = (1 + x^2)^2 y, \text{ with } y(0) = 0,\ y(1) = 1.$$
    Plot asymptotic and numerical solutions for $\epsilon = 0.11$.

50. Find the lowest order approximate solution to
$$\epsilon^2 y'' + \epsilon y^2 - y + 1 = 0, \text{ with } y(0) = 1,\ y(1) = 2,$$
    where ϵ is small. Plot asymptotic and numerical solutions for $\epsilon = 0.23$.

51. A pendulum is used to measure the earth's gravity. The frequency of oscillation is measured, and the gravity calculated assuming a small amplitude of motion and knowing the length of the pendulum. What must the maximum initial angular displacement of the pendulum be if the error in gravity is to be less than 1%. Neglect air resistance.

52. Find two terms of an approximate solution of

$$y'' + \frac{\lambda}{\lambda + x} y = 0,$$

with $y(0) = 0, y(1) = 1$, where λ is a large parameter.

53. Find all solutions of $e^{\epsilon x} = x^2$ through $O(\epsilon^2)$, where ϵ is a small parameter.

54. Solve

$$(1 + \epsilon)y'' + \epsilon y^2 = 1,$$

with $y(0) = 0, y(1) = 1$ through $O(\epsilon^2)$, where ϵ is a small parameter.

55. Solve to lowest order

$$\epsilon y'' + y' + \epsilon y^2 = 1,$$

with $y(0) = -1, y(1) = 1$, where ϵ is a small parameter. For $\epsilon = 0.2$, plot asymptotic and numerical solutions to the full equation.

56. Find the series solution of the differential equation

$$y'' + xy = 0,$$

around $x = 0$ up to four terms.

57. Find the local solution of the equation

$$y'' = \sqrt{x} y,$$

near $x \to 0^+$.

58. Find the solution of the transcendental equation

$$\sin x = \epsilon \cos 2x,$$

near $x = \pi$ for small positive ϵ.

59. Solve

$$\epsilon y'' - y' = 1,$$

with $y(0) = 0, y(1) = 2$ for small ϵ. Plot asymptotic and numerical solutions for $\epsilon = 0.04$.

60. Find two terms of the perturbation solution of

$$(1 + \epsilon y)y'' + \epsilon y'^2 - N^2 y = 0,$$

with $y'(0) = 0, y(1) = 1.$ for small ϵ. N is a constant. Plot the asymptotic and numerical solution for $\epsilon = 0.12$, $N = 10$.

61. Solve

$$\epsilon y'' + y' = \frac{1}{2},$$

with $y(0) = 0, y(1) = 1$ for small ϵ. Plot asymptotic and numerical solutions for $\epsilon = 0.12$.

62. Find if the van der Pol equation

$$\ddot{y} - \epsilon(1 - y^2)\dot{y} + k^2 y = 0,$$

has a limit cycle of the form $y = A \cos \omega t$.

63. Solve $y' = e^{-2xy}$ for large x where y is positive. Plot $y(x)$.

Chapter 5

Orthogonal functions and Fourier series

see Kaplan, Chapter 7,
see Lopez, Chapters 10, 16,
see Riley, Hobson, and Bence, Chapter 15.4, 15.5.

Solution of linear differential equations gives rise to complementary functions. Some of these are well known, such as sine and cosine. This chapter will consider these and other functions which arise from the solution of a variety of linear second order differential equations with constant and non-constant coefficients. The notion of *eigenvalues*, *eigenfunctions*, *orthogonal*, and *orthonormal* functions will be introduced; a stronger foundation will be built in Chapter 7 on linear analysis. A key result of the present chapter will be to show how one can expand an arbitrary function in terms of infinite sums of the product of scalar amplitudes with orthogonal basis functions. Such a summation is known as a *Fourier*[1] series.

5.1 Sturm-Liouville equations

Consider on the domain $x \in [x_0, x_1]$ the following general linear homogeneous second order differential equation with general homogeneous boundary conditions:

$$a(x)\frac{d^2y}{dx^2} + b(x)\frac{dy}{dx} + c(x)y + \lambda y = 0, \tag{5.1}$$

$$\alpha_1 y(x_0) + \alpha_2 y'(x_0) = 0, \tag{5.2}$$

$$\beta_1 y(x_1) + \beta_2 y'(x_1) = 0. \tag{5.3}$$

Define the following functions:

$$p(x) = \exp\left(\int_{x_o}^{x} \frac{b(s)}{a(s)} ds\right), \tag{5.4}$$

[1] Jean Baptiste Joseph Fourier, 1768-1830, French mathematician.

$$r(x) = \frac{1}{a(x)} \exp\left(\int_{x_o}^{x} \frac{b(s)}{a(s)} ds \right), \tag{5.5}$$

$$q(x) = \frac{c(x)}{a(x)} \exp\left(\int_{x_o}^{x} \frac{b(s)}{a(s)} ds \right). \tag{5.6}$$

With these definitions, Eq. (5.1) is transformed to the type known as a *Sturm-Liouville*[2] equation:

$$\frac{d}{dx}\left(p(x) \frac{dy}{dx} \right) + (q(x) + \lambda r(x))\ y(x) = 0, \tag{5.7}$$

$$\underbrace{\left(\frac{1}{r(x)} \left(\frac{d}{dx}\left(p(x) \frac{d}{dx} \right) + q(x) \right) \right)}_{\mathbf{L}_s} y(x) = -\lambda\ y(x). \tag{5.8}$$

Here the Sturm-Liouville linear operator $\mathbf{L_s}$ is

$$\mathbf{L}_s = \frac{1}{r(x)} \left(\frac{d}{dx}\left(p(x) \frac{d}{dx} \right) + q(x) \right), \tag{5.9}$$

so we have Eq. (5.8) compactly stated as

$$\mathbf{L}_s y(x) = -\lambda\ y(x). \tag{5.10}$$

It can be shown that $\mathbf{L}_s$ is what is known as a *self-adjoint* linear operator; see Sec. 7.4.2. What has been shown then is that all systems of the form of Eqs. (5.1-5.3) can be transformed into a self-adjoint form.

Now the trivial solution $y(x) = 0$ will satisfy the differential equation and boundary conditions, Eqs. (5.1-5.3). In addition, for special real values of λ, known as *eigenvalues*, there are special non-trivial functions, known as *eigenfunctions* which also satisfy Eqs. (5.1-5.3). Eigenvalues and eigenfunctions will be discussed in more general terms in Sec. 7.4.4.

Now it can be shown that if we have for $x \in [x_0, x_1]$

$$p(x) > 0, \tag{5.11}$$

$$r(x) > 0, \tag{5.12}$$

$$q(x) \geq 0, \tag{5.13}$$

then an infinite number of real positive eigenvalues λ and corresponding eigenfunctions $y_n(x)$ exist for which Eqs. (5.1-5.3) are satisfied. Moreover, it can also be shown (Hildebrand, p. 204) that a consequence of the homogeneous boundary conditions is the *orthogonality condition*:

$$<y_n, y_m> = \int_{x_0}^{x_1} r(x) y_n(x) y_m(x)\ dx = 0, \text{ for } n \neq m, \tag{5.14}$$

$$<y_n, y_n> = \int_{x_0}^{x_1} r(x) y_n(x) y_n(x)\ dx = K^2. \tag{5.15}$$

[2] Jacques Charles François Sturm, 1803-1855, Swiss-born French mathematician and Joseph Liouville, 1809-1882, French mathematician.

Consequently, in the same way that in ordinary vector mechanics $\mathbf{i}\cdot\mathbf{j}=0$, $\mathbf{i}\cdot\mathbf{k}=0$, $\mathbf{i}\cdot\mathbf{i}=1$ implies $\mathbf{i}$ is orthogonal to $\mathbf{j}$ and $\mathbf{k}$, the eigenfunctions of a Sturm-Liouville operator $\mathbf{L}_s$ are said to be orthogonal to each other. The so-called inner product notation, $<\cdot,\cdot>$, will be explained in detail in Sec. 7.3.2. Here $K\in\mathbb{R}^1$ is a real constant. This can be written compactly using the Kronecker delta function, δ_{nm} as

$$\int_{x_0}^{x_1} r(x)y_n(x)y_m(x)\,dx \;=\; K^2\delta_{nm}. \tag{5.16}$$

Sturm-Liouville theory shares many more analogies with vector algebra. In the same sense that the dot product of a vector with itself is guaranteed positive, we have defined a "product" for the eigenfunctions in which the "product" of a Sturm-Liouville eigenfunction with itself is guaranteed positive.

Motivated by Eq. (5.16), we can define functions $\varphi_n(x)$:

$$\varphi_n(x)=\frac{\sqrt{r(x)}}{K}\,y_n(x), \tag{5.17}$$

so that

$$<\varphi_n,\varphi_m>=\int_{x_0}^{x_1}\varphi_n(x)\varphi_m(x)\,dx=\delta_{nm}. \tag{5.18}$$

Such functions are said to be *orthonormal*, in the same way that $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are orthonormal. While orthonormal functions have great utility, note that in the context of our Sturm-Liouville nomenclature, that $\varphi_n(x)$ does not in general satisfy the Sturm-Liouville equation: $\mathbf{L}_s\varphi_n(x)\neq-\lambda_n\varphi_n(x)$. If, however, $r(x)=C$, where C is a scalar constant, then in fact $\mathbf{L}_s\varphi_n(x)=-\lambda_n\varphi_n(x)$. Whatever the case, we are guaranteed $\mathbf{L}_s y_n(x)=-\lambda_n y_n(x)$. The $y_n(x)$ functions are orthogonal under the influence of the weighting function $r(x)$, but not necessarily orthonormal. The following sections give special cases of the Sturm-Liouville equation with general homogeneous boundary conditions.

5.1.1 Linear oscillator

A linear oscillator gives perhaps the simplest example of a Sturm-Liouville problem. We will consider the domain $x\in[0,1]$. For other domains, we could easily transform coordinates; e.g. if $x\in[x_0,x_1]$, then the linear mapping $\tilde{x}=(x-x_0)/(x_1-x_0)$ lets us consider $\tilde{x}\in[0,1]$.

The equations governing a linear oscillator with general homogeneous boundary conditions are

$$\frac{d^2y}{dx^2}+\lambda y=0,\qquad \alpha_1 y(0)+\alpha_2\frac{dy}{dx}(0)=0,\qquad \beta_1 y(1)+\beta_2\frac{dy}{dx}(1)=0. \tag{5.19}$$

Here we have

$$a(x) \;=\; 1, \tag{5.20}$$
$$b(x) \;=\; 0, \tag{5.21}$$
$$c(x) \;=\; 0, \tag{5.22}$$

so

$$p(x) = \exp\left(\int_{x_o}^{x} \frac{0}{1}\, ds\right) = e^0 = 1, \tag{5.23}$$

$$r(x) = \frac{1}{1}\exp\left(\int_{x_o}^{x} \frac{0}{1}\, ds\right) = e^0 = 1, \tag{5.24}$$

$$q(x) = \frac{0}{1}\exp\left(\int_{x_o}^{x} \frac{0}{1}\, ds\right) = 0. \tag{5.25}$$

So, we can consider the domain $x \in (-\infty, \infty)$. In practice it is more common to consider the finite domain in which $x \in [0, 1]$. The Sturm-Liouville operator is

$$\mathbf{L}_s = \frac{d^2}{dx^2}. \tag{5.26}$$

The eigenvalue problem is

$$\underbrace{\frac{d^2}{dx^2}}_{\mathbf{L}_s}\, y(x) = -\lambda\, y(x). \tag{5.27}$$

We can find a series solution by assuming $y = \sum_{n=0}^{\infty} a_n x^n$. This leads us to the recursion relationship

$$a_{n+2} = \frac{-\lambda a_n}{(n+1)(n+2)}. \tag{5.28}$$

So, given two seed values, a_0 and a_1, detailed analysis of the type considered in Sec. 4.1.2 reveals the solution can be expressed as the infinite series

$$y(x) = a_0 \underbrace{\left(1 - \frac{(\sqrt{\lambda}x)^2}{2!} + \frac{(\sqrt{\lambda}x)^4}{4!} - \ldots\right)}_{\cos(\sqrt{\lambda}x)} + a_1 \underbrace{\left(\sqrt{\lambda}x - \frac{(\sqrt{\lambda}x)^3}{3!} + \frac{(\sqrt{\lambda}x)^5}{5!} - \ldots\right)}_{\sin(\sqrt{\lambda}x)}. \tag{5.29}$$

The series is recognized as being composed of linear combinations of the Taylor series for $\cos(\sqrt{\lambda}x)$ and $\sin(\sqrt{\lambda}x)$ about $x = 0$. Letting $a_0 = C_1$ and $a_1 = C_2$, we can express the general solution in terms of these two complementary functions as

$$y(x) = C_1 \cos(\sqrt{\lambda}x) + C_2 \sin(\sqrt{\lambda}x). \tag{5.30}$$

Applying the general homogeneous boundary conditions from Eq. (5.19) leads to a challenging problem for determining admissible eigenvalues λ. To apply the boundary conditions, we need dy/dx, which is

$$\frac{dy}{dx} = -C_1\sqrt{\lambda}\sin(\sqrt{\lambda}x) + C_2\sqrt{\lambda}\cos(\sqrt{\lambda}x). \tag{5.31}$$

Enforcing the boundary conditions at $x = 0$ and $x = 1$ leads us to two equations:

$$\alpha_1 C_1 + \alpha_2 \sqrt{\lambda} C_2 = 0, \tag{5.32}$$

$$C_1 \left(\beta_1 \cos\sqrt{\lambda} - \beta_2 \sqrt{\lambda} \sin\sqrt{\lambda} \right) + C_2 \left(\beta_1 \sin\sqrt{\lambda} + \beta_2 \sqrt{\lambda} \cos\sqrt{\lambda} \right) = 0. \tag{5.33}$$

This can be posed as the linear system

$$\begin{pmatrix} \alpha_1 & \alpha_2\sqrt{\lambda} \\ \left(\beta_1 \cos\sqrt{\lambda} - \beta_2 \sqrt{\lambda} \sin\sqrt{\lambda} \right) & \left(\beta_1 \sin\sqrt{\lambda} + \beta_2 \sqrt{\lambda} \cos\sqrt{\lambda} \right) \end{pmatrix} \cdot \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5.34}$$

For non-trivial solutions, the determinant of the coefficient matrix must be zero, which leads to the transcendental equation

$$\alpha_1 \left(\beta_1 \sin\sqrt{\lambda} + \beta_2 \sqrt{\lambda} \cos\sqrt{\lambda} \right) - \alpha_2 \sqrt{\lambda} \left(\beta_1 \cos\sqrt{\lambda} - \beta_2 \sqrt{\lambda} \sin\sqrt{\lambda} \right) = 0. \tag{5.35}$$

For known values of α_1, α_2, β_2, and β_1, one seeks values of λ which satisfy Eq. (5.35). This is a solution which in general must be done numerically, except for the simplest of cases.

One important simple case is for $\alpha_1 = 1$, $\alpha_2 = 0$, $\beta_1 = 1$, $\beta_2 = 0$. This gives the boundary conditions to be $y(0) = y(1) = 0$. Boundary conditions where the function values are specified are known as *Dirichlet*[3] conditions. In this case, Eq. (5.35) reduces to $\sin\sqrt{\lambda} = 0$, which is easily solved as $\sqrt{\lambda} = n\pi$, with $n = 0, \pm 1, \pm 2, \ldots$. We also get $C_1 = 0$; consequently, $y = C_2 \sin(n\pi x)$. Note that for $n = 0$, the solution is the trivial $y = 0$.

Another set of conditions also leads to a similarly simple result. Taking $\alpha_1 = 0$, $\alpha_2 = 1$, $\beta_1 = 0$, $\beta_2 = 1$, the boundary conditions are $y'(0) = y'(1) = 0$. Boundary conditions where the function's derivative values are specified are known as *Neumann*[4] conditions. In this case, Eq. (5.35) reduces to $-\lambda \sin\sqrt{\lambda} = 0$, which is easily solved as $\sqrt{\lambda} = n\pi$, with $n = 0, \pm 1, \pm 2, \ldots$. We also get $C_2 = 0$; consequently, $y = C_1 \cos(n\pi x)$. Here, for $n = 0$, the solution is the non-trivial $y = C_1$.

Some of the eigenfunctions for Dirichlet and Neumann boundary conditions are plotted in Fig. 5.1. Note these two families form the linearly independent complementary functions of Eq. (5.19). Also note that as n rises, the number of zero-crossings within the domain rises. This will be seen to be characteristic of all sets of eigenfunctions for Sturm-Liouville equations.

Example 5.1

Find the eigenvalues and eigenfunctions for a linear oscillator equation with Dirichlet boundary conditions:

$$\frac{d^2y}{dx^2} + \lambda y = 0, \qquad y(0) = y(\ell) = 0. \tag{5.36}$$

[3]Johann Peter Gustav Lejeune Dirichlet, 1805-1859, German mathematician who formally defined a function in the modern sense.

[4]Carl Gottfried Neumann, 1832-1925, German mathematician.

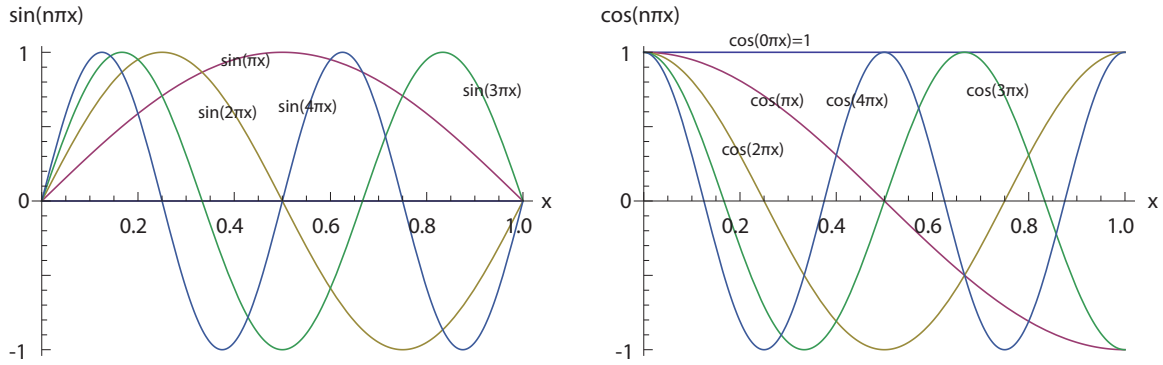

Figure 5.1: Solutions to the linear oscillator equation, Eq. (5.19), in terms of two sets of complementary functions, $\sin(n\pi x)$ and $\cos(n\pi x)$.

We could transform the domain via $\tilde{x} = x/\ell$ so that $\tilde{x} \in [0, 1]$, but this problem is sufficiently straightforward to allow us to deal with the original domain. We know by inspection that the general solution is

$$y(x) = C_1 \cos(\sqrt{\lambda}x) + C_2 \sin(\sqrt{\lambda}x). \tag{5.37}$$

For $y(0) = 0$, we get

$$\begin{aligned} y(0) = 0 &= C_1 \cos(\sqrt{\lambda}(0)) + C_2 \sin(\sqrt{\lambda}(0)), & (5.38)\\ 0 &= C_1(1) + C_2(0), & (5.39)\\ C_1 &= 0. & (5.40) \end{aligned}$$

So

$$y(x) = C_2 \sin(\sqrt{\lambda}x). \tag{5.41}$$

At the boundary at $x = \ell$ we have

$$y(\ell) = 0 = C_2 \sin(\sqrt{\lambda}\,\ell). \tag{5.42}$$

For non-trivial solutions we need $C_2 \neq 0$, which then requires that

$$\sqrt{\lambda}\ell = n\pi \qquad n = \pm 1, \pm 2, \pm 3, \ldots, \tag{5.43}$$

so

$$\lambda = \left(\frac{n\pi}{\ell}\right)^2. \tag{5.44}$$

The eigenvalues and eigenfunctions are

$$\lambda_n = \frac{n^2\pi^2}{\ell^2}, \tag{5.45}$$

and

$$y_n(x) = \sin\left(\frac{n\pi x}{\ell}\right), \tag{5.46}$$

respectively.

Check orthogonality for $y_2(x)$ and $y_3(x)$.

$$I = \int_0^\ell \sin\left(\frac{2\pi x}{\ell}\right) \sin\left(\frac{3\pi x}{\ell}\right) dx, \tag{5.47}$$

$$= \frac{\ell}{2\pi}\left(\sin\left(\frac{\pi x}{\ell}\right) - \frac{1}{5}\sin\left(\frac{5\pi x}{\ell}\right)\right)\Bigg|_0^\ell, \tag{5.48}$$

$$= 0. \tag{5.49}$$

Check orthogonality for $y_4(x)$ and $y_4(x)$.

$$I = \int_0^\ell \sin\left(\frac{4\pi x}{\ell}\right) \sin\left(\frac{4\pi x}{\ell}\right) dx, \tag{5.50}$$

$$= \left(\frac{x}{2} - \frac{\ell}{16\pi}\sin\left(\frac{8\pi x}{\ell}\right)\right)\Bigg|_0^\ell, \tag{5.51}$$

$$= \frac{\ell}{2}. \tag{5.52}$$

In fact

$$\int_0^\ell \sin\left(\frac{n\pi x}{\ell}\right) \sin\left(\frac{n\pi x}{\ell}\right) dx = \frac{\ell}{2}, \tag{5.53}$$

so the orthonormal functions $\varphi_n(x)$ for this problem are

$$\varphi_n(x) = \sqrt{\frac{2}{\ell}} \sin\left(\frac{n\pi x}{\ell}\right). \tag{5.54}$$

With this choice, we recover the orthonormality condition

$$\int_0^\ell \varphi_n(x)\varphi_m(x)\, dx = \delta_{nm}, \tag{5.55}$$

$$\frac{2}{\ell}\int_0^\ell \sin\left(\frac{n\pi x}{\ell}\right) \sin\left(\frac{m\pi x}{\ell}\right)\, dx = \delta_{nm}. \tag{5.56}$$

5.1.2 Legendre's differential equation

Legendre's[5] differential equation is given next. Here, it is convenient to let the term $n(n+1)$ play the role of λ.

$$(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \underbrace{n(n+1)}_{\lambda} y = 0. \tag{5.57}$$

[5] Adrien-Marie Legendre, 1752-1833, French/Parisian mathematician.

Here

$$a(x) = 1 - x^2, \tag{5.58}$$
$$b(x) = -2x, \tag{5.59}$$
$$c(x) = 0. \tag{5.60}$$

Then, taking $x_o = -1$, we have

$$p(x) = \exp \int_{-1}^{x} \frac{-2s}{1-s^2}\, ds, \tag{5.61}$$
$$= \exp\left(\ln\left(1-s^2\right)\right)\Big|_{-1}^{x}, \tag{5.62}$$
$$= \left(1-s^2\right)\Big|_{-1}^{x}, \tag{5.63}$$
$$= 1 - x^2. \tag{5.64}$$

We find then that

$$r(x) = 1, \tag{5.65}$$
$$q(x) = 0. \tag{5.66}$$

Thus, we require $x \in (-1, 1)$. In Sturm-Liouville form, Eq. (5.57) reduces to

$$\frac{d}{dx}\left((1-x^2)\,\frac{dy}{dx}\right) + n(n+1)\,y = 0, \tag{5.67}$$
$$\underbrace{\frac{d}{dx}\left((1-x^2)\,\frac{d}{dx}\right)}_{\mathbf{L}_s} y(x) = -n(n+1)\,y(x). \tag{5.68}$$

So

$$\mathbf{L}_s = \frac{d}{dx}\left((1-x^2)\,\frac{d}{dx}\right). \tag{5.69}$$

Now $x = 0$ is a regular point, so we can expand in a power series around this point. Let

$$y = \sum_{m=0}^{\infty} a_m x^m. \tag{5.70}$$

Substituting into Eq. (5.57), we find after detailed analysis that

$$a_{m+2} = a_m \frac{(m+n+1)(m-n)}{(m+1)(m+2)}. \tag{5.71}$$

With a_0 and a_1 as given seeds, we can thus generate all values of a_m for $m \geq 2$. We find

$$\begin{aligned} y(x) &= a_0 \underbrace{\left(1 - n(n+1)\frac{x^2}{2!} + n(n+1)(n-2)(n+3)\frac{x^4}{4!} - \ldots\right)}_{y_1(x)} \\ &\quad + a_1 \underbrace{\left(x - (n-1)(n+2)\frac{x^3}{3!} + (n-1)(n+2)(n-3)(n+4)\frac{x^5}{5!} - \ldots\right)}_{y_2(x)}. \end{aligned} \tag{5.72}$$

Thus, the general solution takes the form

$$y(x) = a_0 y_1(x) + a_1 y_2(x), \tag{5.73}$$

with complementary functions $y_1(x)$ and $y_2(x)$ defined as

$$y_1(x) = 1 - n(n+1)\frac{x^2}{2!} + n(n+1)(n-2)(n+3)\frac{x^4}{4!} - \ldots, \tag{5.74}$$

$$y_2(x) = x - (n-1)(n+2)\frac{x^3}{3!} + (n-1)(n+2)(n-3)(n+4)\frac{x^5}{5!} - \ldots. \tag{5.75}$$

This solution holds for arbitrary real values of n. However, for $n = 0, 2, 4, \ldots$, $y_1(x)$ is a finite polynomial, while $y_2(x)$ is an infinite series which diverges at $|x| = 1$. For $n = 1, 3, 5, \ldots$, it is the other way around. Thus, for integer, non-negative n either 1) y_1 is a polynomial of degree n, and y_2 is a polynomial of infinite degree, or 2) y_1 is a polynomial of infinite degree, and y_2 is a polynomial of degree n.

We could in fact treat y_1 and y_2 as the complementary functions for Eq. (5.57). However, the existence of finite degree polynomials in special cases has led to an alternate definition of the standard complementary functions for Eq. (5.57). The finite polynomials (y_1 for even n, and y_2 for odd n) can be normalized by dividing through by their values at $x = 1$ to give the *Legendre polynomials*, $P_n(x)$:

$$P_n(x) = \begin{cases} \frac{y_1(x)}{y_1(1)}, & \text{for } n \text{ even}, \\ \frac{y_2(x)}{y_2(1)}, & \text{for } n \text{ odd}. \end{cases} \tag{5.76}$$

The Legendre polynomials are thus

$$n = 0, \qquad P_0(x) = 1, \tag{5.77}$$

$$n = 1, \qquad P_1(x) = x, \tag{5.78}$$

$$n = 2, \qquad P_2(x) = \frac{1}{2}(3x^2 - 1), \tag{5.79}$$

$$n = 3, \qquad P_3(x) = \frac{1}{2}(5x^3 - 3x), \tag{5.80}$$

$$n = 4, \qquad P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3), \tag{5.81}$$

$$\vdots$$

$$n, \qquad P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n, \qquad \text{Rodrigues' formula.} \tag{5.82}$$

The Rodrigues[6] formula gives a generating formula for general n.

The orthogonality condition is

$$\int_{-1}^{1} P_n(x)P_m(x)\ dx \ = \ \frac{2}{2n+1}\delta_{nm}. \tag{5.83}$$

Direct substitution shows that $P_n(x)$ satisfies both the differential equation, Eq. (5.57), and the orthogonality condition. It is then easily shown that the following functions are orthonormal on the interval $x \in (-1, 1)$:

$$\varphi_n(x) = \sqrt{n+\frac{1}{2}}P_n(x), \tag{5.84}$$

giving

$$\int_{-1}^{1} \varphi_n(x)\varphi_m(x)dx = \delta_{nm}. \tag{5.85}$$

The total solution, Eq. (5.73), can be recast as the sum of the finite sum of polynomials $P_n(x)$ (Legendre functions of the first kind and degree n) and the infinite sum of polynomials $Q_n(x)$ (Legendre functions of the second kind and degree n):

$$y(x) = C_1P_n(x) + C_2Q_n(x). \tag{5.86}$$

Here $Q_n(x)$, the infinite series portion of the solution, is obtained by

$$Q_n(x) = \begin{cases} y_1(1)y_2(x), & \text{for } n \text{ even}, \\ -y_2(1)y_1(x), & \text{for } n \text{ odd}. \end{cases} \tag{5.87}$$

One can also show the Legendre functions of the second kind, $Q_n(x)$, satisfy a similar orthogonality condition. Additionally, $Q_n(\pm 1)$ is singular. One can further show that the infinite series of polynomials which form $Q_n(x)$ can be recast as a finite series of polynomials along with a logarithmic function. The first few values of $Q_n(x)$ are in fact

$$n = 0, \qquad Q_0(x) \ = \ \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right), \tag{5.88}$$

$$n = 1, \qquad Q_1(x) \ = \ \frac{x}{2}\ln\left(\frac{1+x}{1-x}\right) - 1, \tag{5.89}$$

$$n = 2, \qquad Q_2(x) \ = \ \frac{3x^2-1}{4}\ln\left(\frac{1+x}{1-x}\right) - \frac{3}{2}x, \tag{5.90}$$

$$n = 3, \qquad Q_3(x) \ = \ \frac{5x^3-3x}{4}\ln\left(\frac{1+x}{1-x}\right) - \frac{5}{2}x^2 + \frac{2}{3}, \tag{5.91}$$

$$\vdots$$

The first few eigenfunctions of Eq. (5.57) for the two families of complementary functions are plotted in Fig. 5.2.

[6]Benjamin Olinde Rodrigues, 1794-1851, obscure French mathematician, of Portuguese and perhaps Spanish roots.

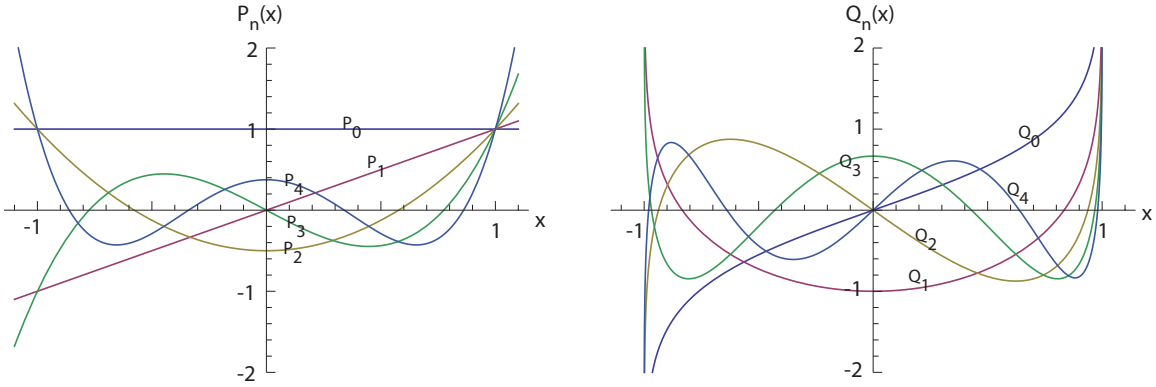

Figure 5.2: Solutions to the Legendre equation, Eq. (5.57), in terms of two sets of complementary functions, $P_n(x)$ and $Q_n(x)$.

5.1.3 Chebyshev equation

The Chebyshev[7] equation is

$$(1-x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + \lambda y = 0. \tag{5.92}$$

Let's get this into Sturm-Liouville form.

$$a(x) = 1-x^2, \tag{5.93}$$
$$b(x) = -x, \tag{5.94}$$
$$c(x) = 0. \tag{5.95}$$

Now, taking $x_0 = -1$,

$$p(x) = \exp\left(\int_{-1}^{x} \frac{b(s)}{a(s)}ds\right), \tag{5.96}$$
$$= \exp\left(\int_{-1}^{x} \frac{-s}{1-s^2}ds\right), \tag{5.97}$$
$$= \exp\left(\frac{1}{2}\ln(1-s^2)\right)\Big|_{-1}^{x}, \tag{5.98}$$
$$= \sqrt{1-s^2}\Big|_{-1}^{x}, \tag{5.99}$$
$$= \sqrt{1-x^2}, \tag{5.100}$$
$$r(x) = \frac{\exp\left(\int_{-1}^{x}\frac{b(s)}{a(s)}ds\right)}{a(x)} = \frac{1}{\sqrt{1-x^2}}, \tag{5.101}$$
$$q(x) = 0. \tag{5.102}$$

[7]Pafnuty Lvovich Chebyshev, 1821-1894, Russian mathematician.

Thus, for $p(x) > 0$, we require $x \in (-1, 1)$. The Chebyshev equation, Eq. (5.92), in Sturm-Liouville form is

$$\frac{d}{dx}\left(\sqrt{1-x^2}\,\frac{dy}{dx}\right) + \frac{\lambda}{\sqrt{1-x^2}}\,y = 0, \tag{5.103}$$

$$\underbrace{\sqrt{1-x^2}\frac{d}{dx}\left(\sqrt{1-x^2}\,\frac{d}{dx}\right)}_{\mathbf{L}_s}\,y(x) = -\lambda\,y(x). \tag{5.104}$$

Thus,

$$\mathbf{L}_s = \sqrt{1-x^2}\frac{d}{dx}\left(\sqrt{1-x^2}\,\frac{d}{dx}\right). \tag{5.105}$$

That the two forms are equivalent can be easily checked by direct expansion.

Series solution techniques reveal for eigenvalues of λ one family of complementary functions of Eq. (5.92) can be written in terms of the so-called Chebyshev polynomials, $T_n(x)$. These are also known as Chebyshev polynomials of the first kind. These polynomials can be obtained by a regular series expansion of the original differential equation. These eigenvalues and eigenfunctions are listed next:

$$\lambda = 0, \qquad T_0(x) = 1, \tag{5.106}$$

$$\lambda = 1, \qquad T_1(x) = x, \tag{5.107}$$

$$\lambda = 4, \qquad T_2(x) = -1 + 2x^2, \tag{5.108}$$

$$\lambda = 9, \qquad T_3(x) = -3x + 4x^3, \tag{5.109}$$

$$\lambda = 16, \qquad T_4(x) = 1 - 8x^2 + 8x^4, \tag{5.110}$$

$$\vdots$$

$$\lambda = n^2, \qquad T_n(x) = \cos(n\cos^{-1}x), \qquad \text{Rodrigues' formula.} \tag{5.111}$$

The orthogonality condition is

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} \pi\delta_{nm}, & \text{if } n = 0, \\ \frac{\pi}{2}\delta_{nm}, & \text{if } n = 1, 2, \ldots. \end{cases}. \tag{5.112}$$

Direct substitution shows that $T_n(x)$ satisfies both the differential equation, Eq. (5.92), and the orthogonality condition. We can deduce then that the functions $\varphi_n(x)$

$$\varphi_n(x) = \begin{cases} \sqrt{\frac{1}{\pi\sqrt{1-x^2}}}\,T_n(x), & \text{if } n = 0, \\ \sqrt{\frac{2}{\pi\sqrt{1-x^2}}}\,T_n(x), & \text{if } n = 1, 2, \ldots \end{cases}. \tag{5.113}$$

are an orthonormal set of functions on the interval $x \in (-1, 1)$. That is,

$$\int_{-1}^{1} \varphi_n(x)\varphi_m(x)dx = \delta_{nm}. \tag{5.114}$$

The Chebyshev polynomials of the first kind, $T_n(x)$ form one set of complementary functions which satisfy Eq. (5.92). The other set of complementary functions are $V_n(x)$, and can be shown to be

$$\begin{aligned}
\lambda = 0, \qquad V_0(x) &= 0, && (5.115)\\
\lambda = 1, \qquad V_1(x) &= \sqrt{1-x^2}, && (5.116)\\
\lambda = 4, \qquad V_2(x) &= \sqrt{1-x^2}(2x), && (5.117)\\
\lambda = 9, \qquad V_3(x) &= \sqrt{1-x^2}(-1+4x^2), && (5.118)\\
\lambda = 16, \qquad V_4(x) &= \sqrt{1-x^2}(-4x^2+8x^3), && (5.119)\\
&\vdots\\
\lambda = n^2, \qquad V_n(x) &= \sin(n\cos^{-1}x), \qquad \text{Rodrigues' formula.} && (5.120)
\end{aligned}$$

The general solution to Eq. (5.214) is a linear combination of the two complementary functions:

$$y(x) = C_1 T_n(x) + C_2 V_n(x). \tag{5.121}$$

One can also show that $V_n(x)$ satisfies an orthogonality condition:

$$\int_{-1}^{1} \frac{V_n(x)V_m(x)}{\sqrt{1-x^2}}\, dx = \frac{\pi}{2}\delta_{nm}. \tag{5.122}$$

The first few eigenfunctions of Eq. (5.92) for the two families of complementary functions are plotted in Fig. 5.3.

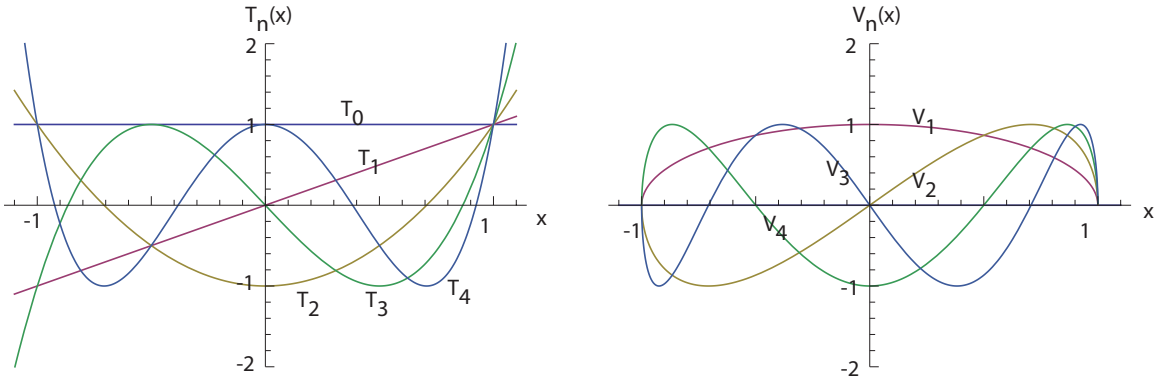

Figure 5.3: Solutions to the Chebyshev equation, Eq. (5.92), in terms of two sets of complementary functions, $T_n(x)$ and $V_n(x)$.

5.1.4 Hermite equation

The Hermite[8] equation is discussed next. There are two common formulations, the physicists' and the probabilists'. We will focus on the first and briefly discuss the second.

5.1.4.1 Physicists'

The physicists' Hermite equation is

$$\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \lambda y = 0. \tag{5.123}$$

We find that

$$p(x) = e^{-x^2}, \tag{5.124}$$
$$r(x) = e^{-x^2}, \tag{5.125}$$
$$q(x) = 0. \tag{5.126}$$

Thus, we allow $x \in (-\infty, \infty)$. In Sturm-Liouville form, Eq. (5.123) becomes

$$\frac{d}{dx}\left(e^{-x^2}\ \frac{dy}{dx}\right) + \lambda e^{-x^2}\ y = 0, \tag{5.127}$$
$$\underbrace{e^{x^2}\frac{d}{dx}\left(e^{-x^2}\ \frac{d}{dx}\right)}_{\mathbf{L}_s}\ y(x) = -\lambda\ y(x). \tag{5.128}$$

So

$$\mathbf{L}_s = e^{x^2}\frac{d}{dx}\left(e^{-x^2}\ \frac{d}{dx}\right). \tag{5.129}$$

One set of complementary functions can be expressed in terms of polynomials known as the Hermite polynomials, $H_n(x)$. These polynomials can be obtained by a regular series expansion of the original differential equation. The eigenvalues and eigenfunctions corresponding to the physicists' Hermite polynomials are listed next:

$$\lambda = 0, \qquad H_0(x) = 1, \tag{5.130}$$
$$\lambda = 2, \qquad H_1(x) = 2x, \tag{5.131}$$
$$\lambda = 4, \qquad H_2(x) = -2 + 4x^2, \tag{5.132}$$
$$\lambda = 6, \qquad H_3(x) = -12x + 8x^3, \tag{5.133}$$
$$\lambda = 8, \qquad H_4(x) = 12 - 48x^2 + 16x^4, \tag{5.134}$$
$$\vdots \tag{5.135}$$
$$\lambda = 2n, \qquad H_n(x) = (-1)^n e^{x^2}\frac{d^n e^{-x^2}}{dx^n}, \qquad \text{Rodrigues' formula.} \tag{5.136}$$

[8] Charles Hermite, 1822-1901, Lorraine-born French mathematician.

The orthogonality condition is

$$\int_{-\infty}^{\infty} e^{-x^2} H_n(x) H_m(x)\ dx \quad = \quad 2^n n! \sqrt{\pi} \delta_{nm} \tag{5.137}$$

Direct substitution shows that $H_n(x)$ satisfies both the differential equation, Eq. (5.123), and the orthogonality condition. It is then easily shown that the following functions are orthonormal on the interval $x \in (-\infty, \infty)$:

$$\varphi_n(x) = \frac{e^{-x^2/2} H_n(x)}{\sqrt{\sqrt{\pi} 2^n n!}}, \tag{5.138}$$

giving

$$\int_{-\infty}^{\infty} \varphi_n(x) \varphi_m(x) dx = \delta_{mn}. \tag{5.139}$$

The general solution to Eq. (5.123) is

$$y(x) = C_1 H_n(x) + C_2 \hat{H}_n(x), \tag{5.140}$$

where the other set of complementary functions is $\hat{H}_n(x)$. For general n, $\hat{H}_n(x)$ is a version of the so-called Kummer confluent hypergeometric function of the first kind $\hat{H}_n(x) = {}_1F_1\left(-n/2; 1/2; x^2\right)$. Note, this general solution should be treated carefully, especially as the second complementary function, $\hat{H}_n(x)$, is rarely discussed in the literature, and notation is often non-standard. For our eigenvalues of n, somewhat simpler results can be obtained in terms of the imaginary error function, erfi(x); see Sec. 10.7.4. The first few of these functions are

$$\lambda = 0, \quad n = 0, \qquad \hat{H}_0(x) \quad = \quad \frac{\sqrt{\pi}}{2} \text{erfi}(x), \tag{5.141}$$

$$\lambda = 2, \quad n = 1, \qquad \hat{H}_1(x) \quad = \quad e^{x^2} - \sqrt{\pi x^2}\ \text{erfi}(\sqrt{x^2}), \tag{5.142}$$

$$\lambda = 4, \quad n = 2, \qquad \hat{H}_2(x) \quad = \quad -x e^{x^2} + \sqrt{\pi}\ \text{erfi}(x) \left(x^2 - \frac{1}{2}\right), \tag{5.143}$$

$$\lambda = 6, \quad n = 3, \qquad \hat{H}_3(x) \quad = \quad e^{x^2} \left(1 - x^2\right) + \sqrt{\pi x^2}\ \text{erfi}(x) \left(x^2 - \frac{3}{2}\right). \tag{5.144}$$

The first few eigenfunctions of the Hermite equation, Eq. (5.123), for the two families of complementary functions are plotted in Fig. 5.4.

5.1.4.2 Probabilists'

The probabilists' Hermite equation is

$$\frac{d^2 y}{dx^2} - x \frac{dy}{dx} + \lambda y = 0. \tag{5.145}$$

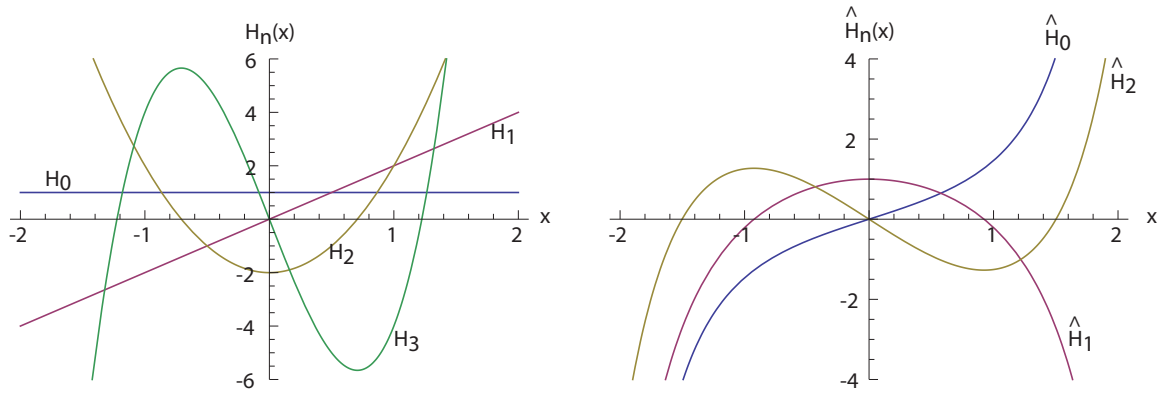

Figure 5.4: Solutions to the physicists' Hermite equation, Eq. (5.123), in terms of two sets of complementary functions $H_n(x)$ and $\hat{H}_n(x)$.

We find that

$$p(x) = e^{-x^2/2}, \tag{5.146}$$
$$r(x) = e^{-x^2/2}, \tag{5.147}$$
$$q(x) = 0. \tag{5.148}$$

Thus, we allow $x \in (-\infty, \infty)$. In Sturm-Liouville form, Eq. (5.145) becomes

$$\frac{d}{dx}\left(e^{-x^2/2}\,\frac{dy}{dx}\right) + \lambda e^{-x^2/2}\,y = 0, \tag{5.149}$$
$$\underbrace{e^{x^2/2}\frac{d}{dx}\left(e^{-x^2/2}\,\frac{d}{dx}\right)}_{\mathbf{L}_s}\,y(x) = -\lambda\,y(x). \tag{5.150}$$

So

$$\mathbf{L}_s = e^{x^2/2}\frac{d}{dx}\left(e^{-x^2/2}\,\frac{d}{dx}\right). \tag{5.151}$$

One set of complementary functions can be expressed in terms of polynomials known as the probabilists' Hermite polynomials, $He_n(x)$. These polynomials can be obtained by a regular series expansion of the original differential equation. The eigenvalues and eigenfunctions corresponding to the probabilists' Hermite polynomials are listed next:

$$\lambda = 0, \qquad He_0(x) = 1, \tag{5.152}$$
$$\lambda = 1, \qquad He_1(x) = x, \tag{5.153}$$
$$\lambda = 2, \qquad He_2(x) = -1 + x^2, \tag{5.154}$$
$$\lambda = 3, \qquad He_3(x) = -3x + x^3, \tag{5.155}$$

$$\lambda = 4, \qquad He_4(x) = 3 - 6x^2 + x^4, \tag{5.156}$$

$$\vdots \tag{5.157}$$

$$\lambda = n, \qquad He_n(x) = (-1)^n e^{x^2/2} \frac{d^n e^{-x^2/2}}{dx^n}, \qquad \text{Rodrigues' formula.} \tag{5.158}$$

The orthogonality condition is

$$\int_{-\infty}^{\infty} e^{-x^2/2} He_n(x) He_m(x)\, dx = n! \sqrt{2\pi} \delta_{nm} \tag{5.159}$$

Direct substitution shows that $He_n(x)$ satisfies both the differential equation, Eq. (5.145), and the orthogonality condition. It is then easily shown that the following functions are orthonormal on the interval $x \in (-\infty, \infty)$:

$$\varphi_n(x) = \frac{e^{-x^2/4} He_n(x)}{\sqrt{\sqrt{2\pi} n!}}, \tag{5.160}$$

giving

$$\int_{-\infty}^{\infty} \varphi_n(x) \varphi_m(x) dx = \delta_{mn}. \tag{5.161}$$

Plots and the second set of complementary functions for the probabilists' Hermite equation are obtained in a similar manner to those for the physicists'. One can easily show the relation between the two to be

$$He_n(x) = 2^{-n/2} H_n\left(\frac{x}{\sqrt{2}}\right). \tag{5.162}$$

5.1.5 Laguerre equation

The Laguerre[9] equation is

$$x \frac{d^2y}{dx^2} + (1 - x) \frac{dy}{dx} + \lambda y = 0. \tag{5.163}$$

We find that

$$p(x) = xe^{-x}, \tag{5.164}$$

$$r(x) = e^{-x}, \tag{5.165}$$

$$q(x) = 0. \tag{5.166}$$

Thus, we require $x \in (0, \infty)$.

[9]Edmond Nicolas Laguerre, 1834-1886, French mathematician.

In Sturm-Liouville form, Eq. (5.163) becomes

$$\frac{d}{dx}\left(xe^{-x}\,\frac{dy}{dx}\right)+\lambda e^{-x}\,y \;=\; 0, \tag{5.167}$$

$$\underbrace{e^{x}\frac{d}{dx}\left(xe^{-x}\,\frac{d}{dx}\right)}_{\mathbf{L}_s}\,y(x) \;=\; -\lambda\,y(x). \tag{5.168}$$

So

$$\mathbf{L}_s = e^{x}\frac{d}{dx}\left(xe^{-x}\,\frac{d}{dx}\right). \tag{5.169}$$

One set of the complementary functions can be expressed in terms of polynomials of finite order known as the Laguerre polynomials, $L_n(x)$. These polynomials can be obtained by a regular series expansion of Eq. (5.163). Eigenvalues and eigenfunctions corresponding to the Laguerre polynomials are listed next:

$$\lambda=0, \qquad L_0(x) \;=\; 1, \tag{5.170}$$

$$\lambda=1, \qquad L_1(x) \;=\; 1-x, \tag{5.171}$$

$$\lambda=2, \qquad L_2(x) \;=\; 1-2x+\frac{1}{2}x^2, \tag{5.172}$$

$$\lambda=3, \qquad L_3(x) \;=\; 1-3x+\frac{3}{2}x^2-\frac{1}{6}x^3, \tag{5.173}$$

$$\lambda=4, \qquad L_4(x) \;=\; 1-4x+3x^2-\frac{2}{3}x^3+\frac{1}{24}x^4, \tag{5.174}$$

$$\vdots \tag{5.175}$$

$$\lambda=n, \qquad L_n(x) \;=\; \frac{1}{n!}e^{x}\frac{d^n\left(x^n e^{-x}\right)}{dx^n}, \qquad \text{Rodrigues' formula.} \tag{5.176}$$

The orthogonality condition reduces to

$$\int_0^\infty e^{-x}L_n(x)L_m(x)\,dx \;=\; \delta_{nm}. \tag{5.177}$$

Direct substitution shows that $L_n(x)$ satisfies both the differential equation, Eq. (5.163), and the orthogonality condition. It is then easily shown that the following functions are orthonormal on the interval $x\in(0,\infty)$:

$$\varphi_n(x)=e^{-x/2}L_n(x), \tag{5.178}$$

so that

$$\int_0^\infty \varphi_n(x)\varphi_m(x)dx=\delta_{mn}. \tag{5.179}$$

The general solution to Eq. (5.163) is

$$y(x) = C_1 L_n(x) + C_2 \hat{L}_n(x), \tag{5.180}$$

where the other set of complementary functions is $\hat{L}_n(x)$. For general n, $\hat{L}_n(x) = U(-n, 1, x)$, one of the so-called Tricomi confluent hypergeometric functions. Again the literature is not extensive on these functions. For integer eigenvalues n, $\hat{L}_n(x)$ reduces somewhat and can be expressed in terms of the exponential integral function, $\mathrm{Ei}(x)$, see Sec. 10.7.6. The first few of these functions are

$$\lambda = n = 0, \qquad \hat{L}_0(x) = \mathrm{Ei}(x), \tag{5.181}$$

$$\lambda = n = 1, \qquad \hat{L}_1(x) = -e^x - \mathrm{Ei}(x)(1-x), \tag{5.182}$$

$$\lambda = n = 2, \qquad \hat{L}_2(x) = \frac{1}{4}\left(e^x(3-x) + \mathrm{Ei}(x)\left(2 - 4x + x^2\right)\right), \tag{5.183}$$

$$\lambda = n = 3, \qquad \hat{L}_3(x) = \frac{1}{36}\left(e^x\left(-11 + 8x - x^2\right) + \mathrm{Ei}(x)\left(-6 + 18x - 9x^2 + x^3\right)\right), \tag{5.184}$$

The first few eigenfunctions of the Laguerre equation, Eq. (5.163), for the two families of complementary functions are plotted in Fig. 5.5.

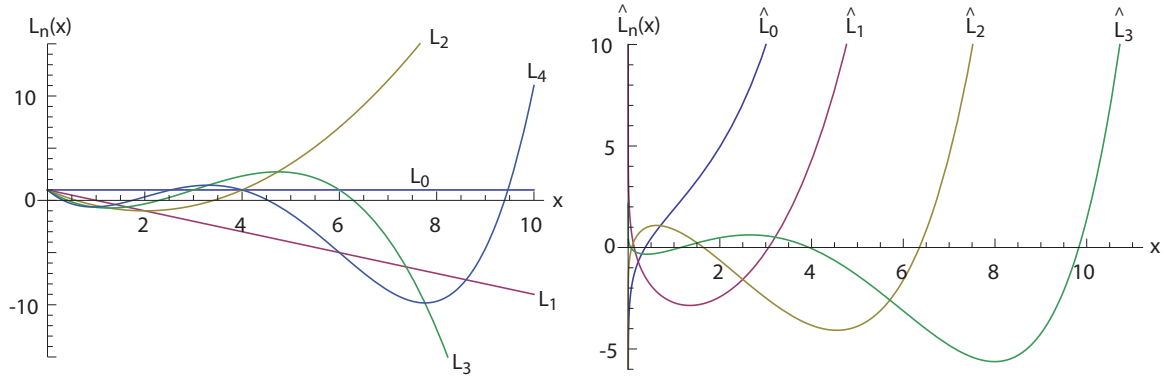

Figure 5.5: Solutions to the Laguerre equation, Eq. (5.163), in terms of two sets of complementary functions, $L_n(x)$ and $\hat{L}_n(x)$.

5.1.6 Bessel's differential equation

5.1.6.1 First and second kind

Bessel's[10] differential equation is as follows, with it being convenient to define $\lambda = -\nu^2$.

$$x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + (\mu^2 x^2 - \nu^2)y = 0. \tag{5.185}$$

[10]Friedrich Wilhelm Bessel, 1784-1846, Westphalia-born German mathematician.

We find that

$$p(x) = x, \tag{5.186}$$
$$r(x) = \frac{1}{x}, \tag{5.187}$$
$$q(x) = \mu^2 x. \tag{5.188}$$

We thus require $x \in (0, \infty)$, though in practice, it is more common to employ a finite domain such as $x \in (0, \ell)$. In Sturm-Liouville form, we have

$$\frac{d}{dx}\left(x\,\frac{dy}{dx}\right) + \left(\mu^2 x - \frac{\nu^2}{x}\right)\,y = 0, \tag{5.189}$$
$$\underbrace{\left(x\left(\frac{d}{dx}\left(x\,\frac{d}{dx}\right) + \mu^2 x\right)\right)}_{\mathbf{L}_s}\,y(x) = \nu^2\,y(x). \tag{5.190}$$

The Sturm-Liouville operator is

$$\mathbf{L}_s = x\left(\frac{d}{dx}\left(x\,\frac{d}{dx}\right) + \mu^2 x\right). \tag{5.191}$$

In some other cases it is more convenient to take $\lambda = \mu^2$ in which case we get

$$p(x) = x, \tag{5.192}$$
$$r(x) = x, \tag{5.193}$$
$$q(x) = -\frac{\nu^2}{x}, \tag{5.194}$$

and the Sturm-Liouville form and operator are:

$$\underbrace{\left(\frac{1}{x}\left(\frac{d}{dx}\left(x\,\frac{d}{dx}\right) - \frac{\nu^2}{x}\right)\right)}_{\mathbf{L}_s}\,y(x) = -\mu^2\,y(x), \tag{5.195}$$
$$\mathbf{L}_s = \frac{1}{x}\left(\frac{d}{dx}\left(x\,\frac{d}{dx}\right) - \frac{\nu^2}{x}\right). \tag{5.196}$$

The general solution is

$$y(x) = C_1 J_\nu(\mu x) + C_2 Y_\nu(\mu x), \qquad \text{if } \nu \text{ is an integer}, \tag{5.197}$$
$$y(x) = C_1 J_\nu(\mu x) + C_2 J_{-\nu}(\mu x), \qquad \text{if } \nu \text{ is not an integer}, \tag{5.198}$$

where $J_\nu(\mu x)$ and $Y_\nu(\mu x)$ are called the Bessel and Neumann functions of order ν. Often $J_\nu(\mu x)$ is known as a *Bessel function of the first kind* and $Y_\nu(\mu x)$ is known as a *Bessel*

function of the second kind. Both J_ν and Y_ν are represented by infinite series rather than finite series such as the series for Legendre polynomials.

The Bessel function of the first kind of order ν, $J_\nu(\mu x)$, is represented by

$$J_\nu(\mu x) = \left(\frac{1}{2}\mu x\right)^\nu \sum_{k=0}^{\infty} \frac{\left(-\frac{1}{4}\mu^2 x^2\right)^k}{k!\Gamma(\nu+k+1)}. \tag{5.199}$$

The Neumann function $Y_\nu(\mu x)$ has a complicated series representation (see Hildebrand).

The representations for $J_0(\mu x)$ and $Y_0(\mu x)$ are

$$J_0(\mu x) = 1 - \frac{\left(\frac{1}{4}\mu^2 x^2\right)^1}{(1!)^2} + \frac{\left(\frac{1}{4}\mu^2 x^2\right)^2}{(2!)^2} + \ldots + \frac{\left(-\frac{1}{4}\mu^2 x^2\right)^n}{(n!)^2}, \tag{5.200}$$

$$Y_0(\mu x) = \frac{2}{\pi}\left(\ln\left(\frac{1}{2}\mu x\right) + \gamma\right) J_0(\mu x) \tag{5.201}$$

$$+\frac{2}{\pi}\left(\frac{\left(\frac{1}{4}\mu^2 x^2\right)^1}{(1!)^2} - \left(1+\frac{1}{2}\right)\frac{\left(\frac{1}{4}\mu^2 x^2\right)^2}{(2!)^2}\ldots\right). \tag{5.202}$$

It can be shown using term by term differentiation that

$$\frac{dJ_\nu(\mu x)}{dx} = \mu\frac{J_{\nu-1}(\mu x) - J_{\nu+1}(\mu x)}{2}, \qquad \frac{dY_\nu(\mu x)}{dx} = \mu\frac{Y_{\nu-1}(\mu x) - Y_{\nu+1}(\mu x)}{2}, \tag{5.203}$$

$$\frac{d}{dx}\left(x^\nu J_\nu\left(\mu x\right)\right) = \mu x^\nu J_{\nu-1}\left(\mu x\right), \qquad \frac{d}{dx}\left(x^\nu Y_\nu\left(\mu x\right)\right) = \mu x^\nu Y_{\nu-1}\left(\mu x\right). \tag{5.204}$$

The Bessel functions $J_0(\mu_0 x)$, $J_0(\mu_1 x)$, $J_0(\mu_2 x)$, $J_0(\mu_3 x)$ are plotted in Fig. 5.6. Here the eigenvalues μ_n can be determined from trial and error. The first four are found to be $\mu_0 = 2.40483$, $\mu_1 = 5.52008$, $\mu_2 = 8.65373$, and $\mu_3 = 11.7915$. In general, one can say

$$\lim_{n\to\infty} \mu_n = n\pi + O(1). \tag{5.205}$$

The Bessel functions $J_0(x), J_1(x), J_2(x), J_3(x)$, and $J_4(x)$ along with the Neumann functions $Y_0(x), Y_1(x), Y_2(x), Y_3(x)$, and $Y_4(x)$ are plotted in Fig. 5.7 (so here $\mu = 1$).

The orthogonality condition for a domain $x \in (0,1)$, taken here for the case in which the eigenvalue is μ_n, can be shown to be

$$\int_0^1 x J_\nu(\mu_n x) J_\nu(\mu_m x)\, dx = \frac{1}{2}\left(J_{\nu+1}(\mu_n)\right)^2 \delta_{nm}. \tag{5.206}$$

Here we must choose μ_n such that $J_\nu(\mu_n) = 0$, which corresponds to a vanishing of the function at the outer limit $x = 1$; see Hildebrand, p. 226. So the orthonormal Bessel function is

$$\varphi_n(x) = \frac{\sqrt{2x}J_\nu(\mu_n x)}{|J_{\nu+1}(\mu_n)|}. \tag{5.207}$$

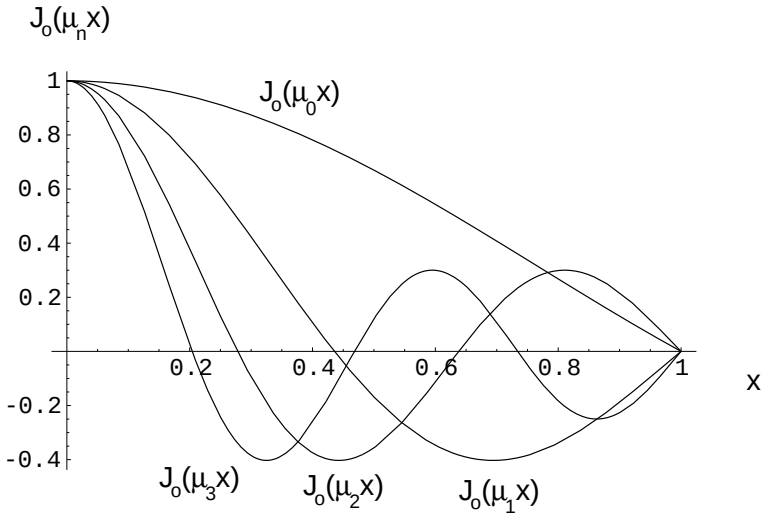

Figure 5.6: Bessel functions $J_0(\mu_0 x)$, $J_0(\mu_1 x)$, $J_0(\mu_2 x)$, $J_0(\mu_3 x)$.

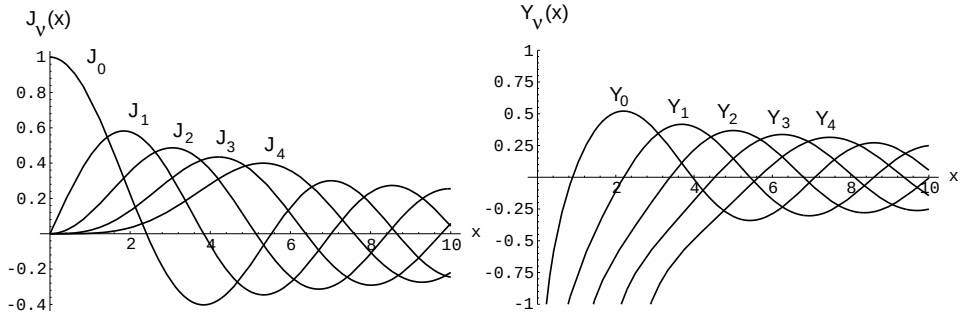

Figure 5.7: Bessel functions $J_0(x)$, $J_1(x)$, $J_2(x)$, $J_3(x)$, $J_4(x)$ and Neumann functions $Y_0(x)$, $Y_1(x)$, $Y_2(x)$, $Y_3(x)$, $Y_4(x)$.

5.1.6.2 Third kind

Hankel[11] functions, also known as *Bessel functions of the third kind* are defined by

$$H_\nu^{(1)}(x) = J_\nu(x) + iY_\nu(x), \tag{5.208}$$
$$H_\nu^{(2)}(x) = J_\nu(x) - iY_\nu(x). \tag{5.209}$$

5.1.6.3 Modified Bessel functions

The modified Bessel equation is

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} - (x^2+\nu^2)y = 0, \tag{5.210}$$

the solutions of which are the modified Bessel functions. The *modified Bessel function of the first kind of order* ν is

$$I_\nu(x) = i^{-\nu}J_\nu(ix). \tag{5.211}$$

The *modified Bessel function of the second kind of order* ν is

$$K_\nu(x) = \frac{\pi}{2}i^{\nu+1}H_n^{(1)}(ix). \tag{5.212}$$

5.1.6.4 Ber and bei functions

The real and imaginary parts of the solutions of

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} - (p^2+ix^2)y = 0, \tag{5.213}$$

where p is a real constant, are called the ber and bei functions.

5.2 Fourier series representation of arbitrary functions

It is often useful, especially when solving partial differential equations, to be able to represent an arbitrary function $f(x)$ in the domain $x \in [x_0, x_1]$ with an appropriately weighted sum of orthonormal functions $\varphi_n(x)$:

$$f(x) = \sum_{n=0}^{\infty} \alpha_n\varphi_n(x). \tag{5.214}$$

We generally truncate the infinite series to a finite number of N terms so that $f(x)$ is approximated by

$$f(x) \simeq \sum_{n=1}^{N} \alpha_n\varphi_n(x). \tag{5.215}$$

[11]Hermann Hankel, 1839-1873, German mathematician.

We can better label an N-term approximation of a function as a *projection* of the function from an infinite dimensional space onto an N-dimensional function space. This will be discussed further in Sec. 7.3.2.6. The projection is useful only if the infinite series converges so that the error incurred in neglecting terms past N is small relative to the terms included.

The problem is to determine what the coefficients α_n must be. They can be found in the following manner. We first assume the expansion exists and multiply both sides by $\varphi_k(x)$:

$$f(x)\varphi_k(x) = \sum_{n=0}^{\infty} \alpha_n \varphi_n(x)\varphi_k(x), \tag{5.216}$$

$$\int_{x_0}^{x_1} f(x)\varphi_k(x)\ dx = \int_{x_0}^{x_1} \sum_{n=0}^{\infty} \alpha_n \varphi_n(x)\varphi_k(x)\ dx, \tag{5.217}$$

$$= \sum_{n=0}^{\infty} \alpha_n \underbrace{\int_{x_0}^{x_1} \varphi_n(x)\varphi_k(x)\ dx}_{\delta_{nk}}, \tag{5.218}$$

$$= \sum_{n=0}^{\infty} \alpha_n \delta_{nk}, \tag{5.219}$$

$$= \alpha_0 \underbrace{\delta_{0k}}_{=0} + \alpha_1 \underbrace{\delta_{1k}}_{=0} + \ldots + \alpha_k \underbrace{\delta_{kk}}_{=1} + \ldots + \alpha_\infty \underbrace{\delta_{\infty k}}_{=0}, \tag{5.220}$$

$$= \alpha_k. \tag{5.221}$$

So trading k and n

$$\alpha_n = \int_{x_0}^{x_1} f(x)\varphi_n(x)\ dx. \tag{5.222}$$

The series is known as a *Fourier series*. Depending on the expansion functions, the series is often specialized as Fourier-sine, Fourier-cosine, Fourier-Legendre, Fourier-Bessel, etc. We have inverted Eq. (5.214) to solve for the unknown α_n. The inversion was aided greatly by the fact that the basis functions were orthonormal. For non-orthonormal, as well as non-orthogonal bases, more general techniques exist for the determination of α_n.

Example 5.2

Represent

$$f(x) = x^2, \qquad \text{on} \qquad x \in [0, 3], \tag{5.223}$$

with a series of

- trigonometric functions,
- Legendre polynomials,
- Chebyshev polynomials, and
- Bessel functions.

Trigonometric Series

For the trigonometric series, let's try a Fourier sine series. The orthonormal functions in this case are, from Eq. (5.54),

$$\varphi_n(x) = \sqrt{\frac{2}{3}} \sin\left(\frac{n\pi x}{3}\right). \tag{5.224}$$

The coefficients from Eq. (5.222) are thus

$$\alpha_n = \int_0^3 \underbrace{x^2}_{f(x)} \underbrace{\left(\sqrt{\frac{2}{3}} \sin\left(\frac{n\pi x}{3}\right)\right)}_{\varphi_n(x)} \, dx, \tag{5.225}$$

so

$$\begin{aligned}
\alpha_0 &= 0, && (5.226)\\
\alpha_1 &= 4.17328, && (5.227)\\
\alpha_2 &= -3.50864, && (5.228)\\
\alpha_3 &= 2.23376, && (5.229)\\
\alpha_4 &= -1.75432, && (5.230)\\
\alpha_5 &= 1.3807. && (5.231)
\end{aligned}$$

Note that the magnitude of the coefficient on the orthonormal function, α_n, decreases as n increases. From this, one can loosely infer that the higher frequency modes contain less "energy."

$$\begin{aligned}
f(x) &= \sqrt{\frac{2}{3}} \left(4.17328 \sin\left(\frac{\pi x}{3}\right) - 3.50864 \sin\left(\frac{2\pi x}{3}\right)\right. && (5.232)\\
&\quad \left. +2.23376 \sin\left(\frac{3\pi x}{3}\right) - 1.75432 \sin\left(\frac{4\pi x}{3}\right) + 1.3807 \sin\left(\frac{5\pi x}{3}\right) + \ldots\right). && (5.233)
\end{aligned}$$

The function $f(x) = x^2$ and five terms of the approximation are plotted in Fig. 5.8.

Legendre polynomials

Next, let's try the Legendre polynomials. The Legendre polynomials are orthogonal on $x \in [-1, 1]$, and we have $x \in [0, 3]$, so let's define

$$\tilde{x} = \frac{2}{3}x - 1, \tag{5.234}$$

$$x = \frac{3}{2}(\tilde{x} + 1), \tag{5.235}$$

so that the domain $x \in [0, 3]$ maps into $\tilde{x} \in [-1, 1]$. So, expanding x^2 on the domain $x \in [0, 3]$ is equivalent to expanding

$$\underbrace{\left(\frac{3}{2}\right)^2 (\tilde{x} + 1)^2}_{x^2} = \frac{9}{4}(\tilde{x} + 1)^2, \qquad \tilde{x} \in [-1, 1]. \tag{5.236}$$

Now from Eq. (5.84),

$$\varphi_n(\tilde{x}) = \sqrt{n + \frac{1}{2}} P_n(\tilde{x}). \tag{5.237}$$

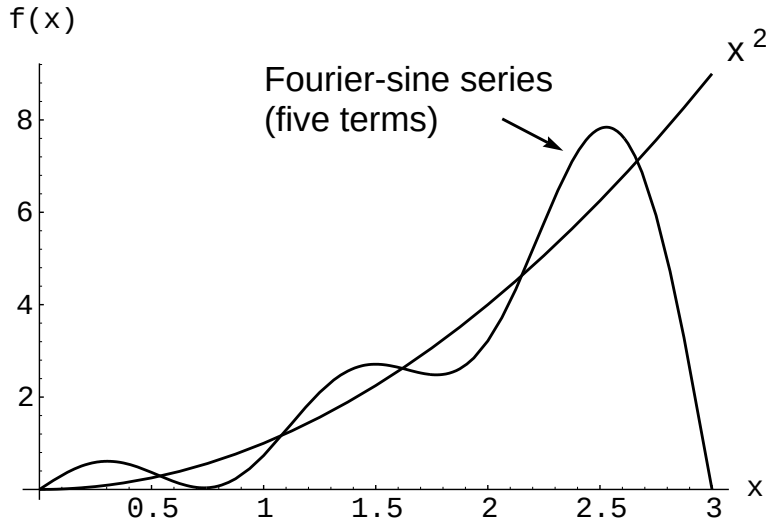

Figure 5.8: Five term Fourier-sine series approximation to $f(x) = x^2$.

So from Eq. (5.222)

$$\alpha_n = \int_{-1}^{1} \underbrace{\left(\frac{9}{4}(\tilde{x}+1)^2\right)}_{f(\tilde{x})} \underbrace{\left(\sqrt{n+\frac{1}{2}}P_n(\tilde{x})\right)}_{\varphi_n(\tilde{x})}\, d\tilde{x}. \tag{5.238}$$

Evaluating, we get

$$\begin{aligned}
\alpha_0 &= 3\sqrt{2} = 4.24264, && (5.239)\\
\alpha_1 &= 3\sqrt{\frac{3}{2}} = 3.67423, && (5.240)\\
\alpha_2 &= \frac{3}{\sqrt{10}} = 0.948683, && (5.241)\\
\alpha_3 &= 0, && (5.242)\\
&\vdots && (5.243)\\
\alpha_n &= 0, \qquad n > 3. && (5.244)
\end{aligned}$$

Once again, the fact the $\alpha_0 > \alpha_1 > \alpha_2$ indicates the bulk of the "energy" is contained in the lower frequency modes. Carrying out the multiplication and returning to x space gives the *finite* series, which can be expressed in a variety of forms:

$$\begin{aligned}
x^2 &= \alpha_0\varphi_0(\tilde{x}) + \alpha_1\varphi_1(\tilde{x}) + \alpha_2\varphi_2(\tilde{x}), && (5.245)\\
&= 3\sqrt{2}\underbrace{\left(\sqrt{\frac{1}{2}}P_0\left(\frac{2}{3}x-1\right)\right)}_{=\varphi_0(\tilde{x})} + 3\sqrt{\frac{3}{2}}\underbrace{\left(\sqrt{\frac{3}{2}}P_1\left(\frac{2}{3}x-1\right)\right)}_{=\varphi_1(\tilde{x})} + \frac{3}{\sqrt{10}}\underbrace{\left(\sqrt{\frac{5}{2}}P_2\left(\frac{2}{3}x-1\right)\right)}_{=\varphi_2(\tilde{x})}, && (5.246)\\
&= 3P_0\left(\frac{2}{3}x-1\right) + \frac{9}{2}P_1\left(\frac{2}{3}x-1\right) + \frac{3}{2}P_2\left(\frac{2}{3}x-1\right), && (5.247)\\
&= 3(1) + \frac{9}{2}\left(\frac{2}{3}x-1\right) + \frac{3}{2}\left(-\frac{1}{2} + \frac{3}{2}\left(\frac{2}{3}x-1\right)^2\right), && (5.248)
\end{aligned}$$

$$= 3 + \left(-\frac{9}{2} + 3x\right) + \left(\frac{3}{2} - 3x + x^2\right), \tag{5.249}$$

$$= x^2. \tag{5.250}$$

Thus, the Fourier-Legendre representation is *exact* over the entire domain. This is because the function which is being expanded has the same general functional form as the Legendre polynomials; both are polynomials.

Chebyshev polynomials

Let's now try the Chebyshev polynomials. These are orthogonal on the same domain as the Legendre polynomials, so let's use the same transformation as before. Now from Eq. (5.113)

$$\varphi_0(\tilde{x}) = \sqrt{\frac{1}{\pi\sqrt{1-\tilde{x}^2}}} T_0(\tilde{x}), \tag{5.251}$$

$$\varphi_n(\tilde{x}) = \sqrt{\frac{2}{\pi\sqrt{1-\tilde{x}^2}}} T_n(\tilde{x}), \qquad n > 0. \tag{5.252}$$

So

$$\alpha_0 = \int_{-1}^{1} \underbrace{\frac{9}{4}(\tilde{x}+1)^2}_{f(\tilde{x})} \underbrace{\sqrt{\frac{1}{\pi\sqrt{1-\tilde{x}^2}}} T_0(\tilde{x})}_{\varphi_0(\tilde{x})} \, d\tilde{x}, \tag{5.253}$$

$$\alpha_n = \int_{-1}^{1} \underbrace{\frac{9}{4}(\tilde{x}+1)^2}_{f(\tilde{x})} \underbrace{\sqrt{\frac{2}{\pi\sqrt{1-\tilde{x}^2}}} T_n(\tilde{x})}_{\varphi_n(\tilde{x})} \, d\tilde{x}. \tag{5.254}$$

Evaluating, we get

$$\alpha_0 = 4.2587, \tag{5.255}$$

$$\alpha_1 = 3.4415, \tag{5.256}$$

$$\alpha_2 = -0.28679, \tag{5.257}$$

$$\alpha_3 = -1.1472, \tag{5.258}$$

$$\vdots$$

With this representation, we see that $|\alpha_3| > |\alpha_2|$, so it is not yet clear that the "energy" is concentrated in the high frequency modes. Consideration of more terms would verify that in fact it is the case that the "energy " of high frequency modes is decaying; in fact $\alpha_4 = -0.683$, $\alpha_5 = -0.441$, $\alpha_6 = -0.328$, $\alpha_7 = -0.254$. So

$$f(x) = x^2 = \sqrt{\frac{2}{\pi\sqrt{1-\left(\frac{2}{3}x-1\right)^2}}} \left(\frac{4.2587}{\sqrt{2}} \, T_0\left(\frac{2}{3}x-1\right) + 3.4415 \, T_1\left(\frac{2}{3}x-1\right) \right. \tag{5.259}$$

$$\left. -0.28679 \, T_2\left(\frac{2}{3}x-1\right) - 1.1472 \, T_3\left(\frac{2}{3}x-1\right) + \ldots \right). \tag{5.260}$$

The function $f(x) = x^2$ and four terms of the approximation are plotted in Fig. 5.9.

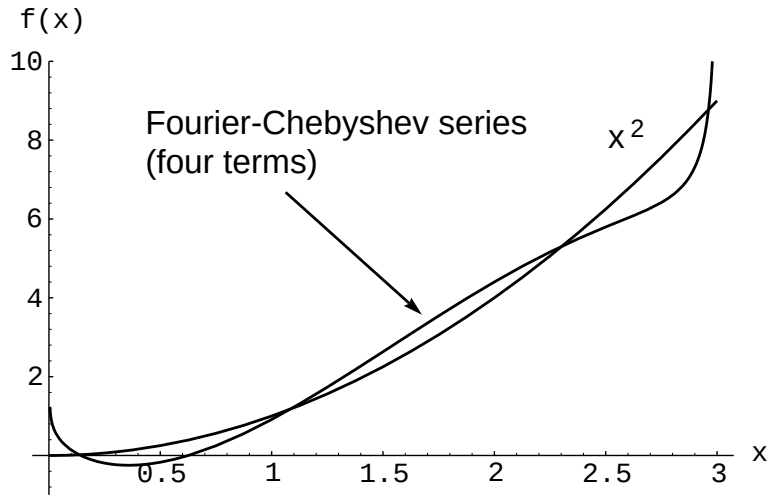

Figure 5.9: Four term Fourier-Chebyshev series approximation to $f(x) = x^2$.

Bessel functions

Now let's expand in terms of Bessel functions. The Bessel functions have been defined such that they are orthogonal on a domain between zero and unity when the eigenvalues are the zeros of the Bessel function. To achieve this we adopt the transformation (and inverse):

$$\tilde{x} = \frac{x}{3}, \qquad x = 3\tilde{x}. \tag{5.261}$$

With this transformation our domain transforms as follows:

$$x \in [0,3] \longrightarrow \tilde{x} \in [0,1]. \tag{5.262}$$

So in the transformed space, we seek an expansion

$$\underbrace{9\tilde{x}^2}_{f(\tilde{x})} = \sum_{n=0}^{\infty} \alpha_n J_\nu(\mu_n \tilde{x}). \tag{5.263}$$

Let's choose to expand on J_0, so we take

$$9\tilde{x}^2 = \sum_{n=0}^{\infty} \alpha_n J_0(\mu_n \tilde{x}). \tag{5.264}$$

Now, the eigenvalues μ_n are such that $J_0(\mu_n) = 0$. We find using trial and error methods that solutions for all the zeros can be found:

$$\mu_0 = 2.4048, \tag{5.265}$$
$$\mu_1 = 5.5201, \tag{5.266}$$
$$\mu_2 = 8.6537, \tag{5.267}$$
$$\vdots$$

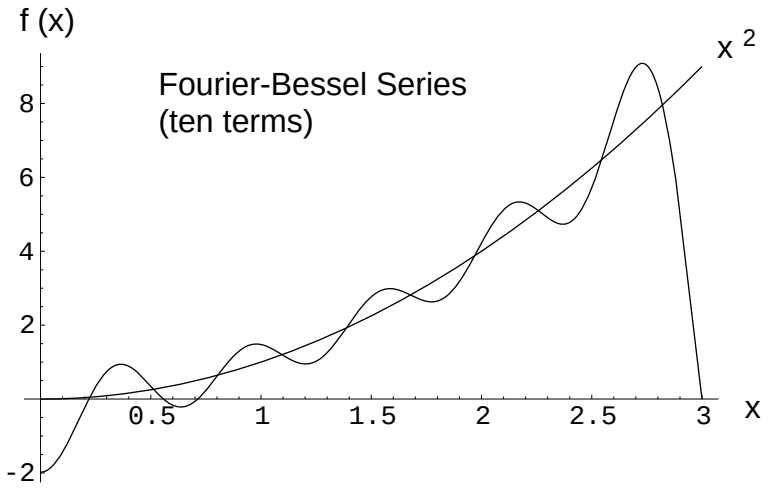

Figure 5.10: Ten term Fourier-Bessel series approximation to $f(x) = x^2$.

Similar to the other functions, we could expand in terms of the orthonormalized Bessel functions, $\varphi_n(x)$. Instead, for variety, let's directly operate on Eq. (5.264) to determine the values for α_n.

$$9\tilde{x}^2\tilde{x}J_0(\mu_k\tilde{x}) = \sum_{n=0}^{\infty}\alpha_n\tilde{x}J_0(\mu_n\tilde{x})J_0(\mu_k\tilde{x}), \tag{5.268}$$

$$\int_0^1 9\tilde{x}^3J_0(\mu_k\tilde{x})\ d\tilde{x} = \int_0^1\sum_{n=0}^{\infty}\alpha_n\tilde{x}J_0(\mu_n\tilde{x})J_0(\mu_k\tilde{x})\ d\tilde{x}, \tag{5.269}$$

$$9\int_0^1 \tilde{x}^3J_0(\mu_k\tilde{x})\ d\tilde{x} = \sum_{n=0}^{\infty}\alpha_n\int_0^1 \tilde{x}J_0(\mu_n\tilde{x})J_0(\mu_k\tilde{x})\ d\tilde{x}, \tag{5.270}$$

$$= \alpha_k\int_0^1 \tilde{x}J_0(\mu_k\tilde{x})J_0(\mu_k\tilde{x})\ d\tilde{x}. \tag{5.271}$$

So replacing k by n and dividing we get

$$\alpha_n = \frac{9\int_0^1 \tilde{x}^3J_0(\mu_n\tilde{x})\ d\tilde{x}}{\int_0^1 \tilde{x}J_0(\mu_n\tilde{x})J_0(\mu_n\tilde{x})\ d\tilde{x}}. \tag{5.272}$$

Evaluating the first three terms we get

$$\alpha_0 = 4.446, \tag{5.273}$$

$$\alpha_1 = -8.325, \tag{5.274}$$

$$\alpha_2 = 7.253, \tag{5.275}$$

$$\vdots$$

Because the basis functions are not normalized, it is difficult to infer how the amplitude is decaying by looking at α_n alone. The function $f(x) = x^2$ and ten terms of the Fourier-Bessel series approximation are plotted in Fig. 5.10 The Fourier-Bessel approximation is

$$f(x) = x^2 = 4.446\ J_0\left(2.4048\left(\frac{x}{3}\right)\right) - 8.325\ J_0\left(5.5201\left(\frac{x}{3}\right)\right) + 7.253\ J_0\left(8.6537\left(\frac{x}{3}\right)\right) + \ldots. \tag{5.276}$$

Note that other Fourier-Bessel expansions exist. Also note that even though the Bessel function does not match the function itself at either boundary point, that the series still appears to be converging.

Problems

1. Show that oscillatory solutions of the delay equation

$$\frac{dx}{dt}(t) + x(t) + bx(t-1) = 0,$$

are possible only when $b = 2.2617$. Find the frequency.

2. Show that $x^a J_\nu(bx^c)$ is a solution of

$$y'' - \frac{2a-1}{x}y' + \left(b^2c^2x^{2c-2} + \frac{a^2 - \nu^2c^2}{x^2}\right)y = 0.$$

Hence solve in terms of Bessel functions:

(a) $\frac{d^2y}{dx^2} + k^2xy = 0$,

(b) $\frac{d^2y}{dx^2} + x^4y = 0$.

3. Laguerre's differential equation is

$$xy'' + (1-x)y' + \lambda y = 0.$$

Show that when $\lambda = n$, a nonnegative integer, there is a polynomial solution $L_n(x)$ (called a Laguerre polynomial) of degree n with coefficient of x^n equal to 1. Determine L_0 through L_4.

4. Consider the function $y(x) = x^2 - 2x + 1$ defined for $x \in [0, 4]$. Find eight term expansions in terms of a) Fourier-Sine, b) Fourier-Legendre, c) Fourier-Hermite (physicists'), d) Fourier-Bessel series and plot your results on a single graph.

5. Consider the function $y(x) = 0, x \in [0, 1)$, $y(x) = 2x - 2, x \in [1, 2]$. Find an eight term Fourier-Legendre expansion of this function. Plot the function and the eight term expansion for $x \in [0, 2]$.

6. Consider the function $y(x) = 2x, x \in [0, 6]$. Find an eight term a) Fourier-Chebyshev and b) Fourier-sine expansion of this function. Plot the function and the eight term expansions for $x \in [0, 6]$. Which expansion minimizes the error in representation of the function?

7. Consider the function $y(x) = \cos^2(x^2)$. Find an eight term a) Fourier-Laguerre, ($x \in [0, \infty)$), and b) Fourier-sine ($x \in [0, 10]$) expansion of this function. Plot the function and the eight term expansions for $x \in [0, 10]$. Which expansion minimizes the error in representation of the function?

Chapter 6

Vectors and tensors

see Kaplan, Chapters 3, 4, 5,
see Lopez, Chapters 17-23,
see Aris,
see Borisenko and Tarapov,
see McConnell,
see Schey,
see Riley, Hobson, and Bence, Chapters 6, 8, 19.

This chapter will outline many topics considered in traditional vector calculus and include an introduction to differential geometry.

6.1 Cartesian index notation

Here we will consider what is known as Cartesian index notation as a way to represent vectors and tensors. In contrast to Sec. 1.3, which considered general coordinate transformations, when we restrict our transformations to rotations about the origin, many simplifications result. For such transformations, the distinction between contravariance and covariance disappears, as does the necessity for Christoffel symbols, and also the need for an "upstairs-downstairs" index notation.

Many vector relations can be written in a compact form by using Cartesian index notation. Let x_1, x_2, x_3 represent the three coordinate directions and $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ the unit vectors in those directions. Then a vector $\mathbf{u}$ may be written as

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3 = \sum_{i=1}^{3} u_i\mathbf{e}_i = u_i\mathbf{e}_i = u_i, \tag{6.1}$$

where u_1, u_2, and u_3 are the three Cartesian components of $\mathbf{u}$. Note that we do not need to use the summation sign every time if we use the Einstein convention to sum from 1 to 3 if

an index is repeated. The single free index on the right side of Eq. (6.1) indicates that an $\mathbf{e}_i$ is assumed.

Two additional symbols are needed for later use. They are the Kronecker delta, as specialized from Eq. (1.63),

$$\delta_{ij} \equiv \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases} \tag{6.2}$$

and the alternating symbol (or Levi-Civita[1] symbol)

$$\epsilon_{ijk} \equiv \begin{cases} 1, & \text{if indices are in cyclical order 1,2,3,1,2,}\cdots, \\ -1, & \text{if indices are not in cyclical order,} \\ 0, & \text{if two or more indices are the same.} \end{cases} \tag{6.3}$$

The identity

$$\epsilon_{ijk}\epsilon_{lmn} = \delta_{il}\delta_{jm}\delta_{kn} + \delta_{im}\delta_{jn}\delta_{kl} + \delta_{in}\delta_{jl}\delta_{km} - \delta_{il}\delta_{jn}\delta_{km} - \delta_{im}\delta_{jl}\delta_{kn} - \delta_{in}\delta_{jm}\delta_{kl}, \tag{6.4}$$

relates the two. The following identities are also easily shown:

$$\begin{aligned} \delta_{ii} &= 3, & (6.5)\\ \delta_{ij} &= \delta_{ji}, & (6.6)\\ \delta_{ij}\delta_{jk} &= \delta_{ik}, & (6.7)\\ \epsilon_{ijk}\epsilon_{ilm} &= \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}, & (6.8)\\ \epsilon_{ijk}\epsilon_{ljk} &= 2\delta_{il}, & (6.9)\\ \epsilon_{ijk}\epsilon_{ijk} &= 6, & (6.10)\\ \epsilon_{ijk} &= -\epsilon_{ikj}, & (6.11)\\ \epsilon_{ijk} &= -\epsilon_{jik}, & (6.12)\\ \epsilon_{ijk} &= -\epsilon_{kji}, & (6.13)\\ \epsilon_{ijk} = \epsilon_{kij} &= \epsilon_{jki}. & (6.14) \end{aligned}$$

Regarding index notation:

- a repeated index indicates summation on that index,
- a non-repeated index is known as a free index,
- the number of free indices give the order of the tensor:
  - $u,\ uv,\ u_iv_iw,\ u_{ii},\ u_{ij}v_{ij}$, zeroth order tensor–scalar,
  - $u_i,\ u_iv_{ij}$, first order tensor–vector,
  - $u_{ij},\ u_{ij}v_{jk},\ u_iv_j$, second order tensor,

[1] Tullio Levi-Civita, 1883-1941, Italian mathematician.

  - $u_{ijk},\ u_i v_j w_k,\ u_{ij} v_{km} w_m$, third order tensor,
  - $u_{ijkl},\ u_{ij} v_{kl}$, fourth order tensor.

- indices cannot be repeated more than once:
  - $u_{iik},\ u_{ij},\ u_{iijj},\ v_i u_{jk}$ are proper.
  - $u_i v_i w_i,\ u_{iiij},\ u_{ij} v_{ii}$ are improper!
- Cartesian components commute: $u_{ij} v_i w_{klm} = v_i w_{klm} u_{ij}$,
- Cartesian indices *do not* commute: $u_{ijkl} \neq u_{jlik}$.

Example 6.1

Let us consider, using generalized coordinates described earlier in Sec. 1.3, a trivial identity transformation from the Cartesian ξ^i coordinates to the transformed coordinates x^i:

$$x^1 = \xi^1, \qquad x^2 = \xi^2, \qquad x^3 = \xi^3. \tag{6.15}$$

Here, we are returning to the more general "upstairs-downstairs" index notation of Sec. 1.3. Recalling Eq. (1.78), the Jacobian of the transformation is

$$\mathbf{J} = \frac{\partial \xi^i}{\partial x^j} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \delta^i_j = \mathbf{I}. \tag{6.16}$$

From Eq. (1.85), the metric tensor then is

$$g_{ij} = \mathbf{G} = \mathbf{J}^T \cdot \mathbf{J} = \mathbf{I} \cdot \mathbf{I} = \mathbf{I} = \delta_{ij}. \tag{6.17}$$

Then we find by the transformation rules that for this transformation, the covariant and contravariant representations of a general vector $\mathbf{u}$ are one and the same:

$$u_i = g_{ij} u^j = \delta_{ij} u^j = \delta^i_j u^j = u^i. \tag{6.18}$$

Consequently, for Cartesian vectors, there is no need to use a notation which distinguishes covariant and contravariant representations. We will hereafter write all Cartesian vectors with only a subscript notation.

6.2 Cartesian tensors

6.2.1 Direction cosines

Consider the alias transformation of the (x_1, x_2) Cartesian coordinate system by rotation of each coordinate axes by angle α to the rotated Cartesian coordinate system $\overline{x}_1, \overline{x}_2$ as sketched

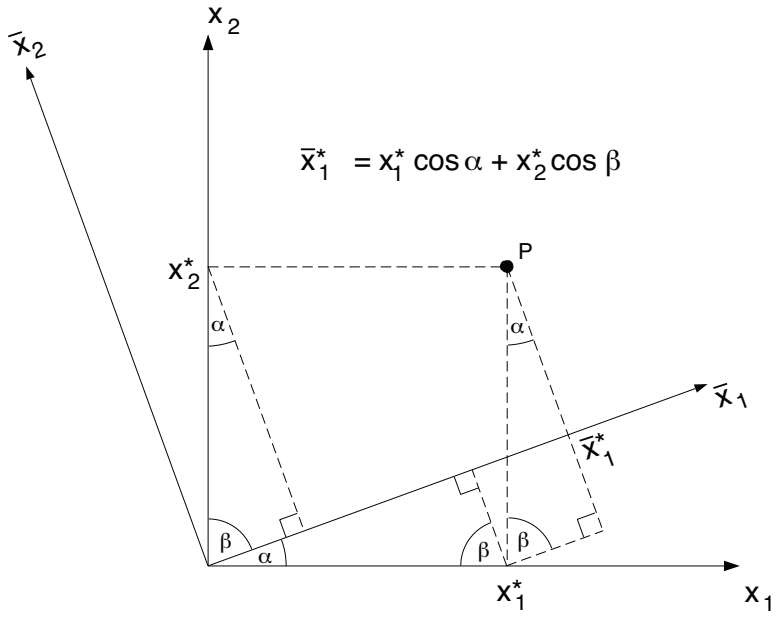

Figure 6.1: Rotation of axes in a two-dimensional Cartesian system.

in Fig. 6.1. Relative to our earlier notation for general non-Cartesian systems, Sec. 1.3, in this chapter, x plays the role of the earlier ξ, and $\overline{x}$ plays the role of the earlier x. We define the angle between the x_1 and $\overline{x}_1$ axes as α:

$$\alpha \equiv [x_1, \overline{x}_1]. \tag{6.19}$$

With $\beta = \pi/2 - \alpha$, the angle between the $\overline{x}_1$ and x_2 axes is

$$\beta \equiv [x_2, \overline{x}_1]. \tag{6.20}$$

The point P can be represented in both coordinate systems. In the unrotated system, P is represented by the coordinates:

$$P : (x_1^*, x_2^*). \tag{6.21}$$

In the rotated coordinate system, P is represented by

$$P : (\overline{x}_1^*, \overline{x}_2^*). \tag{6.22}$$

Trigonometry shows us that

$$\overline{x}_1^* = x_1^* \cos\alpha + x_2^* \cos\beta, \tag{6.23}$$

$$\overline{x}_1^* = x_1^* \cos[x_1, \overline{x}_1] + x_2^* \cos[x_2, \overline{x}_1]. \tag{6.24}$$

Dropping the stars, and extending to three dimensions, we find that

$$\overline{x}_1 = x_1 \cos[x_1, \overline{x}_1] + x_2 \cos[x_2, \overline{x}_1] + x_3 \cos[x_3, \overline{x}_1]. \tag{6.25}$$

Extending to expressions for $\overline{x}_2$ and $\overline{x}_3$ and writing in matrix form, we get

$$\underbrace{\begin{pmatrix}\overline{x}_1 & \overline{x}_2 & \overline{x}_3\end{pmatrix}}_{=\overline{x}_j=\overline{\mathbf{x}}^T} = \underbrace{\begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}}_{=x_i=\mathbf{x}^T} \cdot \underbrace{\begin{pmatrix}\cos[x_1,\overline{x}_1] & \cos[x_1,\overline{x}_2] & \cos[x_1,\overline{x}_3]\\ \cos[x_2,\overline{x}_1] & \cos[x_2,\overline{x}_2] & \cos[x_2,\overline{x}_3]\\ \cos[x_3,\overline{x}_1] & \cos[x_3,\overline{x}_2] & \cos[x_3,\overline{x}_3]\end{pmatrix}}_{=\ell_{ij}=\mathbf{Q}}. \tag{6.26}$$

Using the notation

$$\ell_{ij} = \cos[x_i, \overline{x}_j], \tag{6.27}$$

Eq. (6.26) is written as

$$\underbrace{\begin{pmatrix}\overline{x}_1 & \overline{x}_2 & \overline{x}_3\end{pmatrix}}_{=\overline{x}_j=\overline{\mathbf{x}}^T} = \underbrace{\begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}}_{=x_i=\mathbf{x}^T} \cdot \underbrace{\begin{pmatrix}\ell_{11} & \ell_{12} & \ell_{13}\\ \ell_{21} & \ell_{22} & \ell_{23}\\ \ell_{31} & \ell_{32} & \ell_{33}\end{pmatrix}}_{=\mathbf{Q}}. \tag{6.28}$$

Here ℓ_{ij} are known as the *direction cosines*. Expanding the first term we find

$$\overline{x}_1 = x_1\ell_{11} + x_2\ell_{21} + x_3\ell_{31}. \tag{6.29}$$

More generally, we have

$$\begin{aligned}\overline{x}_j &= x_1\ell_{1j} + x_2\ell_{2j} + x_3\ell_{3j}, &(6.30)\\ &= \sum_{i=1}^{3} x_i\ell_{ij}, &(6.31)\\ &= x_i\ell_{ij}. &(6.32)\end{aligned}$$

Here we have employed Einstein's convention that repeated indices implies a summation over that index.

What amounts to the *law of cosines*,

$$\ell_{ij}\ell_{kj} = \delta_{ik}, \tag{6.33}$$

can easily be proven by direct substitution. Direction cosine matrices applied to geometric entities such as polygons have the property of being *volume- and orientation-preserving* because $\det \ell_{ij} = 1$. General volume-preserving transformations have determinant of ± 1. For right-handed coordinate systems, transformations which have positive determinants are orientation-preserving, and those which have negative determinants are orientation-reversing. Transformations which are volume-preserving but orientation-reversing have determinant of -1, and involve a reflection.

Example 6.2

Show for the two-dimensional system described in Fig. 6.1 that $\ell_{ij}\ell_{kj} = \delta_{ik}$ holds.

Expanding for the two-dimensional system, we get

$$\ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} = \delta_{ik}. \tag{6.34}$$

First, take $i = 1, k = 1$. We get then

$$\begin{aligned} \ell_{11}\ell_{11} + \ell_{12}\ell_{12} = \delta_{11} &= 1, & (6.35)\\ \cos\alpha\cos\alpha + \cos(\alpha+\pi/2)\cos(\alpha+\pi/2) &= 1, & (6.36)\\ \cos\alpha\cos\alpha + (-\sin(\alpha))(-\sin(\alpha)) &= 1, & (6.37)\\ \cos^2\alpha + \sin^2\alpha &= 1. & (6.38) \end{aligned}$$

This is obviously true. Next, take $i = 1, k = 2$. We get then

$$\begin{aligned} \ell_{11}\ell_{21} + \ell_{12}\ell_{22} = \delta_{12} &= 0, & (6.39)\\ \cos\alpha\cos(\pi/2-\alpha) + \cos(\alpha+\pi/2)\cos(\alpha) &= 0, & (6.40)\\ \cos\alpha\sin\alpha - \sin\alpha\cos\alpha &= 0. & (6.41) \end{aligned}$$

This is obviously true. Next, take $i = 2, k = 1$. We get then

$$\begin{aligned} \ell_{21}\ell_{11} + \ell_{22}\ell_{12} = \delta_{21} &= 0, & (6.42)\\ \cos(\pi/2-\alpha)\cos\alpha + \cos\alpha\cos(\pi/2+\alpha) &= 0, & (6.43)\\ \sin\alpha\cos\alpha + \cos\alpha(-\sin\alpha) &= 0. & (6.44) \end{aligned}$$

This is obviously true. Next, take $i = 2, k = 2$. We get then

$$\begin{aligned} \ell_{21}\ell_{21} + \ell_{22}\ell_{22} = \delta_{22} &= 1, & (6.45)\\ \cos(\pi/2-\alpha)\cos(\pi/2-\alpha) + \cos\alpha\cos\alpha &= 1, & (6.46)\\ \sin\alpha\sin\alpha + \cos\alpha\cos\alpha &= 1. & (6.47) \end{aligned}$$

Again, this is obviously true.

Using the law of cosines, Eq. (6.33), we can easily find the inverse transformation back to the unprimed coordinates via the following operations. First operate on Eq. (6.32) with ℓ_{kj}.

$$\begin{aligned} \ell_{kj}\overline{x}_j &= \ell_{kj}x_i\ell_{ij}, & (6.48)\\ &= \ell_{ij}\ell_{kj}x_i, & (6.49)\\ &= \delta_{ik}x_i, & (6.50)\\ &= x_k, & (6.51)\\ \ell_{ij}\overline{x}_j &= x_i, & (6.52)\\ x_i &= \ell_{ij}\overline{x}_j. & (6.53) \end{aligned}$$

Note that the Jacobian matrix of the transformation is $\mathbf{J} = \partial x_i/\partial\overline{x}_j = \ell_{ij}$. It can be shown that the metric tensor is $\mathbf{G} = \mathbf{J}^T \cdot \mathbf{J} = \ell_{ji}\ell_{ki} = \delta_{jk} = \mathbf{I}$, so $g = 1$, and the transformation is volume-preserving. Moreover, since $\mathbf{J}^T \cdot \mathbf{J} = \mathbf{I}$, we see that $\mathbf{J}^T = \mathbf{J}^{-1}$. As such, it is

precisely the type of matrix for which the gradient takes on the same form in original and transformed coordinates, as presented in the discussion surrounding Eq. (1.95). As will be discussed in detail in Sec. 8.6, matrices which have these properties are known as *orthogonal* are often denoted by $\mathbf{Q}$. So for this class of transformations, $\mathbf{J} = \mathbf{Q} = \partial x_i/\partial \overline{x}_j = \ell_{ij}$. Note that $\mathbf{Q}^T \cdot \mathbf{Q} = \mathbf{I}$ and that $\mathbf{Q}^T = \mathbf{Q}^{-1}$. The matrix $\mathbf{Q}$ is a rotation matrix when its elements are composed of the direction cosines ℓ_{ij}. Note then that $\mathbf{Q}^T = \ell_{ji}$. For a coordinate system which obeys the right-hand rule, we require $\det \mathbf{Q} = 1$ so that it is also orientation-preserving.

Example 6.3

Consider the previous two-dimensional example of a matrix which rotates a vector through an angle α using matrix methods.

We have

$$\mathbf{J} = \frac{\partial x_i}{\partial \overline{x}_j} = \ell_{ij} = \mathbf{Q} = \begin{pmatrix} \cos\alpha & \cos\left(\alpha + \frac{\pi}{2}\right) \\ \cos\left(\frac{\pi}{2} - \alpha\right) & \cos\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \tag{6.54}$$

We get the rotated coordinates via Eq. (6.26):

$$\begin{aligned} \overline{\mathbf{x}}^T &= \mathbf{x}^T \cdot \mathbf{Q}, && (6.55)\\ \begin{pmatrix} \overline{x}_1 & \overline{x}_2 \end{pmatrix} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \cdot \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, && (6.56)\\ &= \begin{pmatrix} x_1\cos\alpha + x_2\sin\alpha & -x_1\sin\alpha + x_2\cos\alpha \end{pmatrix}, && (6.57)\\ \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \end{pmatrix} &= \begin{pmatrix} x_1\cos\alpha + x_2\sin\alpha \\ -x_1\sin\alpha + x_2\cos\alpha \end{pmatrix}. && (6.58) \end{aligned}$$

We can also rearrange to say

$$\begin{aligned} \overline{\mathbf{x}} &= \mathbf{Q}^T \cdot \mathbf{x}, && (6.59)\\ \mathbf{Q} \cdot \overline{\mathbf{x}} &= \underbrace{\mathbf{Q} \cdot \mathbf{Q}^T}_{\mathbf{I}} \cdot \mathbf{x}, && (6.60)\\ \mathbf{Q} \cdot \overline{\mathbf{x}} &= \mathbf{I} \cdot \mathbf{x}, && (6.61)\\ \mathbf{x} &= \mathbf{Q} \cdot \overline{\mathbf{x}}. && (6.62) \end{aligned}$$

The law of cosines holds because

$$\begin{aligned} \mathbf{Q} \cdot \mathbf{Q}^T &= \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, && (6.63)\\ &= \begin{pmatrix} \cos^2\alpha + \sin^2\alpha & 0 \\ 0 & \sin^2\alpha + \cos^2\alpha \end{pmatrix}, && (6.64)\\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, && (6.65)\\ &= \mathbf{I} = \delta_{ij}. && (6.66) \end{aligned}$$

Consider the determinant of $\mathbf{Q}$:

$$\det \mathbf{Q} = \cos^2\alpha - (-\sin^2\alpha) = \cos^2\alpha + \sin^2\alpha = 1. \tag{6.67}$$

Thus, the transformation is volume- and orientation-preserving; hence, it is a rotation. The rotation is through an angle α.

Example 6.4

Consider the so-called *reflection matrix* in two dimensions:

$$\mathbf{Q} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}. \tag{6.68}$$

Note the reflection matrix is obtained by multiplying the second column of the rotation matrix of Eq. (6.54) by -1. We see that

$$\begin{aligned} \mathbf{Q}\cdot\mathbf{Q}^T &= \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix} \cdot \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}, & (6.69)\\ &= \begin{pmatrix} \cos^2\alpha + \sin^2\alpha & 0 \\ 0 & \sin^2\alpha + \cos^2\alpha \end{pmatrix}, & (6.70)\\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I} = \delta_{ij}. & (6.71) \end{aligned}$$

The determinant of the reflection matrix is

$$\det\mathbf{Q} = -\cos^2\alpha - \sin^2\alpha = -1. \tag{6.72}$$

Thus, the transformation is volume-preserving, but not orientation-preserving. One can show by considering its action on vectors $\mathbf{x}$ is that it reflects them about a line passing through the origin inclined at an angle of $\alpha/2$ to the horizontal.

6.2.1.1 Scalars

An entity ϕ is a scalar if it is invariant under a rotation of coordinate axes.

6.2.1.2 Vectors

A set of three scalars $(v_1, v_2, v_3)^T$ is defined as a *vector* if under a rotation of coordinate axes, the triple also transforms according to

$$\overline{v}_j = v_i \ell_{ij}, \qquad \overline{\mathbf{v}}^T = \mathbf{v}^T \cdot \mathbf{Q}. \tag{6.73}$$

We could also transpose both sides and have

$$\overline{\mathbf{v}} = \mathbf{Q}^T \cdot \mathbf{v}. \tag{6.74}$$

A vector associates a scalar with a chosen direction in space by an expression which is linear in the direction cosines of the chosen direction.

Example 6.5

Returning to generalized coordinate notation, show the equivalence between covariant and contravariant representations for pure rotations of a vector $\mathbf{v}$.

Consider then a transformation from a Cartesian space ξ^j to a transformed space x^i via a pure rotation:

$$\xi^i = \ell^i_j x^j. \tag{6.75}$$

Here ℓ^i_j is simply a matrix of direction cosines as we have previously defined; we employ the upstairs-downstairs index notation for consistency. The Jacobian is

$$\frac{\partial \xi^i}{\partial x^j} = \ell^i_j. \tag{6.76}$$

From Eq. (1.85), the metric tensor is

$$g_{kl} = \frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^i}{\partial x^l} = \ell^i_k \ell^i_l = \delta_{kl}. \tag{6.77}$$

Here we have employed the law of cosines, which is easily extensible to the "upstairs-downstairs" notation.

So a vector $\mathbf{v}$ has the same covariant and contravariant components since

$$v_i = g_{ij} v^j = \delta_{ij} v^j = \delta^i_j v^j = v^i. \tag{6.78}$$

Note the vector itself has components that do transform under rotation:

$$v^i = \ell^i_j V^j. \tag{6.79}$$

Here V^j is the contravariant representation of the vector $\mathbf{v}$ in the unrotated coordinate system. One could also show that $V_j = V^j$, as always for a Cartesian system.

6.2.1.3 Tensors

A set of nine scalars is defined as a second order *tensor* if under a rotation of coordinate axes, they transform as

$$\overline{T}_{ij} = \ell_{ki}\ell_{lj}T_{kl}, \qquad \overline{\mathbf{T}} = \mathbf{Q}^T \cdot \mathbf{T} \cdot \mathbf{Q}. \tag{6.80}$$

A tensor associates a vector with each direction in space by an expression that is linear in the direction cosines of the chosen transformation. It will be seen that

- the first subscript gives associated direction (or face; hence first–face), and
- the second subscript gives the vector components for that face.

Graphically, one can use the sketch in Fig. 6.2 to visualize a second order tensor. In Fig. 6.2, $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)}$, and $\mathbf{q}^{(3)}$, are the vectors associated with the 1, 2, and 3 faces, respectively.

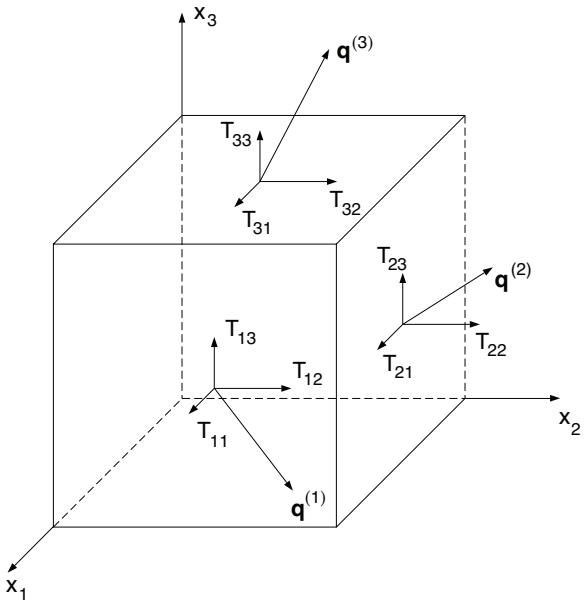

Figure 6.2: Tensor visualization.

6.2.2 Matrix representation

Tensors can be represented as matrices (but all matrices are not tensors!):

$$T_{ij} = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} \begin{matrix} \text{–vector associated with 1 direction,} \\ \text{–vector associated with 2 direction,} \\ \text{–vector associated with 3 direction.} \end{matrix} \tag{6.81}$$

A simple way to choose a vector q_j associated with a plane of arbitrary orientation is to form the inner product of the tensor T_{ij} and the unit normal associated with the plane n_i:

$$q_j = n_i T_{ij}, \qquad \mathbf{q}^T = \mathbf{n}^T \cdot \mathbf{T}. \tag{6.82}$$

Here n_i has components which are the direction cosines of the chosen direction. For example to determine the vector associated with face 2, we choose

$$n_i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{6.83}$$

Thus, in Gibbs notation we have

$$\mathbf{n}^T \cdot \mathbf{T} = (0,1,0) \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} \quad = \quad (T_{21}, T_{22}, T_{23}). \tag{6.84}$$

In Einstein notation, we arrive at the same conclusion via

$$\begin{aligned} n_i T_{ij} &= n_1 T_{1j} + n_2 T_{2j} + n_3 T_{3j}, && (6.85)\\ &= (0) T_{1j} + (1) T_{2j} + (0) T_{3j}, && (6.86)\\ &= (T_{21}, T_{22}, T_{23}). && (6.87) \end{aligned}$$

6.2.3 Transpose of a tensor, symmetric and anti-symmetric tensors

The *transpose* T_{ij}^T of a tensor T_{ij} is found by trading elements across the diagonal

$$T_{ij}^T \equiv T_{ji}, \tag{6.88}$$

so

$$T_{ij}^T = \begin{pmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{32} \\ T_{13} & T_{23} & T_{33} \end{pmatrix}. \tag{6.89}$$

A tensor is *symmetric* if it is equal to its transpose, *i.e.*

$$T_{ij} = T_{ji}, \qquad \mathbf{T} = \mathbf{T}^T, \qquad \text{if symmetric.} \tag{6.90}$$

A tensor is *anti-symmetric* if it is equal to the additive inverse of its transpose, *i.e.*

$$T_{ij} = -T_{ji}, \qquad \mathbf{T} = -\mathbf{T}^T, \qquad \text{if anti-symmetric.} \tag{6.91}$$

A tensor is *asymmetric* if it is neither symmetric nor anti-symmetric.

The tensor inner product of a symmetric tensor S_{ij} and anti-symmetric tensor A_{ij} can be shown to be 0:

$$S_{ij} A_{ij} = 0, \qquad \mathbf{S} : \mathbf{A} = 0. \tag{6.92}$$

Here the ":" notation indicates a tensor inner product.

Example 6.6

Show $S_{ij} A_{ij} = 0$ for a two-dimensional space.

Take a general symmetric tensor to be

$$S_{ij} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}. \tag{6.93}$$

Take a general anti-symmetric tensor to be

$$A_{ij} = \begin{pmatrix} 0 & d \\ -d & 0 \end{pmatrix}. \tag{6.94}$$

So

$$\begin{aligned} S_{ij} A_{ij} &= S_{11} A_{11} + S_{12} A_{12} + S_{21} A_{21} + S_{22} A_{22}, && (6.95)\\ &= a(0) + bd - bd + c(0), && (6.96)\\ &= 0. && (6.97) \end{aligned}$$

An arbitrary tensor can be represented as the sum of a symmetric and anti-symmetric tensor:

$$T_{ij} = \underbrace{\frac{1}{2}T_{ij} + \frac{1}{2}T_{ij}}_{=T_{ij}} + \underbrace{\frac{1}{2}T_{ji} - \frac{1}{2}T_{ji}}_{=0}, \tag{6.98}$$

$$= \underbrace{\frac{1}{2}\left(T_{ij} + T_{ji}\right)}_{\equiv T_{(ij)}} + \underbrace{\frac{1}{2}\left(T_{ij} - T_{ji}\right)}_{\equiv T_{[ij]}}. \tag{6.99}$$

So with

$$T_{(ij)} \equiv \frac{1}{2}\left(T_{ij} + T_{ji}\right), \tag{6.100}$$

$$T_{[ij]} \equiv \frac{1}{2}\left(T_{ij} - T_{ji}\right), \tag{6.101}$$

we arrive at

$$T_{ij} = \underbrace{T_{(ij)}}_{\text{symmetric}} + \underbrace{T_{[ij]}}_{\text{anti-symmetric}}. \tag{6.102}$$

The first term, $T_{(ij)}$, is called the symmetric part of T_{ij}; the second term, $T_{[ij]}$, is called the anti-symmetric part of T_{ij}.

6.2.4 Dual vector of an anti-symmetric tensor

As the anti-symmetric part of a three by three tensor has only three independent components, we might expect a three-component vector can be associated with this. Let us define the *dual vector* to be

$$d_i \equiv \frac{1}{2}\epsilon_{ijk}T_{jk} = \frac{1}{2}\underbrace{\epsilon_{ijk}T_{(jk)}}_{=0} + \frac{1}{2}\epsilon_{ijk}T_{[jk]}. \tag{6.103}$$

For fixed i, ϵ_{ijk} is anti-symmetric. So the first term is zero, being for fixed i the tensor inner product of an anti-symmetric and symmetric tensor. Thus,

$$d_i = \frac{1}{2}\epsilon_{ijk}T_{[jk]}. \tag{6.104}$$

Let us find the inverse. Apply ϵ_{ilm} to both sides of Eq. (6.103) to get

$$\epsilon_{ilm}d_i = \frac{1}{2}\epsilon_{ilm}\epsilon_{ijk}T_{jk}, \tag{6.105}$$

$$
\begin{aligned}
&= \frac{1}{2}(\delta_{lj}\delta_{mk} - \delta_{lk}\delta_{mj})T_{jk}, && (6.106)\\
&= \frac{1}{2}(T_{lm} - T_{ml}), && (6.107)\\
&= T_{[lm]}, && (6.108)\\
T_{[lm]} &= \epsilon_{ilm}d_i, && (6.109)\\
T_{[ij]} &= \epsilon_{kij}d_k, && (6.110)\\
T_{[ij]} &= \epsilon_{ijk}d_k. && (6.111)
\end{aligned}
$$

Expanding, we can see that

$$
T_{[ij]} = \epsilon_{ijk}d_k = \epsilon_{ij1}d_1 + \epsilon_{ij2}d_2 + \epsilon_{ij3}d_3 = \begin{pmatrix} 0 & d_3 & -d_2 \\ -d_3 & 0 & d_1 \\ d_2 & -d_1 & 0 \end{pmatrix}. \tag{6.112}
$$

The matrix form realized is obvious when one considers that an individual term, such as $\epsilon_{ij1}d_1$ only has a value when $i, j = 2, 3$ or $i, j = 3, 2$, and takes on values of $\pm d_1$ in those cases. In summary, the general dimension three tensor can be written as

$$
T_{ij} = T_{(ij)} + \epsilon_{ijk}d_k. \tag{6.113}
$$

6.2.5 Principal axes and tensor invariants

Given a tensor T_{ij}, find the associated direction such that the vector components in this associated direction are parallel to the direction. So we want

$$
n_i T_{ij} = \lambda n_j. \tag{6.114}
$$

This defines an eigenvalue problem; this will be discussed further in Sec. 7.4.4. Linear algebra gives us the eigenvalues and associated eigenvectors.

$$
\begin{aligned}
n_i T_{ij} &= \lambda n_i \delta_{ij}, && (6.115)\\
n_i(T_{ij} - \lambda\delta_{ij}) &= 0, && (6.116)\\
(n_1, n_2, n_3)\begin{pmatrix} T_{11}-\lambda & T_{12} & T_{13} \\ T_{21} & T_{22}-\lambda & T_{23} \\ T_{31} & T_{32} & T_{33}-\lambda \end{pmatrix} &= (0,0,0). && (6.117)
\end{aligned}
$$

This is equivalent to $\mathbf{n}^T \cdot (\mathbf{T} - \lambda\mathbf{I}) = \mathbf{0}^T$ or $(\mathbf{T} - \lambda\mathbf{I})^T \cdot \mathbf{n} = \mathbf{0}$. We get non-trivial solutions if

$$
\begin{vmatrix} T_{11}-\lambda & T_{12} & T_{13} \\ T_{21} & T_{22}-\lambda & T_{23} \\ T_{31} & T_{32} & T_{33}-\lambda \end{vmatrix} = 0. \tag{6.118}
$$

We are actually finding the so-called *left* eigenvectors of T_{ij}. These arise with less frequency than the right eigenvectors, which are defined by $T_{ij}u_j = \lambda\delta_{ij}u_j$. Right and left eigenvalue problems are discussed later in Sec. 7.4.4.

We know from linear algebra that such an equation for a third order matrix gives rise to a characteristic polynomial for λ of the form

$$\lambda^3 - I_T^{(1)}\lambda^2 + I_T^{(2)}\lambda - I_T^{(3)} = 0, \tag{6.119}$$

where $I_T^{(1)}, I_T^{(2)}, I_T^{(3)}$ are scalars which are functions of all the scalars T_{ij}. The I_T's are known as the *invariants* of the tensor T_{ij}. The invariants will not change if the coordinate axes are rotated; in contrast, the scalar components T_{ij} will change under rotation. The invariants can be shown to be given by

$$\begin{aligned}
I_T^{(1)} &= T_{ii} = T_{11} + T_{22} + T_{33} = \text{tr } \mathbf{T}, && (6.120)\\
I_T^{(2)} &= \frac{1}{2}\left(T_{ii}T_{jj} - T_{ij}T_{ji}\right) = \frac{1}{2}\left((\text{tr } \mathbf{T})^2 - \text{tr}(\mathbf{T}\cdot\mathbf{T})\right) = (\det \mathbf{T})(\text{tr } \mathbf{T}^{-1}), && (6.121)\\
&= \frac{1}{2}\left(T_{(ii)}T_{(jj)} + T_{[ij]}T_{[ij]} - T_{(ij)}T_{(ij)}\right), && (6.122)\\
I_T^{(3)} &= \epsilon_{ijk}T_{1i}T_{2j}T_{3k} = \det \mathbf{T}. && (6.123)
\end{aligned}$$

Here, "tr" denotes the trace. It can also be shown that if $\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}$ are the three eigenvalues, then the invariants can also be expressed as

$$\begin{aligned}
I_T^{(1)} &= \lambda^{(1)} + \lambda^{(2)} + \lambda^{(3)}, && (6.124)\\
I_T^{(2)} &= \lambda^{(1)}\lambda^{(2)} + \lambda^{(2)}\lambda^{(3)} + \lambda^{(3)}\lambda^{(1)}, && (6.125)\\
I_T^{(3)} &= \lambda^{(1)}\lambda^{(2)}\lambda^{(3)}. && (6.126)
\end{aligned}$$

If T_{ij} is real and symmetric, it can be shown that

- the eigenvalues are real,
- eigenvectors corresponding to distinct eigenvalues are real and orthogonal, and
- the left and right eigenvectors are identical.

A sketch of a volume element rotated to be aligned with a set of orthogonal principal axes is shown in Figure 6.3.

If the matrix is asymmetric, the eigenvalues could be complex, and the eigenvectors are not orthogonal. It is often most physically relevant to decompose a tensor into symmetric and anti-symmetric parts and find the orthogonal basis vectors and real eigenvalues associated with the symmetric part and the dual vector associated with the anti-symmetric part.

In continuum mechanics,

- the symmetric part of a tensor can be associated with deformation along principal axes, and
- the anti-symmetric part of a tensor can be associated with rotation of an element.

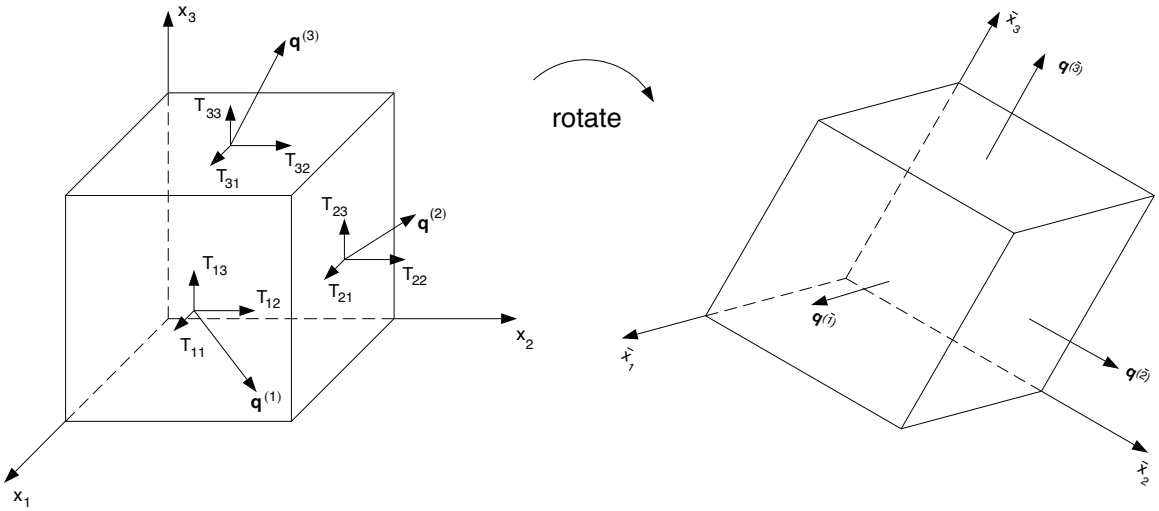

Figure 6.3: Sketch depicting rotation of volume element to be aligned with principal axes. Tensor T_{ij} must be symmetric to guarantee existence of orthogonal principal directions.

Example 6.7

Decompose the tensor given here into a combination of orthogonal basis vectors and a dual vector.

$$T_{ij} = \begin{pmatrix} 1 & 1 & -2 \\ 3 & 2 & -3 \\ -4 & 1 & 1 \end{pmatrix}. \tag{6.127}$$

First

$$T_{(ij)} = \frac{1}{2}(T_{ij} + T_{ji}) = \begin{pmatrix} 1 & 2 & -3 \\ 2 & 2 & -1 \\ -3 & -1 & 1 \end{pmatrix}, \tag{6.128}$$

$$T_{[ij]} = \frac{1}{2}(T_{ij} - T_{ji}) = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -2 \\ -1 & 2 & 0 \end{pmatrix}. \tag{6.129}$$

First, get the dual vector d_i:

$$d_i = \frac{1}{2}\epsilon_{ijk}T_{[jk]}, \tag{6.130}$$

$$d_1 = \frac{1}{2}\epsilon_{1jk}T_{[jk]} = \frac{1}{2}(\epsilon_{123}T_{[23]} + \epsilon_{132}T_{[32]}) = \frac{1}{2}((1)(-2) + (-1)(2)) = -2, \tag{6.131}$$

$$d_2 = \frac{1}{2}\epsilon_{2jk}T_{[jk]} = \frac{1}{2}(\epsilon_{213}T_{[13]} + \epsilon_{231}T_{[31]}) = \frac{1}{2}((-1)(1) + (1)(-1)) = -1, \tag{6.132}$$

$$d_3 = \frac{1}{2}\epsilon_{3jk}T_{[jk]} = \frac{1}{2}(\epsilon_{312}T_{[12]} + \epsilon_{321}T_{[21]}) = \frac{1}{2}((1)(-1) + (-1)(1)) = -1, \tag{6.133}$$

$$d_i = (-2, -1, -1)^T. \tag{6.134}$$

Note that Eq. (6.112) is satisfied.

Now find the eigenvalues and eigenvectors for the symmetric part.

$$\begin{vmatrix} 1-\lambda & 2 & -3 \\ 2 & 2-\lambda & -1 \\ -3 & -1 & 1-\lambda \end{vmatrix} = 0. \tag{6.135}$$

We get the characteristic polynomial,

$$\lambda^3 - 4\lambda^2 - 9\lambda + 9 = 0. \tag{6.136}$$

The eigenvalue and associated normalized eigenvector for each root is

$$\lambda^{(1)} = 5.36488, \qquad n_i^{(1)} = (-0.630537, -0.540358, 0.557168)^T, \tag{6.137}$$

$$\lambda^{(2)} = -2.14644, \qquad n_i^{(2)} = (-0.740094, 0.202303, -0.641353)^T, \tag{6.138}$$

$$\lambda^{(3)} = 0.781562, \qquad n_i^{(3)} = (-0.233844, 0.816754, 0.527476)^T. \tag{6.139}$$

It is easily verified that each eigenvector is orthogonal. When the coordinates are transformed to be aligned with the principal axes, the magnitude of the vector associated with each face is the eigenvalue; this vector points in the same direction of the unit normal associated with the face.

Example 6.8

For a given tensor, which we will take to be symmetric, though the theory applies to non-symmetric tensors as well,

$$T_{ij} = \mathbf{T} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & -1 \\ 4 & -1 & 1 \end{pmatrix}, \tag{6.140}$$

find the three basic tensor invariants, $I_T^{(1)}$, $I_T^{(2)}$, and $I_T^{(3)}$, and show they are truly invariant when the tensor is subjected to a rotation with direction cosine matrix of

$$\ell_{ij} = \mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \end{pmatrix}. \tag{6.141}$$

Calculation shows that $\det \mathbf{Q} = 1$, and $\mathbf{Q} \cdot \mathbf{Q}^T = \mathbf{I}$, so the matrix $\mathbf{Q}$ is volume- and orientation-preserving, and thus a rotation matrix. As an aside, the construction of an orthogonal matrix, such as our $\mathbf{Q}$ is non-trivial. One method of construction involves determining a set of orthogonal vectors via a process to be described later, see Sec. 7.3.2.5.

The eigenvalues of $\mathbf{T}$, which are the principal values, are easily calculated to be

$$\lambda^{(1)} = 5.28675, \qquad \lambda^{(2)} = -3.67956, \qquad \lambda^{(3)} = 3.39281. \tag{6.142}$$

The three invariants of T_{ij} are

$$I_T^{(1)} = \mathrm{tr}(\mathbf{T}) = \mathrm{tr} \begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & -1 \\ 4 & -1 & 1 \end{pmatrix} = 1 + 3 + 1 = 5, \tag{6.143}$$

$$
\begin{aligned}
I_T^{(2)} &= \frac{1}{2}\left((\mathrm{tr}(\mathbf{T}))^2 - \mathrm{tr}(\mathbf{T}\cdot\mathbf{T})\right) \\
&= \frac{1}{2}\left(\left(\mathrm{tr}\begin{pmatrix}1&2&4\\2&3&-1\\4&-1&1\end{pmatrix}\right)^2 - \mathrm{tr}\left(\begin{pmatrix}1&2&4\\2&3&-1\\4&-1&1\end{pmatrix}\cdot\begin{pmatrix}1&2&4\\2&3&-1\\4&-1&1\end{pmatrix}\right)\right), \\
&= \frac{1}{2}\left(5^2 - \mathrm{tr}\begin{pmatrix}21&4&6\\4&14&4\\6&4&18\end{pmatrix}\right), \\
&= \frac{1}{2}(25-21-14-18), \\
&= -14, \qquad (6.144)\\
I_T^{(3)} &= \det\mathbf{T} = \det\begin{pmatrix}1&2&4\\2&3&-1\\4&-1&1\end{pmatrix} = -66. \qquad (6.145)
\end{aligned}
$$

Now when we rotate the tensor $\mathbf{T}$, we get a transformed tensor given by

$$
\begin{aligned}
\overline{\mathbf{T}} = \mathbf{Q}^T\cdot\mathbf{T}\cdot\mathbf{Q} &= \begin{pmatrix}\frac{1}{\sqrt{6}}&\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{2}}\\ \sqrt{\frac{2}{3}}&-\frac{1}{\sqrt{3}}&0\\ \frac{1}{\sqrt{6}}&\frac{1}{\sqrt{3}}&-\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}1&2&4\\2&3&-1\\4&-1&1\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{6}}&\sqrt{\frac{2}{3}}&\frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{3}}&-\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{2}}&0&-\frac{1}{\sqrt{2}}\end{pmatrix}, \qquad (6.146)\\
&= \begin{pmatrix}4.10238&2.52239&1.60948\\2.52239&-0.218951&-2.91291\\1.60948&-2.91291&1.11657\end{pmatrix}. \qquad (6.147)
\end{aligned}
$$

We then seek the tensor invariants of $\overline{\mathbf{T}}$. Leaving out some of the details, which are the same as those for calculating the invariants of $\mathbf{T}$, we find the invariants indeed are invariant:

$$
\begin{aligned}
I_T^{(1)} &= 4.10238 - 0.218951 + 1.11657 = 5, \qquad (6.148)\\
I_T^{(2)} &= \frac{1}{2}(5^2 - 53) = -14, \qquad (6.149)\\
I_T^{(3)} &= -66. \qquad (6.150)
\end{aligned}
$$

Finally, we verify that the tensor invariants are indeed related to the principal values (the eigenvalues of the tensor) as follows

$$
\begin{aligned}
I_T^{(1)} &= \lambda^{(1)} + \lambda^{(2)} + \lambda^{(3)} = 5.28675 - 3.67956 + 3.39281 = 5, \qquad (6.151)\\
I_T^{(2)} &= \lambda^{(1)}\lambda^{(2)} + \lambda^{(2)}\lambda^{(3)} + \lambda^{(3)}\lambda^{(1)}, \\
&= (5.28675)(-3.67956) + (-3.67956)(3.39281) + (3.39281)(5.28675) = -14, \qquad (6.152)\\
I_T^{(3)} &= \lambda^{(1)}\lambda^{(2)}\lambda^{(3)} = (5.28675)(-3.67956)(3.39281) = -66. \qquad (6.153)
\end{aligned}
$$

6.3 Algebra of vectors

Here we will primarily use bold letters for vectors, such as in $\mathbf{u}$. At times we will use the notation u_i to represent a vector.

6.3.1 Definition and properties

Null vector: A vector with zero components.

Multiplication by a scalar α: $\alpha\mathbf{u} = \alpha u_1\mathbf{e}_1 + \alpha u_2\mathbf{e}_2 + \alpha u_3\mathbf{e}_3 = \alpha u_i$,

Sum of vectors: $\mathbf{u} + \mathbf{v} = (u_1 + v_1)\mathbf{e}_1 + (u_2 + v_2)\mathbf{e}_2 + (u_3 + v_3)\mathbf{e}_3 = (u_i + v_i)$,

Magnitude, length, or norm of a vector: $||\mathbf{u}||_2 = \sqrt{u_1^2 + u_2^2 + u_3^2} = \sqrt{u_i u_i}$,

Triangle inequality: $||\mathbf{u} + \mathbf{v}||_2 \leq ||\mathbf{u}||_2 + ||\mathbf{v}||_2$.

Here the subscript 2 in $||\cdot||_2$ indicates we are considering a Euclidean norm. In many sources in the literature this subscript is omitted, and the norm is understood to be the Euclidean norm. In a more general sense, we can still retain the property of a norm for a more general p-norm for a three-dimensional vector:

$$||\mathbf{u}||_p = \left(|u_1|^p + |u_2|^p + |u_3|^p\right)^{1/p}, \qquad 1 \leq p < \infty. \tag{6.154}$$

For example the 1-norm of a vector is the sum of the absolute values of its components:

$$||\mathbf{u}||_1 = \left(|u_1| + |u_2| + |u_3|\right). \tag{6.155}$$

The ∞-norm selects the largest component:

$$||\mathbf{u}||_\infty = \lim_{p\to\infty} \left(|u_1|^p + |u_2|^p + |u_3|^p\right)^{1/p} = \max_{i=1,2,3}|u_i|. \tag{6.156}$$

6.3.2 Scalar product (dot product, inner product)

The scalar product of $\mathbf{u}$ and $\mathbf{v}$ is defined for vectors with real components as

$$<\mathbf{u}, \mathbf{v}> = \mathbf{u}^T \cdot \mathbf{v} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = u_1v_1 + u_2v_2 + u_3v_3 = u_iv_i. \tag{6.157}$$

Note that the term u_iv_i is a scalar, which explains the nomenclature "scalar product."

The vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be *orthogonal* if $\mathbf{u}^T \cdot \mathbf{v} = 0$. Also

$$<\mathbf{u}, \mathbf{u}> = \mathbf{u}^T \cdot \mathbf{u} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1^2 + u_2^2 + u_3^2 = u_iu_i = (||\mathbf{u}||_2)^2. \tag{6.158}$$

We will consider important modifications for vectors with complex components later in Sec. 7.3.2. In the same section, we will consider the generalized notion of an inner product, denoted here by $<.,.>$.

6.3.3 Cross product

The cross product of $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \epsilon_{ijk} u_j v_k. \tag{6.159}$$

Note the cross product of two vectors is a vector.

Property: $\mathbf{u} \times \alpha\mathbf{u} = \mathbf{0}$. Let's use Cartesian index notation to prove this

$$\begin{aligned} \mathbf{u} \times \alpha\mathbf{u} &= \epsilon_{ijk} u_j \alpha u_k, && (6.160)\\ &= \alpha \epsilon_{ijk} u_j u_k, && (6.161)\\ &= \alpha(\epsilon_{i11} u_1 u_1 + \epsilon_{i12} u_1 u_2 + \epsilon_{i13} u_1 u_3, && (6.162)\\ &\quad + \epsilon_{i21} u_2 u_1 + \epsilon_{i22} u_2 u_2 + \epsilon_{i23} u_2 u_3 && (6.163)\\ &\quad + \epsilon_{i31} u_3 u_1 + \epsilon_{i32} u_3 u_2 + \epsilon_{i33} u_3 u_3) && (6.164)\\ &= 0, \qquad \text{for } i = 1, 2, 3, && (6.165) \end{aligned}$$

since $\epsilon_{i11} = \epsilon_{i22} = \epsilon_{i33} = 0$ and $\epsilon_{i12} = -\epsilon_{i21}$, $\epsilon_{i13} = -\epsilon_{i31}$, and $\epsilon_{i23} = -\epsilon_{i32}$.

6.3.4 Scalar triple product

The scalar triple product of three vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ is defined by

$$\begin{aligned} [\mathbf{u}, \mathbf{v}, \mathbf{w}] &= \mathbf{u}^T \cdot (\mathbf{v} \times \mathbf{w}), && (6.166)\\ &= \epsilon_{ijk} u_i v_j w_k. && (6.167) \end{aligned}$$

The scalar triple product is a scalar. Geometrically, it represents the volume of the parallelepiped with edges parallel to the three vectors.

6.3.5 Identities

$$\begin{aligned} [\mathbf{u}, \mathbf{v}, \mathbf{w}] &= -[\mathbf{u}, \mathbf{w}, \mathbf{v}], && (6.168)\\ \mathbf{u} \times (\mathbf{v} \times \mathbf{w}) &= (\mathbf{u}^T \cdot \mathbf{w})\mathbf{v} - (\mathbf{u}^T \cdot \mathbf{v})\mathbf{w}, && (6.169)\\ (\mathbf{u} \times \mathbf{v}) \times (\mathbf{w} \times \mathbf{x}) &= [\mathbf{u}, \mathbf{w}, \mathbf{x}]\mathbf{v} - [\mathbf{v}, \mathbf{w}, \mathbf{x}]\mathbf{u}, && (6.170)\\ (\mathbf{u} \times \mathbf{v})^T \cdot (\mathbf{w} \times \mathbf{x}) &= (\mathbf{u}^T \cdot \mathbf{w})(\mathbf{v}^T \cdot \mathbf{x}) - (\mathbf{u}^T \cdot \mathbf{x})(\mathbf{v}^T \cdot \mathbf{w}). && (6.171) \end{aligned}$$

Example 6.9

Prove Eq. (6.169) using Cartesian index notation.

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \epsilon_{ijk} u_j \left(\epsilon_{klm} v_l w_m\right), \tag{6.172}$$

$$
\begin{aligned}
&= \epsilon_{ijk}\epsilon_{klm}u_j v_l w_m, && (6.173)\\
&= \epsilon_{kij}\epsilon_{klm}u_j v_l w_m, && (6.174)\\
&= (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})\, u_j v_l w_m, && (6.175)\\
&= u_j v_i w_j - u_j v_j w_i, && (6.176)\\
&= u_j w_j v_i - u_j v_j w_i, && (6.177)\\
&= (\mathbf{u}^T \cdot \mathbf{w})\mathbf{v} - (\mathbf{u}^T \cdot \mathbf{v})\mathbf{w}. && (6.178)
\end{aligned}
$$

6.4 Calculus of vectors

6.4.1 Vector function of single scalar variable

If we have the scalar function $\phi(\tau)$ and vector functions $\mathbf{u}(\tau)$ and $\mathbf{v}(\tau)$, some useful identities, based on the product rule, which can be proved include

$$
\begin{aligned}
\frac{d}{d\tau}(\phi\mathbf{u}) = \phi\frac{d\mathbf{u}}{d\tau} + \frac{d\phi}{d\tau}\mathbf{u}, \qquad \frac{d}{d\tau}(\phi u_i) &= \phi\frac{du_i}{d\tau} + \frac{d\phi}{d\tau}u_i, && (6.179)\\
\frac{d}{d\tau}(\mathbf{u}^T \cdot \mathbf{v}) = \mathbf{u}^T \cdot \frac{d\mathbf{v}}{d\tau} + \frac{d\mathbf{u}^T}{d\tau} \cdot \mathbf{v}, \qquad \frac{d}{d\tau}(u_i v_i) &= u_i\frac{dv_i}{d\tau} + \frac{du_i}{d\tau}v_i, && (6.180)\\
\frac{d}{d\tau}(\mathbf{u} \times \mathbf{v}) = \mathbf{u} \times \frac{d\mathbf{v}}{d\tau} + \frac{d\mathbf{u}}{d\tau} \times \mathbf{v}, \qquad \frac{d}{d\tau}(\epsilon_{ijk}u_j v_k) &= \epsilon_{ijk}u_j\frac{dv_k}{d\tau} + \epsilon_{ijk}v_k\frac{du_j}{d\tau}. && (6.181)
\end{aligned}
$$

Here τ is a general scalar parameter, which may or may not have a simple physical interpretation.

6.4.2 Differential geometry of curves

Now let us consider a general discussion of curves in space. If

$$\mathbf{r}(\tau) = x_i(\tau)\mathbf{e}_i = x_i(\tau), \tag{6.182}$$

then $\mathbf{r}(\tau)$ describes a curve in three-dimensional space. If we require that the basis vectors be constants (this will not be the case in most general coordinate systems, but is for ordinary Cartesian systems), the derivative of Eq. (6.182) is

$$\frac{d\mathbf{r}(\tau)}{d\tau} = \mathbf{r}'(\tau) = x_i'(\tau)\mathbf{e}_i = x_i'(\tau). \tag{6.183}$$

Now $\mathbf{r}'(\tau)$ is a vector that is tangent to the curve. A unit vector in this direction is

$$\mathbf{t} = \frac{\mathbf{r}'(\tau)}{||\mathbf{r}'(\tau)||_2}, \tag{6.184}$$

where

$$||\mathbf{r}'(\tau)||_2 = \sqrt{x_i' x_i'}. \tag{6.185}$$

In the special case in which τ is time t, we denote the derivative by a dot ($\dot{}$) notation rather than a prime ($'$) notation; $\dot{\mathbf{r}}$ is the velocity vector, $\dot{x}_i$ its components, and $||\dot{\mathbf{r}}||_2$ the magnitude. Note that the unit tangent vector $\mathbf{t}$ *is not* the scalar parameter for time, t. Also we will occasionally use the scalar components of $\mathbf{t}$: t_i, which again are not related to time t.

Take $s(t)$ to be the distance along the curve. Pythagoras' theorem tells us for differential distances that

$$\begin{aligned} ds^2 &= dx_1^2 + dx_2^2 + dx_3^2, & (6.186)\\ ds &= \sqrt{dx_1^2 + dx_2^2 + dx_3^2}, & (6.187)\\ ds &= ||dx_i||_2, & (6.188)\\ \frac{ds}{dt} &= \left|\left|\frac{dx_i}{dt}\right|\right|_2, & (6.189)\\ &= ||\dot{\mathbf{r}}(t)||_2, & (6.190) \end{aligned}$$

so that

$$\mathbf{t} = \frac{\dot{\mathbf{r}}}{||\dot{\mathbf{r}}||_2} = \frac{\frac{d\mathbf{r}}{dt}}{\frac{ds}{dt}} = \frac{d\mathbf{r}}{ds}, \qquad t_i = \frac{dr_i}{ds}. \tag{6.191}$$

Also integrating Eq. (6.190) with respect to t gives

$$s = \int_a^b ||\dot{\mathbf{r}}(t)||_2 dt = \int_a^b \sqrt{\frac{dx_i}{dt}\frac{dx_i}{dt}}\, dt = \int_a^b \sqrt{\frac{dx_1}{dt}\frac{dx_1}{dt} + \frac{dx_2}{dt}\frac{dx_2}{dt} + \frac{dx_3}{dt}\frac{dx_3}{dt}}\, dt, \tag{6.192}$$

to be the distance along the curve between $t = a$ and $t = b$.

Example 6.10

If

$$\mathbf{r}(t) = 2t^2\mathbf{i} + t^3\mathbf{j}, \tag{6.193}$$

find the unit tangent at $t = 1$, and the length of the curve from $t = 0$ to $t = 1$.

The derivative is

$$\dot{\mathbf{r}}(t) = 4t\mathbf{i} + 3t^2\mathbf{j}. \tag{6.194}$$

At $t = 1$,

$$\dot{\mathbf{r}}(t = 1) = 4\mathbf{i} + 3\mathbf{j} \tag{6.195}$$

so that the unit vector in this direction is

$$\mathbf{t} = \frac{4}{5}\mathbf{i} + \frac{3}{5}\mathbf{j}. \tag{6.196}$$

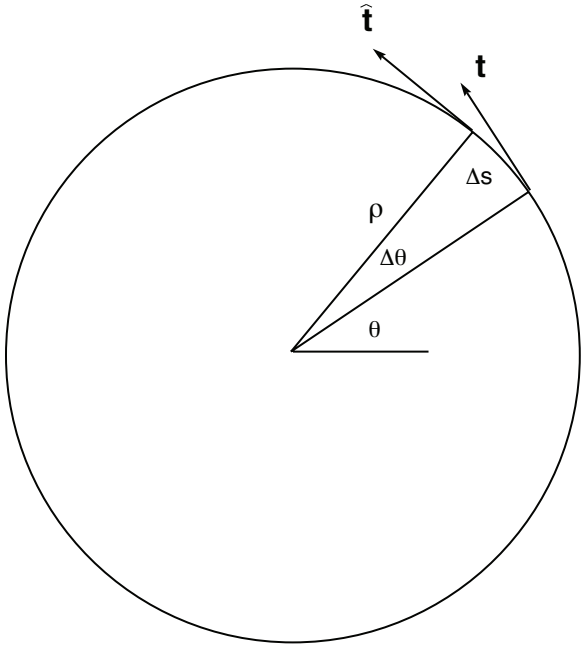

Figure 6.4: Sketch for determination of radius of curvature.

The length of the curve from $t = 0$ to $t = 1$ is

$$s = \int_0^1 \sqrt{16t^2 + 9t^4}dt, \tag{6.197}$$

$$= \frac{1}{27}(16 + 9t^2)^{3/2} \,|_0^1, \tag{6.198}$$

$$= \frac{61}{27}. \tag{6.199}$$

In Fig. 6.4, $\mathbf{r}(t)$ describes a circle. Two unit tangents, $\mathbf{t}$ and $\hat{\mathbf{t}}$ are drawn at times t and $t + \Delta t$. At time t we have

$$\mathbf{t} = -\sin\theta \ \mathbf{i} + \cos\theta \ \mathbf{j}. \tag{6.200}$$

At time $t + \Delta t$ we have

$$\hat{\mathbf{t}} = -\sin(\theta + \Delta\theta) \ \mathbf{i} + \cos(\theta + \Delta\theta) \ \mathbf{j}. \tag{6.201}$$

Expanding Eq. (6.201) in a Taylor series about $\Delta\theta = 0$, we get

$$\hat{\mathbf{t}} = \left(-\sin\theta - \Delta\theta\cos\theta + O(\Delta\theta)^2\right) \ \mathbf{i} + \left(\cos\theta - \Delta\theta\sin\theta + O(\Delta\theta)^2\right) \ \mathbf{j}, \tag{6.202}$$

so as $\Delta\theta \to 0$,

$$\hat{\mathbf{t}} - \mathbf{t} = -\Delta\theta\cos\theta \ \mathbf{i} - \Delta\theta\sin\theta \ \mathbf{j}, \tag{6.203}$$

$$\Delta\mathbf{t} = \Delta\theta \underbrace{\left(-\cos\theta \ \mathbf{i} - \sin\theta \ \mathbf{j}\right)}_{\text{unit vector}}. \tag{6.204}$$

It is easily verified that $\Delta\mathbf{t}^T \cdot \mathbf{t} = 0$, so $\Delta\mathbf{t}$ is normal to $\mathbf{t}$. Furthermore, since $-\cos\theta\mathbf{i} - \sin\theta\mathbf{j}$ is a unit vector,

$$||\Delta\mathbf{t}||_2 = \Delta\theta. \tag{6.205}$$

Now for $\Delta\theta \to 0$,

$$\Delta s = \rho\Delta\theta. \tag{6.206}$$

where ρ is the radius of curvature. So

$$||\Delta\mathbf{t}||_2 = \frac{\Delta s}{\rho} \tag{6.207}$$

Thus,

$$\left|\left|\frac{\Delta\mathbf{t}}{\Delta s}\right|\right|_2 = \frac{1}{\rho}. \tag{6.208}$$

Taking all limits to zero, we get

$$\left|\left|\frac{d\mathbf{t}}{ds}\right|\right|_2 = \frac{1}{\rho}. \tag{6.209}$$

The term on the right side of Eq. (6.209) is often defined as the curvature, κ:

$$\kappa = \frac{1}{\rho}. \tag{6.210}$$

Thus, the curvature κ is the magnitude of $d\mathbf{t}/ds$; it gives a measure of how the unit tangent changes as one moves along the curve.

6.4.2.1 Curves on a plane

The plane curve $y = f(x)$ in the x-y plane can be represented as

$$\mathbf{r}(t) = x(t)\ \mathbf{i} + y(t)\ \mathbf{j}, \tag{6.211}$$

where $x(t) = t$ and $y(t) = f(t)$. Differentiating, we have

$$\dot{\mathbf{r}}(t) = \dot{x}(t)\ \mathbf{i} + \dot{y}(t)\ \mathbf{j}. \tag{6.212}$$

The unit vector from Eq. (6.184) is

$$\mathbf{t} = \frac{\dot{x}\mathbf{i} + \dot{y}\mathbf{j}}{(\dot{x}^2 + \dot{y}^2)^{1/2}}, \tag{6.213}$$

$$= \frac{\mathbf{i} + y'\mathbf{j}}{(1 + (y')^2)^{1/2}}, \tag{6.214}$$

where the primes are derivatives with respect to x. Since

$$ds^2 = dx^2 + dy^2, \tag{6.215}$$

$$ds = \left(dx^2 + dy^2\right)^{1/2}, \tag{6.216}$$

$$\frac{ds}{dx} = \frac{1}{dx}\left(dx^2 + dy^2\right)^{1/2}, \tag{6.217}$$

$$\frac{ds}{dx} = (1 + (y')^2)^{1/2}, \tag{6.218}$$

we have, by first expanding $d\mathbf{t}/ds$ with the chain rule, then applying the quotient rule to expand the derivative of Eq. (6.214) along with the use of Eq. (6.218),

$$\frac{d\mathbf{t}}{ds} = \frac{\frac{d\mathbf{t}}{dx}}{\frac{ds}{dx}}, \tag{6.219}$$

$$= \underbrace{\frac{(1 + (y')^2)^{1/2} y''\mathbf{j} - (\mathbf{i} + y'\mathbf{j})(1 + (y')^2)^{-1/2} y' y''}{1 + (y')^2}}_{d\mathbf{t}/dx} \underbrace{\frac{1}{(1 + (y')^2)^{1/2}}}_{1/(ds/dx)}, \tag{6.220}$$

$$= \underbrace{\frac{y''}{(1 + (y')^2)^{3/2}}}_{=\kappa} \underbrace{\frac{-y'\mathbf{i} + \mathbf{j}}{(1 + (y')^2)^{1/2}}}_{\mathbf{n}}. \tag{6.221}$$

As the second factor of Eq. (6.221) is a unit vector, the leading scalar factor must be the magnitude of $d\mathbf{t}/ds$. We define this unit vector to be $\mathbf{n}$, and note that it is orthogonal to the unit tangent vector $\mathbf{t}$:

$$\mathbf{n}^T \cdot \mathbf{t} = \frac{-y'\mathbf{i} + \mathbf{j}}{(1 + (y')^2)^{1/2}} \cdot \frac{\mathbf{i} + y'\mathbf{j}}{(1 + (y')^2)^{1/2}}, \tag{6.222}$$

$$= \frac{-y' + y'}{1 + (y')^2}, \tag{6.223}$$

$$= 0. \tag{6.224}$$

Expanding our notion of curvature and radius of curvature, we define $d\mathbf{t}/ds$ such that

$$\frac{d\mathbf{t}}{ds} = \kappa\mathbf{n}, \tag{6.225}$$

$$\left|\left|\frac{d\mathbf{t}}{ds}\right|\right|_2 = \kappa = \frac{1}{\rho}. \tag{6.226}$$

Thus,

$$\kappa = \frac{y''}{(1 + (y')^2)^{3/2}}, \tag{6.227}$$

$$\rho = \frac{(1 + (y')^2)^{3/2}}{y''}, \tag{6.228}$$

for curves on a plane.

6.4.2.2 Curves in three-dimensional space

We next expand these notions to three-dimensional space. A set of local, right-handed, orthogonal coordinates can be defined at a point on a curve $\mathbf{r}(t)$. The unit vectors at this point are the tangent $\mathbf{t}$, the principal normal $\mathbf{n}$, and the binormal $\mathbf{b}$, where

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} \tag{6.229}$$

$$\mathbf{n} = \frac{1}{\kappa}\frac{d\mathbf{t}}{ds}, \tag{6.230}$$

$$\mathbf{b} = \mathbf{t} \times \mathbf{n}. \tag{6.231}$$

We will first show that $\mathbf{t}$, $\mathbf{n}$, and $\mathbf{b}$ form an orthogonal system of unit vectors. We have already seen that $\mathbf{t}$ is a unit vector tangent to the curve. By the product rule for vector differentiation, we have the identity

$$\mathbf{t}^T \cdot \frac{d\mathbf{t}}{ds} = \frac{1}{2}\frac{d}{ds}(\underbrace{\mathbf{t}^T \cdot \mathbf{t}}_{=1}). \tag{6.232}$$

Since $\mathbf{t}^T \cdot \mathbf{t} = ||\mathbf{t}||_2^2 = 1$, we recover

$$\mathbf{t}^T \cdot \frac{d\mathbf{t}}{ds} = 0. \tag{6.233}$$

Thus, $\mathbf{t}$ is orthogonal to $d\mathbf{t}/ds$. Since $\mathbf{n}$ is parallel to $d\mathbf{t}/ds$, it is orthogonal to $\mathbf{t}$ also. From Eqs. (6.209) and (6.230), we see that $\mathbf{n}$ is a unit vector. Furthermore, $\mathbf{b}$ is a unit vector orthogonal to both $\mathbf{t}$ and $\mathbf{n}$ because of its definition in terms of a cross product of those vectors in Eq. (6.231).

Next, we will derive some basic relations involving the unit vectors and the characteristics of the curve. Take d/ds of Eq. (6.231):

$$\frac{d\mathbf{b}}{ds} = \frac{d}{ds}\left(\mathbf{t} \times \mathbf{n}\right), \tag{6.234}$$

$$= \frac{d\mathbf{t}}{ds} \times \underbrace{\mathbf{n}}_{(1/\kappa)d\mathbf{t}/ds} + \mathbf{t} \times \frac{d\mathbf{n}}{ds}, \tag{6.235}$$

$$= \frac{d\mathbf{t}}{ds} \times \frac{1}{\kappa}\frac{d\mathbf{t}}{ds} + \mathbf{t} \times \frac{d\mathbf{n}}{ds}, \tag{6.236}$$

$$= \frac{1}{\kappa}\underbrace{\frac{d\mathbf{t}}{ds} \times \frac{d\mathbf{t}}{ds}}_{=0} + \mathbf{t} \times \frac{d\mathbf{n}}{ds}, \tag{6.237}$$

$$= \mathbf{t} \times \frac{d\mathbf{n}}{ds}. \tag{6.238}$$

So we see that $d\mathbf{b}/ds$ is orthogonal to $\mathbf{t}$. In addition, since $||\mathbf{b}||_2 = 1$,

$$\begin{aligned} \mathbf{b}^T \cdot \frac{d\mathbf{b}}{ds} &= \frac{1}{2}\frac{d}{ds}(\mathbf{b}^T \cdot \mathbf{b}), & (6.239)\\ &= \frac{1}{2}\frac{d}{ds}(||\mathbf{b}||_2^2), & (6.240)\\ &= \frac{1}{2}\frac{d}{ds}(1^2), & (6.241)\\ &= 0. & (6.242) \end{aligned}$$

So $d\mathbf{b}/ds$ is orthogonal to $\mathbf{b}$ also. Since $d\mathbf{b}/ds$ is orthogonal to both $\mathbf{t}$ and $\mathbf{b}$, it must be aligned with the only remaining direction, $\mathbf{n}$. So, we can write

$$\frac{d\mathbf{b}}{ds} = \tau\mathbf{n}, \tag{6.243}$$

where τ is the magnitude of $d\mathbf{b}/ds$, which we call the *torsion* of the curve.

From Eq. (6.231) it is easily deduced that $\mathbf{n} = \mathbf{b} \times \mathbf{t}$,. Differentiating this with respect to s, we get

$$\begin{aligned} \frac{d\mathbf{n}}{ds} &= \frac{d\mathbf{b}}{ds} \times \mathbf{t} + \mathbf{b} \times \frac{d\mathbf{t}}{ds}, & (6.244)\\ &= \tau\mathbf{n} \times \mathbf{t} + \mathbf{b} \times \kappa\mathbf{n}, & (6.245)\\ &= -\tau\mathbf{b} - \kappa\mathbf{t}. & (6.246) \end{aligned}$$

Summarizing

$$\begin{aligned} \frac{d\mathbf{t}}{ds} &= \kappa\mathbf{n}, & (6.247)\\ \frac{d\mathbf{n}}{ds} &= -\kappa\mathbf{t} - \tau\mathbf{b}, & (6.248)\\ \frac{d\mathbf{b}}{ds} &= \tau\mathbf{n}. & (6.249) \end{aligned}$$

These are the Frenet-Serret[2] relations. In matrix form, we can say that

$$\frac{d}{ds}\begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & -\tau \\ 0 & \tau & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}. \tag{6.250}$$

Note the coefficient matrix is anti-symmetric.

Example 6.11

Find the local coordinates, the curvature, and the torsion for the helix

$$\mathbf{r}(t) = a\cos t\ \mathbf{i} + a\sin t\ \mathbf{j} + bt\ \mathbf{k}. \tag{6.251}$$

[2]Jean Frédéric Frenet, 1816-1900, French mathematician, and Joseph Alfred Serret, 1819-1885, French mathematician.

Taking the derivative and finding its magnitude we get

$$\frac{d\mathbf{r}(t)}{dt} = -a\sin t\ \mathbf{i} + a\cos t\ \mathbf{j} + b\ \mathbf{k}, \tag{6.252}$$

$$\left|\left|\frac{d\mathbf{r}(t)}{dt}\right|\right|_2 = \sqrt{a^2\sin^2 t + a^2\cos^2 t + b^2}, \tag{6.253}$$

$$= \sqrt{a^2+b^2}. \tag{6.254}$$

This gives us the unit tangent vector $\mathbf{t}$:

$$\mathbf{t} = \frac{\frac{d\mathbf{r}}{dt}}{\left|\left|\frac{d\mathbf{r}}{dt}\right|\right|_2} = \frac{-a\sin t\ \mathbf{i} + a\cos t\ \mathbf{j} + b\ \mathbf{k}}{\sqrt{a^2+b^2}}. \tag{6.255}$$

We also have

$$\frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}, \tag{6.256}$$

$$= \sqrt{a^2\sin^2 t + a^2\cos^2 t + b^2}, \tag{6.257}$$

$$= \sqrt{a^2+b^2}. \tag{6.258}$$

Continuing, we have

$$\frac{d\mathbf{t}}{ds} = \frac{\frac{d\mathbf{t}}{dt}}{\frac{ds}{dt}}, \tag{6.259}$$

$$= -a\frac{\cos t\ \mathbf{i} + \sin t\ \mathbf{j}}{\sqrt{a^2+b^2}}\ \frac{1}{\sqrt{a^2+b^2}}, \tag{6.260}$$

$$= \underbrace{\frac{a}{a^2+b^2}}_{\kappa}\underbrace{(-\cos t\ \mathbf{i} - \sin t\ \mathbf{j})}_{\mathbf{n}}, \tag{6.261}$$

$$= \kappa\mathbf{n}. \tag{6.262}$$

Thus, the unit principal normal is

$$\mathbf{n} = -(\cos t\ \mathbf{i} + \sin t\ \mathbf{j}). \tag{6.263}$$

The curvature is

$$\kappa = \frac{a}{a^2+b^2}. \tag{6.264}$$

The radius of curvature is

$$\rho = \frac{a^2+b^2}{a}. \tag{6.265}$$

We also find the unit binormal

$$\mathbf{b} = \mathbf{t}\times\mathbf{n}, \tag{6.266}$$

$$= \frac{1}{\sqrt{a^2+b^2}}\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a\sin t & a\cos t & b \\ -\cos t & -\sin t & 0 \end{vmatrix}, \tag{6.267}$$

$$= \frac{b\sin t\ \mathbf{i} - b\cos t\ \mathbf{j} + a\ \mathbf{k}}{\sqrt{a^2+b^2}}. \tag{6.268}$$

The torsion is determined from

$$\tau \mathbf{n} = \frac{\frac{d\mathbf{b}}{dt}}{\frac{ds}{dt}}, \tag{6.269}$$

$$= b\frac{\cos t\ \mathbf{i} + \sin t\ \mathbf{j}}{a^2+b^2}, \tag{6.270}$$

$$= \underbrace{\frac{-b}{a^2+b^2}}_{\tau}\ \underbrace{(-\cos t\ \mathbf{i} - \sin t\ \mathbf{j})}_{\mathbf{n}}, \tag{6.271}$$

from which

$$\tau = -\frac{b}{a^2+b^2}. \tag{6.272}$$

Further identities which can be proved relate directly to the time parameterization of $\mathbf{r}$:

$$\frac{d\mathbf{r}}{dt} \times \frac{d^2\mathbf{r}}{dt^2} = \kappa v^3 \mathbf{b}, \tag{6.273}$$

$$\left(\frac{d\mathbf{r}}{dt} \times \frac{d^2\mathbf{r}}{dt^2}\right)^T \cdot \frac{d^3\mathbf{r}}{dt^3} = -\kappa^2 v^6 \tau, \tag{6.274}$$

$$\frac{\sqrt{||\ddot{\mathbf{r}}||_2^2\ ||\dot{\mathbf{r}}||_2^2 - (\dot{\mathbf{r}}^T \cdot \ddot{\mathbf{r}})^2}}{||\dot{\mathbf{r}}||_2^3} = \kappa, \tag{6.275}$$

where $v = ds/dt$.

6.5 Line and surface integrals

If $\mathbf{r}$ is a position vector,

$$\mathbf{r} = x_i \mathbf{e}_i, \tag{6.276}$$

then $\phi(\mathbf{r})$ is a scalar field, and $\mathbf{u}(\mathbf{r})$ is a vector field.

6.5.1 Line integrals

A line integral is of the form

$$I = \int_C \mathbf{u}^T \cdot d\mathbf{r}, \tag{6.277}$$

where $\mathbf{u}$ is a vector field, and $d\mathbf{r}$ is an element of curve C. If $\mathbf{u} = u_i$, and $d\mathbf{r} = dx_i$, then we can write

$$I = \int_C u_i\ dx_i. \tag{6.278}$$

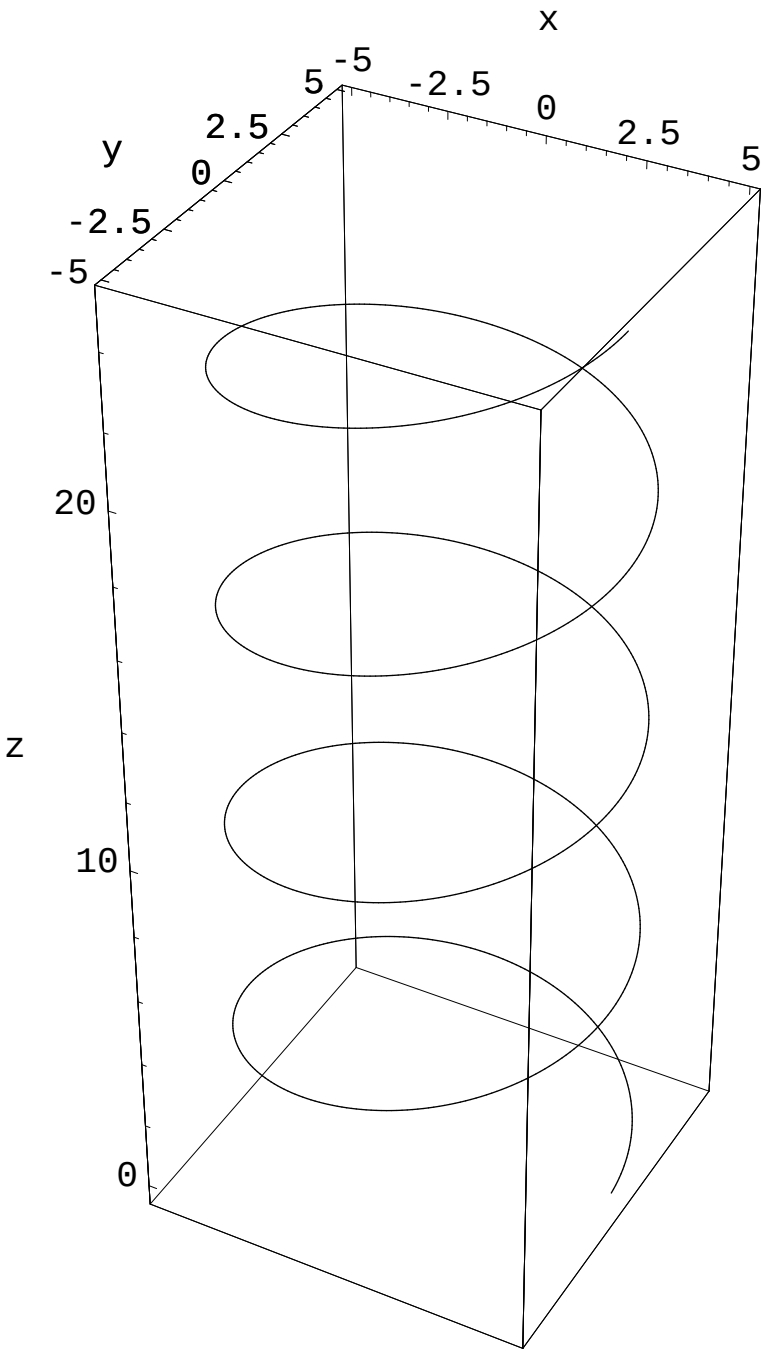

Figure 6.5: Three-dimensional curve parameterized by $x(t) = a\cos t$, $y(t) = a\sin t$, $z(t) = bt$, with $a = 5$, $b = 1$, for $t \in [0, 25]$.

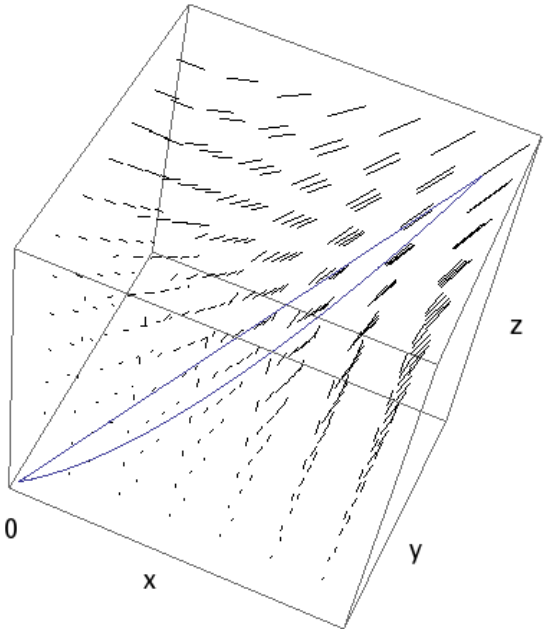

Figure 6.6: The vector field $\mathbf{u} = yz\mathbf{i} + xy\mathbf{j} + xz\mathbf{k}$ and the curves a) $x = y^2 = z$; b) $x = y = z$.

Example 6.12

Find

$$I = \int_C \mathbf{u}^T \cdot d\mathbf{r}, \tag{6.279}$$

if

$$\mathbf{u} = yz\ \mathbf{i} + xy\ \mathbf{j} + xz\ \mathbf{k}, \tag{6.280}$$

and C goes from $(0,0,0)$ to $(1,1,1)$ along

(a) the curve $x = y^2 = z$,
(b) the straight line $x = y = z$.

The vector field and two paths are sketched in Fig. 6.6. We have

$$\int_C \mathbf{u}^T \cdot d\mathbf{r} = \int_C (yz\ dx + xy\ dy + xz\ dz). \tag{6.281}$$

(a) Substituting $x = y^2 = z$, and thus $dx = 2ydy$, $dx = dz$, we get

$$\begin{aligned} I &= \int_0^1 y^3(2y\ dy) + y^3\ dy + y^4(2y\ dy), && (6.282)\\ &= \int_0^1 (2y^4 + y^3 + 2y^5)dy, && (6.283)\\ &= \frac{2y^5}{5} + \frac{y^4}{4} + \frac{y^6}{3}\Big|_0^1, && (6.284)\\ &= \frac{59}{60}. && (6.285) \end{aligned}$$

We can achieve the same result in an alternative way that is often more useful for more curves whose representation is more complicated. Let us parameterize C by taking $x = t$, $y = t^2$, $z = t$. Thus $dx = dt$, $dy = 2tdt$, $dz = dt$. The end points of C are at $t = 0$ and $t = 1$. So the integral is

$$\begin{aligned} I &= \int_0^1 (t^2 t\ dt + t t^2 (2t)\ dt + t(t)\ dt, && (6.286)\\ &= \int_0^1 (t^3 + 2t^4 + t^2)\ dt, && (6.287)\\ &= \left.\frac{t^4}{4} + \frac{2t^5}{5} + \frac{t^3}{3}\right|_0^1, && (6.288)\\ &= \frac{59}{60}. && (6.289) \end{aligned}$$

(b) Substituting $x = y = z$, and thus $dx = dy = dz$, we get

$$I = \int_0^1 (x^2 dx + x^2 dx + x^2 dx) = \int_0^1 3x^2 dx = x^3|_0^1 = 1. \quad (6.290)$$

Note a different value for I was obtained on path (b) relative to that found on path (a); thus, the integral here is path-dependent.

In general the value of a line integral depends on the path. If, however, we have the special case in which we can form $\mathbf{u} = \nabla\phi$ in Eq. (6.277), where ϕ is a scalar field, then

$$\begin{aligned} I &= \int_C (\nabla\phi)^T \cdot d\mathbf{r}, && (6.291)\\ &= \int_C \frac{\partial\phi}{\partial x_i} dx_i, && (6.292)\\ &= \int_C d\phi, && (6.293)\\ &= \phi(\mathbf{b}) - \phi(\mathbf{a}), && (6.294) \end{aligned}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the beginning and end of curve C. The integral I is then independent of path. $\mathbf{u}$ is then called a *conservative* field, and ϕ is its *potential*.

6.5.2 Surface integrals

A surface integral is of the form

$$I = \int_S \mathbf{u}^T \cdot \mathbf{n}\ dS = \int_S u_i n_i\ dS \quad (6.295)$$

where $\mathbf{u}$ (or u_i) is a vector field, S is an open or closed surface, dS is an element of this surface, and $\mathbf{n}$ (or n_i) is a unit vector normal to the surface element.

6.6 Differential operators

Surface integrals can be used for coordinate-independent definitions of differential operators. Beginning with some well-known theorems: the divergence theorem for a scalar, the divergence theorem, and a little known theorem, which is possible to demonstrate, we have, where S is a surface enclosing volume V,

$$\int_V \nabla\phi \; dV = \int_S \mathbf{n}\phi \; dS, \tag{6.296}$$

$$\int_V \nabla^T \cdot \mathbf{u} \; dV = \int_S \mathbf{n}^T \cdot \mathbf{u} \; dS, \tag{6.297}$$

$$\int_V (\nabla \times \mathbf{u}) \; dV = \int_S \mathbf{n} \times \mathbf{u} \; dS. \tag{6.298}$$

Now we invoke the mean value theorem, which asserts that somewhere within the limits of integration, the integrand takes on its mean value, which we denote with an overline, so that, for example, $\int_V \alpha \; dV = \overline{\alpha} V$. Thus, we get

$$\overline{(\nabla\phi)} \; V = \int_S \mathbf{n}\phi \; dS, \tag{6.299}$$

$$\overline{(\nabla^T \cdot \mathbf{u})} \; V = \int_S \mathbf{n}^T \cdot \mathbf{u} \; dS, \tag{6.300}$$

$$\overline{(\nabla \times \mathbf{u})} \; V = \int_S \mathbf{n} \times \mathbf{u} \; dS. \tag{6.301}$$

As we let $V \to 0$, mean values approach local values, so we get

$$\nabla\phi \equiv \text{grad } \phi = \lim_{V\to 0} \frac{1}{V} \int_S \mathbf{n}\phi \; dS, \tag{6.302}$$

$$\nabla^T \cdot \mathbf{u} \equiv \text{div } \mathbf{u} = \lim_{V\to 0} \frac{1}{V} \int_S \mathbf{n}^T \cdot \mathbf{u} \; dS, \tag{6.303}$$

$$\nabla \times \mathbf{u} \equiv \text{curl } \mathbf{u} = \lim_{V\to 0} \frac{1}{V} \int_S \mathbf{n} \times \mathbf{u} \; dS, \tag{6.304}$$

where $\phi(\mathbf{r})$ is a scalar field, and $\mathbf{u}(\mathbf{r})$ is a vector field. V is the region enclosed within a closed surface S, and $\mathbf{n}$ is the unit normal to an element of the surface dS. Here "grad" is the gradient operator, "div" is the divergence operator, and "curl" is the curl operator.

Consider the element of volume in Cartesian coordinates shown in Fig. 6.7. The differential operations in this coordinate system can be deduced from the definitions and written in terms of the vector operator ∇:

$$\nabla = \mathbf{e}_1 \frac{\partial}{\partial x_1} + \mathbf{e}_2 \frac{\partial}{\partial x_2} + \mathbf{e}_3 \frac{\partial}{\partial x_3} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{pmatrix} = \frac{\partial}{\partial x_i}. \tag{6.305}$$

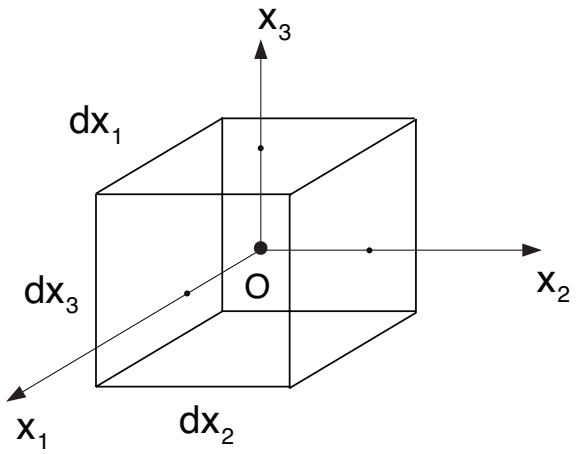

Figure 6.7: Element of volume.

We also adopt the unconventional, row vector operator

$$\nabla^T = \left(\frac{\partial}{\partial x_1} \quad \frac{\partial}{\partial x_2} \quad \frac{\partial}{\partial x_3} \right). \tag{6.306}$$

The operator ∇^T is well-defined for Cartesian coordinate systems, but does not extend to non-orthogonal systems.

6.6.1 Gradient of a scalar

Let's evaluate the gradient of a scalar function of a vector

$$\text{grad } (\phi(x_i)). \tag{6.307}$$

We take the reference value of ϕ to be at the origin O. Consider first the x_1 variation. At O, $x_1 = 0$, and our function takes the value of ϕ. At the faces a distance $x_1 = \pm\, dx_1/2$ away from O in the x_1-direction, our function takes a value of

$$\phi \pm \frac{\partial \phi}{\partial x_1} \frac{dx_1}{2}. \tag{6.308}$$

Writing $V = dx_1 dx_2 dx_3$, Eq. (6.302) gives

$$\begin{aligned}
\text{grad } \phi &= \lim_{V \to 0} \frac{1}{V} \left(\left(\phi + \frac{\partial \phi}{\partial x_1} \frac{dx_1}{2} \right) \mathbf{e}_1 dx_2 dx_3 - \left(\phi - \frac{\partial \phi}{\partial x_1} \frac{dx_1}{2} \right) \mathbf{e}_1 dx_2 dx_3 \right. && (6.309) \\
&\quad \left. + \text{similar terms from the } x_2 \text{ and } x_3 \text{ faces} \right), \\
&= \frac{\partial \phi}{\partial x_1} \mathbf{e}_1 + \frac{\partial \phi}{\partial x_2} \mathbf{e}_2 + \frac{\partial \phi}{\partial x_3} \mathbf{e}_3, && (6.310) \\
&= \frac{\partial \phi}{\partial x_i} \mathbf{e}_i = \frac{\partial \phi}{\partial x_i}, && (6.311) \\
&= \nabla \phi. && (6.312)
\end{aligned}$$

The derivative of ϕ on a particular path is called the directional derivative. If the path has a unit tangent $\mathbf{t}$, the derivative in this direction is

$$(\nabla\phi)^T \cdot \mathbf{t} = t_i \frac{\partial \phi}{\partial x_i}. \tag{6.313}$$

If $\phi(x, y, z) =$ constant is a surface, then $d\phi = 0$ on this surface. Also

$$\begin{aligned} d\phi &= \frac{\partial \phi}{\partial x_i} dx_i, & (6.314)\\ &= (\nabla\phi)^T \cdot d\mathbf{r}. & (6.315) \end{aligned}$$

Since $d\mathbf{r}$ is tangent to the surface, $\nabla\phi$ must be normal to it. The tangent plane at $\mathbf{r} = \mathbf{r}_0$ is defined by the position vector $\mathbf{r}$ such that

$$(\nabla\phi)^T \cdot (\mathbf{r} - \mathbf{r}_0) = 0. \tag{6.316}$$

Example 6.13

At the point (1,1,1), find the unit normal to the surface

$$z^3 + xz = x^2 + y^2. \tag{6.317}$$

Define

$$\phi(x, y, z) = z^3 + xz - x^2 - y^2 = 0. \tag{6.318}$$

A normal at (1,1,1) is

$$\begin{aligned} \nabla\phi &= (z - 2x)\ \mathbf{i} - 2y\ \mathbf{j} + (3z^2 + x)\mathbf{k}, & (6.319)\\ &= -1\ \mathbf{i} - 2\ \mathbf{j} + 4\ \mathbf{k}. & (6.320) \end{aligned}$$

The unit normal is

$$\begin{aligned} \mathbf{n} &= \frac{\nabla\phi}{||\nabla\phi||_2}, & (6.321)\\ &= \frac{1}{\sqrt{21}}(-1\ \mathbf{i} - 2\ \mathbf{j} + 4\ \mathbf{k}). & (6.322) \end{aligned}$$

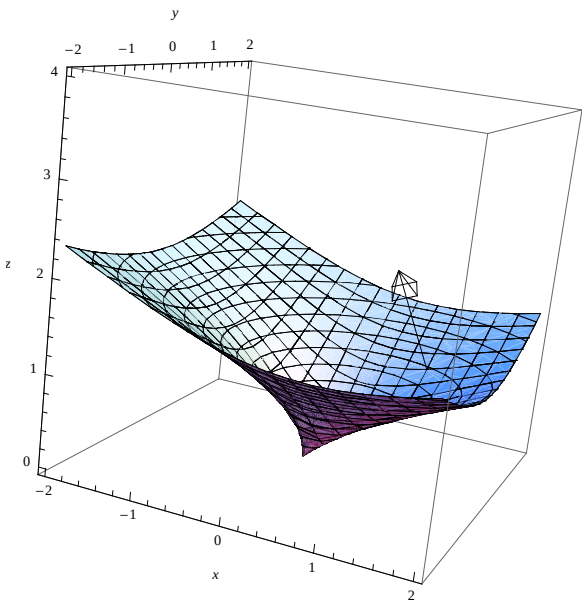

Figure 6.8: Plot of surface $z^3 + xz = x^2 + y^2$ and normal vector at $(1, 1, 1)$.

6.6.2 Divergence

6.6.2.1 Vectors

Equation (6.303) becomes

$$\begin{aligned}
\text{div } \mathbf{u} &= \lim_{V\to 0} \frac{1}{V} \left(\left(u_1 + \frac{\partial u_1}{\partial x_1}\frac{dx_1}{2} \right) dx_2 dx_3 - \left(u_1 - \frac{\partial u_1}{\partial x_1}\frac{dx_1}{2} \right) dx_2 dx_3 \right. \qquad (6.323)\\
&\quad \left. + \text{similar terms from the } x_2 \text{ and } x_3 \text{ faces} \right),\\
&= \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}, \qquad (6.324)\\
&= \frac{\partial u_i}{\partial x_i}, \qquad (6.325)\\
&= \nabla^T \cdot \mathbf{u} = \begin{pmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}. \qquad (6.326)
\end{aligned}$$

6.6.2.2 Tensors

The extension to tensors is straightforward

$$\text{div}\mathbf{T} = \nabla^T \cdot \mathbf{T}, \qquad (6.327)$$

$$= \frac{\partial T_{ij}}{\partial x_i}. \tag{6.328}$$

Notice that this yields a vector quantity.

6.6.3 Curl of a vector

The application of Eq. (6.304) is not obvious here. Consider just one of the faces: the face whose outer normal is $\mathbf{e}_1$. For that face, one needs to evaluate

$$\int_S \mathbf{n} \times \mathbf{u} \; dS. \tag{6.329}$$

On this face, one has $\mathbf{n} = \mathbf{e}_1$, and

$$\mathbf{u} = \left(u_1 + \frac{\partial u_1}{\partial x_1} dx_1\right) \mathbf{e}_1 + \left(u_2 + \frac{\partial u_2}{\partial x_1} dx_1\right) \mathbf{e}_2 + \left(u_3 + \frac{\partial u_3}{\partial x_1} dx_1\right) \mathbf{e}_3. \tag{6.330}$$

So, on this face the integrand is

$$\mathbf{n} \times \mathbf{u} = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ 1 & 0 & 0 \\ \left(u_1 + \frac{\partial u_1}{\partial x_1} dx_1\right) & \left(u_2 + \frac{\partial u_2}{\partial x_1} dx_1\right) & \left(u_3 + \frac{\partial u_3}{\partial x_1} dx_1\right) \end{vmatrix}, \tag{6.331}$$

$$= \left(u_2 + \frac{\partial u_2}{\partial x_1} dx_1\right) \mathbf{e}_3 - \left(u_3 + \frac{\partial u_3}{\partial x_1} dx_1\right) \mathbf{e}_2. \tag{6.332}$$

Two similar terms appear on the opposite face, whose unit vector points in the $-\mathbf{e}_1$ direction.

Carrying out the integration then for equation (6.304), one gets

$$\begin{aligned} \text{curl } \mathbf{u} = \lim_{V \to 0} \frac{1}{V} \Bigg(& \left(u_2 + \frac{\partial u_2}{\partial x_1} \frac{dx_1}{2}\right) \mathbf{e}_3 dx_2 dx_3 - \left(u_3 + \frac{\partial u_3}{\partial x_1} \frac{dx_1}{2}\right) \mathbf{e}_2 dx_2 dx_3 \\ & - \left(u_2 - \frac{\partial u_2}{\partial x_1} \frac{dx_1}{2}\right) \mathbf{e}_3 dx_2 dx_3 + \left(u_3 - \frac{\partial u_3}{\partial x_1} \frac{dx_1}{2}\right) \mathbf{e}_2 dx_2 dx_3 \\ & + \text{similar terms from the } x_2 \text{ and } x_3 \text{ faces} \Bigg), \end{aligned} \tag{6.333}$$

$$= \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} \\ u_1 & u_2 & u_3 \end{vmatrix}, \tag{6.334}$$

$$= \epsilon_{ijk} \frac{\partial u_k}{\partial x_j}, \tag{6.335}$$

$$= \nabla \times \mathbf{u}. \tag{6.336}$$

The curl of a tensor does not arise often in practice.

6.6.4 Laplacian

6.6.4.1 Scalar

The Laplacian[3] is simply div grad, and can be written, when operating on ϕ, as

$$\text{div grad } \phi = \nabla^T \cdot (\nabla \phi) = \nabla^2 \phi = \frac{\partial^2 \phi}{\partial x_i \partial x_i}. \tag{6.337}$$

6.6.4.2 Vector

Equation (6.346) is used to evaluate the Laplacian of a vector:

$$\nabla^2 \mathbf{u} = \nabla^T \cdot \nabla \mathbf{u} = \nabla(\nabla^T \cdot \mathbf{u}) - \nabla \times (\nabla \times \mathbf{u}). \tag{6.338}$$

6.6.5 Identities

$$\begin{aligned}
\nabla \times (\nabla \phi) &= \mathbf{0}, && (6.339)\\
\nabla^T \cdot (\nabla \times \mathbf{u}) &= 0 && (6.340)\\
\nabla^T \cdot (\phi \mathbf{u}) &= \phi \nabla^T \cdot \mathbf{u} + (\nabla \phi)^T \cdot \mathbf{u}, && (6.341)\\
\nabla \times (\phi \mathbf{u}) &= \phi \nabla \times \mathbf{u} + \nabla \phi \times \mathbf{u}, && (6.342)\\
\nabla^T \cdot (\mathbf{u} \times \mathbf{v}) &= \mathbf{v}^T \cdot (\nabla \times \mathbf{u}) - \mathbf{u}^T \cdot (\nabla \times \mathbf{v}), && (6.343)\\
\nabla \times (\mathbf{u} \times \mathbf{v}) &= (\mathbf{v}^T \cdot \nabla)\mathbf{u} - (\mathbf{u}^T \cdot \nabla)\mathbf{v} + \mathbf{u}(\nabla^T \cdot \mathbf{v}) - \mathbf{v}(\nabla^T \cdot \mathbf{u}), && (6.344)\\
\nabla(\mathbf{u}^T \cdot \mathbf{v}) &= (\mathbf{u}^T \cdot \nabla)\mathbf{v} + (\mathbf{v}^T \cdot \nabla)\mathbf{u} + \mathbf{u} \times (\nabla \times \mathbf{v}) + \mathbf{v} \times (\nabla \times \mathbf{u}), && (6.345)\\
\nabla \cdot \nabla^T \mathbf{u} &= \nabla(\nabla^T \cdot \mathbf{u}) - \nabla \times (\nabla \times \mathbf{u}). && (6.346)
\end{aligned}$$

Example 6.14

Show that Eq. (6.346)

$$\nabla \cdot \nabla^T \mathbf{u} = \nabla(\nabla^T \cdot \mathbf{u}) - \nabla \times (\nabla \times \mathbf{u}). \tag{6.347}$$

is true.

Going from right to left

$$\begin{aligned}
\nabla(\nabla^T \cdot \mathbf{u}) - \nabla \times (\nabla \times \mathbf{u}) &= \frac{\partial}{\partial x_i}\frac{\partial u_j}{\partial x_j} - \epsilon_{ijk}\frac{\partial}{\partial x_j}\left(\epsilon_{klm}\frac{\partial u_m}{\partial x_l}\right), && (6.348)\\
&= \frac{\partial}{\partial x_i}\frac{\partial u_j}{\partial x_j} - \epsilon_{kij}\epsilon_{klm}\frac{\partial}{\partial x_j}\left(\frac{\partial u_m}{\partial x_l}\right), && (6.349)\\
&= \frac{\partial^2 u_j}{\partial x_i \partial x_j} - (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})\frac{\partial^2 u_m}{\partial x_j \partial x_l}, && (6.350)
\end{aligned}$$

[3]Pierre-Simon Laplace, 1749-1827, Normandy-born French mathematician.

$$= \frac{\partial^2 u_j}{\partial x_i \partial x_j} - \frac{\partial^2 u_j}{\partial x_j \partial x_i} + \frac{\partial^2 u_i}{\partial x_j \partial x_j}, \tag{6.351}$$

$$= \frac{\partial}{\partial x_j}\left(\frac{\partial u_i}{\partial x_j}\right), \tag{6.352}$$

$$= \nabla^T \cdot \nabla \mathbf{u}. \tag{6.353}$$

6.6.6 Curvature revisited

If a curve in two-dimensional space is given implicitly by the function

$$\phi(x, y) = 0, \tag{6.354}$$

it can be shown that the curvature is given by the formula

$$\kappa = \nabla \cdot \left(\frac{\nabla \phi}{||\nabla \phi||_2}\right), \tag{6.355}$$

provided one takes precautions to preserve the sign as will be demonstrated in the following example. Note that $\nabla\phi$ is a gradient vector which must be normal to any so-called *level set* curve for which ϕ is constant; moreover, it points in the direction of most rapid change of ϕ. The corresponding vector $\nabla\phi/||\nabla\phi||_2$ must be a unit normal vector to level sets of ϕ.

Example 6.15

Show Eq. (6.355) is equivalent to Eq. (6.227) if $y = f(x)$.

Let us take

$$\phi(x, y) = f(x) - y = 0. \tag{6.356}$$

Then, with $'$ denoting a derivative with respect to x, we get

$$\nabla\phi = \frac{\partial \phi}{\partial x}\mathbf{i} + \frac{\partial \phi}{\partial y}\mathbf{j}, \tag{6.357}$$

$$= f'(x)\mathbf{i} - \mathbf{j}. \tag{6.358}$$

We then see that

$$||\nabla\phi||_2 = \sqrt{f'(x)^2 + 1}, \tag{6.359}$$

so that

$$\frac{\nabla\phi}{||\nabla\phi||_2} = \frac{f'(x)\mathbf{i} - \mathbf{j}}{\sqrt{1 + f'(x)^2}}. \tag{6.360}$$

Then we see that by applying Eq. (6.355), we get

$$\kappa = \nabla \cdot \left(\frac{\nabla \phi}{||\nabla \phi||_2} \right), \tag{6.361}$$

$$= \nabla \cdot \left(\frac{f'(x)\mathbf{i} - \mathbf{j}}{\sqrt{1 + f'(x)^2}} \right), \tag{6.362}$$

$$= \frac{\partial}{\partial x} \left(\frac{f'(x)}{\sqrt{1 + f'(x)^2}} \right) + \underbrace{\frac{\partial}{\partial y} \left(\frac{-1}{\sqrt{1 + f'(x)^2}} \right)}_{=0}, \tag{6.363}$$

$$= \frac{\sqrt{1 + f'(x)^2} f''(x) - f'(x) f'(x) f''(x) \left(1 + f'(x)^2\right)^{-1/2}}{1 + f'(x)^2}, \tag{6.364}$$

$$= \frac{\left(1 + f'(x)^2\right) f''(x) - f'(x) f'(x) f''(x)}{\left(1 + f'(x)^2\right)^{3/2}}, \tag{6.365}$$

$$= \frac{f''(x)}{\left(1 + f'(x)^2\right)^{3/2}}. \tag{6.366}$$

Equation (6.366) is fully equivalent to the earlier developed Eq. (6.227). Note however that if we had chosen $\phi(x, y) = y - f(x) = 0$, we would have recovered a formula for curvature with the opposite sign.

Considering now surfaces embedded in a three dimensional space described parametrically by

$$\phi(x, y, z) = 0. \tag{6.367}$$

It can be shown that the so-called *mean curvature* of the surface κ_M is given by Eq. (6.355):

$$\kappa_M = \nabla \cdot \left(\frac{\nabla \phi}{||\nabla \phi||_2} \right) \tag{6.368}$$

Note that their are many other measures of curvature of surfaces.

Lastly, let us return to consider one-dimensional curves embedded within a high dimensional space. The curves may be considered to be defined as solutions to the differential equations of the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}(\mathbf{x}). \tag{6.369}$$

We can consider $\mathbf{v}(\mathbf{x})$ to be a velocity field which is dependent on position $\mathbf{x}$, but independent of time. A particle with a known initial condition will move through the field, acquiring a new velocity at each new spatial point it encounters, and thus tracing a non-trivial trajectory. We now take the velocity gradient tensor to be $\mathbf{F}$, with

$$\mathbf{F} = \nabla \mathbf{v}^T. \tag{6.370}$$

With this, it can then be shown after detailed analysis that the curvature of the trajectory is given by

$$\kappa = \frac{\sqrt{(\mathbf{v}^T \cdot \mathbf{F} \cdot \mathbf{F}^T \cdot \mathbf{v})(\mathbf{v}^T \cdot \mathbf{v}) - (\mathbf{v}^T \cdot \mathbf{F}^T \cdot \mathbf{v})^2}}{(\mathbf{v}^T \cdot \mathbf{v})^{3/2}} \tag{6.371}$$

In terms of the unit tangent vector, $\mathbf{t} = \mathbf{v}/||\mathbf{v}||_2$, Eq. (6.371) reduces to

$$\kappa = \frac{\sqrt{(\mathbf{t}^T \cdot \mathbf{F} \cdot \mathbf{F}^T \cdot \mathbf{t}) - (\mathbf{t}^T \cdot \mathbf{F}^T \cdot \mathbf{t})^2}}{||\mathbf{v}||_2} \tag{6.372}$$

Example 6.16

Find the curvature of the curve given by

$$\frac{dx}{dt} = -y, \qquad x(0) = 0, \tag{6.373}$$

$$\frac{dy}{dt} = x, \qquad y(0) = 2. \tag{6.374}$$

We can of course solve this exactly by first dividing one equation by the other to get

$$\frac{dy}{dx} = -\frac{x}{y}, \qquad y(x = 0) = 2. \tag{6.375}$$

Separating variables, we get

$$ydy = -xdx, \tag{6.376}$$

$$\frac{y^2}{2} = -\frac{x^2}{2} + C, \tag{6.377}$$

$$\frac{2^2}{2} = -\frac{0^2}{2} + C, \tag{6.378}$$

$$C = 2. \tag{6.379}$$

Thus,

$$x^2 + y^2 = 4, \tag{6.380}$$

is the curve of interest. It is a circle whose radius is 2 and thus whose radius of curvature $\rho = 2$; thus, its curvature $\kappa = 1/\rho = 1/2$.

Let us reproduce this result using Eq. (6.371). We can think of the two-dimensional velocity vector as

$$\mathbf{v} = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}. \tag{6.381}$$

The velocity gradient is then

$$\mathbf{F} = \nabla \mathbf{v}^T = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} \begin{pmatrix} u(x,y) & v(x,y) \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{6.382}$$

Now, let us use Eq. (6.371) to directly compute the curvature. The simple nature of our velocity field induces several simplifications. First, because the velocity gradient tensor here is antisymmetric, we have

$$\mathbf{v}^T \cdot \mathbf{F}^T \cdot \mathbf{v} = \begin{pmatrix} -y & x \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -y \\ x \end{pmatrix} = \begin{pmatrix} -y & x \end{pmatrix} \begin{pmatrix} -x \\ -y \end{pmatrix} = xy - xy = 0. \tag{6.383}$$

Second, we see that

$$\mathbf{F}\cdot\mathbf{F}^T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\cdot\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}. \tag{6.384}$$

So for this problem, Eq. (6.371) reduces to

$$\begin{aligned}
\kappa &= \frac{\sqrt{(\mathbf{v}^T\cdot\underbrace{\mathbf{F}\cdot\mathbf{F}^T}_{\mathbf{I}}\cdot\mathbf{v})(\mathbf{v}^T\cdot\mathbf{v}) - (\underbrace{\mathbf{v}^T\cdot\mathbf{F}^T\cdot\mathbf{v}}_{=0})^2}}{(\mathbf{v}^T\cdot\mathbf{v})^{3/2}}, && (6.385)\\
&= \frac{\sqrt{(\mathbf{v}^T\cdot\mathbf{v})(\mathbf{v}^T\cdot\mathbf{v})}}{(\mathbf{v}^T\cdot\mathbf{v})^{3/2}}, && (6.386)\\
&= \frac{(\mathbf{v}^T\cdot\mathbf{v})}{(\mathbf{v}^T\cdot\mathbf{v})^{3/2}}, && (6.387)\\
&= \frac{1}{\sqrt{\mathbf{v}^T\cdot\mathbf{v}}}, && (6.388)\\
&= \frac{1}{||\mathbf{v}||_2}, && (6.389)\\
&= \frac{1}{\sqrt{x^2+y^2}}, && (6.390)\\
&= \frac{1}{\sqrt{4}}, && (6.391)\\
&= \frac{1}{2}. && (6.392)
\end{aligned}$$

6.7 Special theorems

6.7.1 Green's theorem

Let $\mathbf{u} = u_x\,\mathbf{i} + u_y\,\mathbf{j}$ be a vector field, C a closed curve, and D the region enclosed by C, all in the x-y plane. Then

$$\oint_C \mathbf{u}^T\cdot d\mathbf{r} = \int\int_D \left(\frac{\partial u_y}{\partial x} - \frac{\partial u_x}{\partial y}\right)\,dx\,dy. \tag{6.393}$$

Example 6.17

Show that Green's theorem is valid if $\mathbf{u} = y\,\mathbf{i} + 2xy\,\mathbf{j}$, and C consists of the straight lines (0,0) to (1,0) to (1,1) to (0,0).

$$\oint_C \mathbf{u}^T\cdot d\mathbf{r} = \int_{C_1}\mathbf{u}^T\cdot d\mathbf{r} + \int_{C_2}\mathbf{u}^T\cdot d\mathbf{r} + \int_{C_3}\mathbf{u}^T\cdot d\mathbf{r}, \tag{6.394}$$

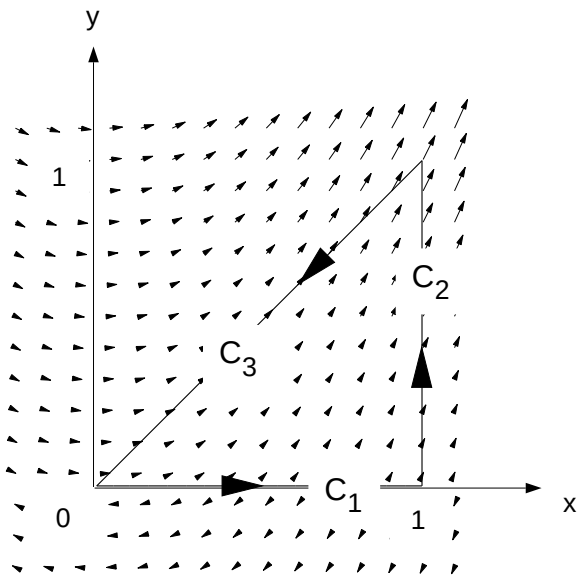

Figure 6.9: Sketch of vector field $\mathbf{u} = y\mathbf{i} + 2xy\mathbf{j}$ and closed contour integral C.

where C_1, C_2, and C_3 are the straight lines (0,0) to (1,0), (1,0) to (1,1), and (1,1) to (0,0), respectively. This is sketched in Figure 6.9.

For this problem we have

$$C_1: \quad y = 0, \quad dy = 0, \quad x \in [0,1], \quad \mathbf{u} = 0\ \mathbf{i} + 0\ \mathbf{j}, \tag{6.395}$$

$$C_2: \quad x = 1, \quad dx = 0, \quad y \in [0,1], \quad \mathbf{u} = y\ \mathbf{i} + 2y\ \mathbf{j}, \tag{6.396}$$

$$C_3: \quad x = y, \quad dx = dy, \quad x \in [1,0], \quad y \in [1,0], \quad \mathbf{u} = x\ \mathbf{i} + 2x^2\ \mathbf{j}. \tag{6.397}$$

Thus,

$$\oint_C \mathbf{u} \cdot d\mathbf{r} = \underbrace{\int_0^1 (0\ \mathbf{i} + 0\ \mathbf{j}) \cdot (dx\ \mathbf{i})}_{C_1} + \underbrace{\int_0^1 (y\ \mathbf{i} + 2y\ \mathbf{j}) \cdot (dy\ \mathbf{j})}_{C_2} + \underbrace{\int_1^0 (x\ \mathbf{i} + 2x^2\ \mathbf{j}) \cdot (dx\ \mathbf{i} + dx\ \mathbf{j})}_{C_3}, \tag{6.398}$$

$$= \int_0^1 2y\ dy + \int_1^0 (x + 2x^2)\ dx, \tag{6.399}$$

$$= y^2\big|_0^1 + \left(\frac{1}{2}x^2 + \frac{2}{3}x^3\right)\Big|_1^0 = 1 - \frac{1}{2} - \frac{2}{3}, \tag{6.400}$$

$$= -\frac{1}{6}. \tag{6.401}$$

On the other hand,

$$\iint_D \left(\frac{\partial u_y}{\partial x} - \frac{\partial u_x}{\partial y}\right)\ dx\ dy = \int_0^1 \int_0^x (2y - 1)\ dy\ dx, \tag{6.402}$$

$$= \int_0^1 \left((y^2 - y)\big|_0^x \right)\ dx, \tag{6.403}$$

$$= \int_0^1 (x^2 - x)\ dx, \tag{6.404}$$

$$= \left(\frac{x^3}{3} - \frac{x^2}{2}\right)\Big|_0^1, \tag{6.405}$$

$$= \frac{1}{3} - \frac{1}{2}, \tag{6.406}$$

$$= -\frac{1}{6}. \tag{6.407}$$

6.7.2 Divergence theorem

Let us consider Eq. (6.300) in more detail. Let S be a closed surface, and V the region enclosed within it, then the divergence theorem is

$$\int_S \mathbf{u}^T \cdot \mathbf{n}\, dS = \int_V \nabla^T \cdot \mathbf{u}\, dV, \tag{6.408}$$

$$\int_S u_i n_i dS = \int_V \frac{\partial u_i}{\partial x_i} dV, \tag{6.409}$$

where dV an element of volume, dS is an element of the surface, and $\mathbf{n}$ (or n_i) is the outward unit normal to it. The divergence theorem is also known as Gauss's theorem. It extends to tensors of arbitrary order:

$$\int_S T_{ijk\ldots} n_i dS = \int_V \frac{\partial T_{ijk\ldots}}{\partial x_i} dV. \tag{6.410}$$

Note if $T_{ijk\ldots} = C$, then we get

$$\int_S n_i dS = 0. \tag{6.411}$$

The divergence theorem can be thought of as an extension of the familiar one-dimensional scalar result:

$$\phi(b) - \phi(a) = \int_a^b \frac{d\phi}{dx}\, dx. \tag{6.412}$$

Here the end points play the role of the surface integral, and the integral on x plays the role of the volume integral.

Example 6.18

Show that the divergence theorem is valid if

$$\mathbf{u} = x\, \mathbf{i} + y\, \mathbf{j} + 0\mathbf{k}, \tag{6.413}$$

and S is the closed surface which consists of a circular base and the hemisphere of unit radius with center at the origin and $z \geq 0$, that is,

$$x^2 + y^2 + z^2 = 1. \tag{6.414}$$

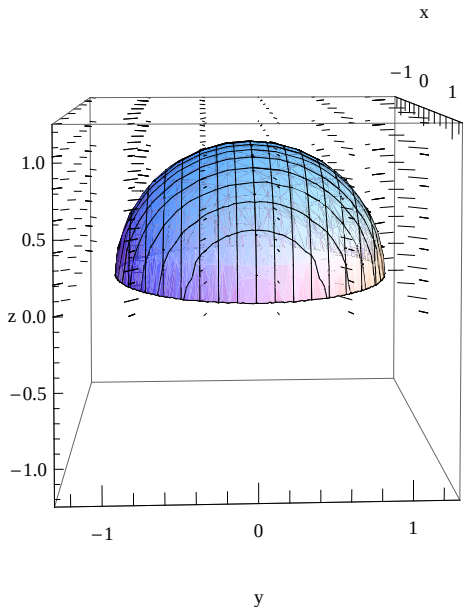

Figure 6.10: Sketch depicting $x^2 + y^2 + z^1 = 1,\ z \geq 0$ and vector field $\mathbf{u} = x\mathbf{i} + y\mathbf{j} + 0\mathbf{k}$.

In spherical coordinates, defined by

$$x = r\sin\theta\cos\phi, \tag{6.415}$$
$$y = r\sin\theta\sin\phi, \tag{6.416}$$
$$z = r\cos\theta, \tag{6.417}$$

the hemispherical surface is described by

$$r = 1. \tag{6.418}$$

A sketch of the surface of interest along with the vector field is shown in Figure 6.10.

We split the surface integral into two parts

$$\int_S \mathbf{u}^T \cdot \mathbf{n}\ dS = \int_B \mathbf{u}^T \cdot \mathbf{n}\ dS + \int_H \mathbf{u}^T \cdot \mathbf{n}\ dS, \tag{6.419}$$

where B is the base and H the curved surface of the hemisphere.

The first term on the right is zero since $\mathbf{n} = -\mathbf{k}$, and $\mathbf{u}^T \cdot \mathbf{n} = 0$ on B. In general, the unit normal pointing in the r direction can be shown to be

$$\mathbf{e}_r = \mathbf{n} = \sin\theta\cos\phi\mathbf{i} + \sin\theta\sin\phi\mathbf{j} + \cos\theta\mathbf{k}. \tag{6.420}$$

This is in fact the unit normal on H. Thus, on H, where $r = 1$, we have

$$\mathbf{u}^T \cdot \mathbf{n} = (x\mathbf{i} + y\mathbf{j} + 0\mathbf{k})^T \cdot (\sin\theta\cos\phi\mathbf{i} + \sin\theta\sin\phi\mathbf{j} + \cos\theta\mathbf{k}), \tag{6.421}$$
$$= (r\sin\theta\cos\phi\mathbf{i} + r\sin\theta\sin\phi\mathbf{j} + 0\mathbf{k})^T \cdot (\sin\theta\cos\phi\mathbf{i} + \sin\theta\sin\phi\mathbf{j} + \cos\theta\mathbf{k}), \tag{6.422}$$
$$= \underbrace{r}_{1}\sin^2\theta\cos^2\phi + \underbrace{r}_{1}\sin^2\theta\sin^2\phi, \tag{6.423}$$
$$= \sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi, \tag{6.424}$$
$$= \sin^2\theta, \tag{6.425}$$

$$\int_H \mathbf{u}^T \cdot \mathbf{n}\ dS = \int_0^{2\pi} \int_0^{\pi/2} \underbrace{\sin^2\theta}_{\mathbf{u}^T\cdot\mathbf{n}} \underbrace{(\sin\theta\ d\theta\ d\phi)}_{dS}, \tag{6.426}$$

$$= \int_0^{2\pi} \int_0^{\pi/2} \sin^3\theta\ d\theta\ d\phi, \tag{6.427}$$

$$= \int_0^{2\pi} \int_0^{\pi/2} \left(\frac{3}{4}\sin\theta - \frac{1}{4}\sin 3\theta\right)\ d\theta\ d\phi, \tag{6.428}$$

$$= 2\pi \int_0^{\pi/2} \left(\frac{3}{4}\sin\theta - \frac{1}{4}\sin 3\theta\right)\ d\theta, \tag{6.429}$$

$$= 2\pi\left(\frac{3}{4} - \frac{1}{12}\right), \tag{6.430}$$

$$= \frac{4}{3}\pi. \tag{6.431}$$

On the other hand, if we use the divergence theorem, we find that

$$\nabla^T \cdot \mathbf{u} = \frac{\partial}{\partial x}(x) + \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial z}(0) = 2, \tag{6.432}$$

so that

$$\int_V \nabla^T \cdot \mathbf{u}\ dV = 2\int_V dV = 2\frac{2}{3}\pi = \frac{4}{3}\pi, \tag{6.433}$$

since the volume of the hemisphere is $(2/3)\pi$.

6.7.3 Green's identities

Applying the divergence theorem, Eq. (6.409), to the vector $\mathbf{u} = \phi\nabla\psi$, we get

$$\int_S \phi(\nabla\psi)^T \cdot \mathbf{n}\ dS = \int_V \nabla^T \cdot (\phi\nabla\psi)\ dV, \tag{6.434}$$

$$\int_S \phi\frac{\partial\psi}{\partial x_i} n_i\ dS = \int_V \frac{\partial}{\partial x_i}\left(\phi\frac{\partial\psi}{\partial x_i}\right)\ dV. \tag{6.435}$$

From this, we get Green's first identity

$$\int_S \phi(\nabla\psi)^T \cdot \mathbf{n}\ dS = \int_V (\phi\nabla^2\psi + (\nabla\phi)^T \cdot \nabla\psi)\ dV, \tag{6.436}$$

$$\int_S \phi\frac{\partial\psi}{\partial x_i} n_i\ dS = \int_V \left(\phi\frac{\partial^2\psi}{\partial x_i\partial x_i} + \frac{\partial\phi}{\partial x_i}\frac{\partial\psi}{\partial x_i}\right)\ dV. \tag{6.437}$$

Interchanging ϕ and ψ in Eq. (6.436), we get

$$\int_S \psi(\nabla\phi)^T \cdot \mathbf{n}\ dS = \int_V (\psi\nabla^2\phi + (\nabla\psi)^T \cdot \nabla\phi)\ dV, \tag{6.438}$$

$$\int_S \psi\frac{\partial\phi}{\partial x_i} n_i\ dS = \int_V \left(\psi\frac{\partial^2\phi}{\partial x_i\partial x_i} + \frac{\partial\psi}{\partial x_i}\frac{\partial\phi}{\partial x_i}\right)\ dV. \tag{6.439}$$

Subtracting Eq. (6.438) from Eq. (6.436), we get Green's second identity

$$\int_S (\phi \nabla \psi - \psi \nabla \phi)^T \cdot \mathbf{n} \; dS = \int_V (\phi \nabla^2 \psi - \psi \nabla^2 \phi) \; dV, \tag{6.440}$$

$$\int_S \left(\phi \frac{\partial \psi}{\partial x_i} - \psi \frac{\partial \phi}{\partial x_i} \right) n_i \; dS = \int_V \left(\phi \frac{\partial^2 \psi}{\partial x_i \partial x_i} - \psi \frac{\partial^2 \phi}{\partial x_i \partial x_i} \right) \; dV \tag{6.441}$$

6.7.4 Stokes' theorem

Consider Stokes'[4] theorem. Let S be an open surface, and the curve C its boundary. Then

$$\int_S (\nabla \times \mathbf{u})^T \cdot \mathbf{n} \; dS = \oint_C \mathbf{u}^T \cdot d\mathbf{r}, \tag{6.442}$$

$$\int_S \epsilon_{ijk} \frac{\partial u_k}{\partial x_j} \; n_i \; dS = \oint_C u_i \; dr_i, \tag{6.443}$$

where $\mathbf{n}$ is the unit vector normal to the element dS, and $d\mathbf{r}$ an element of curve C.

Example 6.19

Evaluate

$$I = \int_S (\nabla \times \mathbf{u})^T \cdot \mathbf{n} \; dS, \tag{6.444}$$

using Stokes's theorem, where

$$\mathbf{u} = x^3 \; \mathbf{j} - (z+1) \; \mathbf{k}, \tag{6.445}$$

and S is the surface $z = 4 - 4x^2 - y^2$ for $z \geq 0$.

Using Stokes's theorem, the surface integral can be converted to a line integral along the boundary C which is the curve $4 - 4x^2 - y^2 = 0$.

$$I = \oint_C \mathbf{u}^T \cdot d\mathbf{r}, \tag{6.446}$$

$$= \oint \underbrace{(x^3 \; \mathbf{j} - (z+1) \; \mathbf{k})}_{\mathbf{u}^T} \cdot \underbrace{(dx \; \mathbf{i} + dy \; \mathbf{j})}_{d\mathbf{r}}, \tag{6.447}$$

$$= \int_C x^3 \; dy. \tag{6.448}$$

C can be represented by the parametric equations $x = \cos t$, $y = 2 \sin t$. This is easily seen by direct substitution on C:

$$4 - 4x^2 - y^2 = 4 - 4\cos^2 t - (2 \sin t)^2 = 4 - 4(\cos^2 t + \sin^2 t) = 4 - 4 = 0. \tag{6.449}$$

Thus, $dy = 2 \cos t \; dt$, so that

$$I = \int_0^{2\pi} \underbrace{\cos^3 t}_{x^3} \underbrace{(2 \cos t \; dt)}_{dy}, \tag{6.450}$$

[4] George Gabriel Stokes, 1819-1903, Irish-born English mathematician.

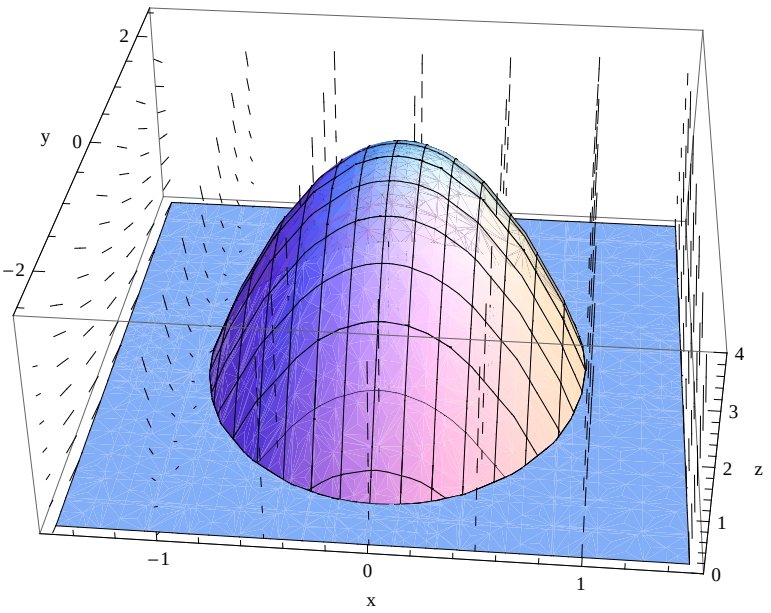

Figure 6.11: Sketch depicting $z = 4 - 4x^2 - y^2$ and vector field $\mathbf{u} = x^3\mathbf{j} - (z+1)\mathbf{k}$.

$$= 2\int_0^{2\pi} \cos^4 t \; dt, \tag{6.451}$$

$$= 2\int_0^{2\pi} \left(\frac{1}{8}\cos 4t + \frac{1}{2}\cos 2t + \frac{3}{8}\right) \; dt, \tag{6.452}$$

$$= 2\left(\frac{1}{32}\sin 4t + \frac{1}{4}\sin 2t + \frac{3}{8}t\right)\Big|_0^{2\pi}, \tag{6.453}$$

$$= \frac{3}{2}\pi. \tag{6.454}$$

A sketch of the surface of interest along with the vector field is shown in Figure 6.11. The curve C is on the boundary $z = 0$.

6.7.5 Leibniz's rule

If we consider an arbitrary moving volume $V(t)$ with a corresponding surface area $S(t)$ with surface volume elements moving at velocity w_k, Leibniz's rule, extended from the earlier Eq. (1.293), gives us a means to calculate the time derivatives of integrated quantities. For an arbitrary order tensor, it is

$$\frac{d}{dt}\int_{V(t)} T_{jk\ldots}(x_i,t)\; dV = \int_{V(t)} \frac{\partial T_{jk\ldots}(x_i,t)}{\partial t}\; dV + \int_{S(t)} n_m w_m T_{jk\ldots}(x_i,t)\; dS. \tag{6.455}$$

Note if $T_{jk\ldots}(x_i, t) = 1$, we get

$$\frac{d}{dt}\int_{V(t)} (1)\ dV = \int_{V(t)} \frac{\partial}{\partial t}(1)\ dV + \int_{S(t)} n_m w_m (1)\ dS, \tag{6.456}$$

$$\frac{dV}{dt} = \int_{S(t)} n_m w_m\ dS. \tag{6.457}$$

Here the volume changes due to the net surface motion. In one dimension $T_{jk\ldots}(x_i, t) = f(x, t)$ we get

$$\frac{d}{dt}\int_{x=a(t)}^{x=b(t)} f(x,t)\ dx = \int_{x=a(t)}^{x=b(t)} \frac{\partial f}{\partial t}\ dx + \frac{db}{dt} f(b(t), t) - \frac{da}{dt} f(a(t), t). \tag{6.458}$$

Problems

1. Find the angle between the planes

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

2. Find the curve of intersection of the cylinders $x^2 + y^2 = 1$ and $y^2 + z^2 = 1$. Determine also the radius of curvature of this curve at the points (0,1,0) and (1,0,1).

3. Show that for a curve $\mathbf{r}(t)$

$$\begin{aligned} \mathbf{t}^T \cdot \frac{d\mathbf{t}}{ds} \times \frac{d^2\mathbf{t}}{ds^2} &= \kappa^2 \tau, \\ \frac{\frac{d\mathbf{r}^T}{ds} \cdot \frac{d^2\mathbf{r}}{ds^2} \times \frac{d^3\mathbf{r}}{ds^3}}{\frac{d^2\mathbf{r}^T}{ds^2} \cdot \frac{d^2\mathbf{r}}{ds^2}} &= \tau, \end{aligned}$$

   where $\mathbf{t}$ is the unit tangent, s is the length along the curve, κ is the curvature, and τ is the torsion.

4. Find the equation for the tangent to the curve of intersection of $x = 2$ and $y = 1 + xz \sin y^2 z$ at the point $(2, 1, \pi)$.

5. Find the curvature and torsion of the curve $\mathbf{r}(t) = 2t\mathbf{i} + t^2\mathbf{j} + 2t^3\mathbf{k}$ at the point $(2, 1, 2)$.

6. Apply Stokes's theorem to the plane vector field $\mathbf{u}(x, y) = u_x\mathbf{i} + u_y\mathbf{j}$ and a closed curve enclosing a plane region. What is the result called? Use this result to find $\oint_C \mathbf{u}^T \cdot d\mathbf{r}$, where $\mathbf{u} = -y\mathbf{i} + x\mathbf{j}$ and the integration is counterclockwise along the sides C of the trapezoid with corners at (0,0), (2,0), (2,1), and (1,1).

7. Orthogonal bipolar coordinates (u, v, w) are defined by

$$\begin{aligned} x &= \frac{\alpha \sinh v}{\cosh v - \cos u}, \\ y &= \frac{\alpha \sin u}{\cosh v - \cos u}, \\ z &= w. \end{aligned}$$

   For $\alpha = 1$, plot some of the surfaces of constant x and y in the $u - v$ plane.

8. Using Cartesian index notation, show that

$$\nabla \times (\mathbf{u} \times \mathbf{v}) = (\mathbf{v}^T \cdot \nabla)\mathbf{u} - (\mathbf{u}^T \cdot \nabla)\mathbf{v} + \mathbf{u}(\nabla^T \cdot \mathbf{v}) - \mathbf{v}(\nabla^T \cdot \mathbf{u}),$$

where $\mathbf{u}$ and $\mathbf{v}$ are vector fields.

9. Consider two Cartesian coordinate systems: S with unit vectors $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, and S' with $(\mathbf{i}', \mathbf{j}', \mathbf{k}')$, where $\mathbf{i}' = \mathbf{i}$, $\mathbf{j}' = (\mathbf{j} - \mathbf{k})/\sqrt{2}$, $\mathbf{k}' = (\mathbf{j} + \mathbf{k})/\sqrt{2}$. The tensor $\mathbf{T}$ has the following components in S:

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

Find its components in S'.

10. Find the matrix $\mathbf{A}$ that operates on any vector of unit length in the x-y plane and turns it through an angle θ around the z-axis without changing its length. Show that $\mathbf{A}$ is orthogonal; that is that all of its columns are mutually orthogonal vectors of unit magnitude.

11. What is the unit vector normal to the plane passing through the points (1,0,0), (0,1,0) and (0,0,2)?

12. Prove the following identities using Cartesian index notation:

    (a) $(\mathbf{a} \times \mathbf{b})^{\mathrm{T}} \cdot \mathbf{c} = \mathbf{a}^{\mathrm{T}} \cdot (\mathbf{b} \times \mathbf{c})$,

    (b) $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a}^{\mathrm{T}} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a}^{\mathrm{T}} \cdot \mathbf{b})$,

    (c) $(\mathbf{a} \times \mathbf{b})^{\mathrm{T}} \cdot (\mathbf{c} \times \mathbf{d}) = ((\mathbf{a} \times \mathbf{b}) \times \mathbf{c})^{\mathrm{T}} \cdot \mathbf{d}$.

13. The position of a point is given by $\mathbf{r} = \mathbf{i}a\cos\omega t + \mathbf{j}b\sin\omega t$. Show that the path of the point is an ellipse. Find its velocity $\mathbf{v}$ and show that $\mathbf{r} \times \mathbf{v} =$ constant. Show also that the acceleration of the point is directed towards the origin and its magnitude is proportional to the distance from the origin.

14. System S is defined by the unit vectors $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. Another Cartesian system S' is defined by unit vectors $\mathbf{e}'_1$, $\mathbf{e}'_2$, and $\mathbf{e}'_3$ in directions $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ where

$$\begin{aligned} \mathbf{a} &= \mathbf{e}_1, \\ \mathbf{b} &= \mathbf{e}_2 - \mathbf{e}_3. \end{aligned}$$

    (a) Find $\mathbf{e}'_1$, $\mathbf{e}'_2$, $\mathbf{e}'_3$, (b) find the transformation array A_{ij}, (c) show that $\delta_{ij} = A_{ki}A_{kj}$ is satisfied, and (d) find the components of the vector $\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3$ in S'.

15. Use Green's theorem to calculate $\oint_C \mathbf{u}^T \cdot d\mathbf{r}$, where $\mathbf{u} = x^2\mathbf{i} + 2xy\mathbf{j}$, and C is the counterclockwise path around a rectangle with vertices at (0,0), (2,0), (0,4) and (2,4).

16. Derive an expression for the gradient, divergence, curl, and Laplacian operators in orthogonal paraboloidal coordinates

$$\begin{aligned} x &= uv\cos\theta, \\ y &= uv\sin\theta, \\ z &= \frac{1}{2}(u^2 - v^2). \end{aligned}$$

Determine the scale factors. Find $\nabla\phi$, $\nabla^T \cdot \mathbf{u}$, $\nabla \times \mathbf{u}$, and $\nabla^2\phi$ in this coordinate system.

17. Derive an expression for the gradient, divergence, curl and Laplacian operators in orthogonal parabolic cylindrical coordinates (u, v, w) where

$$\begin{aligned} x &= uv, \\ y &= \frac{1}{2}(u^2 - v^2), \\ z &= w, \end{aligned}$$

where $u \in [0, \infty)$, $v \in (-\infty, \infty)$, and $w \in (-\infty, \infty)$.

18. Consider orthogonal elliptic cylindrical coordinates (u, v, z) which are related to Cartesian coordinates (x, y, z) by

$$\begin{aligned} x &= a \cosh u \cos v \\ y &= a \sinh u \sin v \\ z &= z \end{aligned}$$

where $u \in [0, \infty)$, $v \in [0, 2\pi)$ and $z \in (-\infty, \infty)$. Determine $\nabla f, \nabla^T \cdot \mathbf{u}, \nabla \times \mathbf{u}$ and $\nabla^2 f$ in this system, where f is a scalar field and $\mathbf{u}$ is a vector field.

19. Determine a unit vector in the plane of the vectors $\mathbf{i} - \mathbf{j}$ and $\mathbf{j} + \mathbf{k}$ and perpendicular to the vector $\mathbf{i} - \mathbf{j} + \mathbf{k}$.

20. Determine a unit vector perpendicular to the plane of the vectors $\mathbf{a} = \mathbf{i} + 2\mathbf{j} - \mathbf{k}$, $\mathbf{b} = 2\mathbf{i} + \mathbf{j} + 0\mathbf{k}$.

21. Find the curvature and the radius of curvature of $y = a \sin x$ at the peaks and valleys.

22. Determine the unit vector normal to the surface $x^3 - 2xyz + z^3 = 0$ at the point (1,1,1).

23. Show using indicial notation that

$$\begin{aligned} \nabla \times \nabla \phi &= = 0, \\ \nabla^T \cdot \nabla \times \mathbf{u} &= 0 \\ \nabla(\mathbf{u}^T \cdot \mathbf{v}) &= (\mathbf{u}^T \cdot \nabla)\mathbf{v} + (\mathbf{v}^T \cdot \nabla)\mathbf{u} + \mathbf{u} \times (\nabla \times \mathbf{v}) + \mathbf{v} \times (\nabla \times \mathbf{u}), \\ \frac{1}{2}\nabla(\mathbf{u}^T \cdot \mathbf{u}) &= (\mathbf{u}^T \cdot \nabla)\mathbf{u} + \mathbf{u} \times (\nabla \times \mathbf{u}), \\ \nabla^T \cdot (\mathbf{u} \times \mathbf{v}) &= \mathbf{v}^T \cdot \nabla \times \mathbf{u} - \mathbf{u}^T \cdot \nabla \times \mathbf{v}, \\ \nabla \times (\nabla \times \mathbf{u}) &= \nabla(\nabla^T \cdot \mathbf{u}) - \nabla^2 \mathbf{u}, \\ \nabla \times (\mathbf{u} \times \mathbf{v}) &= (\mathbf{v}^T \cdot \nabla)\mathbf{u} - (\mathbf{u}^T \cdot \nabla)\mathbf{v} + \mathbf{u}(\nabla^T \cdot \mathbf{v}) - \mathbf{v}(\nabla^T \cdot \mathbf{u}). \end{aligned}$$

24. Show that the Laplacian operator $\frac{\partial^2}{\partial x_i \partial x_i}$ has the same form in S and S'.

25. If

$$T_{ij} = \begin{pmatrix} x_1 x_2^2 & 3x_3 & x_1 - x_2 \\ x_2 x_1 & x_1 x_3 & x_3^2 + 1 \\ 0 & 4 & 2x_2 - x_3 \end{pmatrix},$$

a) Evaluate T_{ij} at $P : (3, 1, 2)$,

b) find $T_{(ij)}$ and $T_{[ij]}$ at P,

c) find the associated dual vector d_i,

d) find the principal values and the orientations of each associated normal vector for the symmetric part of T_{ij} evaluated at P,

e) evaluate the divergence of T_{ij} at P,

f) evaluate the curl of the divergence of T_{ij} at P.

26. Consider the tensor
$$T_{ij} = \begin{pmatrix} 2 & -1 & 2 \\ 3 & 1 & 0 \\ 0 & 1 & 4 \end{pmatrix},$$
defined in a Cartesian coordinate system. Consider the vector associated with the plane whose normal points in the direction $(2, 5, -1)$. What is the magnitude of the component of the associated vector that is aligned with the normal to the plane?

27. Find the invariants of the tensor
$$T_{ij} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}.$$

28. Find the tangent to the curve of intersection of the surfaces $y^2 = x$ and $y = xy$ at $(x, y, z) = (1, 1, 1)$.

Chapter 7

Linear analysis

see Kaplan, Chapter 1,
see Friedman, Chapter 1, 2,
see Riley, Hobson, and Bence, Chapters 7, 10, 15,
see Lopez, Chapters 15, 31,
see Greenberg, Chapters 17 and 18,
see Wylie and Barrett, Chapter 13,
see Michel and Herget,
see Zeidler,
see Riesz and Nagy,
see Debnath and Mikusinski.

This chapter will introduce some more formal notions of what is known as linear analysis. We will generalize our notion of a vector; in addition to traditional vectors which exist within a space of finite dimension, we will see how what is known as function space can be thought of a vector space of infinite dimension. This chapter will also introduce some of the more formal notation of modern mathematics.

7.1 Sets

Consider two sets $\mathbb{A}$ and $\mathbb{B}$. We use the following notation

$x \in \mathbb{A}$,	x is an element of $\mathbb{A}$,
$x \notin \mathbb{A}$,	x is not an element of $\mathbb{A}$,
$\mathbb{A} = \mathbb{B}$,	$\mathbb{A}$ and $\mathbb{B}$ have the same elements,
$\mathbb{A} \subset \mathbb{B}$,	the elements of $\mathbb{A}$ also belong to $\mathbb{B}$,
$\mathbb{A} \cup \mathbb{B}$,	set of elements that belong to $\mathbb{A}$ or $\mathbb{B}$,
$\mathbb{A} \cap \mathbb{B}$,	set of elements that belong to $\mathbb{A}$ and $\mathbb{B}$, and
$\mathbb{A} - \mathbb{B}$,	set of elements that belong to $\mathbb{A}$ but not to $\mathbb{B}$.

If $\mathbb{A} \subset \mathbb{B}$, then $\mathbb{B} - \mathbb{A}$ is the *complement* of $\mathbb{A}$ in $\mathbb{B}$.

Some sets that are commonly used are:

$\mathbb{Z}$,	set of all integers,
$\mathbb{N}$,	set of all positive integers,
$\mathbb{Q}$,	set of all rational numbers,
$\mathbb{R}$,	set of all real numbers,
$\mathbb{R}_+$,	set of all non-negative real numbers, and
$\mathbb{C}$,	set of all complex numbers.

- An *interval* is a portion of the real line.
- An *open* interval (a, b) does not include the end points, so that if $x \in (a, b)$, then $a < x < b$. In set notation this is $\{x \in \mathbb{R} : a < x < b\}$ if x is real.
- A *closed* interval $[a, b]$ includes the end points. If $x \in [a, b]$, then $a \leq x \leq b$. In set notation this is $\{x \in \mathbb{R} : a \leq x \leq b\}$ if x is real.
- The complement of any open subset of $[a, b]$ is a closed set.
- A set $\mathbb{A} \subset \mathbb{R}$ is bounded from above if there exists a real number, called the *upper bound*, such that every $x \in \mathbb{A}$ is less than or equal to that number.
- The *least upper bound* or *supremum* is the minimum of all upper bounds.
- In a similar fashion, a set $\mathbb{A} \subset \mathbb{R}$ can be bounded from below, in which case it will have a *greatest lower bound* or *infimum*.
- A set which has no elements is the empty set $\{\}$, also known as the null set $\emptyset$. Note the set with 0 as the only element, 0, is not empty.
- A set that is either finite, or for which each element can be associated with a member of $\mathbb{N}$ is said to be *countable*. Otherwise the set is *uncountable*.
- An ordered pair is $P = (x, y)$, where $x \in \mathbb{A}$, and $y \in \mathbb{B}$. Then $P \in \mathbb{A} \times \mathbb{B}$, where the symbol $\times$ represents a Cartesian product. If $x \in \mathbb{A}$ and $y \in \mathbb{A}$ also, then we write $P = (x, y) \in \mathbb{A}^2$.
- A real *function* of a single variable can be written as $f : \mathbb{X} \to \mathbb{Y}$ or $y = f(x)$ where f maps $x \in \mathbb{X} \subset \mathbb{R}$ to $y \in \mathbb{Y} \subset \mathbb{R}$. For each x, there is only one y, though there may be more than one x that maps to a given y. The set $\mathbb{X}$ is called the *domain* of f, y the *image* of x, and the *range* the set of all images.

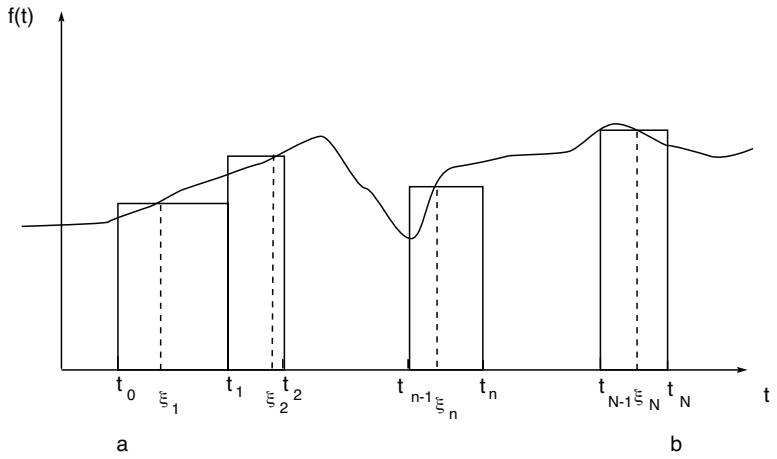

Figure 7.1: Riemann integration process.

7.2 Differentiation and integration

7.2.1 Fréchet derivative

An example of a Fréchet[1] derivative is the Jacobian derivative. It is a generalization of the ordinary derivative.

7.2.2 Riemann integral

Consider a function $f(t)$ defined in the interval $[a, b]$. Choose $t_1, t_2, \cdots, t_{N-1}$ such that

$$a = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = b. \tag{7.1}$$

Let $\xi_n \in [t_{n-1}, t_n]$, and

$$I_N = f(\xi_1)(t_1 - t_0) + f(\xi_2)(t_2 - t_1) + \cdots + f(\xi_N)(t_N - t_{N-1}). \tag{7.2}$$

Also let $\max_n |t_n - t_{n-1}| \to 0$ as $N \to \infty$. Then $I_N \to I$, where

$$I = \int_a^b f(t)\ dt. \tag{7.3}$$

If I exists and is independent of the manner of subdivision, then $f(t)$ is Riemann[2] integrable in $[a, b]$. The Riemann integration process is sketched in Fig. 7.1.

Example 7.1

Determine if the function $f(t)$ is Riemann integrable in $[0, 1]$ where

$$f(t) = \begin{cases} 0, & \text{if } t \text{ is rational,} \\ 1, & \text{if } t \text{ is irrational.} \end{cases} \tag{7.4}$$

[1]Maurice René Fréchet, 1878-1973, French mathematician.

[2]Georg Friedrich Bernhard Riemann, 1826-1866, Hanover-born German mathematician.

On choosing ξ_n rational, $I = 0$, but if ξ_n is irrational, then $I = 1$. So $f(t)$ is not Riemann integrable.

7.2.3 Lebesgue integral

Let us consider sets belonging to the interval $[a, b]$ where a and b are real scalars. The *covering* of a set is an open set which contains the given set; the covering will have a certain length. The outer measure of a set is the length of the smallest covering possible. The inner measure of the set is $(b - a)$ minus the outer measure of the complement of the set. If the two measures are the same, then the value is the *measure* and the set is *measurable*.

For the set $I = (a, b)$, the *measure* is $m(I) = |b - a|$. If there are two disjoint intervals $I_1 = (a, b)$ and $I_2 = (c, d)$. Then the measure of $I = I_1 \cup I_2$ is $m(I) = |b - a| + |c - d|$.

Consider again a function $f(t)$ defined in the interval $[a, b]$. Let the set

$$e_n = \{t \ : \ y_{n-1} \leq f(t) \leq y_n\}, \tag{7.5}$$

(e_n is the set of all t's for which $f(t)$ is bounded between two values, y_{n-1} and y_n). Also let the sum I_N be defined as

$$I_N = y_1 m(e_1) + y_2 m(e_2) + \cdots + y_N m(e_N). \tag{7.6}$$

Let $\max_n |y_n - y_{n-1}| \to 0$ as $N \to \infty$. Then $I_N \to I$, where

$$I = \int_a^b f(t) \ dt. \tag{7.7}$$

Here I is said to be the *Lebesgue*[3] integral of $f(t)$. The Lebesgue integration process is sketched in Fig. 7.2.

Example 7.2

To integrate the function in the previous example, we observe first that the set of rational and irrational numbers in [0,1] has measure zero and 1 respectively. Thus, from Eq. (7.6) the Lebesgue integral exists, and is equal to 1. Loosely speaking, the reason is that the rationals are not dense in $[0, 1]$ while the irrationals are dense in $[0, 1]$. That is to say every rational number exists in isolation from other rational numbers and surrounded by irrationals. Thus, the rationals exist as isolated points on the real line; these points have measure 0; The irrationals have measure 1 over the same interval; hence the integral is $I_N = y_1 m(e_1) + y_2 m(e_2) = 1(1) + 0(0) = 1$.

[3]Henri Lèon Lebesgue, 1875-1941, French mathematician.

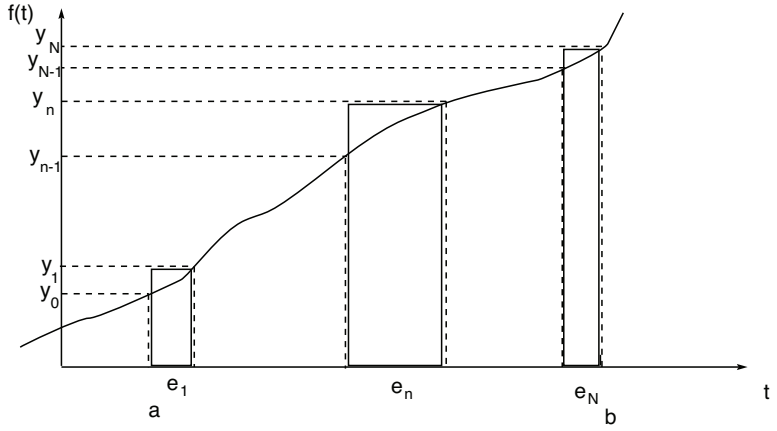

Figure 7.2: Lebesgue integration process.

The Riemann integral is based on the concept of the length of an interval, and the Lebesgue integral on the measure of a set. When both integrals exist, their values are the same. If the Riemann integral exists, the Lebesgue integral also exists. The converse is not necessarily true.

The importance of the distinction is subtle. It can be shown that certain integral operators which operate on Lebesgue integrable functions are guaranteed to generate a function which is also Lebesgue integrable. In contrast, certain operators operating on functions which are at most Riemann integrable can generate functions which are not Riemann integrable.

7.2.4 Cauchy principal value

If the integrand $f(x)$ of a definite integral contains a singularity at $x = x_o$ with $x_o \in (a, b)$, then the *Cauchy principal value* is

$$\oint_a^b f(x)dx = PV \int_a^b f(x)dx = \lim_{\epsilon \to 0} \left(\int_a^{x_o - \epsilon} f(x)dx + \int_{x_o + \epsilon}^b f(x)dx \right). \tag{7.8}$$

7.3 Vector spaces

A *field* $\mathbb{F}$ is typically a set of numbers which contains the sum, difference, product, and quotient (excluding division by zero) of any two numbers in the field.[4] Examples are the sets of rational numbers $\mathbb{Q}$, real numbers, $\mathbb{R}$, or complex numbers, $\mathbb{C}$. We will usually use only $\mathbb{R}$ or $\mathbb{C}$. Note the integers $\mathbb{Z}$ are not a field as the quotient of two integers is not necessarily an integer.

Consider a set $\mathbb{S}$ with two operations defined: addition of two elements (denoted by $+$) both belonging to the set, and multiplication of a member of the set by a scalar belonging

[4]More formally a field is what is known as a commutative ring with some special properties, not discussed here. What is known as function fields can also be defined.

to a field $\mathbb{F}$ (indicated by juxtaposition). Let us also require the set to be closed under the operations of addition and multiplication by a scalar, i.e. if $x \in \mathbb{S}$, $y \in \mathbb{S}$, and $\alpha \in \mathbb{F}$ then $x + y \in \mathbb{S}$, and $\alpha x \in \mathbb{S}$. Furthermore:

1. $\forall\ x, y \in \mathbb{S} :\ x + y = y + x$. For all elements x and y in $\mathbb{S}$, the addition operator on such elements is commutative.

2. $\forall\ x, y, z \in \mathbb{S} :\ (x + y) + z = x + (y + z)$. For all elements x and y in $\mathbb{S}$, the addition operator on such elements is associative.

3. $\exists\ 0 \in \mathbb{S} \mid \forall\ x \in \mathbb{S},\ x + 0 = x$: there exists a 0, which is an element of $\mathbb{S}$, such that for all x in $\mathbb{S}$ when the addition operator is applied to 0 and x, the original element x is yielded.

4. $\forall\ x \in \mathbb{S},\ \exists - x \in \mathbb{S} \mid x + (-x) = 0$. For all x in $\mathbb{S}$ there exists an element $-x$, also in $\mathbb{S}$, such that when added to x, yields the 0 element.

5. $\exists\ 1 \in \mathbb{F} \mid \forall\ x \in \mathbb{S},\ 1x = x$. There exists an element 1 in $\mathbb{F}$ such that for all x in $\mathbb{S}$,1 multiplying the element x yields the element x.

6. $\forall\ a, b \in \mathbb{F}, \forall x \in \mathbb{S},\ (a + b)x = ax + bx$. For all a and b which are in $\mathbb{F}$ and for all x which are in $\mathbb{S}$, the addition operator distributes onto multiplication.

7. $\forall\ a \in \mathbb{F}, \forall\ x, y \in \mathbb{S},\ a(x + y) = ax + ay$.

8. $\forall\ a, b \in \mathbb{F}, \forall\ x \in \mathbb{S},\ a(bx) = (ab)x$.

Such a set is called a *linear space* or *vector space* over the field $\mathbb{F}$, and its elements are called *vectors*. We will see that our definition is inclusive enough to include elements which are traditionally thought of as vectors (in the sense of a directed line segment), and some which are outside of this tradition. Note that typical vector elements x and y are no longer indicated in bold. However, they are in general *not* scalars, though in special cases, they can be.

The element $0 \in \mathbb{S}$ is called the *null vector*. Examples of vector spaces $\mathbb{S}$ over the field of real numbers (i.e. $\mathbb{F} : \mathbb{R}$) are:

1. $\mathbb{S} : \mathbb{R}^1$. Set of real numbers, $x = x_1$, with addition and scalar multiplication defined as usual; also known as $\mathbb{S} : \mathbb{R}$.

2. $\mathbb{S} : \mathbb{R}^2$. Set of ordered pairs of real numbers, $x = (x_1, x_2)^T$, with addition and scalar multiplication defined as:

$$x + y = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix} = (x_1 + y_1, x_2 + y_2)^T, \tag{7.9}$$

$$\alpha x = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \end{pmatrix} = (\alpha x_1, \alpha x_2)^T, \tag{7.10}$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1, x_2)^T \in \mathbb{R}^2, \qquad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (y_1, y_2)^T \in \mathbb{R}^2, \qquad \alpha \in \mathbb{R}^1. \quad (7.11)$$

Note $\mathbb{R}^2 = \mathbb{R}^1 \times \mathbb{R}^1$, where the symbol $\times$ represents a Cartesian product.

3. $\mathbb{S} : \mathbb{R}^N$. Set of N real numbers, $x = (x_1, \cdots, x_N)^T$, with addition and scalar multiplication defined similar to that just defined in $\mathbb{R}^2$.

4. $\mathbb{S} : \mathbb{R}^\infty$. Set of an infinite number of real numbers, $x = (x_1, x_2, \cdots)^T$, with addition and scalar multiplication defined similar to those defined for $\mathbb{R}^N$. Note, one can interpret functions, e.g. $x = 3t^2 + t, t \in \mathbb{R}^1$ to generate vectors $x \in \mathbb{R}^\infty$.

5. $\mathbb{S} : \mathbb{C}$. Set of all complex numbers $z = z_1$, with $z_1 = a_1 + ib_1; a_1, b_1 \in \mathbb{R}^1$.

6. $\mathbb{S} : \mathbb{C}^2$. Set of all ordered pairs of complex numbers $z = (z_1, z_2)^T$, with $z_1 = a_1 + ib_1, z_2 = a_2 + ib_2; a_1, a_2, b_1, b_2 \in \mathbb{R}^1$.

7. $\mathbb{S} : \mathbb{C}^N$. Set of N complex numbers, $z = (z_1, \cdots, z_N)^T$.

8. $\mathbb{S} : \mathbb{C}^\infty$. Set of an infinite number of complex numbers, $z = (z_1, z_2, \cdots)^T$. Scalar complex functions give rise to sets in $\mathbb{C}^\infty$.

9. $\mathbb{S} : \mathbb{M}$. Set of all $M \times N$ matrices with addition and multiplication by a scalar defined as usual, and $M \in \mathbb{N}$, $N \in \mathbb{N}$.

10. $\mathbb{S} : C[a, b]$ Set of real-valued continuous functions, $x(t)$ for $t \in [a, b] \in \mathbb{R}^1$ with addition and scalar multiplication defined as usual.

11. $\mathbb{S} : C^N[a, b]$ Set of real-valued functions $x(t)$ for $t \in [a, b]$ with continuous N^{th} derivative with addition and scalar multiplication defined as usual; $N \in \mathbb{N}$.

12. $\mathbb{S} : \mathbb{L}_2[a, b]$ Set of real-valued functions $x(t)$ such that $x(t)^2$ is Lebesgue integrable in $t \in [a, b] \in \mathbb{R}^1, a < b$, with addition and multiplication by a scalar defined as usual. Note that the integral must be finite.

13. $\mathbb{S} : \mathbb{L}_p[a, b]$ Set of real-valued functions $x(t)$ such that $|x(t)|^p$, $p \in [1, \infty)$, is Lebesgue integrable in $t \in [a, b] \in \mathbb{R}^1, a < b$, with addition and multiplication by a scalar defined as usual. Note that the integral must be finite.

14. $\mathbb{S} : \overline{\mathbb{L}}_p[a, b]$ Set of complex-valued functions $x(t)$ such that $|x(t)|^p$, $p \in [1, \infty) \in \mathbb{R}^1$, is Lebesgue integrable in $t \in [a, b] \in \mathbb{R}^1, a < b$, with addition and multiplication by a scalar defined as usual.

15. $\mathbb{S} : \mathbb{W}_2^1(G)$, Set of real-valued functions $u(x)$ such that $u(x)^2$ and $\sum_{n=1}^{N} (\partial u/\partial x_n)^2$ are Lebesgue integrable in G, where $x \in G \in \mathbb{R}^N, N \in \mathbb{N}$. This is an example of a Sobolov[5] space, which is useful in variational calculus and the finite element method. Sobolov space $W_2^1(G)$ is to Lebesgue space $\mathbb{L}_2[a,b]$ as the real space $\mathbb{R}^1$ is to the rational space $\mathbb{Q}^1$. That is Sobolov space allows a broader class of functions to be solutions to physical problems. See Zeidler.

16. $\mathbb{S} : \mathbb{P}^N$ Set of all polynomials of degree $\leq N$ with addition and multiplication by a scalar defined as usual; $N \in \mathbb{N}$.

Some examples of sets that are *not* vector spaces are $\mathbb{Z}$ and $\mathbb{N}$ over the field $\mathbb{R}$ for the same reason that they do not form a field, namely that they are not closed over the multiplication operation.

- $\mathbb{S}'$ is a *subspace* of $\mathbb{S}$ if $\mathbb{S}' \subset \mathbb{S}$, and $\mathbb{S}'$ is itself a vector space. For example $\mathbb{R}^2$ is a subspace of $\mathbb{R}^3$.

- If $\mathbb{S}_1$ and $\mathbb{S}_2$ are subspaces of $\mathbb{S}$, then $\mathbb{S}_1 \cap \mathbb{S}_2$ is also a subspace. The set $\mathbb{S}_1 + \mathbb{S}_2$ of all $x_1 + x_2$ with $x_1 \in \mathbb{S}_1$ and $x_2 \in \mathbb{S}_2$ is also a subspace of $\mathbb{S}$.

- If $\mathbb{S}_1 + \mathbb{S}_2 = \mathbb{S}$, and $\mathbb{S}_1 \cap \mathbb{S}_2 = \{0\}$, then $\mathbb{S}$ is the *direct sum* of $\mathbb{S}_1$ and $\mathbb{S}_2$, written as $\mathbb{S} = \mathbb{S}_1 \oplus \mathbb{S}_2$.

- If $x_1, x_2, \cdots, x_N$ are elements of a vector space $\mathbb{S}$ and $\alpha_1, \alpha_2, \cdots, \alpha_N$ belong to the field $\mathbb{F}$, then $x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_N x_N \in \mathbb{S}$ is a *linear combination*.

- Vectors $x_1, x_2, \cdots, x_N$ for which it is possible to have $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_N x_N = 0$ where the scalars α_n are not all zero, are said to be *linearly dependent*. Otherwise they are *linearly independent*.

- For $M \leq N$, the set of all linear combinations of M vectors $\{x_1, x_2, \cdots, x_M\}$ of a vector space constitute a subspace of an N-dimensional vector space.

- A set of N linearly independent vectors in an N-dimensional vector space is said to *span* the space.

- If the vector space $\mathbb{S}$ contains a set of N linearly independent vectors, and any set with $(N+1)$ elements is linearly dependent, then the space is said to be *finite dimensional*, and N is the *dimension* of the space. If N does not exist, the space is *infinite dimensional*.

- A *basis* of a finite dimensional space of dimension N is a set of N linearly independent vectors $\{u_1, u_2, \ldots, u_N\}$. All elements of the vector space can be represented as linear combinations of the basis vectors.

[5]Sergei Lvovich Sobolev, 1908-1989, St. Petersburg-born Russian physicist and mathematician.

- A set of vectors in a linear space $\mathbb{S}$ is *convex* iff $\forall x, y \in \mathbb{S}$ and $\alpha \in [0,1] \in \mathbb{R}^1$ implies $\alpha x + (1-\alpha)y \in \mathbb{S}$. For example if we consider $\mathbb{S}$ to be a subspace of $\mathbb{R}^2$, that is a region of the x, y plane, $\mathbb{S}$ is convex if for any two points in $\mathbb{S}$, all points on the line segment between them also lie in $\mathbb{S}$. Spaces with lobes are not convex. Functions f are convex iff the space on which they operate are convex and if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)\ \forall\ x, y \in \mathbb{S}, \alpha \in [0,1] \in \mathbb{R}^1$.

7.3.1 Normed spaces

The *norm* $||x||$ of a vector $x \in \mathbb{S}$ is a real number that satisfies the following properties:

1. $||x|| \geq 0$,
2. $||x|| = 0$ if and only if $x = 0$,
3. $||\alpha x|| = |\alpha|\ ||x||, \quad \alpha \in \mathbb{C}^1$, and
4. $||x + y|| \leq ||x|| + ||y||$, (triangle or Minkowski[6] inequality).

The norm is a natural generalization of the length of a vector. All properties of a norm can be cast in terms of ordinary finite dimensional Euclidean vectors, and thus have geometrical interpretations. The first property says length is greater than or equal to zero. The second says the only vector with zero length is the zero vector. The third says the length of a scalar multiple of a vector is equal to the magnitude of the scalar times the length of the original vector. The Minkowski inequality is easily understood in terms of vector addition. If we add vectorially two vectors x and y, we will get a third vector whose length is less than or equal to the sum of the lengths of the original two vectors. We will get equality when x and y point in the same direction. The interesting generalization is that these properties hold for the norms of *functions* as well as ordinary geometric vectors.

Examples of norms are:

1. $x \in \mathbb{R}^1$, $||x|| = |x|$. This space is also written as $\ell_1(\mathbb{R}^1)$ or in abbreviated form ℓ_1^1. The subscript on ℓ in either case denotes the type of norm; the superscript in the second form denotes the dimension of the space. Another way to denote this norm is $||x||_1$.
2. $x \in \mathbb{R}^2$, $x = (x_1, x_2)^T$, the Euclidean norm $||x|| = ||x||_2 = +\sqrt{x_1^2 + x_2^2} = +\sqrt{x^T x}$. We can call this normed space $\mathbb{E}^2$, or $\ell_2(\mathbb{R}^2)$, or ℓ_2^2.
3. $x \in \mathbb{R}^N$, $x = (x_1, x_2, \cdots, x_N)^T$, $||x|| = ||x||_2 = +\sqrt{x_1^2 + x_2^2 + \cdots + x_N^2} = +\sqrt{x^T x}$. We can call this norm the Euclidean norm and the normed space Euclidean $\mathbb{E}^N$, or $\ell_2(\mathbb{R}^N)$ or ℓ_2^N.
4. $x \in \mathbb{R}^N$, $x = (x_1, x_2, \cdots, x_N)^T$, $||x|| = ||x||_1 = |x_1| + |x_2| + \cdots + |x_N|$. This is also $\ell_1(\mathbb{R}^N)$ or ℓ_1^N.

[6] Hermann Minkowski, 1864-1909, Russian/Lithuanian-born German-based mathematician and physicist.

5. $x \in \mathbb{R}^N$, $x = (x_1, x_2, \cdots, x_N)^T$, $||x|| = ||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_N|^p)^{1/p}$, where $1 \leq p < \infty$. This space is called or $\ell_p(\mathbb{R}^N)$ or ℓ_p^N.

6. $x \in \mathbb{R}^N$, $x = (x_1, x_2, \cdots, x_N)^T$, $||x|| = ||x||_\infty = \max_{1 \leq n \leq N} |x_n|$. This space is called $\ell_\infty(\mathbb{R}^N)$ or ℓ_∞^N.

7. $x \in \mathbb{C}^N$, $x = (x_1, x_2, \cdots, x_N)^T$, $||x|| = ||x||_2 = +\sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_N|^2} = +\sqrt{\overline{x}^T x}$. This space is described as $\ell_2(\mathbb{C}^N)$.

8. $x \in C[a,b]$, $||x|| = \max_{a \leq t \leq b} |x(t)|$; $t \in [a,b] \in \mathbb{R}^1$.

9. $x \in C^1[a,b]$, $||x|| = \max_{a \leq t \leq b} |x(t)| + \max_{a \leq t \leq b} |x'(t)|$; $t \in [a,b] \in \mathbb{R}^1$.

10. $x \in \mathbb{L}_2[a,b]$, $||x|| = ||x||_2 = +\sqrt{\int_a^b x(t)^2 \, dt}$; $t \in [a,b] \in \mathbb{R}^1$.

11. $x \in \mathbb{L}_p[a,b]$, $||x|| = ||x||_p = + \left(\int_a^b |x(t)|^p \, dt \right)^{1/p}$; $t \in [a,b] \in \mathbb{R}^1$.

12. $x \in \overline{\mathbb{L}}_2[a,b]$, $||x|| = ||x||_2 = +\sqrt{\int_a^b |x(t)|^2 \, dt} = +\sqrt{\int_a^b \overline{x(t)} x(t) \, dt}$; $t \in [a,b] \in \mathbb{R}^1$.

13. $x \in \overline{\mathbb{L}}_p[a,b]$, $||x|| = ||x||_p = + \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} = + \left(\int_a^b \left(\overline{x(t)} x(t) \right)^{p/2} \, dt \right)^{1/p}$; $t \in [a,b] \in \mathbb{R}^1$.

14. $u \in \mathbb{W}_2^1(G), ||u|| = ||u||_{1,2} = +\sqrt{\int_G \left(u(x)u(x) + \sum_{n=1}^N (\partial u / \partial x_n)(\partial u / \partial x_n) \right) \, dx}$; $x \in G \in \mathbb{R}^N, u \in \mathbb{L}_2(G)$, $\partial u / \partial x_n \in \mathbb{L}_2(G)$. This is an example of a Sobolov space which is useful in variational calculus and the finite element method.

Some additional notes on properties of norms include

- A vector space in which a norm is defined is called a *normed vector space*.
- The *metric* or *distance* between x and y is defined by $d(x,y) = ||x - y||$. This a natural metric induced by the norm. Thus, $||x||$ is the distance between x and the null vector.
- The diameter of a set of vectors is the supremum (i.e. least upper bound) of the distance between any two vectors of the set.
- Let $\mathbb{S}_1$ and $\mathbb{S}_2$ be subsets of a normed vector space $\mathbb{S}$ such that $\mathbb{S}_1 \subset \mathbb{S}_2$. Then $\mathbb{S}_1$ is *dense* in $\mathbb{S}_2$ if for every $x^{(2)} \in \mathbb{S}_2$ and every $\epsilon > 0$, there is a $x^{(1)} \in \mathbb{S}_1$ for which $||x^{(2)} - x^{(1)}|| < \epsilon$.
- A *sequence* $x^{(1)}, x^{(2)}, \cdots \in \mathbb{S}$, where $\mathbb{S}$ is a normed vector space, is a *Cauchy*[7] sequence if for every $\epsilon > 0$ there exists a number N_ϵ such that $||x^{(m)} - x^{(n)}|| < \epsilon$ for every m and n greater than N_ϵ.

[7] Augustin-Louis Cauchy, 1789-1857, French mathematician and physicist.

- The sequence $x^{(1)}, x^{(2)}, \cdots \in \mathbb{S}$, where $\mathbb{S}$ is a normed vector space, converges if there exists an $x \in \mathbb{S}$ such that $\lim_{n\to\infty} ||x^{(n)} - x|| = 0$. Then x is the *limit point* of the sequence, and we write $\lim_{n\to\infty} x^{(n)} = x$ or $x^{(n)} \to x$.
- Every convergent sequence is a Cauchy sequence, but the converse is not true.
- A normed vector space $\mathbb{S}$ is *complete* if every Cauchy sequence in $\mathbb{S}$ is convergent, i.e. if $\mathbb{S}$ contains all the limit points.
- A complete normed vector space is also called a *Banach*[8] space.
- It can be shown that every finite dimensional normed vector space is complete.
- Norms $||\cdot||_n$ and $||\cdot||_m$ in $\mathbb{S}$ are *equivalent* if there exist $a, b > 0$ such that, for any $x \in \mathbb{S}$,
$$a||x||_m \leq ||x||_n \leq b||x||_m. \tag{7.12}$$
- In a finite dimensional vector space, any norm is equivalent to any other norm. So, the convergence of a sequence in such a space does not depend on the choice of norm.

We recall that if $z \in \mathbb{C}^1$, then we can represent z as $z = a + ib$ where $a \in \mathbb{R}^1, b \in \mathbb{R}^1$; further, the complex conjugate of z is represented as $\overline{z} = a - ib$. It can be easily shown for $z_1 \in \mathbb{C}^1, z_2 \in \mathbb{C}^1$ that

- $\overline{(z_1 + z_2)} = \overline{z_1} + \overline{z_2}$,
- $\overline{(z_1 - z_2)} = \overline{z_1} - \overline{z_2}$,
- $\overline{z_1 z_2} = \overline{z_1}\ \overline{z_2}$, and
- $\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\overline{z_1}}{\overline{z_2}}$.

We also recall that the modulus of z, $|z|$ has the following properties:

$$\begin{aligned} |z|^2 &= z\overline{z}, &(7.13)\\ &= (a+ib)(a-ib), &(7.14)\\ &= a^2 + iab - iab - i^2b^2, &(7.15)\\ &= a^2 + b^2 \geq 0. &(7.16)\end{aligned}$$

Example 7.3

Consider $x \in \mathbb{R}^3$ and take

$$x = \begin{pmatrix} 1 \\ -4 \\ 2 \end{pmatrix}. \tag{7.17}$$

[8]Stefan Banach, 1892-1945, Polish mathematician.

Find the norm if $x \in \ell_1^3$ (absolute value norm), $x \in \ell_2^3$ (Euclidean norm), if $x = \ell_3^3$ (another norm), and if $x \in \ell_\infty^3$ (maximum norm).

By the definition of the absolute value norm for $x \in \ell_1^3$,

$$||x|| = ||x||_1 = |x_1| + |x_2| + |x_3|, \tag{7.18}$$

we get

$$||x||_1 = |1| + |-4| + |2| = 1 + 4 + 2 = 7. \tag{7.19}$$

Now consider the Euclidean norm for $x \in \ell_2^3$. By the definition of the Euclidean norm,

$$||x|| = ||x||_2 + \sqrt{x_1^2 + x_2^2 + x_3^2}, \tag{7.20}$$

we get

$$||x||_2 = +\sqrt{1^2 + (-4)^2 + 2^2} = \sqrt{1 + 16 + 4} = +\sqrt{21} \sim 4.583. \tag{7.21}$$

Since the norm is Euclidean, this is the ordinary length of the vector.

For the norm, $x \in \ell_3^3$, we have

$$||x|| = ||x||_3 = + \left(|x_1|^3 + |x_2|^3 + |x_3|^3 \right)^{1/3}, \tag{7.22}$$

so

$$||x||_3 = + \left(|1|^3 + |-4|^3 + |2|^3 \right)^{1/3} = (1 + 64 + 8)^{1/3} \sim 4.179 \tag{7.23}$$

For the maximum norm, $x \in \ell_\infty^3$, we have

$$||x|| = ||x||_\infty = \lim_{p \to \infty} + \left(|x_1|^p + |x_2|^p + |x_3|^p \right)^{1/p}, \tag{7.24}$$

so

$$||x||_\infty = \lim_{p \to \infty} + \left(|1|^p + |-4|^p + |2|^p \right)^{1/p} = 4. \tag{7.25}$$

This selects the magnitude of the component of x whose magnitude is maximum. Note that as p increases the norm of the vector decreases.

Example 7.4

For $x \in \ell_2(\mathbb{C}^2)$, find the norm of

$$x = \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} 0 + 1i \\ 1 + 0i \end{pmatrix}. \tag{7.26}$$

The definition of the space defines the norm is a 2 norm ("Euclidean"):

$$||x|| = ||x||_2 = +\sqrt{\overline{x}^T x} = +\sqrt{\overline{x_1} x_1 + \overline{x_2} x_2} = \sqrt{|x_1|^2 + |x_2|^2}, \tag{7.27}$$

so

$$||x||_2 = +\sqrt{\begin{pmatrix} \overline{0 + 1i} & \overline{1 + 0i} \end{pmatrix} \begin{pmatrix} 0 + 1i \\ 1 + 0i \end{pmatrix}}, \tag{7.28}$$

$$||x||_2 = +\sqrt{\overline{(0+1i)}(0+1i) + \overline{(1+0i)}(1+0i)} = +\sqrt{(0-1i)(0+1i) + (1-0i)(1+0i)}, \tag{7.29}$$

$$||x||_2 = +\sqrt{-i^2+1} = +\sqrt{2}. \tag{7.30}$$

Note that if we were negligent in the use of the conjugate and defined the norm as $||x||_2 = +\sqrt{x^T x}$, we would obtain

$$||x||_2 = +\sqrt{x^T x} = +\sqrt{\begin{pmatrix} i & 1 \end{pmatrix}\begin{pmatrix} i \\ 1 \end{pmatrix}} = +\sqrt{i^2+1} = +\sqrt{-1+1} = 0! \tag{7.31}$$

This violates the property of the norm that $||x|| > 0$ if $x \neq 0$!

Example 7.5

Consider $x \in \mathbb{L}_2[0,1]$ where $x(t) = 2t; \quad t \in [0,1] \in \mathbb{R}^1$. Find $||x||$.

By the definition of the norm for this space, we have

$$||x|| = ||x||_2 = +\sqrt{\int_0^1 x^2(t)\ dt}, \tag{7.32}$$

$$||x||_2^2 = \int_0^1 x(t)x(t)\ dt = \int_0^1 (2t)(2t)\ dt = 4\int_0^1 t^2\ dt = 4\left(\frac{t^3}{3}\right)\Big|_0^1, \tag{7.33}$$

$$||x||_2^2 = 4\left(\frac{1^3}{3} - \frac{0^3}{3}\right) = \frac{4}{3}, \tag{7.34}$$

$$||x||_2 = \frac{2\sqrt{3}}{3} \sim 1.1547. \tag{7.35}$$

Example 7.6

Consider $x \in \overline{\mathbb{L}}_3[-2,3]$ where $x(t) = 1 + 2it; t \in [-2,3] \in \mathbb{R}^1$. Find $||x||$.

By the definition of the norm we have

$$||x|| = ||x||_3 = +\left(\int_{-2}^3 |1+2it|^3\ dt\right)^{1/3}, \tag{7.36}$$

$$||x||_3 = +\left(\int_{-2}^3 \left(\overline{(1+2it)}\,(1+2it)\right)^{3/2}\ dt\right)^{1/3}, \tag{7.37}$$

$$||x||_3^3 = \int_{-2}^3 \left(\overline{(1+2it)}\,(1+2it)\right)^{3/2}\ dt, \tag{7.38}$$

$$||x||_3^3 = \int_{-2}^3 \left((1-2it)\,(1+2it)\right)^{3/2}\ dt, \tag{7.39}$$

$$||x||_3^3 = \int_{-2}^{3} \left(1+4t^2\right)^{3/2}\, dt, \tag{7.40}$$

$$||x||_3^3 = \left(\sqrt{1+4t^2}\left(\frac{5t}{8}+t^3\right)+\frac{3}{16}\sinh^{-1}(2t)\right)\Bigg|_{-2}^{3}, \tag{7.41}$$

$$||x||_3^3 = \frac{37\sqrt{17}}{4}+\frac{3\sinh^{-1}(4)}{16}+\frac{3}{16}\left(154\sqrt{17}+\sinh^{-1}(6)\right) \sim 214.638, \tag{7.42}$$

$$||x||_3 \sim 5.98737. \tag{7.43}$$

Example 7.7

Consider $x \in \overline{\mathbb{L}}_p[a,b]$ where $x(t) = c;\quad t \in [a,b] \in \mathbb{R}^1, c \in \mathbb{C}^1$. Find $||x||$.

Let us take the complex constant $c = \alpha + i\beta$, $\alpha \in \mathbb{R}^1, \beta \in \mathbb{R}^1$. Then

$$|c| = \left(\alpha^2+\beta^2\right)^{1/2}. \tag{7.44}$$

Now

$$||x|| = ||x||_p = \left(\int_a^b |x(t)|^p\, dt\right)^{1/p}, \tag{7.45}$$

$$||x||_p = \left(\int_a^b \left(\alpha^2+\beta^2\right)^{p/2}\, dt\right)^{1/p}, \tag{7.46}$$

$$||x||_p = \left(\left(\alpha^2+\beta^2\right)^{p/2}\int_a^b dt\right)^{1/p}, \tag{7.47}$$

$$||x||_p = \left(\left(\alpha^2+\beta^2\right)^{p/2}(b-a)\right)^{1/p}, \tag{7.48}$$

$$||x||_p = \left(\alpha^2+\beta^2\right)^{1/2}(b-a)^{1/p}, \tag{7.49}$$

$$||x||_p = |c|(b-a)^{1/p}. \tag{7.50}$$

Note the norm is proportional to the magnitude of the complex constant c. For finite p, it also increases with the extent of the domain $b-a$. For infinite p, it is independent of the length of the domain, and simply selects the value $|c|$. This is consistent with the norm in $\overline{\mathbb{L}}_\infty$ selecting the maximum value of the function.

Example 7.8

Consider $x \in \mathbb{L}_p[0,b]$ where $x(t) = 2t^2;\quad t \in [0,b] \in \mathbb{R}^1$. Find $||x||$.

Now

$$||x|| = ||x||_p = \left(\int_0^b |x(t)|^p \, dt \right)^{1/p}, \tag{7.51}$$

$$||x||_p = \left(\int_0^b |2t^2|^p \, dt \right)^{1/p}, \tag{7.52}$$

$$||x||_p = \left(\int_0^b 2^p t^{2p} \, dt \right)^{1/p}, \tag{7.53}$$

$$||x||_p = \left(\left(\frac{2^p t^{2p+1}}{2p+1} \right) \Big|_0^b \right)^{1/p}, \tag{7.54}$$

$$||x||_p = \left(\frac{2^p b^{2p+1}}{2p+1} \right)^{1/p}, \tag{7.55}$$

$$||x||_p = \frac{2b^{\frac{2p+1}{p}}}{(2p+1)^{1/p}} \tag{7.56}$$

Note as $p \to \infty$ that $(2p+1)^{1/p} \to 1$, and $(2p+1)/p \to 2$, so

$$\lim_{p \to \infty} ||x|| = 2b^2. \tag{7.57}$$

This is the maximum value of $x(t) = 2t^2$ in $t \in [0, b]$, as expected.

Example 7.9

Consider $u \in \mathbb{W}_2^1(G)$ with $u(x) = 2x^4$; $x \in [0, 3] \in \mathbb{R}^1$. Find $||u||$.

Here we require $u \in \mathbb{L}_2[0, 3]$ and $\partial u / \partial x \in \mathbb{L}_2[0, 3]$, which for our choice of u, is satisfied. The formula for the norm in $\mathbb{W}_2^1[0, 3]$ is

$$||u|| = ||u||_{1,2} = +\sqrt{\int_0^3 \left(u(x)u(x) + \frac{du}{dx}\frac{du}{dx} \right) \, dx}, \tag{7.58}$$

$$||u||_{1,2} = +\sqrt{\int_0^3 \left((2x^4)(2x^4) + (8x^3)(8x^3) \right) \, dx}, \tag{7.59}$$

$$||u||_{1,2} = +\sqrt{\int_0^3 \left(4x^8 + 64x^6 \right) \, dx}, \tag{7.60}$$

$$||u||_{1,2} = +\sqrt{\left(\frac{4x^9}{9} + \frac{64x^7}{7} \right) \Big|_0^3} = 54\sqrt{\frac{69}{7}} \sim 169.539. \tag{7.61}$$

Example 7.10

Consider the sequence of vectors $\{x_{(1)},\ x_{(2)}, \ldots\} \in \mathbb{Q}^3$, where $\mathbb{Q}^3$ is the space of rational numbers over the field of rational numbers, and

$$\begin{aligned} x_{(1)} &= (1,3,0) = \left(x_{(1)1},\ x_{(1)2},\ x_{(1)3}\right), && (7.62)\\ x_{(2)} &= \left(\frac{1}{1+1},3,0\right) = \left(\frac{1}{2},3,0\right), && (7.63)\\ x_{(3)} &= \left(\frac{1}{1+\frac{1}{2}},3,0\right) = \left(\frac{2}{3},3,0\right), && (7.64)\\ x_{(4)} &= \left(\frac{1}{1+\frac{2}{3}},3,0\right) = \left(\frac{3}{5},3,0\right), && (7.65)\\ &\vdots && (7.66)\\ x_{(n)} &= \left(\frac{1}{1+x_{(n-1)1}},3,0\right), && (7.67)\\ &\vdots \end{aligned}$$

for $n \geq 2$. Does this sequence have a limit point in $\mathbb{Q}^3$? Is this a Cauchy sequence?

Consider the first term only; the other two are trivial. The series has converged when the n^{th} term is equal to the $(n-1)^{th}$ term:

$$x_{(n-1)1} = \frac{1}{1+x_{(n-1)1}}. \tag{7.68}$$

Rearranging, it is found that

$$x_{(n-1)1}^2 + x_{(n-1)1} - 1 = 0. \tag{7.69}$$

Solving, one finds that

$$x_{(n-1)1} = \frac{-1 \pm \sqrt{5}}{2}. \tag{7.70}$$

We find from numerical experimentation that it is the "+" root to which x_1 converges:

$$\lim_{n\to\infty} x_{(n-1)1} = \frac{\sqrt{5}-1}{2}. \tag{7.71}$$

As $n \to \infty$,

$$x_{(n)} \to \left(\frac{\sqrt{5}-1}{2},3,0\right). \tag{7.72}$$

Thus, the limit point for this sequence is *not* in $\mathbb{Q}^3$; hence the sequence is not convergent. Had the set been defined in $\mathbb{R}^3$, it would have been convergent.

However, the sequence *is a Cauchy sequence*. Consider, say $\epsilon = .01$. If we choose, we then find by numerical experimentation that $N_\epsilon = 4$. Choosing, for example $m = 5 > N_\epsilon$ and $n = 21 > N_\epsilon$, we get

$$\begin{aligned} x_{(5)} &= \left(\frac{5}{8},3,0\right), && (7.73)\\ x_{(21)} &= \left(\frac{10946}{17711},3,0\right), && (7.74)\\ ||x_{(5)} - x_{(21)}||_2 &= \left|\left|\left(\frac{987}{141688},0,0\right)\right|\right|_2 = 0.00696 < 0.01. && (7.75) \end{aligned}$$

This could be generalized for arbitrary ϵ, so the sequence can be shown to be a Cauchy sequence.

Example 7.11

Does the infinite sequence of functions

$$v = \{v_1(t), v_2(t), \cdots, v_n(t), \cdots\} = \left\{t(t), t(t^2), t(t^3), \cdots, t(t^n), \cdots\right\}, \tag{7.76}$$

converge in $\mathbb{L}_2[0,1]$? Does the sequence converge in $C[0,1]$?

First, check if the sequence is a Cauchy sequence:

$$\lim_{n,m\to\infty} ||v_n(t) - v_m(t)||_2 = \sqrt{\int_0^1 \left(t^{n+1} - t^{m+1}\right)^2 \, dt} = \sqrt{\frac{1}{2n+3} - \frac{2}{m+n+3} + \frac{1}{2m+3}} = 0. \tag{7.77}$$

As this norm approaches zero, it will be possible for any $\epsilon > 0$ to find an integer N_ϵ such that $||v_n(t) - v_m(t)||_2 < \epsilon$. So, the sequence is a Cauchy sequence. We also have

$$\lim_{n\to\infty} v_n(t) = \begin{cases} 0, & t \in [0,1), \\ 1, & t = 1. \end{cases} \tag{7.78}$$

The function given in Eq. (7.78), the "limit point" to which the sequence converges, is in $\mathbb{L}_2[0,1]$, which is sufficient condition for convergence of the sequence of functions in $\mathbb{L}_2[0,1]$. However the "limit point" is not a continuous function, so despite the fact that the sequence is a Cauchy sequence and elements of the sequence are in $C[0,1]$, the sequence does not converge in $C[0,1]$.

Example 7.12

Analyze the sequence of functions

$$v = \{v_1, v_2, \ldots, v_n, \ldots\} = \left\{\sqrt{2}\sin(\pi t), \sqrt{2}\sin(2\pi t), \ldots, \sqrt{2}\sin(n\pi t), \ldots\right\}, \tag{7.79}$$

in $\mathbb{L}_2[0,1]$.

This is simply a set of sine functions, which can be shown to form a basis; such a proof will not be given here. Each element of the set is orthonormal to other elements:

$$||v_n(t)||_2 = \left(\int_0^1 \left(\sqrt{2}\sin(n\pi t)\right)^2 \, dt\right)^{1/2} = 1. \tag{7.80}$$

It is also easy to show that $\int_0^1 v_n(t) v_m(t) \, dt = 0$, so the basis is orthonormal. As $n \to \infty$, the norm of the basis function remains bounded, and is, in fact, unity.

Consider the norm of the difference of the m^{th} and n^{th} functions:

$$||v_n(t) - v_m(t)||_2 = \left(\int_0^1 \left(\sqrt{2}\sin(n\pi t) - \sqrt{2}\sin(m\pi t)\right)^2 \, dt\right)^{\frac{1}{2}} = \sqrt{2}. \tag{7.81}$$

This is valid for all m and n. Since we can find a value of $\epsilon > 0$ which violates the conditions for a Cauchy sequence, this series of functions is not a Cauchy sequence.

7.3.2 Inner product spaces

The inner product $<x, y>$ is, in general, a complex scalar ($<x, y> \in \mathbb{C}^1$) associated with two elements x and y of a normed vector space satisfying the following rules. For $x, y, z \in \mathbb{S}$ and $\alpha, \beta \in \mathbb{C}$,

1. $<x, x> \ > 0$ if $x \neq 0$,
2. $<x, x> = 0$ if and only if $x = 0$,
3. $<x, \alpha y + \beta z> = \alpha <x, y> + \beta <x, z>$, $\qquad \alpha \in \mathbb{C}^1, \beta \in \mathbb{C}^1$, and
4. $<x, y> = \overline{<y, x>}$, where $\overline{<\cdot>}$ indicates the complex conjugate of the inner product.

Inner product spaces are subspaces of linear vector spaces and are sometimes called *pre-Hilbert*[9] *spaces*. A pre-Hilbert space is not necessarily complete, so it may or may not form a Banach space.

Example 7.13

Show

$$<\alpha x, y> = \overline{\alpha} <x, y>. \tag{7.82}$$

Using the properties of the inner product and the complex conjugate we have

$$\begin{aligned} <\alpha x, y> &= \overline{<y, \alpha x>}, & (7.83)\\ &= \overline{\alpha <y, x>}, & (7.84)\\ &= \overline{\alpha}\ \overline{<y, x>}, & (7.85)\\ &= \overline{\alpha}\ <x, y>. & (7.86) \end{aligned}$$

Note that in a real vector space we have

$$\begin{aligned} <x, \alpha y> = <\alpha x, y> &= \alpha <x, y>, \qquad \text{and also that,} & (7.87)\\ <x, y> &= <y, x>, & (7.88) \end{aligned}$$

since every scalar is equal to its complex conjugate.

Note that some authors use $<\alpha y + \beta z, x> = \alpha <y, x> + \beta <z, x>$ instead of Property 3 that we have chosen.

[9] David Hilbert, 1862-1943, German mathematician of great influence.

7.3.2.1 Hilbert space

A Banach space (i.e. a complete normed vector space) on which an inner product is defined is also called a *Hilbert* space. While Banach spaces allow for the definition of several types of norms, Hilbert spaces are more restrictive: we *must* define the norm such that

$$||x|| = ||x||_2 = +\sqrt{<x,x>}. \tag{7.89}$$

As a counterexample if $x \in \mathbb{R}^2$, and we take $||x|| = ||x||_3 = (|x_1|^3 + |x_2|^3)^{1/3}$ (thus $x \in \ell_3^2$ which is a Banach space), we *cannot find* a definition of the inner product which satisfies all its properties. Thus, the space ℓ_3^2 cannot be a Hilbert space! Unless specified otherwise the unsubscripted norm $||\cdot||$ can be taken to represent the Hilbert space norm $||\cdot||_2$. It is common for both sub-scripted and unscripted versions of the norm to appear in the literature.

The *Cauchy-Schwarz*[10] *inequality* is embodied in the following:

Theorem

For x and y which are elements of a Hilbert space,

$$||x||_2 \, ||y||_2 \geq |<x,y>|. \tag{7.90}$$

If $y = 0$, both sides are zero and the equality holds. Let us take $y \neq 0$. Then, we have

$$\begin{aligned}
||x - \alpha y||_2^2 &= <x - \alpha y, x - \alpha y>, \text{ where } \alpha \text{ is any scalar,} && (7.91)\\
&= <x,x> - <x,\alpha y> - <\alpha y, x> + <\alpha y, \alpha y>, && (7.92)\\
&= <x,x> - \alpha <x,y> - \overline{\alpha} <y,x> + \alpha\overline{\alpha} <y,y>, && (7.93)\\
&\quad \text{on choosing } \alpha = \frac{<y,x>}{<y,y>} = \frac{\overline{<x,y>}}{<y,y>}, && (7.94)\\
&= <x,x> - \frac{\overline{<x,y>}}{<y,y>}<x,y> \\
&\quad \underbrace{-\frac{<x,y>}{<y,y>}<y,x> + \frac{<y,x><x,y>}{<y,y>^2}<y,y>}_{=0}, && (7.95)\\
&= ||x||_2^2 - \frac{|<x,y>|^2}{||y||_2^2}, && (7.96)\\
||x - \alpha y||_2^2 \, ||y||_2^2 &= ||x||_2^2 \, ||y||_2^2 - |<x,y>|^2. && (7.97)
\end{aligned}$$

Since $||x - \alpha y||_2^2 \, ||y||_2^2 \geq 0$,

$$\begin{aligned}
||x||_2^2 \, ||y||_2^2 - |<x,y>|^2 &\geq 0, && (7.98)\\
||x||_2^2 \, ||y||_2^2 &\geq |<x,y>|^2, && (7.99)\\
||x||_2 \, ||y||_2 &\geq |<x,y>|, \qquad QED. && (7.100)
\end{aligned}$$

[10] Karl Hermann Amandus Schwarz, 1843-1921, Silesia-born German mathematician, deeply influenced by Weierstrass, on the faculty at Berlin, captain of the local volunteer fire brigade, and assistant to railway stationmaster.

Note that this effectively defines the angle between two vectors. Because of the inequality, we have

$$\frac{||x||_2 \; ||y||_2}{|<x,y>|} \geq 1, \tag{7.101}$$

$$\frac{|<x,y>|}{||x||_2 \; ||y||_2} \leq 1. \tag{7.102}$$

Defining α to be the angle between the vectors x and y, we recover the familiar result from vector analysis

$$\cos\alpha = \frac{<x,y>}{||x||_2 \; ||y||_2}. \tag{7.103}$$

This reduces to the ordinary relationship we find in Euclidean geometry when $x, y \in \mathbb{R}^3$. The Cauchy-Schwarz inequality is actually a special case of the so-called Hölder[11] inequality:

$$||x||_p ||y||_q \geq |<x,y>|, \qquad \text{with} \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{7.104}$$

The Hölder inequality reduces to the Cauchy-Schwarz inequality when $p = q = 2$.

Examples of Hilbert spaces include

- Finite dimensional vector spaces
  - $x \in \mathbb{R}^3, y \in \mathbb{R}^3$ with $<x,y> = x^T y = x_1 y_1 + x_2 y_2 + x_3 y_3$, where $x = (x_1, x_2, x_3)^T$, and $y = (y_1, y_2, y_3)^T$. This is the ordinary dot product for three-dimensional Cartesian vectors. With this definition of the inner product $<x,x> = ||x||^2 = x_1^2 + x_2^2 + x_3^2$, so the space is the Euclidean space, $\mathbb{E}^3$. The space is also $\ell_2(\mathbb{R}^3)$ or ℓ_2^3.
  - $x \in \mathbb{R}^N, y \in \mathbb{R}^N$ with $<x,y> = x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$, where $x = (x_1, x_2, \cdots, x_N)^T$, and $y = (y_1, y_2, \cdots, y_N)^T$. This is the ordinary dot product for N-dimensional Cartesian vectors; the space is the Euclidean space, $\mathbb{E}^N$, or $\ell_2(\mathbb{R}^N)$, or ℓ_2^N.
  - $x \in \mathbb{C}^N, y \in \mathbb{C}^N$ with $<x,y> = \overline{x}^T y = \overline{x}_1 y_1 + \overline{x}_2 y_2 + \cdots + \overline{x}_N y_N$, where $x = (x_1, x_2, \cdots, x_N)^T$, and $y = (y_1, y_2, \cdots, y_N)^T$. This space is also $\ell_2(\mathbb{C}^N)$. Note that
    - $<x,x> = \overline{x}_1 x_1 + \overline{x}_2 x_2 + \cdots + \overline{x}_N x_N = |x_1|^2 + |x_2|^2 + \ldots + |x_N|^2 = ||x||_2^2$.
    - $<x,y> = \overline{x}_1 y_1 + \overline{x}_2 y_2 + \ldots + \overline{x}_N y_N$.
    - It is easily shown that this definition guarantees $||x||_2 \geq 0$ and $<x,y> = \overline{<y,x>}$.

- Lebesgue spaces
  - $x \in \mathbb{L}_2[a,b]$, $y \in \mathbb{L}_2[a,b]$, $t \in [a,b] \in \mathbb{R}^1$ with $<x,y> = \int_a^b x(t) y(t) \; dt$.

[11] Otto Hölder, 1859-1937, Stuttgart-born German mathematician.

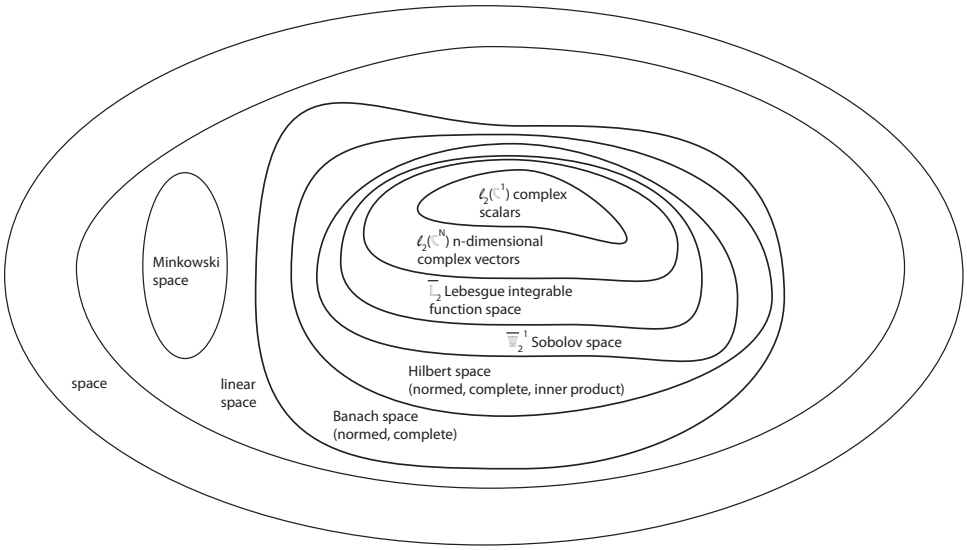

Figure 7.3: Venn diagram showing relationship between various classes of spaces.

- $x \in \overline{\mathbb{L}}_2[a,b]$, $y \in \overline{\mathbb{L}}_2[a,b]$, $t \in [a,b] \in \mathbb{R}^1$ with $<x,y> = \int_a^b \overline{x(t)} y(t)\, dt$.

- Sobolov spaces
  - $u \in \mathbb{W}_2^1(G), v \in \mathbb{W}_2^1(G), x \in G \in \mathbb{R}^N, N \in \mathbb{N}, u \in \mathbb{L}_2(G), \partial u/\partial x_n \in \mathbb{L}_2(G), v \in \mathbb{L}_2(G), \partial v/\partial x_n \in \mathbb{L}_2(G)$ with

$$<u,v> = \int_G \left(u(x)v(x) + \sum_{n=1}^{N} \frac{\partial u}{\partial x_n} \frac{\partial v}{\partial x_n} \right) dx. \tag{7.105}$$

A Venn[12] diagram of some of the common spaces is shown in Fig. 7.3.

7.3.2.2 Non-commutation of the inner product

By the fourth property of inner products, we see that the inner product operation is not commutative in general. Specifically when the vectors are complex, $<x,y> \neq <y,x>$. When the vectors x and y are real, the inner product is real, and the inner product commutes, e.g. $\forall x \in \mathbb{R}^N, y \in \mathbb{R}^N, <x,y> = <y,x>$. At first glance one may wonder why one would define a non-commutative operation. It is done to preserve the positive definite character of the norm. If, for example, we had instead defined the inner product to commute for complex vectors, we might have taken $<x,y> = x^T y$. Then if we had taken $x = (i,1)^T$ and $y = (1,1)^T$, we would have $<x,y> = <y,x> = 1+i$. However, we would also have $<x,x> = ||x||_2^2 = (i,1)(i,1)^T = 0$! Obviously, this would violate the property of the norm since we must have $||x||_2^2 > 0$ for $x \neq 0$.

[12] John Venn, 1834-1923, English mathematician.

Interestingly, one can interpret the Heisenberg[13] uncertainty principle to be entirely consistent with our definition of an inner product which does not commute in a complex space. In quantum mechanics, the superposition of physical states of a system is defined by a complex-valued vector field. Position is determined by application of a position operator, and momentum is determined by application of a momentum operator. If one wants to know both position and momentum, both operators are applied. However, they do not commute, and application of them in different orders leads to a result which varies by a factor related to Planck's[14] constant.

Matrix multiplication is another example of an inner product that does not commute, in general. Such topics are considered in the more general group theory. Operators that commute are known as Abelian[15] and those that do not are known as non-Abelian.

7.3.2.3 Minkowski space

While non-relativistic quantum mechanics, as well as classical mechanics, works well in complex Hilbert spaces, the situation becomes more difficult when one considers Einstein's theories of special and general relativity. In those theories, which are developed to be consistent with experimental observations of 1) systems moving at velocities near the speed of light, 2) systems involving vast distances and gravitation, or 3) systems involving minute length scales, the relevant linear vector space is known as Minkowski space. The vectors have four components, describing the three space-like and one time-like location of an event in space-time, given for example by $x = (x_0, x_1, x_2, x_3)^T$, where $x_0 = ct$, with c as the speed of light. Unlike Hilbert or Banach spaces, however, norms and inner products in the sense that we have defined do not exist! While so-called Minkowski norms and Minkowski inner products are defined in Minkowski space, they are defined in such a fashion that the inner product of a space-time vector with itself can be negative! From the theory of special relativity, the inner product which renders the equations invariant under a Lorentz[16] transformation (necessary so that the speed of light measures the same in all frames and, moreover, not the Galilean[17] transformation of Newtonian theory) is

$$<x, x> = x_0^2 - x_1^2 - x_2^2 - x_3^2. \tag{7.106}$$

Obviously, this inner product can take on negative values. The theory goes on to show that when relativistic effects are important, ordinary concepts of Euclidean geometry become meaningless, and a variety of non-intuitive results can be obtained. In the Venn diagram, we see that Minkowski spaces certainly are not Banach, but there are also linear spaces that are not Minkowski, so it occupies an island in the diagram.

[13] Werner Karl Heisenberg, 1901-1976, German physicist.

[14] Max Karl Ernst Ludwig Planck, 1858-1947, German physicist.

[15] Niels Henrick Abel, 1802-1829, Norwegian mathematician, considered solution of quintic equations by elliptic functions, proved impossibility of solving quintic equations with radicals, gave first solution of an integral equation, famously ignored by Gauss.

[16] Hendrik Antoon Lorentz, 1853-1928, Dutch physicist.

[17] after Galileo Galilei, 1564-1642, Italian polymath.

Example 7.14

For x and y belonging to a Hilbert space, prove the parallelogram equality:

$$||x+y||_2^2 + ||x-y||_2^2 = 2||x||_2^2 + 2||y||_2^2. \tag{7.107}$$

The left side is

$$\begin{aligned} <x+y, x+y> + <x-y, x-y> &= (<x,x> + <x,y> + <y,x> + <y,y>), & (7.108)\\ &\quad + (<x,x> - <x,y> - <y,x> + <y,y>), & (7.109)\\ &= 2<x,x> + 2<y,y>, & (7.110)\\ &= 2||x||_2^2 + 2||y||_2^2. & (7.111) \end{aligned}$$

Example 7.15

For $x, y \in \ell_2(\mathbb{R}^2)$, find $<x,y>$ if

$$x = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad y = \begin{pmatrix} 2 \\ -2 \end{pmatrix}. \tag{7.112}$$

The solution is

$$<x,y> = x^T y = \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} = (1)(2) + (3)(-2) = -4. \tag{7.113}$$

Note that the inner product yields a real scalar, but in contrast to the norm, it can be negative. Note also that the Cauchy-Schwarz inequality holds as $||x||_2 \, ||y||_2 = \sqrt{10}\sqrt{8} \sim 8.944 > |-4|$. Also the Minkowski inequality holds as $||x+y||_2 = ||(3,1)^T||_2 = +\sqrt{10} < ||x||_2 + ||y||_2 = \sqrt{10} + \sqrt{8}$.

Example 7.16

For $x, y \in \ell_2(\mathbb{C}^2)$, find $<x,y>$ if

$$x = \begin{pmatrix} -1+i \\ 3-2i \end{pmatrix}, \qquad y = \begin{pmatrix} 1-2i \\ -2 \end{pmatrix}. \tag{7.114}$$

The solution is

$$<x,y> = \overline{x}^T y = \begin{pmatrix} -1-i & 3+2i \end{pmatrix} \begin{pmatrix} 1-2i \\ -2 \end{pmatrix} = (-1-i)(1-2i) + (3+2i)(-2) = -9-3i. \tag{7.115}$$

Note that the inner product is a complex scalar which has negative components. It is easily shown that $||x||_2 = 3.870$ and $||y||_2 = 3$ and $||x+y||_2 = 2.4495$. Also $|<x,y>| = 9.4868$. The Cauchy-Schwarz inequality holds as $(3.870)(3) = 11.61 > 9.4868$. The Minkowski inequality holds as $2.4495 < 3.870 + 3 = 6.870$.

Example 7.17

For $x, y \in \mathbb{L}_2[0,1]$, find $<x,y>$ if

$$x(t) = 3t + 4, \qquad y(t) = -t - 1. \tag{7.116}$$

The solution is

$$<x,y> = \int_0^1 (3t+4)(-t-1)\ dt = \left(-4t - \frac{7t^2}{2} - t^3\right)\Big|_0^1 = -\frac{17}{2} = -8.5. \tag{7.117}$$

Once more the inner product is a negative scalar. It is easily shown that $||x||_2 = 5.56776$ and $||y||_2 = 1.52753$ and $||x+y||_2 = 4.04145$. Also $|<x,y>| = 8.5$. It is easily seen that the Cauchy-Schwarz inequality holds as $(5.56776)(1.52753) = 8.505 > 8.5$. The Minkowski inequality holds as $4.04145 < 5.56776 + 1.52753 = 7.095$.

Example 7.18

For $x, y \in \overline{\mathbb{L}}_2[0,1]$, find $<x,y>$ if

$$x(t) = it, \qquad y(t) = t + i. \tag{7.118}$$

We recall that

$$<x,y> = \int_0^1 \overline{x(t)} y(t)\ dt. \tag{7.119}$$

The solution is

$$<x,y> = \int_0^1 (-it)(t+i)\ dt = \left(\frac{t^2}{2} - \frac{it^3}{3}\right)\Big|_0^1 = \frac{1}{2} - \frac{i}{3}. \tag{7.120}$$

The inner product is a complex scalar. It is easily shown that $||x||_2 = 0.5776$ and $||y||_2 = 1.1547$ and $||x+y||_2 = 1.6330$. Also $|<x,y>| = 0.601$. The Cauchy-Schwarz inequality holds as $(0.57735)(1.1547) = 0.6667 > 0.601$. The Minkowski inequality holds as $1.63299 < 0.57735 + 1.1547 = 1.7321$.

Example 7.19

For $u, v \in W_2^1(G))$, find $<u,v>$ if

$$u(x) = x_1 + x_2, \qquad v(x) = -x_1 x_2, \tag{7.121}$$

and G is the square region in the x_1, x_2 plane $x_1 \in [0,1], x_2 \in [0,1]$.

We recall that

$$<u, v> = \int_G \left(u(x)v(x) + \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} \right) \, dx, \tag{7.122}$$

$$<u, v> = \int_0^1 \int_0^1 \left((x_1 + x_2)(-x_1 x_2) + (1)(-x_2) + (1)(-x_1) \right) \, dx_1 \, dx_2 = -\frac{4}{3} = -1.33333. \tag{7.123}$$

The inner product here is negative real scalar. It is easily shown that $||u||_{1,2} = 1.77951$ and $||v||_{1,2} = 0.881917$ and $||u + v||_{1,2} = 1.13039$. Also $|<u, v>| = 1.33333$. The Cauchy-Schwarz inequality holds as $(1.77951)(0.881917) = 1.56938 > 1.33333$. The Minkowski inequality holds as $1.13039 < 1.77951 + 0.881917 = 2.66143$.

7.3.2.4 Orthogonality

One of the primary advantages of working in Hilbert spaces is that the inner product allows one to utilize of the useful concept of orthogonality:

- x and y are said to be *orthogonal* to each other if

$$<x, y> = 0. \tag{7.124}$$

- In an orthogonal set of vectors $\{v_1, v_2, \cdots\}$ the elements of the set are all orthogonal to each other, so that $<v_n, v_m> = 0$ if $n \neq m$.
- If a set $\{\varphi_1, \varphi_2, \cdots\}$ exists such that $<\varphi_n, \varphi_m> = \delta_{nm}$, then the elements of the set are *orthonormal*.
- A basis $\{v_1, v_2, \cdots, v_N\}$ of a finite-dimensional space that is also orthogonal is an orthogonal basis. On dividing each vector by its norm we get

$$\varphi_n = \frac{v_n}{\sqrt{<v_n, v_n>}}, \tag{7.125}$$

to give us an orthonormal basis $\{\varphi_1, \varphi_2, \cdots, \varphi_N\}$.

Example 7.20

If elements x and y of an inner product space are orthogonal to each other, prove the Pythagorean theorem

$$||x||_2^2 + ||y||_2^2 = ||x + y||_2^2. \tag{7.126}$$

The right side is

$$\begin{aligned} <x + y, x + y> &= <x, x> + \underbrace{<x, y>}_{=0} + \underbrace{<y, x>}_{=0} + <y, y>, & (7.127)\\ &= <x, x> + <y, y>, & (7.128)\\ &= ||x||_2^2 + ||y||_2^2, \qquad QED. & (7.129) \end{aligned}$$

Example 7.21

Show that an orthogonal set of vectors in an inner product space is linearly independent.

Let $\{v_1, v_2, \cdots, v_n, \ldots, v_N\}$ be an orthogonal set of vectors. Then consider

$$\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n + \ldots + \alpha_N v_N = 0. \tag{7.130}$$

Taking the inner product with v_n, we get

$$\begin{aligned} <v_n, (\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n + \ldots + \alpha_N v_N)> &= <v_n, 0>, & (7.131)\\ \alpha_1 \underbrace{<v_n, v_1>}_{0} + \alpha_2 \underbrace{<v_n, v_2>}_{0} + \ldots + \alpha_n \underbrace{<v_n, v_n>}_{\neq 0} + \ldots + \alpha_N \underbrace{<v_n, v_N>}_{0} &= 0, & (7.132)\\ \alpha_n <v_n, v_n> &= 0, & (7.133) \end{aligned}$$

since all the other inner products are zero. Thus, $\alpha_n = 0$, indicating that the set $\{v_1, v_2, \cdots, v_n, \ldots, v_N\}$ is linearly independent.

7.3.2.5 Gram-Schmidt procedure

In a given inner product space, the *Gram-Schmidt*[18] procedure can be used to find an orthonormal set using a linearly independent set of vectors.

Example 7.22

Find an orthonormal set of vectors $\{\varphi_1, \varphi_2, \ldots\}$ in $\mathbb{L}_2[-1, 1]$ using linear combinations of the linearly independent set of vectors $\{1, t, t^2, t^3, \ldots\}$ where $-1 \leq t \leq 1$.

Choose

$$v_1(t) = 1. \tag{7.134}$$

Now choose the second vector linearly independent of v_1 as

$$v_2(t) = a + bt. \tag{7.135}$$

This should be orthogonal to v_1, so that

$$\begin{aligned} \int_{-1}^{1} v_1(t) v_2(t)\, dt &= 0, & (7.136)\\ \int_{-1}^{1} \underbrace{(1)}_{=v_1(t)} \underbrace{(a+bt)}_{=v_2(t)}\, dt &= 0, & (7.137)\\ \left(at + \frac{bt^2}{2}\right)\Big|_{-1}^{1} &= 0, & (7.138)\\ a(1-(-1)) + \frac{b}{2}(1^2 - (-1)^2) &= 0, & (7.139) \end{aligned}$$

[18] Jørgen Pedersen Gram, 1850-1916, Danish actuary and mathematician, and Erhard Schmidt, 1876-1959, German/Estonian-born Berlin mathematician, studied under David Hilbert, founder of modern functional analysis. The Gram-Schmidt procedure was actually first introduced by Laplace.

from which

$$a = 0. \tag{7.140}$$

Taking $b = 1$ arbitrarily, since orthogonality does not depend on the magnitude of $v_2(t)$, we have

$$v_2 = t. \tag{7.141}$$

Choose the third vector linearly independent of $v_1(t)$ and $v_2(t)$, i.e.

$$v_3(t) = a + bt + ct^2. \tag{7.142}$$

For this to be orthogonal to $v_1(t)$ and $v_2(t)$, we get the conditions

$$\int_{-1}^{1} \underbrace{(1)}_{=v_1(t)} \underbrace{(a + bt + ct^2)}_{=v_3(t)}\, dt = 0, \tag{7.143}$$

$$\int_{-1}^{1} \underbrace{t}_{=v_2(t)} \underbrace{(a + bt + ct^2)}_{=v_3(t)}\, dt = 0. \tag{7.144}$$

The first of these gives $c = -3a$. Taking $a = 1$ arbitrarily, we have $c = -3$. The second relation gives $b = 0$. Thus

$$v_3 = 1 - 3t^2. \tag{7.145}$$

In this manner we can find as many orthogonal vectors as we want. We can make them orthonormal by dividing each by its norm, so that we have

$$\varphi_1 = \frac{1}{\sqrt{2}}, \tag{7.146}$$

$$\varphi_2 = \sqrt{\frac{3}{2}}\, t, \tag{7.147}$$

$$\varphi_3 = \sqrt{\frac{5}{8}}\, (1 - 3t^2), \tag{7.148}$$

$$\vdots$$

Scalar multiples of these functions, with the functions set to unity at $t = 1$, are the Legendre polynomials: $P_0(t) = 1,\ P_1(t) = t,\ P_2(t) = (1/2)(3t^2 - 1) \ldots$ As studied earlier in Chapter 5, some other common orthonormal sets can be formed on the foundation of several eigenfunctions to Sturm-Liouville differential equations.

7.3.2.6 Projection of a vector onto a new basis

Here we consider how to project N-dimensional vectors x, first onto general non-orthogonal bases of dimension $M \leq N$, and then specialize for orthogonal bases of dimension $M \leq N$. For ordinary vectors in Euclidean space, N and M will be integers. When $M < N$, we will usually lose information in projecting the N-dimensional x onto a lower M-dimensional basis. When $M = N$, we will lose no information, and the projection can be better characterized as a new *representation*. While much of our discussion is most easily digested when M and N take on finite values, the analysis will be easily extended to infinite dimension, which is appropriate for a space of vectors which are functions.

7.3.2.6.1 Non-orthogonal basis We are given M linearly independent non-orthogonal basis vectors $\{u_1, u_2, \cdots, u_M\}$ on which to project the N-dimensional x, with $M \leq N$. Each of the M basis vectors, u_m, is taken for convenience to be a vector of length N; we must realize that both x and u_m could be functions as well, in which case saying they have length N would be meaningless.

The general task here is to find expressions for the coefficients $\alpha_m,\ m = 1, 2, \ldots M$, to best represent x in the linear combination

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_M u_M = \sum_{m=1}^{M} \alpha_m u_m \simeq x. \tag{7.149}$$

We use the notation for an approximation, $\simeq$, because for $M < N$, x most likely will not be exactly equal to the linear combination of basis vectors. Since $u \in \mathbb{C}^N$, we can define $\mathbf{U}$ as the $N \times M$ matrix whose M columns are populated by the M basis vectors of length N, $u_1, u_2, \ldots, u_M$. We can thus rewrite Eq. (7.149) as

$$\mathbf{U} \cdot \boldsymbol{\alpha} \simeq \mathbf{x}. \tag{7.150}$$

If $M = N$, the approximation would become an equality; thus, we could invert Eq. (7.150) and find simply that $\boldsymbol{\alpha} = \mathbf{U}^{-1} \cdot \mathbf{x}$. However, if $M < N$, $\mathbf{U}^{-1}$ does not exist, and we cannot use this approach to find $\boldsymbol{\alpha}$. We need another strategy.

To get the values of α_m in the most general of cases, we begin by taking inner products of Eq. (7.149) with u_1 to get

$$<u_1, \alpha_1 u_1> + <u_1, \alpha_2 u_2> + \ldots + <u_1, \alpha_M u_M> \ = \ <u_1, x>. \tag{7.151}$$

Using the properties of an inner product and performing the procedure for all $u_m, m = 1, \ldots, M$, we get

$$\begin{aligned}
\alpha_1<u_1, u_1> + \alpha_2<u_1, u_2> + \ldots + \alpha_M<u_1, u_M> &= <u_1, x>, && (7.152)\\
\alpha_1<u_2, u_1> + \alpha_2<u_2, u_2> + \ldots + \alpha_M<u_2, u_M> &= <u_2, x>, && (7.153)\\
&\vdots \\
\alpha_1<u_M, u_1> + \alpha_2<u_M, u_2> + \ldots + \alpha_M<u_M, u_M> &= <u_M, x>. && (7.154)
\end{aligned}$$

Knowing x and $u_1, u_2, \cdots, u_M$, all the inner products can be determined, and Eqs. (7.152-7.154) can be posed as the linear algebraic system:

$$\underbrace{\begin{pmatrix} <u_1, u_1> & <u_1, u_2> & \ldots & <u_1, u_M> \\ <u_2, u_1> & <u_2, u_2> & \ldots & <u_2, u_M> \\ \vdots & \vdots & \ldots & \vdots \\ <u_M, u_1> & <u_M, u_2> & \ldots & <u_M, u_M> \end{pmatrix}}_{\overline{\mathbf{U}}^T \cdot \mathbf{U}} \cdot \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix}}_{\boldsymbol{\alpha}} = \underbrace{\begin{pmatrix} <u_1, x> \\ <u_2, x> \\ \vdots \\ <u_M, x> \end{pmatrix}}_{\overline{\mathbf{U}}^T \cdot \mathbf{x}}. \tag{7.155}$$

Equation (7.155) can also be written compactly as

$$<u_i, u_m>\alpha_m = <u_i, x>. \tag{7.156}$$

In either case, Cramer's rule or Gaussian elimination can be used to determine the unknown coefficients, α_m.

We can understand this in another way by considering an approach using Gibbs notation, valid when each of the M basis vectors $u_m \in \mathbb{C}^N$. Note that the Gibbs notation does not suffice for other classes of basis vectors, e.g. when the vectors are functions, $u_m \in \mathbb{L}_2$. Operate on Eq. (7.150) with $\overline{\mathbf{U}}^T$ to get

$$\left(\overline{\mathbf{U}}^T \cdot \mathbf{U}\right) \cdot \boldsymbol{\alpha} = \overline{\mathbf{U}}^T \cdot \mathbf{x}. \tag{7.157}$$

This is the Gibbs notation equivalent of Eq. (7.155). We cannot expect $\mathbf{U}^{-1}$ to always exist; however, as long as the $M \leq N$ basis vectors are linearly independent, we can expect the $M \times M$ matrix $\left(\overline{\mathbf{U}}^T \cdot \mathbf{U}\right)^{-1}$ to exist. We can then solve for the coefficients $\boldsymbol{\alpha}$ via

$$\boldsymbol{\alpha} = \left(\overline{\mathbf{U}}^T \cdot \mathbf{U}\right)^{-1} \cdot \overline{\mathbf{U}}^T \cdot \mathbf{x}, \qquad M \leq N. \tag{7.158}$$

In this case, one is projecting $\mathbf{x}$ onto a basis of equal or lower dimension than itself, and we recover the $M \times 1$ vector $\boldsymbol{\alpha}$. If one then operates on both sides of Eq. (7.158) with the $N \times M$ operator $\mathbf{U}$, one gets

$$\mathbf{U} \cdot \boldsymbol{\alpha} = \underbrace{\mathbf{U} \cdot \left(\overline{\mathbf{U}}^T \cdot \mathbf{U}\right)^{-1} \cdot \overline{\mathbf{U}}^T}_{\mathbf{P}} \cdot \mathbf{x} = \mathbf{x}_p. \tag{7.159}$$

Here we have defined the $N \times N$ *projection matrix* $\mathbf{P}$ as

$$\mathbf{P} = \mathbf{U} \cdot \left(\overline{\mathbf{U}}^T \cdot \mathbf{U}\right)^{-1} \cdot \overline{\mathbf{U}}^T. \tag{7.160}$$

We have also defined $\mathbf{x}_p = \mathbf{P} \cdot \mathbf{x}$ as the projection of $\mathbf{x}$ onto the basis $\mathbf{U}$. These topics will be considered later in a strictly linear algebraic context in Sec. 8.9. When there are $M = N$ linearly independent basis vectors, Eq. (7.160) can be reduced to show $\mathbf{P} = \mathbf{I}$. In this case $\mathbf{U}^{-1}$ exists, and we get

$$\mathbf{P} = \underbrace{\mathbf{U} \cdot \mathbf{U}^{-1}}_{\mathbf{I}} \cdot \underbrace{\overline{\mathbf{U}}^{T^{-1}} \cdot \overline{\mathbf{U}}^T}_{\mathbf{I}} = \mathbf{I}. \tag{7.161}$$

So with $M = N$ linearly independent basis vectors, we have $\mathbf{U} \cdot \boldsymbol{\alpha} = \mathbf{x}$, and recover the much simpler

$$\boldsymbol{\alpha} = \mathbf{U}^{-1} \cdot \mathbf{x}, \qquad M = N. \tag{7.162}$$

Example 7.23

Project the vector $\mathbf{x} = \begin{pmatrix} 6 \\ -3 \end{pmatrix}$ onto the non-orthogonal basis composed of $u_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $u_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Here we have the length of $\mathbf{x}$ as $N = 2$, and we have $M = N = 2$ linearly independent basis vectors. When the basis vectors are combined into a set of column vectors, they form the matrix

$$\mathbf{U} = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}. \tag{7.163}$$

Because we have a sufficient number of basis vectors to span the space, to get $\boldsymbol{\alpha}$, we can simply apply Eq. (7.162) to get

$$\begin{aligned} \boldsymbol{\alpha} &= \mathbf{U}^{-1} \cdot \mathbf{x}, & (7.164)\\ &= \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 6 \\ -3 \end{pmatrix}, & (7.165)\\ &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} \end{pmatrix} \cdot \begin{pmatrix} 6 \\ -3 \end{pmatrix}, & (7.166)\\ &= \begin{pmatrix} 1 \\ 4 \end{pmatrix}. & (7.167) \end{aligned}$$

Thus

$$\mathbf{x} = \alpha_1 u_1 + \alpha_2 u_2 = 1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 4 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 6 \\ -3 \end{pmatrix}. \tag{7.168}$$

The projection matrix $\mathbf{P} = \mathbf{I}$, and $\mathbf{x}_p = \mathbf{x}$. Thus, the projection is actually a representation, with no lost information.

Example 7.24

Project the vector $\mathbf{x} = \begin{pmatrix} 6 \\ -3 \end{pmatrix}$ on the basis composed of $u_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

Here we have a vector $\mathbf{x}$ with $N = 2$ and an $M = 1$ linearly independent basis vector which, when cast into columns, forms

$$\mathbf{U} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{7.169}$$

This vector does not span the space, so to get the projection, we must use the more general Eq. (7.158), which reduces to

$$\boldsymbol{\alpha} = \left(\underbrace{\begin{pmatrix} 2 & 1 \end{pmatrix}}_{\overline{\mathbf{U}}^T} \cdot \underbrace{\begin{pmatrix} 2 \\ 1 \end{pmatrix}}_{\mathbf{U}} \right)^{-1} \cdot \underbrace{\begin{pmatrix} 2 & 1 \end{pmatrix}}_{\overline{\mathbf{U}}^T} \cdot \underbrace{\begin{pmatrix} 6 \\ -3 \end{pmatrix}}_{\mathbf{x}} = (5)^{-1}(9) = \begin{pmatrix} \frac{9}{5} \end{pmatrix}. \tag{7.170}$$

So the projection is

$$\mathbf{x}_p = \alpha_1 u_1 = \begin{pmatrix} \frac{9}{5} \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{18}{5} \\ \frac{9}{5} \end{pmatrix}. \tag{7.171}$$

Note that the projection is not obtained by simply setting $\alpha_2 = 0$ from the previous example. This is because the component of $\mathbf{x}$ aligned with u_2 itself has a projection onto u_1. Had u_1 been orthogonal to u_2, one could have obtained the projection onto u_1 by setting $\alpha_2 = 0$.

The projection matrix is

$$\mathbf{P} = \underbrace{\begin{pmatrix} 2 \\ 1 \end{pmatrix}}_{\mathbf{U}} \left(\underbrace{\begin{pmatrix} 2 & 1 \end{pmatrix}}_{\overline{\mathbf{U}}^T} \cdot \underbrace{\begin{pmatrix} 2 \\ 1 \end{pmatrix}}_{\mathbf{U}} \right)^{-1} \cdot \underbrace{\begin{pmatrix} 2 & 1 \end{pmatrix}}_{\overline{\mathbf{U}}^T} = \begin{pmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} \end{pmatrix}. \tag{7.172}$$

It is easily verified that $\mathbf{x}_p = \mathbf{P} \cdot \mathbf{x}$.

Example 7.25

Project the function $x(t) = t^3$, $t \in [0, 1]$ onto the space spanned by the non-orthogonal basis functions $u_1 = t$, $u_2 = \sin(4t)$.

This is an unusual projection. The $M = 2$ basis functions are not orthogonal. In fact they bear no clear relation to each other. The success in finding approximations to the original function which are accurate depends on how well the chosen basis functions approximate the original function.

The appropriateness of the basis functions notwithstanding, it is not difficult to calculate the projection. Equation (7.155) reduces to

$$\begin{pmatrix} \int_0^1 (t)(t)\ dt & \int_0^1 (t) \sin 4t\ dt \\ \int_0^1 (\sin 4t)(t)\ dt & \int_0^1 \sin^2 4t\ dt \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \int_0^1 (t)(t^3)\ dt \\ \int_0^1 (\sin 4t)(t^3)\ dt \end{pmatrix}. \tag{7.173}$$

Evaluating the integrals gives

$$\begin{pmatrix} 0.333333 & 0.116111 \\ 0.116111 & 0.438165 \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0.2 \\ -0.0220311 \end{pmatrix}. \tag{7.174}$$

Inverting and solving gives

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0.680311 \\ -0.230558 \end{pmatrix}. \tag{7.175}$$

So our projection of $x(t) = t^3$ onto the basis functions yields the approximation $x_p(t)$:

$$x(t) = t^3 \simeq x_p(t) = \alpha_1 u_1 + \alpha_2 u_2 = 0.680311t - 0.230558 \sin 4t. \tag{7.176}$$

Figure 7.4 shows the original function and its two-term approximation. It seems the approximation is not bad; however, there is no clear path to improvement by adding more basis functions. So one might imagine in a very specialized problem that the ability to project onto an unusual basis could be useful. But in general this is not the approach taken.

Example 7.26

Project the function $x = e^t$, $t \in [0, 1]$ onto the space spanned by the functions $u_m = t^{m-1}, m = 1, \ldots, M$, for $M = 4$.

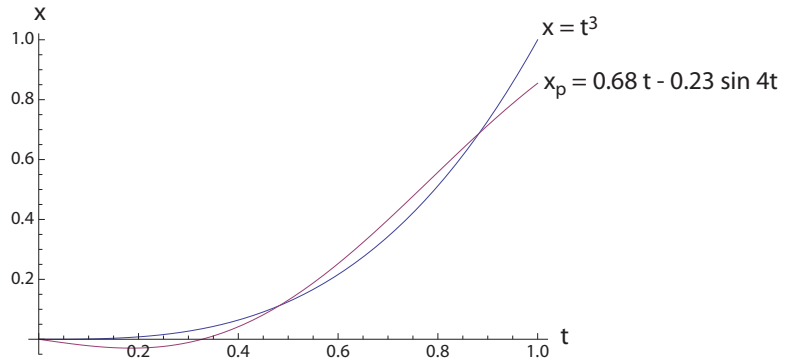

Figure 7.4: Projection of $x(t) = t^3$ onto a two-term non-orthogonal basis composed of functions $u_1 = t$, $u_2 = \sin 4t$.

Similar to the previous example, the basis functions are non-orthogonal. Unlike the previous example, there is a clear way to improve the approximation by increasing M. For $M = 4$, Eq. (7.155) reduces to

$$\begin{pmatrix} \int_0^1 (1)(1)\,dt & \int_0^1 (1)(t)\,dt & \int_0^1 (1)(t^2) & \int_0^1 (1)(t^3) \\ \int_0^1 (t)(1)\,dt & \int_0^1 (t)(t)\,dt & \int_0^1 (t)(t^2) & \int_0^1 (t)(t^3) \\ \int_0^1 (t^2)(1)\,dt & \int_0^1 (t^2)(t)\,dt & \int_0^1 (t^2)(t^2) & \int_0^1 (t^2)(t^3) \\ \int_0^1 (t^3)(1)\,dt & \int_0^1 (t^3)(t)\,dt & \int_0^1 (t^3)(t^2) & \int_0^1 (t^3)(t^3) \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} \int_0^1 (1)(e^t)\,dt \\ \int_0^1 (t)(e^t)\,dt \\ \int_0^1 (t^2)(e^t)\,dt \\ \int_0^1 (t^3)(e^t)\,dt \end{pmatrix}. \tag{7.177}$$

Evaluating the integrals, this becomes

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} -1 + e \\ 1 \\ -2 + e \\ 6 - 2e \end{pmatrix}. \tag{7.178}$$

Solving for α_m, and composing the approximation gives

$$x_p(t) = 0.999060 + 1.01830t + 0.421246t^2 + 0.278625t^3. \tag{7.179}$$

We can compare this to $x_T(t)$, the four-term Taylor series approximation of e^t about $t = 0$:

$$x_T(t) = 1 + t + \frac{t^2}{2} + \frac{t^3}{6} \simeq e^t, \tag{7.180}$$

$$= 1.00000 + 1.00000t - 0.500000t^2 + 0.166667t^3. \tag{7.181}$$

Obviously, the Taylor series approximation is very close to the $M = 4$ projection. The Taylor approximation, $x_T(t)$, gains accuracy as $t \to 0$, while our $x_p(t)$ is better suited to the entire domain $t \in [0, 1]$. We can expect as $M \to \infty$ for the value of each α_m to approach those given by the independent Taylor series approximation. Figure 7.5 shows the original function against its $M = 1, 2, 3, 4$-term approximations, as well as the error. Clearly the approximation improves as M increases; for $M = 4$, the graphs of the original function and its approximation are indistinguishable at this scale.

Also we note that the so-called root-mean-square (rms) error, E_2, is lower for our approximation relative to the Taylor series approximation about $t = 0$. We define rms errors, E_2^p, E_2^T, in terms of a norm, for both our projection and the Taylor approximation, respectively, and find

$$E_2^p = ||x_p(t) - x(t)||_2 = \sqrt{\int_0^1 (x_p(t) - e^t)^2 dt} = 0.000331, \tag{7.182}$$

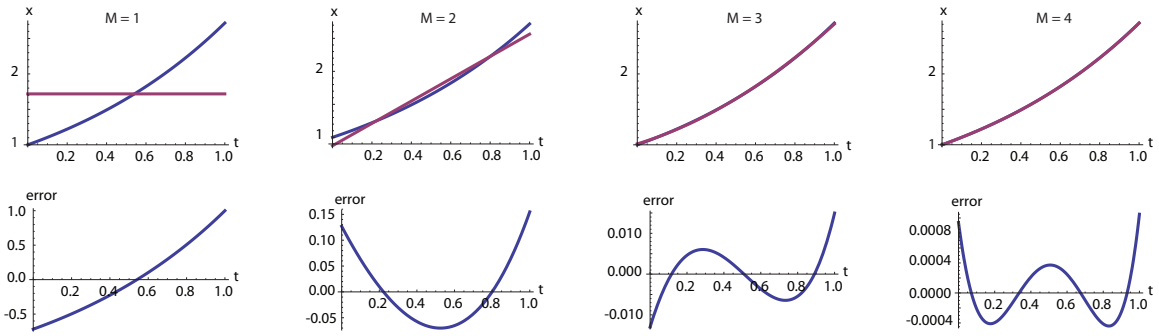

Figure 7.5: The original function $x(t) = e^t$, $t \in [0, 1]$, its projection onto various polynomial basis functions $x(t) \simeq x_p(t) = \sum_{m=1}^{M} \alpha_m t^{m-1}$, and the error, $x - x_p$, for $M = 1, 2, 3, 4$.

$$E_2^T = ||x_T(t) - x(t)||_2 = \sqrt{\int_0^1 (x_T(t) - e^t)^2 dt} \quad = \quad 0.016827. \tag{7.183}$$

Our $M = 4$ approximation is better, when averaged over the entire domain, than the $M = 4$ Taylor series approximation. For larger M, the differences become more dramatic. For example, for $M = 10$, we find $E_2^p = 5.39 \times 10^{-13}$ and $E_2^T = 6.58 \times 10^{-8}$.

7.3.2.6.2 Orthogonal basis The process is simpler if the basis vectors are orthogonal. If orthogonal,

$$<u_i, u_m> = 0, \qquad i \neq m, \tag{7.184}$$

and substituting this into Eq. (7.155), we get

$$\begin{pmatrix} <u_1, u_1> & 0 & \dots & 0 \\ 0 & <u_2, u_2> & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & <u_M, u_M> \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix} = \begin{pmatrix} <u_1, x> \\ <u_2, x> \\ \vdots \\ <u_M, x> \end{pmatrix}. \tag{7.185}$$

Equation (7.185) can be solved directly for the coefficients:

$$\alpha_m = \frac{<u_m, x>}{<u_m, u_m>}. \tag{7.186}$$

So, if the basis vectors are orthogonal, we can write Eq. (7.149) as

$$\frac{<u_1, x>}{<u_1, u_1>} u_1 + \frac{<u_2, x>}{<u_2, u_2>} u_2 + \ldots + \frac{<u_M, x>}{<u_M, u_M>} u_M \quad \simeq \quad x, \tag{7.187}$$

$$\sum_{m=1}^{M} \frac{<u_m, x>}{<u_m, u_m>} u_m = \sum_{m=1}^{M} \alpha_m u_m \quad \simeq \quad x \tag{7.188}$$

If we use an orthonormal basis $\{\varphi_1, \varphi_2, \ldots, \varphi_M\}$, then the projection is even more efficient. We get the generalization of Eq. (5.222):

$$\alpha_m \quad = \quad <\varphi_m, x>, \tag{7.189}$$

which yields

$$\sum_{m=1}^{M} \underbrace{<\varphi_m, x>}_{\alpha_m} \varphi_m \quad \simeq \quad x. \tag{7.190}$$

In all cases, if $M = N$, we can replace the "$\simeq$" by an "$=$", and the approximation becomes in fact a representation.

Similar expansions apply to vectors in infinite-dimensional spaces, except that one must be careful that the orthonormal set is *complete*. Only then is there any guarantee that any vector can be represented as linear combinations of this orthonormal set. If $\{\varphi_1, \varphi_2, \ldots\}$ is a complete orthonormal set of vectors in some domain Ω, then any vector x can be represented as

$$x = \sum_{n=1}^{\infty} \alpha_n \varphi_n, \tag{7.191}$$

where

$$\alpha_n = <\varphi_n, x>. \tag{7.192}$$

This is a *Fourier series* representation, as previously studied in Chapter 5, and the values of α_n are the Fourier coefficients. It is a representation and not just a projection because the summation runs to infinity.

Example 7.27

Expand the top hat function $x(t) = H(t - 1/4) - H(t - 3/4)$ in a Fourier sine series in the domain $t \in [0, 1]$.

Here, the function $x(t)$ is discontinuous at $t = 1/4$ and $t = 3/4$. While $x(t)$ is not a member of $C[0, 1]$, it is a member of $\mathbb{L}_2[0, 1]$. Here we will see that the Fourier sine series projection, composed of functions which are continuous in $[0, 1]$, converges to the discontinuous function $x(t)$.

Building on previous work, we know from Eq. (5.54) that the functions

$$\varphi_n(t) = \sqrt{2}\sin(n\pi t), \qquad n = 1, \ldots, \infty, \tag{7.193}$$

form an orthonormal set for $t \in [0, 1]$. We then find for the Fourier coefficients

$$\alpha_n = \sqrt{2}\int_0^1 \left(H\left(t - \frac{1}{4}\right) - H\left(t - \frac{3}{4}\right)\right)\sin(n\pi t)\ dt = \sqrt{2}\int_{1/4}^{3/4} \sin(n\pi t)\ dt. \tag{7.194}$$

Performing the integration for the first nine terms, we find

$$\alpha_n = \frac{2}{\pi}\left(1, 0, -\frac{1}{3}, 0, -\frac{1}{5}, 0, \frac{1}{7}, 0, \frac{1}{9}, \ldots\right). \tag{7.195}$$

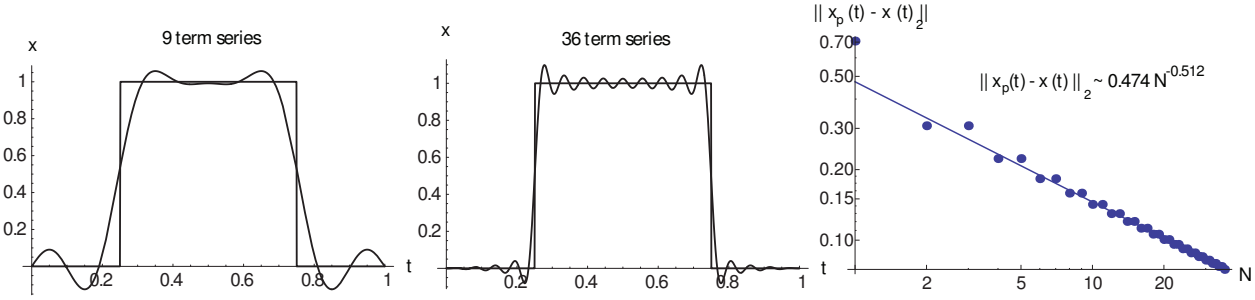

Figure 7.6: Expansion of top hat function $x(t) = H(t - 1/4) - H(t - 3/4)$ in terms of sinusoidal basis functions for two levels of approximation, $N = 9$, $N = 36$ along with a plot of how the error converges as the number of terms increases.

Forming an approximation from these nine terms, we find

$$H\left(t-\frac{1}{4}\right)-H\left(t-\frac{3}{4}\right)=\frac{2\sqrt{2}}{\pi}\left(\sin(\pi t)-\frac{\sin(3\pi t)}{3}-\frac{\sin(5\pi t)}{5}+\frac{\sin(7\pi t)}{7}+\frac{\sin(9\pi t)}{9}+\ldots\right). \quad (7.196)$$

Generalizing, we get

$$H\left(t-\frac{1}{4}\right)-H\left(t-\frac{3}{4}\right)=\frac{2\sqrt{2}}{\pi}\sum_{k=1}^{\infty}(-1)^{k-1}\left(\frac{\sin((4k-3)\pi t)}{4k-3}-\frac{\sin((4k-1)\pi t)}{4k-1}\right). \quad (7.197)$$

The discontinuous function $x(t)$, two continuous approximations to it, and a plot revealing how the error decreases as the number of terms in the approximation increase are shown in Fig. 7.6. Note that as more terms are added, the approximation gets better at most points. But there is always a persistently large error at the discontinuities $t = 1/4$, $t = 3/4$. We say this function is convergent in $\mathbb{L}_2[0, 1]$, but is not convergent in $\mathbb{L}_\infty[0, 1]$. This simply says that the rms error norm converges, while the maximum error norm does not. This is an example of the well-known *Gibbs phenomenon*. Convergence in $\mathbb{L}_2[0, 1]$ is shown in Fig. 7.6. The achieved convergence rate is $||x_p(t) - x_(t)||_2 \sim 0.474088N^{-0.512}$. This suggests that

$$\lim_{N\to\infty}||x_p(t)-x(t)||_2\sim\frac{1}{\sqrt{N}}, \quad (7.198)$$

where N is the number of terms retained in the projection.

The previous example showed one could use continuous functions to approximate a discontinuous function. The converse is also true: discontinuous functions can be used to approximate continuous functions.

Example 7.28

Show that the functions $\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)$ are orthonormal in $\mathbb{L}_2(0, 1]$, where

$$\varphi_n(t)=\begin{cases}\sqrt{N}, & \frac{n-1}{N}<t\leq\frac{n}{N},\\ 0, & \text{otherwise.}\end{cases} \quad (7.199)$$

Expand $x(t) = t^2$ in terms of these functions, and find the error for a finite N.

We note that the basis functions are a set of "top hat" functions whose amplitude increases and width decreases as N increases. For fixed N, the basis functions are a series of top hats that fills the domain $[0,1]$. The area enclosed by a single basis function is $1/\sqrt{N}$. If $n \neq m$, the inner product

$$<\varphi_n, \varphi_m> = \int_0^1 \varphi_n(t)\varphi_m(t)\ dt = 0, \tag{7.200}$$

because the integrand is zero everywhere. If $n = m$, the inner product is

$$\begin{aligned}
\int_0^1 \varphi_n(t)\varphi_n(t)\ dt &= \int_0^{\frac{n-1}{N}} (0)(0)\ dt + \int_{\frac{n-1}{N}}^{\frac{n}{N}} \sqrt{N}\sqrt{N}\ dt + \int_{\frac{n}{N}}^1 (0)(0)\ dt, &(7.201)\\
&= N\left(\frac{n}{N} - \frac{n-1}{N}\right), &(7.202)\\
&= 1. &(7.203)
\end{aligned}$$

So, $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ is an orthonormal set. We can expand the function $f(t) = t^2$ in the form

$$t^2 = \sum_{n=1}^{N} \alpha_n \varphi_n. \tag{7.204}$$

Taking the inner product of both sides with $\varphi_m(t)$, we get

$$\begin{aligned}
\int_0^1 \varphi_m(t)t^2\ dt &= \int_0^1 \varphi_m(t) \sum_{n=1}^{N} \alpha_n \varphi_n(t)\ dt, &(7.205)\\
\int_0^1 \varphi_m(t)t^2\ dt &= \sum_{n=1}^{N} \alpha_n \underbrace{\int_0^1 \varphi_m(t)\varphi_n(t)\ dt}_{=\delta_{nm}}, &(7.206)\\
\int_0^1 \varphi_m(t)t^2\ dt &= \sum_{n=1}^{N} \alpha_n \delta_{nm}, &(7.207)\\
\int_0^1 \varphi_m(t)t^2\ dt &= \alpha_m, &(7.208)\\
\int_0^1 \varphi_n(t)t^2\ dt &= \alpha_n. &(7.209)
\end{aligned}$$

Thus,

$$\alpha_n = 0 + \int_{\frac{n-1}{N}}^{\frac{n}{N}} t^2\sqrt{N}\ dt + 0. \tag{7.210}$$

Thus,

$$\alpha_n = \frac{1}{3N^{5/2}}\left(3n^2 - 3n + 1\right). \tag{7.211}$$

The functions t^2 and the partial sums $f_N(t) = \sum_{n=1}^{N} \alpha_n\varphi_n(t)$ for $N = 5$ and $N = 10$ are shown in Fig. 7.7. Detailed analysis not shown here reveals the $\mathbb{L}_2$ error for the partial sums can be calculated as Δ_N, where

$$\begin{aligned}
\Delta_N^2 &= ||f(t) - f_N(t)||_2^2, &(7.212)\\
&= \int_0^1 \left(t^2 - \sum_{n=1}^{N} \alpha_n\varphi_n(t)\right)^2\ dt, &(7.213)
\end{aligned}$$

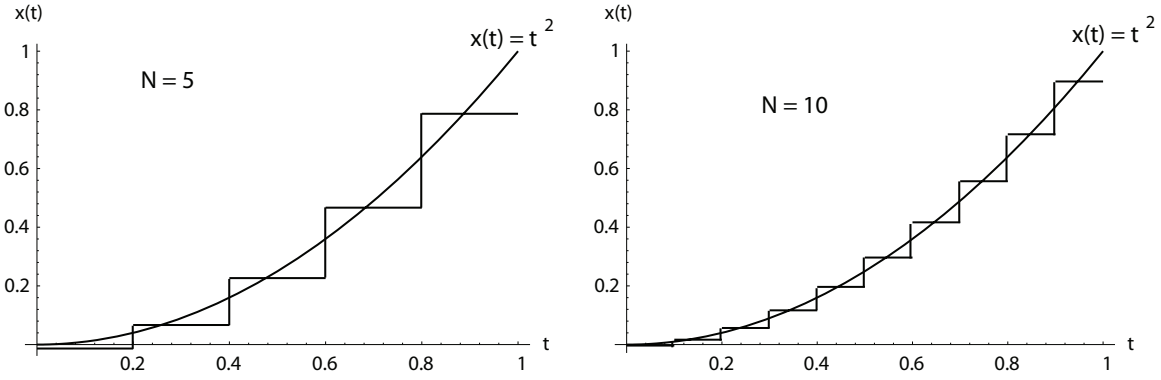

Figure 7.7: Expansion of $x(t) = t^2$ in terms of "top hat" basis functions for two levels of approximation, $N = 5, N = 10$.

$$= \frac{1}{9N^2}\left(1 - \frac{1}{5N^2}\right), \tag{7.214}$$

$$\Delta_N = \frac{1}{3N}\sqrt{1 - \frac{1}{5N^2}}, \tag{7.215}$$

which vanishes as $N \to \infty$ at a rate of convergence proportional to $1/N$.

Example 7.29

Demonstrate the Fourier sine series for $x(t) = 2t$ converges at a rate proportional to $1/\sqrt{N}$, where N is the number of terms used to approximate $x(t)$, in $\mathbb{L}_2[0, 1]$.

Consider the sequence of functions

$$\varphi_n(t) = \left\{\sqrt{2}\sin(\pi t), \sqrt{2}\sin(2\pi t), \ldots, \sqrt{2}\sin(n\pi t), \ldots\right\}. \tag{7.216}$$

It is easy to show linear independence for these functions. They are orthonormal in the Hilbert space $\mathbb{L}_2[0, 1]$, e.g.

$$<\varphi_2, \varphi_3> = \int_0^1 \left(\sqrt{2}\sin(2\pi t)\right)\left(\sqrt{2}\sin(3\pi t)\right)\, dt = 0, \tag{7.217}$$

$$<\varphi_3, \varphi_3> = \int_0^1 \left(\sqrt{2}\sin(3\pi t)\right)\left(\sqrt{2}\sin(3\pi t)\right)\, dt = 1. \tag{7.218}$$

Note that while the basis functions evaluate to 0 at both $t = 0$ and $t = 1$, that the function itself only has value 0 at $t = 0$. We must tolerate a large error at $t = 1$, but hope that this error is confined to an ever collapsing neighborhood around $t = 1$ as more terms are included in the approximation.

The Fourier coefficients are

$$\alpha_n = <2t, \varphi_n(t)> = \int_0^1 (2t)\sqrt{2}\sin(n\pi t)\, dt = \frac{2\sqrt{2}(-1)^{n+1}}{n\pi}. \tag{7.219}$$

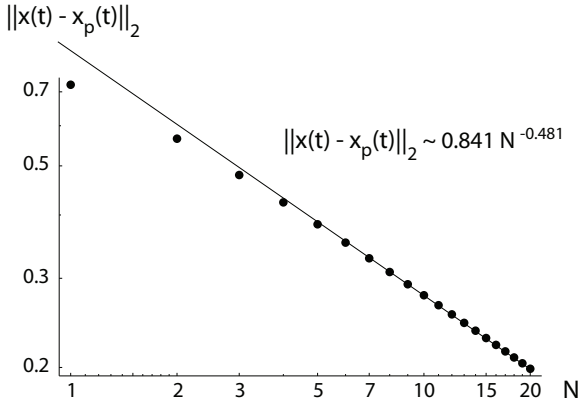

Figure 7.8: Behavior of the error norm of the Fourier sine series approximation to $x(t) = 2t$ on $t \in [0, 1]$ with the number N of terms included in the series.

The approximation then is

$$x_p(t) = \sum_{n=1}^{N} \frac{4(-1)^{n+1}}{n\pi} \sin(n\pi t). \tag{7.220}$$

The norm of the error is then

$$||x(t) - x_p(t)||_2 = \sqrt{\int_0^1 \left(2t - \left(\sum_{n=1}^{N} \frac{4(-1)^{n+1}}{n\pi} \sin(n\pi t)\right)\right)^2 dt}. \tag{7.221}$$

This is difficult to evaluate analytically. It is straightforward to examine this with symbolic calculational software.

A plot of the norm of the error as a function of the number of terms in the approximation, N, is given in the log-log plot of Fig. 7.8. A weighted least squares curve fit, with a weighting factor proportional to N^2 so that priority is given to data as $N \to \infty$, shows that the function

$$||x(t) - x_p(t)||_2 \sim 0.841 \ N^{-0.481}, \tag{7.222}$$

approximates the convergence performance well. In the log-log plot the exponent on N is the slope. It appears from the graph that the slope may be approaching a limit, in which it is likely that

$$||x(t) - x_p(t)||_2 \sim \frac{1}{\sqrt{N}}. \tag{7.223}$$

This indicates convergence of this series. Note that the series converges even though the norm of the n^{th} basis function does not approach zero as $n \to \infty$:

$$\lim_{n\to\infty} ||\varphi_n||_2 = 1, \tag{7.224}$$

since the basis functions are orthonormal. Also note that the behavior of the norm of the final term in the series,

$$||\alpha_N \varphi_N(t)||_2 = \sqrt{\int_0^1 \left(\frac{2\sqrt{2}(-1)^{N+1}}{N\pi} \sqrt{2} \sin(N\pi t)\right)^2 dt} = \frac{2\sqrt{2}}{N\pi}, \tag{7.225}$$

does not tell us how the series actually converges.

Example 7.30

Show the Fourier sine series for $x(t) = t - t^2$ converges at a rate proportional to $1/N^{5/2}$, where N is the number of terms used to approximate $x(t)$, in $\mathbb{L}_2[0,1]$.

Again, consider the sequence of functions

$$\varphi_n(t) = \left\{ \sqrt{2}\sin(\pi t), \sqrt{2}\sin(2\pi t), \ldots, \sqrt{2}\sin(n\pi t), \ldots \right\}. \tag{7.226}$$

which are as before, linearly independent and moreover, orthonormal. Note that in this case, as opposed to the previous example, both the basis functions and the function to be approximated vanish identically at both $t = 0$ and $t = 1$. Consequently, there will be *no error* in the approximation at either end point.

The Fourier coefficients are

$$\alpha_n = \frac{2\sqrt{2}\left(1 + (-1)^{n+1}\right)}{n^3\pi^3}. \tag{7.227}$$

Note that $\alpha_n = 0$ for even values of n. Taking this into account and retaining only the necessary basis functions, we can write the Fourier sine series as

$$x(t) = t(1-t) \sim x_p(t) = \sum_{m=1}^{N} \frac{4\sqrt{2}}{(2m-1)^3\pi^3}\sin((2m-1)\pi t). \tag{7.228}$$

The norm of the error is then

$$||x(t) - x_p(t)||_2 = \sqrt{\int_0^1 \left(t(1-t) - \left(\sum_{m=1}^{N} \frac{4\sqrt{2}}{(2m-1)^3\pi^3}\sin((2m-1)\pi t) \right) \right)^2 dt}. \tag{7.229}$$

Again this is difficult to address analytically, but symbolic computation allows computation of the error norm as a function of N.

A plot of the norm of the error as a function of the number of terms in the approximation, N, is given in the log-log plot of Fig. 7.9. A weighted least squares curve fit, with a weighting factor proportional to N^2 so that priority is given to data as $N \to \infty$, shows that the function

$$||x(t) - x_p(t)||_2 \sim 0.00995\ N^{-2.492}, \tag{7.230}$$

approximates the convergence performance well. Thus, we might suspect that

$$\lim_{n\to\infty} ||x(t) - x_p(t)||_2 \sim \frac{1}{N^{5/2}}. \tag{7.231}$$

Note that the convergence is *much* more rapid than in the previous example! This can be critically important in numerical calculations and demonstrates that a judicious selection of basis functions can have fruitful consequences.

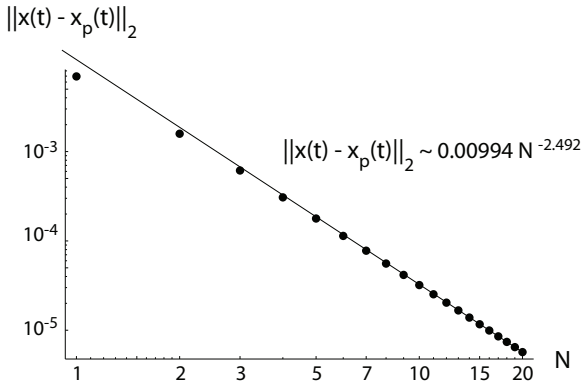

Figure 7.9: Behavior of the error norm of the Fourier sine series approximation to $x(t) = t(1-t)$ on $t \in [0,1]$ with the number N of terms included in the series.

7.3.2.7 Parseval's equation, convergence, and completeness

We consider Parseval's[19] equation and associated issues here. For a basis to be complete, we require that the norm of the difference of the series representation of all functions and the functions themselves converge to zero in $\mathbb{L}_2$ as the number of terms in the series approaches infinity. For an orthonormal basis $\varphi_n(t)$, this is

$$\lim_{N\to\infty} \left|\left| x(t) - \sum_{n=1}^{N} \alpha_n \varphi_n(t) \right|\right|_2 = 0. \tag{7.232}$$

Now for the orthonormal basis, we can show this reduces to a particularly simple form. Consider for instance the error for a one-term Fourier expansion

$$\begin{aligned}
||x-\alpha\varphi||_2^2 &= <x-\alpha\varphi, x-\alpha\varphi>, & (7.233)\\
&= <x,x> - <x,\alpha\varphi> - <\alpha\varphi,x> + <\alpha\varphi,\alpha\varphi>, & (7.234)\\
&= ||x||_2^2 - \alpha<x,\varphi> - \overline{\alpha}<\varphi,x> + \overline{\alpha}\alpha<\varphi,\varphi>, & (7.235)\\
&= ||x||_2^2 - \alpha\overline{<\varphi,x>} - \overline{\alpha}<\varphi,x> + \overline{\alpha}\alpha<\varphi,\varphi>, & (7.236)\\
&= ||x||_2^2 - \alpha\overline{\alpha} - \overline{\alpha}\alpha + \overline{\alpha}\alpha(1), & (7.237)\\
&= ||x||_2^2 - \alpha\overline{\alpha}, & (7.238)\\
&= ||x||_2^2 - |\alpha|^2. & (7.239)
\end{aligned}$$

Here we have used the definition of the Fourier coefficient $<\varphi, x> = \alpha$, and orthonormality $<\varphi,\varphi> = 1$. This is easily extended to multi-term expansions to give

$$\left|\left| x(t) - \sum_{n=1}^{N} \alpha_n \varphi_n(t) \right|\right|_2^2 = ||x(t)||_2^2 - \sum_{n=1}^{N} |\alpha_n|^2. \tag{7.240}$$

So convergence, and thus completeness of the basis, is equivalent to requiring that

$$||x(t)||_2^2 = \lim_{N\to\infty} \sum_{n=1}^{N} |\alpha_n|^2, \tag{7.241}$$

[19] Marc-Antoine Parseval des Chênes, 1755-1835, French mathematician.

for all functions $x(t)$. Note that this requirement is stronger than just requiring that the last Fourier coefficient vanish for large N; also note that it does not address the important question of the rate of convergence, which can be different for different functions $x(t)$, for the same basis.

7.3.3 Reciprocal bases

Let $\{u_1, \cdots, u_N\}$ be a basis of a finite-dimensional inner product space. Also let $\{u_1^R, \cdots, u_N^R\}$ be elements of the same space such that

$$<u_n, u_m^R> = \delta_{nm}. \tag{7.242}$$

Then $\{u_1^R, \cdots, u_N^R\}$ is called the reciprocal (or dual) basis of $\{u_1, \cdots, u_N\}$. Of course an orthonormal basis is its own reciprocal. Since $\{u_1, \cdots, u_N\}$ is a basis, we can write any vector x as

$$x = \sum_{m=1}^{N} \alpha_m u_m. \tag{7.243}$$

Taking the inner product of both sides with u_n^R, we get

$$\begin{aligned} <u_n^R, x> &= <u_n^R, \sum_{m=1}^{N} \alpha_m u_m>, & (7.244)\\ &= \sum_{m=1}^{N} <u_n^R, \alpha_m u_m>, & (7.245)\\ &= \sum_{m=1}^{N} \alpha_m <u_n^R, u_m>, & (7.246)\\ &= \sum_{m=1}^{N} \alpha_m \delta_{nm}, & (7.247)\\ &= \alpha_n, & (7.248) \end{aligned}$$

so that

$$x = \sum_{n=1}^{N} \underbrace{<u_n^R, x>}_{=\alpha_n} u_n. \tag{7.249}$$

The transformation of the representation of a vector x from a basis to a dual basis is a type of alias transformation.

Example 7.31

A vector $\mathbf{v}$ resides in $\mathbb{R}^2$. Its representation in Cartesian coordinates is $\mathbf{v} = \xi = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$. The vectors $u_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $u_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ span the space $\mathbb{R}^2$ and thus can be used as a basis on which to represent $\mathbf{v}$. Find the reciprocal basis u_1^R, u_2^R, and use Eq. (7.249) to represent $\mathbf{v}$ in terms of both the basis u_1, u_2 and then the reciprocal basis u_1^R, u_2^R.

We adopt the dot product as our inner product. Let's get α_1, α_2. To do this we first need the reciprocal basis vectors which are defined by the inner product:

$$<u_n, u_m^R> = \delta_{nm}. \tag{7.250}$$

We take

$$u_1^R = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}, \qquad u_2^R = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}. \tag{7.251}$$

Expanding Eq. (7.250), we get,

$$<u_1, u_1^R> = u_1^T u_1^R = (2,0) \cdot \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = (2)a_{11} + (0)a_{21} = 1, \tag{7.252}$$

$$<u_1, u_2^R> = u_1^T u_2^R = (2,0) \cdot \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = (2)a_{12} + (0)a_{22} = 0, \tag{7.253}$$

$$<u_2, u_1^R> = u_2^T u_1^R = (1,3) \cdot \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = (1)a_{11} + (3)a_{21} = 0, \tag{7.254}$$

$$<u_2, u_2^R> = u_2^T u_2^R = (1,3) \cdot \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = (1)a_{12} + (3)a_{22} = 1. \tag{7.255}$$

Solving, we get

$$a_{11} = \frac{1}{2}, \qquad a_{21} = -\frac{1}{6}, \qquad a_{12} = 0, \qquad a_{22} = \frac{1}{3}, \tag{7.256}$$

so substituting into Eq. (7.251), we get expressions for the reciprocal base vectors:

$$u_1^R = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{6} \end{pmatrix}, \qquad u_2^R = \begin{pmatrix} 0 \\ \frac{1}{3} \end{pmatrix}. \tag{7.257}$$

We can now get the coefficients α_i:

$$\alpha_1 = <u_1^R, \xi> = \left(\frac{1}{2}, -\frac{1}{6}\right) \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \frac{3}{2} - \frac{5}{6} = \frac{2}{3}, \tag{7.258}$$

$$\alpha_2 = <u_2^R, \xi> = \left(0, \frac{1}{3}\right) \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = 0 + \frac{5}{3} = \frac{5}{3}. \tag{7.259}$$

So on the new basis, $\mathbf{v}$ can be represented as

$$\mathbf{v} = \frac{2}{3}\, u_1 + \frac{5}{3}\, u_2. \tag{7.260}$$

The representation is shown geometrically in Fig. 7.10. Note that u_1^R is orthogonal to u_2 and that u_2^R is orthogonal to u_1. Further since $||u_1||_2 > 1, ||u_2||_2 > 1$, we get $||u_1^R||_2 < 1$ and $||u_2^R||_2 < 1$ in order to have $<u_i, u_j^R> = \delta_{ij}$.

In a similar manner it is easily shown that $\mathbf{v}$ can be represented in terms of the reciprocal basis as

$$\mathbf{v} = \sum_{n=1}^{N} \beta_n u_n^R = \beta_1 u_1^R + \beta_2 u_2^R, \tag{7.261}$$

where

$$\beta_n = <u_n, \xi>. \tag{7.262}$$

For this problem, this yields

$$\mathbf{v} = 6u_1^R + 18u_2^R. \tag{7.263}$$

Thus, we see for the non-orthogonal basis that two natural representations of the same vector exist. One of these is actually a covariant representation; the other is contravariant.

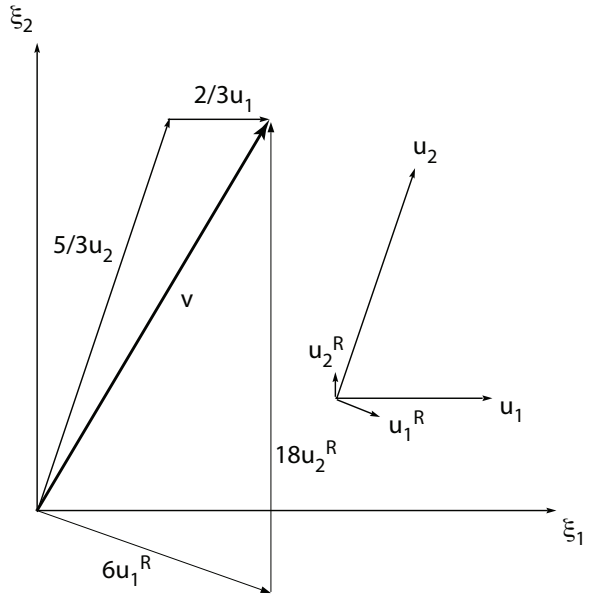

Figure 7.10: Representation of a vector x on a non-orthogonal contravariant basis $u_1,\ u_2$ and its reciprocal covariant basis $u_1^R,\ u_2^R$.

Let us show this is consistent with the earlier described notions using "upstairs-downstairs" notation of Sec. 1.3. Note that our non-orthogonal coordinate system is a transformation of the form

$$\xi^i = \frac{\partial \xi^i}{\partial x^j} x^j, \tag{7.264}$$

where ξ^i is the Cartesian representation, and x^j is the contravariant representation in the transformed system. In Gibbs form, this is

$$\boldsymbol{\xi} = \mathbf{J} \cdot \mathbf{x}. \tag{7.265}$$

Inverting, we also have

$$\mathbf{x} = \mathbf{J}^{-1} \cdot \boldsymbol{\xi}. \tag{7.266}$$

For this problem, we have

$$\frac{\partial \xi^i}{\partial x^j} = \mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} \vdots & \vdots \\ u_1 & u_2 \\ \vdots & \vdots \end{pmatrix}, \tag{7.267}$$

so that

$$\begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}. \tag{7.268}$$

Note that the unit vector in the transformed space

$$\begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{7.269}$$

has representation in Cartesian space of $(2, 0)^T$, and the other unit vector in the transformed space

$$\begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{7.270}$$

has representation in Cartesian space of $(1,3)^T$.

Now the metric tensor is

$$g_{ij} = \mathbf{G} = \mathbf{J}^T \cdot \mathbf{J} = \begin{pmatrix} 2 & 0 \\ 1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 10 \end{pmatrix}. \tag{7.271}$$

The Cartesian vector $\boldsymbol{\xi} = (3,5)^T$, has a contravariant representation in the transformed space of

$$\mathbf{x} = \mathbf{J}^{-1} \cdot \boldsymbol{\xi} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{6} \\ 0 & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix} = x^j. \tag{7.272}$$

This is consistent with our earlier finding.

This vector has a covariant representation as well by the formula

$$x_i = g_{ij} x^j = \begin{pmatrix} 4 & 2 \\ 2 & 10 \end{pmatrix} \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 6 \\ 18 \end{pmatrix}. \tag{7.273}$$

Once again, this is consistent with our earlier finding.

Note further that

$$\mathbf{J}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{6} \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \dots u_1^R \dots \\ \dots u_2^R \dots \end{pmatrix}. \tag{7.274}$$

The *rows* of this matrix describe the reciprocal basis vectors, and is also consistent with our earlier finding. So if we think of the columns of any matrix as forming a basis, the rows of the inverse of that matrix form the reciprocal basis:

$$\underbrace{\begin{pmatrix} \dots u_1^R \dots \\ \dots\dots\dots \\ \dots u_N^R \dots \end{pmatrix}}_{\mathbf{J}^{-1}} \cdot \underbrace{\begin{pmatrix} \vdots & \vdots & \vdots \\ u_1 & \vdots & u_N \\ \vdots & \vdots & \vdots \end{pmatrix}}_{\mathbf{J}} = \mathbf{I}. \tag{7.275}$$

Lastly note that $\det \mathbf{J} = 6$, so the transformation is orientation-preserving, but not volume-preserving. A unit volume element in $\boldsymbol{\xi}$-space is larger than one in $\mathbf{x}$-space. Moreover the mapping $\boldsymbol{\xi} = \mathbf{J} \cdot \mathbf{x}$ can be shown to involve both stretching and rotation.

Example 7.32

For the previous example problem, consider the tensor $\mathbf{A}$, whose representation in the Cartesian space is

$$\mathbf{A} = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}. \tag{7.276}$$

Demonstrate the invariance of the scalar $\boldsymbol{\xi}^T \cdot \mathbf{A} \cdot \boldsymbol{\xi}$ in the non-Cartesian space.

First, in the Cartesian space we have

$$\boldsymbol{\xi}^T \cdot \mathbf{A} \cdot \boldsymbol{\xi} = \begin{pmatrix} 3 & 5 \end{pmatrix} \cdot \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = 152. \tag{7.277}$$

Now $\mathbf{A}$ has a different representation, $\mathbf{A}'$, in the transformed coordinate system via the definition of a tensor, Eq. (1.181), which for this linear alias transformation, reduces to:[20]

$$\mathbf{A}' = \mathbf{J}^{-1} \cdot \mathbf{A} \cdot \mathbf{J}. \tag{7.278}$$

So

$$\mathbf{A}' = \underbrace{\begin{pmatrix} \frac{1}{2} & -\frac{1}{6} \\ 0 & \frac{1}{3} \end{pmatrix}}_{\mathbf{J}^{-1}} \cdot \underbrace{\begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}}_{\mathbf{A}} \cdot \underbrace{\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}}_{\mathbf{J}}, \tag{7.279}$$

$$= \begin{pmatrix} \frac{8}{3} & \frac{19}{3} \\ \frac{2}{3} & \frac{7}{3} \end{pmatrix}, \tag{7.280}$$

$$\tag{7.281}$$

We also see by inversion that

$$\mathbf{A} = \mathbf{J} \cdot \mathbf{A}' \cdot \mathbf{J}^{-1}. \tag{7.282}$$

Since $\boldsymbol{\xi} = \mathbf{J} \cdot \mathbf{x}$, our tensor invariant becomes in the transformed space

$$\boldsymbol{\xi}^T \cdot \mathbf{A} \cdot \boldsymbol{\xi} = (\mathbf{J} \cdot \mathbf{x})^T \cdot (\mathbf{J} \cdot \mathbf{A}' \cdot \mathbf{J}^{-1}) \cdot (\mathbf{J} \cdot \mathbf{x}), \tag{7.283}$$

$$= \mathbf{x}^T \cdot \underbrace{\mathbf{J}^T \cdot \mathbf{J}}_{\mathbf{G}} \cdot \mathbf{A}' \cdot \mathbf{x}, \tag{7.284}$$

$$= \underbrace{\mathbf{x}^T \cdot \mathbf{G}}_{\text{covariant } \mathbf{x}} \cdot \mathbf{A}' \cdot \mathbf{x}, \tag{7.285}$$

$$= \begin{pmatrix} \frac{2}{3} & \frac{5}{3} \end{pmatrix} \cdot \begin{pmatrix} 4 & 2 \\ 2 & 10 \end{pmatrix} \cdot \begin{pmatrix} \frac{8}{3} & \frac{19}{3} \\ \frac{2}{3} & \frac{7}{3} \end{pmatrix} \cdot \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix}, \tag{7.286}$$

$$= \underbrace{\begin{pmatrix} 6 & 18 \end{pmatrix}}_{\text{covariant } \mathbf{x}} \cdot \underbrace{\begin{pmatrix} \frac{8}{3} & \frac{19}{3} \\ \frac{2}{3} & \frac{7}{3} \end{pmatrix}}_{\mathbf{A}'} \cdot \underbrace{\begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix}}_{\text{contravariant } \mathbf{x}}, \tag{7.287}$$

$$= 152. \tag{7.288}$$

Note that $\mathbf{x}^T \cdot \mathbf{G}$ gives the covariant representation of $\mathbf{x}$.

Example 7.33

Given a space spanned by the functions $u_1 = 1$, $u_2 = t$, $u_3 = t^2$, for $t \in [0, 1]$ find a reciprocal basis u_1^R, u_2^R, u_3^R within this space.

We insist that

$$< u_n, u_m^R > = \int_0^1 u_n(t) u_m^R(t) dt = \delta_{nm}. \tag{7.289}$$

[20] If $\mathbf{J}$ had been a rotation matrix $\mathbf{Q}$, for which $\mathbf{Q}^T = \mathbf{Q}^{-1}$ and $\det \mathbf{Q} = 1$, then $\mathbf{A}' = \mathbf{Q}^T \cdot \mathbf{A} \cdot \mathbf{Q}$ from Eq. (6.80). Here our linear transformation has both stretching and rotation associated with it.

If we assume that

$$u_1^R = a_1 + a_2 t + a_3 t^2, \tag{7.290}$$
$$u_2^R = b_1 + b_2 t + b_3 t^2, \tag{7.291}$$
$$u_3^R = c_1 + c_2 t + c_3 t^2, \tag{7.292}$$

and substitute directly into Eq. (7.289), it is easy to find that

$$u_1^R = 9 - 36t + 30t^2, \tag{7.293}$$
$$u_2^R = -36 + 192t - 180t^2, \tag{7.294}$$
$$u_3^R = 30 - 180t + 180t^2. \tag{7.295}$$

7.4 Operators

- For two sets $\mathbb{X}$ and $\mathbb{Y}$, an *operator* (or *mapping*, or *transformation*) f is a rule that associates every $x \in \mathbb{X}$ with an *image* $y \in \mathbb{Y}$. We can write $f : \mathbb{X} \to \mathbb{Y}$, $\mathbb{X} \xrightarrow{f} \mathbb{Y}$ or $x \mapsto y$. $\mathbb{X}$ is the *domain* of the operator, and $\mathbb{Y}$ is the *range*.
- If every element of $\mathbb{Y}$ is not necessarily an image, then $\mathbb{X}$ is mapped *into* $\mathbb{Y}$; this map is called an *injection*.
- If, on the other hand, every element of $\mathbb{Y}$ is an image of some element of $\mathbb{X}$, then $\mathbb{X}$ is mapped *onto* $\mathbb{Y}$ and the map is a *surjection*.
- If, $\forall x \in \mathbb{X}$ there is a unique $y \in \mathbb{Y}$, and for every $y \in \mathbb{Y}$ there is a unique $x \in \mathbb{X}$, the operator is *one-to-one* or *invertible*; it is a *bijection*.
- f and g are inverses of each other, when $\mathbb{X} \xrightarrow{f} \mathbb{Y}$ and $\mathbb{Y} \xrightarrow{g} \mathbb{X}$.
- $f : \mathbb{X} \to \mathbb{Y}$ is continuous at $x_0 \in \mathbb{X}$ if, for every $\epsilon > 0$, there is a $\delta > 0$, such that $||f(x) - f(x_0)|| < \epsilon \ \forall \ x$ satisfying $||x - x_0|| < \delta$.
- If for every bounded sequence x_n in a Hilbert space the sequence $f(x_n)$ contains a convergent subsequence, then f is said to be *compact*.

A Venn diagram showing various classes of operators is given in Fig. 7.11. Examples of continuous operators are:

1. $(x_1, x_2, \cdots, x_N) \mapsto y$, where $y = f(x_1, x_2, \cdots, x_N)$.
2. $f \mapsto g$, where $g = df/dt$.

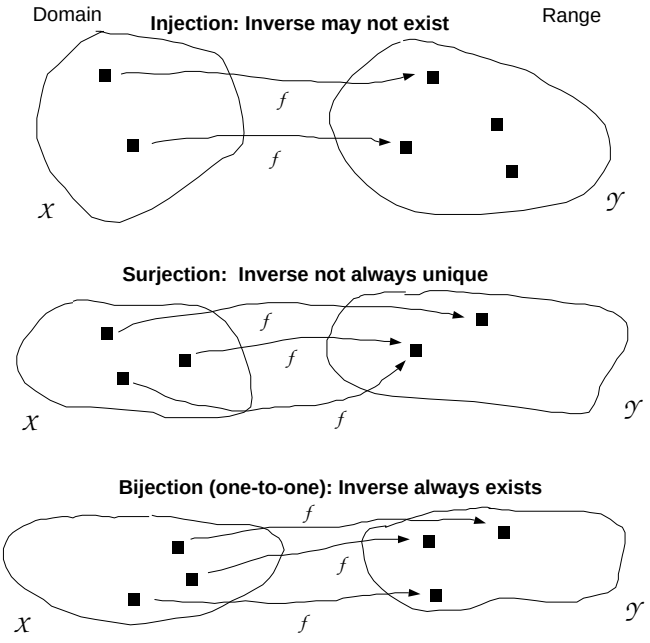

Figure 7.11: Venn diagram showing classes of operators.

3. $f \mapsto g$, where $g(t) = \int_a^b K(s,t) f(s)\ ds$. $K(s,t)$ is called the *kernel* of the integral transformation. If $\int_a^b \int_a^b |K(s,t)|^2\ ds\ dt$ is finite, then f belongs to $\mathbb{L}_2$ if g does.

4. $(x_1, x_2, \cdots, x_M)^T \mapsto (y_1, y_2, \cdots, y_N)^T$, where $y = \mathbf{A}x$ with y, $\mathbf{A}$, and x being $N \times 1$, $N \times M$, and $M \times 1$ matrices, respectively ($y_{N\times 1} = \mathbf{A}_{N\times M} x_{M\times 1}$), and the usual matrix multiplication is assumed. Here $\mathbf{A}$ is a *left operator*, and is the most common type of matrix operator.

5. $(x_1, x_2, \cdots, x_N) \mapsto (y_1, y_2, \cdots, y_M)$, where $y = x\mathbf{A}$ with y, x, and $\mathbf{A}$ being $1 \times M$, $1 \times N$, and $N \times M$ matrices, respectively ($y_{1\times M} = x_{1\times N} \mathbf{A}_{N\times M}$), and the usual matrix multiplication is assumed. Here $\mathbf{A}$ is a *right operator*.

7.4.1 Linear operators

- A *linear* operator $\mathbf{L}$ is one that satisfies

$$\mathbf{L}(x+y) = \mathbf{L}x + \mathbf{L}y, \tag{7.296}$$

$$\mathbf{L}(\alpha x) = \alpha \mathbf{L}x. \tag{7.297}$$

- An operator $\mathbf{L}$ is *bounded* if $\forall x \in \mathbb{X} \exists$ a constant c such that

$$||\mathbf{L}x|| \leq c||x||. \tag{7.298}$$

A derivative is an example of an unbounded linear operator.

- A special operator is the *identity* $\mathbf{I}$, which is defined by $\mathbf{I}x = x$.
- The *null space* or *kernel* of an operator $\mathbf{L}$ is the set of all x such that $\mathbf{L}x = 0$. The null space is a vector space.
- The norm of an operator $\mathbf{L}$ can be defined as

$$||\mathbf{L}|| = \sup_{x \neq 0} \frac{||\mathbf{L}x||}{||x||}. \tag{7.299}$$

- An operator $\mathbf{L}$ is

  positive definite if $<\mathbf{L}x, x> \ > 0$,
  positive semi-definite if $<\mathbf{L}x, x> \ \geq 0$,
  negative definite if $<\mathbf{L}x, x> \ < 0$,
  negative semi-definite if $<\mathbf{L}x, x> \ \leq 0$,

  $\forall\ x \neq 0$.

- For a matrix $\mathbf{A}$, $\mathbb{C}^m \rightarrow \mathbb{C}^N$, the *spectral norm* $||\mathbf{A}||_2$ is defined as

$$||\mathbf{A}||_2 = \sup_{x \neq 0} \frac{||\mathbf{A}x||_2}{||x||_2}. \tag{7.300}$$

  This can be shown to reduce to

$$||\mathbf{A}||_2 = \sqrt{\kappa_{max}}, \tag{7.301}$$

  where κ_{max} is the largest eigenvalue of the matrix $\overline{\mathbf{A}}^T \cdot \mathbf{A}$. It will soon be shown in Sec. 7.4.4 that because $\overline{\mathbf{A}}^T \cdot \mathbf{A}$ is symmetric, that all of its eigenvalues are guaranteed real. Moreover, it can be shown that they are also all greater than or equal to zero. Hence, the definition will satisfy all properties of the norm. This holds only for Hilbert spaces and not for arbitrary Banach spaces. There are also other valid definitions of norms for matrix operators. For example, the *p-norm* of a matrix $\mathbf{A}$ is

$$||\mathbf{A}||_p = \sup_{x \neq 0} \frac{||\mathbf{A}x||_p}{||x||_p}. \tag{7.302}$$

7.4.2 Adjoint operators

The operator $\mathbf{L}^*$ is the *adjoint* of the operator $\mathbf{L}$, if

$$<\mathbf{L}x, y> = <x, \mathbf{L}^* y>. \tag{7.303}$$

If $\mathbf{L}^* = \mathbf{L}$, the operator is self-adjoint.

Example 7.34

Find the adjoint of the real matrix $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$, where

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{7.304}$$

We assume $a_{11}, a_{12}, a_{21}, a_{22}$ are known constants.

Let the adjoint of $\mathbf{A}$ be

$$\mathbf{A}^* = \begin{pmatrix} a_{11}^* & a_{12}^* \\ a_{21}^* & a_{22}^* \end{pmatrix}. \tag{7.305}$$

Here the starred quantities are to be determined. We also have for x and y:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{7.306}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{7.307}$$

We take Eq. (7.303) and expand:

$$<\mathbf{A}x, y> = <x, \mathbf{A}^* y>, \tag{7.308}$$

$$\left(\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right)^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\begin{pmatrix} a_{11}^* & a_{12}^* \\ a_{21}^* & a_{22}^* \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right), \tag{7.309}$$

$$\begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} a_{11}^* y_1 + a_{12}^* y_2 \\ a_{21}^* y_1 + a_{22}^* y_2 \end{pmatrix}, \tag{7.310}$$

$$(\, a_{11}x_1 + a_{12}x_2 \quad a_{21}x_1 + a_{22}x_2 \,) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (\, x_1 \quad x_2 \,) \begin{pmatrix} a_{11}^* y_1 + a_{12}^* y_2 \\ a_{21}^* y_1 + a_{22}^* y_2 \end{pmatrix}, \tag{7.311}$$

$$(a_{11}x_1 + a_{12}x_2)y_1 + (a_{21}x_1 + a_{22}x_2)y_2 = x_1(a_{11}^* y_1 + a_{12}^* y_2) + x_2(a_{21}^* y_1 + a_{22}^* y_2). \tag{7.312}$$

Rearrange and get

$$(a_{11} - a_{11}^*)x_1 y_1 + (a_{21} - a_{12}^*)x_1 y_2 + (a_{12} - a_{21}^*)x_2 y_1 + (a_{22} - a_{22}^*)x_2 y_2 = 0. \tag{7.313}$$

Since this must hold for any x_1, x_2, y_1, y_2, we have

$$a_{11}^* = a_{11}, \tag{7.314}$$

$$a_{12}^* = a_{21}, \tag{7.315}$$

$$a_{21}^* = a_{12}, \tag{7.316}$$

$$a_{22}^* = a_{22}. \tag{7.317}$$

Thus,

$$\mathbf{A}^* = \begin{pmatrix} a_{11} & a_{21}, \\ a_{12} & a_{22} \end{pmatrix}, \tag{7.318}$$

$$= \mathbf{A}^T. \tag{7.319}$$

Thus, a symmetric matrix is self-adjoint. This result is easily extended to complex matrices $\mathbf{A} : \mathbb{C}^N \to \mathbb{C}^m$:

$$\mathbf{A}^* = \overline{\mathbf{A}}^T. \tag{7.320}$$

Example 7.35

Find the adjoint of the differential operator $\mathbf{L} : \mathbb{X} \to \mathbb{X}$, where

$$\mathbf{L} = \frac{d^2}{ds^2} + \frac{d}{ds}, \tag{7.321}$$

and $\mathbb{X}$ is the subspace of $\mathbb{L}_2[0,1]$ with $x(0) = x(1) = 0$ if $x \in \mathbb{X}$.

Using integration by parts on the inner product

$$\begin{aligned}
<\mathbf{L}x, y> &= \int_0^1 (x''(s) + x'(s))\ y(s)\ ds, && (7.322)\\
&= \int_0^1 x''(s)y(s)\ ds + \int_0^1 x'(s)y(s)\ ds, && (7.323)\\
&= \left(x'(1)y(1) - x'(0)y(0) - \int_0^1 x'(s)y'(s)\ ds\right) + \left(\underbrace{x(1)}_{=0} y(1) - \underbrace{x(0)}_{=0} y(0) - \int_0^1 x(s)y'(s)\ ds\right), && (7.324)\\
&= x'(1)y(1) - x'(0)y(0) - \int_0^1 x'(s)y'(s)\ ds - \int_0^1 x(s)y'(s)\ ds, && (7.325)\\
&= x'(1)y(1) - x'(0)y(0) - \left(\underbrace{x(1)}_{=0} y'(1) - \underbrace{x(0)}_{=0} y'(0) - \int_0^1 x(s)y''(s)ds\right) - \int_0^1 x(s)y'(s)ds, && (7.326)\\
&= x'(1)y(1) - x'(0)y(0) + \int_0^1 x(s)y''(s)\ ds - \int_0^1 x(s)y'(s)\ ds, && (7.327)\\
&= x'(1)y(1) - x'(0)y(0) + \int_0^1 x(s)\left(y''(s) - y'(s)\right)\ ds. && (7.328)
\end{aligned}$$

This maintains the form of an inner product in $\mathbb{L}_2[0,1]$ if we require $y(0) = y(1) = 0$; doing this, we get

$$<\mathbf{L}x, y> = \int_0^1 x(s)\left(y''(s) - y'(s)\right)\ ds = <x, \mathbf{L}^* y>. \tag{7.329}$$

We see by inspection that the adjoint operator is

$$\mathbf{L}^* = \frac{d^2}{ds^2} - \frac{d}{ds}. \tag{7.330}$$

Because the adjoint operator is not equal to the operator itself, the operator is not self-adjoint.

Example 7.36

Find the adjoint of the differential operator $\mathbf{L} : \mathbb{X} \to \mathbb{X}$, where $\mathbf{L} = d^2/ds^2$, and $\mathbb{X}$ is the subspace of $\mathbb{L}_2[0,1]$ with $x(0) = x(1) = 0$ if $x \in \mathbb{X}$.

Using integration by parts on the inner product

$$\begin{aligned} <\mathbf{L}x, y> &= \int_0^1 x''(s)y(s)\ ds, && (7.331)\\ &= x'(1)y(1) - x'(0)y(0) - \int_0^1 x'(s)y'(s)\ ds, && (7.332)\\ &= x'(1)y(1) - x'(0)y(0) - \left(\underbrace{x(1)}_{=0} y'(1) - \underbrace{x(0)}_{=0} y'(0) - \int_0^1 x(s)y''(s)\ ds \right), && (7.333)\\ &= x'(1)y(1) - x'(0)y(0) + \int_0^1 x(s)y''(s)\ ds. && (7.334) \end{aligned}$$

If we require $y(0) = y(1) = 0$, then

$$<\mathbf{L}x, y> = \int_0^1 x(s)y''(s)\ dt = <x, \mathbf{L}^* y>. \tag{7.335}$$

In this case, we see that $\mathbf{L} = \mathbf{L}^*$, so the operator is self-adjoint.

Example 7.37

Find the adjoint of the integral operator $\mathbf{L} : \mathbb{L}_2[a, b] \to \mathbb{L}_2[a, b]$, where

$$\mathbf{L}x = \int_a^b K(s, t)x(s)\ ds. \tag{7.336}$$

The inner product

$$\begin{aligned} <\mathbf{L}x, y> &= \int_a^b \left(\int_a^b K(s, t)x(s)\ ds \right) y(t)\ dt, && (7.337)\\ &= \int_a^b \int_a^b K(s, t)x(s)y(t)\ ds\ dt, && (7.338)\\ &= \int_a^b \int_a^b x(s)K(s, t)y(t)\ dt\ ds, && (7.339)\\ &= \int_a^b x(s) \left(\int_a^b K(s, t)y(t)\ dt \right) ds, && (7.340)\\ &= <x, \mathbf{L}^* y, > && (7.341) \end{aligned}$$

where

$$\mathbf{L}^* y = \int_a^b K(s, t)y(t)\ dt, \tag{7.342}$$

or equivalently

$$\mathbf{L}^* y = \int_a^b K(t, s)y(s)\ ds. \tag{7.343}$$

Note in the definition of $\mathbf{L}x$, the second argument of K is a free variable, while in the consequent definition of $\mathbf{L}^* y$, the first argument of K is a free argument. So in general, the operator and its adjoint are different. Note however, that

$$\text{if} \qquad K(s,t) = K(t,s), \qquad \text{then the operator is self-adjoint.} \tag{7.344}$$

That is, a symmetric kernel yields a self-adjoint operator.

Properties:

$$\begin{aligned} ||\mathbf{L}^*|| &= ||\mathbf{L}||, && (7.345)\\ (\mathbf{L}_1 + \mathbf{L}_2)^* &= \mathbf{L}_1^* + \mathbf{L}_2^*, && (7.346)\\ (\alpha \mathbf{L})^* &= \overline{\alpha} \mathbf{L}^*, && (7.347)\\ (\mathbf{L}_1 \mathbf{L}_2)^* &= \mathbf{L}_2^* \mathbf{L}_1^*, && (7.348)\\ (\mathbf{L}^*)^* &= \mathbf{L}, && (7.349)\\ (\mathbf{L}^{-1})^* &= (\mathbf{L}^*)^{-1}, \quad \text{if } \mathbf{L}^{-1} \text{ exists.} && (7.350) \end{aligned}$$

7.4.3 Inverse operators

Let

$$\mathbf{L}x = y. \tag{7.351}$$

If an *inverse* of $\mathbf{L}$ exists, which we will call $\mathbf{L}^{-1}$, then

$$x = \mathbf{L}^{-1} y. \tag{7.352}$$

Using Eq. (7.352) to eliminate x in favor of y in Eq. (7.351), we get

$$\mathbf{L} \underbrace{\mathbf{L}^{-1} y}_{=x} = y, \tag{7.353}$$

so that

$$\mathbf{L}\mathbf{L}^{-1} = \mathbf{I}. \tag{7.354}$$

A property of the inverse operator is

$$(\mathbf{L}_a \mathbf{L}_b)^{-1} = \mathbf{L}_b^{-1} \mathbf{L}_a^{-1} \tag{7.355}$$

Let's show this. Say

$$y = \mathbf{L}_a \mathbf{L}_b x. \tag{7.356}$$

Then

$$\begin{aligned} \mathbf{L}_a^{-1} y &= \mathbf{L}_b x, && (7.357)\\ \mathbf{L}_b^{-1} \mathbf{L}_a^{-1} y &= x. && (7.358) \end{aligned}$$

Consequently, we see that

$$(\mathbf{L}_a\mathbf{L}_b)^{-1} = \mathbf{L}_b^{-1}\mathbf{L}_a^{-1}. \tag{7.359}$$

Example 7.38

Let $\mathbf{L}$ be the operator defined by

$$\mathbf{L}x = \left(\frac{d^2}{dt^2} + k^2\right) x(t) = f(t), \tag{7.360}$$

where x belongs to the subspace of $\mathbb{L}_2[0,\pi]$ with $x(0) = a$ and $x(\pi) = b$. Show that the inverse operator $\mathbf{L}^{-1}$ is given by

$$x(t) = \mathbf{L}^{-1}f(t) = b\frac{\partial g}{\partial \tau}(\pi,t) - a\frac{\partial g}{\partial \tau}(0,t) + \int_0^\pi g(\tau,t)f(\tau)\ d\tau, \tag{7.361}$$

where $g(\tau,t)$ is the Green's function.

From the definition of $\mathbf{L}$ and $\mathbf{L}^{-1}$ in Eqs. (7.360,7.361), we get

$$\mathbf{L}^{-1}(\mathbf{L}x) = b\frac{\partial g}{\partial \tau}(\pi,t) - a\frac{\partial g}{\partial \tau}(0,t) + \int_0^\pi g(\tau,t)\underbrace{\left(\frac{d^2x(\tau)}{d\tau^2} + k^2x(\tau)\right)}_{=f(\tau)}\ d\tau. \tag{7.362}$$

Using integration by parts and the property that $g(0,t) = g(\pi,t) = 0$, the integral on the right side of Eq. (7.362 can be simplified as

$$\begin{aligned}\int_0^\pi g(\tau,t)\underbrace{\left(\frac{d^2x(\tau)}{d\tau^2} + k^2x(\tau)\right)}_{=f(\tau)}\ d\tau &= -\underbrace{x(\pi)}_{=b}\frac{\partial g}{\partial \tau}(\pi,t) + \underbrace{x(0)}_{=a}\frac{\partial g}{\partial \tau}(0,t) \\ &\quad + \int_0^\pi x(\tau)\underbrace{\left(\frac{\partial^2 g}{\partial \tau^2} + k^2 g\right)}_{=\delta(t-\tau)}\ d\tau. \end{aligned} \tag{7.363}$$

Since $x(0) = a$, $x(\pi) = b$, and

$$\frac{\partial^2 g}{\partial \tau^2} + k^2 g = \delta(t-\tau), \tag{7.364}$$

we have

$$\mathbf{L}^{-1}(\mathbf{L}x) = \int_0^\pi x(\tau)\delta(t-\tau)\ d\tau, \tag{7.365}$$

$$= x(t). \tag{7.366}$$

Thus, $\mathbf{L}^{-1}\mathbf{L} = \mathbf{I}$, proving the proposition.

Note, it is easily shown for this problem that the Green's function is

$$g(\tau,t) = -\frac{\sin(k(\pi-\tau))\sin(kt)}{k\sin(k\pi)} \qquad t < \tau, \tag{7.367}$$

$$= -\frac{\sin(k\tau)\sin(k(\pi-t))}{k\sin(k\pi)} \qquad \tau < t, \tag{7.368}$$

so that we can write $x(t)$ explicitly in terms of the forcing function $f(t)$ including the inhomogeneous boundary conditions as follows:

$$x(t) = \frac{b\sin(kt)}{\sin(k\pi)} + \frac{a\sin(k(\pi-t))}{\sin(k\pi)} \tag{7.369}$$

$$-\frac{\sin(k(\pi-t))}{k\sin(k\pi)}\int_0^t f(\tau)\sin(k\tau)\ d\tau - \frac{\sin(kt)}{k\sin(k\pi)}\int_t^\pi f(\tau)\sin(k(\pi-\tau))\ d\tau. \tag{7.370}$$

For linear algebraic systems, the reciprocal or dual basis can be easily formulated in terms of operator notation and is closely related to the inverse operator. If we define $\mathbf{U}$ to be a $N \times N$ matrix which has the N basis vectors u_n, each of length N, which span the N-dimensional space, we seek $\mathbf{U}^R$, the $N \times N$ matrix which has as its columns the vectors u_m^R which form the reciprocal or dual basis. The reciprocal basis is found by enforcing the equivalent of $<u_n, u_m^R> = \delta_{nm}$:

$$\overline{\mathbf{U}}^T \cdot \mathbf{U}^R = \mathbf{I}. \tag{7.371}$$

Solving for $\mathbf{U}^R$,

$$\overline{\overline{\mathbf{U}}^T \cdot \mathbf{U}^R} = \overline{\mathbf{I}}, \tag{7.372}$$

$$\mathbf{U}^T \cdot \overline{\mathbf{U}}^R = \mathbf{I}, \tag{7.373}$$

$$\left(\mathbf{U}^T \cdot \overline{\mathbf{U}}^R\right)^T = \mathbf{I}^T, \tag{7.374}$$

$$\overline{\mathbf{U}}^{R^T} \cdot \mathbf{U} = \mathbf{I}, \tag{7.375}$$

$$\overline{\mathbf{U}}^{R^T} \cdot \mathbf{U} \cdot \mathbf{U}^{-1} = \mathbf{I} \cdot \mathbf{U}^{-1}, \tag{7.376}$$

$$\overline{\mathbf{U}}^{R^T} = \mathbf{U}^{-1}, \tag{7.377}$$

$$\mathbf{U}^R = \overline{\mathbf{U}^{-1}}^T, \tag{7.378}$$

we see that the set of reciprocal basis vectors is given by the conjugate transpose of the inverse of the original matrix of basis vectors. Then the expression for the amplitudes modulating the basis vectors, $\alpha_n = <u_n^R, x>$, is

$$\boldsymbol{\alpha} = \overline{\mathbf{U}^R}^T \cdot \mathbf{x}. \tag{7.379}$$

Substituting for $\mathbf{U}^R$ in terms of its definition, we can also say

$$\boldsymbol{\alpha} = \overline{\overline{\mathbf{U}^{-1}}^T}^T \cdot \mathbf{x} = \mathbf{U}^{-1} \cdot \mathbf{x}. \tag{7.380}$$

Then the expansion for the vector $x = \sum_{n=1}^N \alpha_n u_n = \sum_{n=1}^N <u_n^R, x> u_n$ is written in the alternate notation as

$$\mathbf{x} = \mathbf{U} \cdot \boldsymbol{\alpha} = \mathbf{U} \cdot \mathbf{U}^{-1} \cdot \mathbf{x} = \mathbf{x}. \tag{7.381}$$

Example 7.39

Consider the problem of a previous example with $\mathbf{x} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ and with basis vectors $u_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $u_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$, find the reciprocal basis vectors and an expansion of $\mathbf{x}$ in terms of the basis vectors.

Using the alternate vector and matrix notation, we define the matrix of basis vectors as

$$\mathbf{U} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}. \tag{7.382}$$

Since this matrix is real, the complex conjugation process is not important, but it will be retained for completeness. Using standard techniques, we find that the inverse is

$$\mathbf{U}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{6} \\ 0 & \frac{1}{3} \end{pmatrix}. \tag{7.383}$$

Thus, the matrix with the reciprocal basis vectors in its columns is

$$\mathbf{U}^R = \overline{\mathbf{U}^{-1}}^T = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{6} & \frac{1}{3} \end{pmatrix}. \tag{7.384}$$

This agrees with the earlier analysis. For $\mathbf{x} = (3, 5)^T$, we find the coefficients $\boldsymbol{\alpha}$ to be

$$\boldsymbol{\alpha} = \overline{\mathbf{U}^R}^T \cdot \mathbf{x} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{6} \\ 0 & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix}. \tag{7.385}$$

We see that we do indeed recover $\mathbf{x}$ upon taking the product

$$\mathbf{x} = \mathbf{U} \cdot \boldsymbol{\alpha} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix} = \frac{2}{3} \begin{pmatrix} 2 \\ 0 \end{pmatrix} + \frac{5}{3} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}. \tag{7.386}$$

7.4.4 Eigenvalues and eigenvectors

Let us consider here in a more formal fashion topics that have been previously introduced in Secs. 5.1 and 6.2.5. If $\mathbf{L}$ is a linear operator, its eigenvalue problem consists of finding a nontrivial solution of the equation

$$\mathbf{L}e = \lambda e, \tag{7.387}$$

where e is called an *eigenvector*, and λ an *eigenvalue*.

Theorem

The eigenvalues of an operator and its adjoint are complex conjugates of each other.

Proof: Let λ and λ^* be the eigenvalues of $\mathbf{L}$ and $\mathbf{L}^*$, respectively, and let e and e^* be the corresponding eigenvectors. Consider then,

$$\begin{aligned} <\mathbf{L}e, e^*> &= <e, \mathbf{L}^* e^*>, & (7.388)\\ <\lambda e, e^*> &= <e, \lambda^* e^*>, & (7.389)\\ \overline{\lambda}<e, e^*> &= \lambda^* <e, e^*>, & (7.390)\\ \overline{\lambda} &= \lambda^*. & (7.391) \end{aligned}$$

This holds for $<e, e^*> \neq 0$, which will hold in general.

Theorem

The eigenvalues of a self-adjoint operator are real.

Proof:

Since the operator is self-adjoint, we have

$$\begin{aligned} <\mathbf{L}e, e> &= <e, \mathbf{L}e>, & (7.392)\\ <\lambda e, e> &= <e, \lambda e>, & (7.393)\\ \overline{\lambda}<e, e> &= \lambda <e, e>, & (7.394)\\ \overline{\lambda} &= \lambda, & (7.395)\\ \lambda_R - i\lambda_I &= \lambda_R + i\lambda_I; \qquad \lambda_R, \lambda_I \in \mathbb{R}^2, & (7.396)\\ \lambda_R &= \lambda_R, & (7.397)\\ -\lambda_I &= \lambda_I, & (7.398)\\ \lambda_I &= 0. & (7.399) \end{aligned}$$

Here we note that for non-trivial eigenvectors $<e, e> > 0$, so the division can be performed. The only way a complex number can equal its conjugate is if its imaginary part is zero; consequently, the eigenvalue must be strictly real.

Theorem

The eigenvectors of a self-adjoint operator corresponding to distinct eigenvalues are orthogonal.

Proof: Let λ_i and λ_j be two distinct, $\lambda_i \neq \lambda_j$, real, $\lambda_i, \lambda_j \in \mathbb{R}^1$, eigenvalues of the self-adjoint operator $\mathbf{L}$, and let e_i and e_j be the corresponding eigenvectors. Then,

$$\begin{aligned} <\mathbf{L}e_i, e_j> &= <e_i, \mathbf{L}e_j>, & (7.400)\\ <\lambda_i e_i, e_j> &= <e_i, \lambda_j e_j>, & (7.401)\\ \lambda_i <e_i, e_j> &= \lambda_j <e_i, e_j>, & (7.402)\\ <e_i, e_j> (\lambda_i - \lambda_j) &= 0, & (7.403)\\ <e_i, e_j> &= 0, & (7.404) \end{aligned}$$

since $\lambda_i \neq \lambda_j$.

Theorem

The eigenvectors of any self-adjoint operator on vectors of a finite-dimensional vector space constitute a basis for the space.

As discussed by Friedman, the following conditions are sufficient for the eigenvectors in an infinite-dimensional Hilbert space to be form a complete basis:

- the operator must be self-adjoint,
- the operator is defined on a finite domain, and
- the operator has no singularities in its domain.

If the operator is not self-adjoint, Friedman (p. 204) discusses how the eigenfunctions of the adjoint operator can be used to obtain the coefficients α_k on the eigenfunctions of the operator.

Example 7.40

For $x \in \mathbb{R}^2, \mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$, Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{7.405}$$

The eigenvalue problem is

$$\mathbf{A}x = \lambda x, \tag{7.406}$$

which can be written as

$$\begin{aligned} \mathbf{A}x &= \lambda \mathbf{I}x, & (7.407)\\ (\mathbf{A} - \lambda \mathbf{I})x &= 0, & (7.408) \end{aligned}$$

where the identity matrix is

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{7.409}$$

If we write

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{7.410}$$

then

$$\begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{7.411}$$

By Cramer's rule we could say

$$x_1 = \frac{\det \begin{pmatrix} 0 & 1 \\ 0 & 2-\lambda \end{pmatrix}}{\det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}} = \frac{0}{\det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}}, \tag{7.412}$$

$$x_2 = \frac{\det \begin{pmatrix} 2-\lambda & 0 \\ 1 & 0 \end{pmatrix}}{\det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}} = \frac{0}{\det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}}. \tag{7.413}$$

An obvious, but uninteresting solution is the trivial solution $x_1 = 0, x_2 = 0$. Nontrivial solutions of x_1 and x_2 can be obtained only if

$$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0, \tag{7.414}$$

which gives the characteristic equation

$$(2-\lambda)^2 - 1 = 0. \tag{7.415}$$

Solutions are $\lambda_1 = 1$ and $\lambda_2 = 3$. The eigenvector corresponding to each eigenvalue is found in the following manner. The eigenvalue is substituted in Eq. (7.411). A dependent set of equations in x_1 and x_2 is obtained. The eigenvector solution is thus not unique.

For $\lambda = 1$, Eq. (7.411) gives

$$\begin{pmatrix} 2-1 & 1 \\ 1 & 2-1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{7.416}$$

which are the two identical equations,

$$x_1 + x_2 = 0. \tag{7.417}$$

If we choose $x_1 = \gamma$, then $x_2 = -\gamma$. So the eigenvector corresponding to $\lambda = 1$ is

$$e_1 = \gamma \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{7.418}$$

Since the magnitude of an eigenvector is arbitrary, we will take $\gamma = 1$ and thus

$$e_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{7.419}$$

For $\lambda = 3$, the equations are

$$\begin{pmatrix} 2-3 & 1 \\ 1 & 2-3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{7.420}$$

which yield the two identical equations,

$$-x_1 + x_2 = 0. \tag{7.421}$$

This yields an eigenvector of

$$e_2 = \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{7.422}$$

We take $\beta = 1$, so that

$$e_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{7.423}$$

Comments:

- Since the real matrix is symmetric (thus, self-adjoint), the eigenvalues are real, and the eigenvectors are orthogonal.

- We have actually solved for the *right eigenvectors*. This is the usual set of eigenvectors. The *left eigenvectors* can be found from $\overline{x}^T\mathbf{A} = \overline{x}^T\mathbf{I}\lambda$. Since here $\mathbf{A}$ is equal to its conjugate transpose, $\overline{x}^T\mathbf{A} = \mathbf{A}x$, so the left eigenvectors are the same as the right eigenvectors. More generally, we can say the left eigenvectors of an operator are the right eigenvectors of the adjoint of that operator, $\overline{\mathbf{A}}^T$.
- Multiplication of an eigenvector by any scalar is also an eigenvector.
- The normalized eigenvectors are

$$e_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}, \qquad e_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{7.424}$$

- A natural way to express a vector is on orthonormal basis as given here

$$x = \alpha_1 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} + \alpha_2 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}}_{=\mathbf{Q}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{7.425}$$

- The set of orthonormalized eigenvectors forms an orthogonal matrix $\mathbf{Q}$; see p. 183 or the upcoming Sec. 8.6. Note that it has determinant of unity, so it is a rotation. As suggested by Eq. (6.54), the angle of rotation here is $\alpha = \sin^{-1}(-1/\sqrt{2}) = -\pi/4$.

Example 7.41

For $x \in \mathbb{C}^2, \mathbf{A} : \mathbb{C}^2 \to \mathbb{C}^2$, find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}. \tag{7.426}$$

This matrix is anti-symmetric. We find the eigensystem by solving

$$(\mathbf{A} - \lambda\mathbf{I})\, e = 0. \tag{7.427}$$

The characteristic equation which results is

$$\lambda^2 + 4 = 0, \tag{7.428}$$

which has two imaginary roots which are complex conjugates: $\lambda_1 = 2i$, $\lambda_2 = -2i$. The corresponding eigenvectors are

$$e_1 = \alpha \begin{pmatrix} i \\ 1 \end{pmatrix}, \qquad e_2 = \beta \begin{pmatrix} -i \\ 1 \end{pmatrix}, \tag{7.429}$$

where α and β are arbitrary scalars. Let us take $\alpha = -i$, $\beta = 1$, so

$$e_1 = \begin{pmatrix} 1 \\ -i \end{pmatrix}, \qquad e_2 = \begin{pmatrix} -i \\ 1 \end{pmatrix}. \tag{7.430}$$

Note that

$$<e_1, e_2> = \overline{e_1}^T e_2 = \begin{pmatrix} 1 & i \end{pmatrix} \begin{pmatrix} -i \\ 1 \end{pmatrix} = (-i) + i = 0, \tag{7.431}$$

so this is an orthogonal set of vectors, even though the generating matrix was not self-adjoint. We can render it orthonormal by scaling by the magnitude of each eigenvector. The orthonormal eigenvector set is

$$e_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-i}{\sqrt{2}} \end{pmatrix}, \qquad e_2 = \begin{pmatrix} -\frac{i}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{7.432}$$

These two orthonormalized vectors can form a matrix $\mathbf{Q}$:

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{7.433}$$

It is easy to check that $||\mathbf{Q}||_2 = 1$ and $\det \mathbf{Q} = 1$, so it is a rotation. However, for the complex basis vectors, it is difficult to define an angle of rotation in the traditional sense. Our special choices of α and β were actually made to ensure $\det \mathbf{Q} = 1$.

Example 7.42

For $x \in \mathbb{C}^2, \mathbf{A} : \mathbb{C}^2 \to \mathbb{C}^2$, find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}. \tag{7.434}$$

This matrix is asymmetric. We find the eigensystem by solving

$$(\mathbf{A} - \lambda \mathbf{I})\, e = 0. \tag{7.435}$$

The characteristic equation which results is

$$(1 - \lambda)^2 = 0, \tag{7.436}$$

which has repeated roots $\lambda = 1,\ \lambda = 1$. For this eigenvalue, there is only one ordinary eigenvector

$$e = \alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{7.437}$$

We take arbitrarily $\alpha = 1$ so that

$$e = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{7.438}$$

We can however find a *generalized eigenvector* g such that

$$(\mathbf{A} - \lambda \mathbf{I}) g = e. \tag{7.439}$$

Note then that

$$(\mathbf{A} - \lambda \mathbf{I})(\mathbf{A} - \lambda \mathbf{I}) g = (\mathbf{A} - \lambda \mathbf{I}) e, \tag{7.440}$$

$$(\mathbf{A} - \lambda \mathbf{I})^2 g = 0. \tag{7.441}$$

Now

$$(\mathbf{A} - \lambda\mathbf{I}) = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}. \tag{7.442}$$

So with $g = (\beta, \gamma)^T$, take from Eq. (7.439)

$$\underbrace{\begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}}_{\mathbf{A}-\lambda\mathbf{I}} \underbrace{\begin{pmatrix} \beta \\ \gamma \end{pmatrix}}_{g} = \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{e}. \tag{7.443}$$

We get a solution if $\beta \in \mathbb{R}^1, \gamma = -1$. That is

$$g = \begin{pmatrix} \beta \\ -1 \end{pmatrix}. \tag{7.444}$$

Take $\beta = 0$ to give an orthogonal generalized eigenvector. So

$$g = \begin{pmatrix} 0 \\ -1 \end{pmatrix}. \tag{7.445}$$

Note that the ordinary eigenvector and the generalized eigenvector combine to form a basis, in this case an orthonormal basis.

More properly, we should distinguish the generalized eigenvector we have found as a *generalized eigenvector in the first sense*. There is another common, unrelated generalization in usage which we will study later in Sec. 8.3.2.

Example 7.43

For $x \in \mathbb{C}^2, \mathbf{A} : \mathbb{C}^2 \to \mathbb{C}^2$, find the eigenvalues, right eigenvectors, and left eigenvectors if

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -3 & 1 \end{pmatrix}. \tag{7.446}$$

The right eigenvector problem is the usual

$$\mathbf{A}e_R = \lambda\mathbf{I}e_R. \tag{7.447}$$

The characteristic polynomial is

$$(1 - \lambda)^2 + 6 = 0, \tag{7.448}$$

which has complex roots. The eigensystem is

$$\lambda_1 = 1 - \sqrt{6}i, \quad e_{1R} = \begin{pmatrix} \sqrt{\frac{2}{3}}i \\ 1 \end{pmatrix}, \qquad \lambda_2 = 1 + \sqrt{6}i, \quad e_{2R} = \begin{pmatrix} -\sqrt{\frac{2}{3}}i \\ 1 \end{pmatrix}. \tag{7.449}$$

Note as the operator is not self-adjoint, we are not guaranteed real eigenvalues. The right eigenvectors are not orthogonal as $\overline{e_{1R}}^T e_{2R} = 1/3$.

For the left eigenvectors, we have

$$\overline{e}_L^T\mathbf{A} = \overline{e}_L^T\mathbf{I}\lambda. \tag{7.450}$$

We can put this in a slightly more standard form by taking the conjugate transpose of both sides:

$$\overline{\overline{e}_L^T \mathbf{A}}^T = \overline{\overline{e}_L^T \mathbf{I}\lambda}^T, \tag{7.451}$$

$$\overline{\mathbf{A}}^T e_L = \overline{\mathbf{I}\lambda} e_L, \tag{7.452}$$

$$\overline{\mathbf{A}}^T e_L = \mathbf{I}\overline{\lambda} e_L, \tag{7.453}$$

$$\mathbf{A}^* e_L = \mathbf{I}\lambda^* e_L. \tag{7.454}$$

So the left eigenvectors of $\mathbf{A}$ are the right eigenvectors of the adjoint of $\mathbf{A}$. Now we have

$$\overline{\mathbf{A}}^T = \begin{pmatrix} 1 & -3 \\ 2 & 1 \end{pmatrix}. \tag{7.455}$$

The resulting eigensystem is

$$\lambda_1^* = 1 + \sqrt{6}i, \quad e_{1L} = \begin{pmatrix} \sqrt{\frac{3}{2}}i \\ 1 \end{pmatrix}, \qquad \lambda_2^* = 1 - \sqrt{6}i, \quad e_{2L} = \begin{pmatrix} -\sqrt{\frac{3}{2}}i \\ 1 \end{pmatrix}. \tag{7.456}$$

Note that in addition to being complex conjugates of themselves, which does not hold for general complex matrices, the eigenvalues of the adjoint are complex conjugates of those of the original matrix, which does hold for general complex matrices. That is $\lambda^* = \overline{\lambda}$. The left eigenvectors are not orthogonal as $\overline{e_{1L}}^T e_{2L} = -\frac{1}{2}$. It is easily shown by taking the conjugate transpose of the adjoint eigenvalue problem however that

$$\overline{e}_L^T \mathbf{A} = \overline{e}_L^T \lambda, \tag{7.457}$$

as desired. Note that the eigenvalues for both the left and right eigensystems are the same.

Example 7.44

Consider a small change from the previous example. For $x \in \mathbb{C}^2$, $\mathbf{A} : \mathbb{C}^2 \to \mathbb{C}^2$, find the eigenvalues, right eigenvectors, and left eigenvectors if

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -3 & 1+i \end{pmatrix}. \tag{7.458}$$

The right eigenvector problem is the usual

$$\mathbf{A} e_R = \lambda \mathbf{I} e_R. \tag{7.459}$$

The characteristic polynomial is

$$\lambda^2 - (2+i)\lambda + (7+i) = 0, \tag{7.460}$$

which has complex roots. The eigensystem is

$$\lambda_1 = 1 - 2i, \quad e_{1R} = \begin{pmatrix} i \\ 1 \end{pmatrix}, \qquad \lambda_2 = 1 + 3i, \quad e_{2R} = \begin{pmatrix} -2i \\ 3 \end{pmatrix}. \tag{7.461}$$

Note as the operator is not self-adjoint, we are not guaranteed real eigenvalues. The right eigenvectors are not orthogonal as $\overline{e_{1R}}^T e_{2R} = 1 \neq 0$

For the left eigenvectors, we solve the corresponding right eigensystem for the adjoint of $\mathbf{A}$ which is $\mathbf{A}^* = \overline{\mathbf{A}}^T$.

$$\overline{\mathbf{A}}^T = \begin{pmatrix} 1 & -3 \\ 2 & 1-i \end{pmatrix}. \tag{7.462}$$

The eigenvalue problem is $\overline{\mathbf{A}}^T e_L = \lambda^* e_L$. The eigensystem is

$$\lambda_1^* = 1+2i, \quad e_{1L} = \begin{pmatrix} 3i \\ 2 \end{pmatrix}; \qquad \lambda_2^* = 1-3i, \quad e_{2L} = \begin{pmatrix} -i \\ 1 \end{pmatrix}. \tag{7.463}$$

Note that here, the eigenvalues λ_1^*, λ_2^* have no relation to each other, but they are complex conjugates of the eigenvalues, λ_1, λ_2, of the right eigenvalue problem of the original matrix. The left eigenvectors are not orthogonal as $\overline{e_{1L}}^T e_{2L} = -1$. It is easily shown however that

$$\overline{e}_L^T \mathbf{A} = \overline{e}_L^T \lambda \mathbf{I}, \tag{7.464}$$

as desired.

Example 7.45

For $x \in \mathbb{R}^3$, $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^3$, find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \tag{7.465}$$

From

$$\begin{vmatrix} 2-\lambda & 0 & 0 \\ 0 & 1-\lambda & 1 \\ 0 & 1 & 1-\lambda \end{vmatrix} = 0, \tag{7.466}$$

the characteristic equation is

$$(2-\lambda)\left((1-\lambda)^2 - 1\right) = 0. \tag{7.467}$$

The solutions are $\lambda = 0, 2, 2$. The second eigenvalue is of multiplicity two. Next, we find the eigenvectors

$$e = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \tag{7.468}$$

For $\lambda = 0$, the equations for the components of the eigenvectors are

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{7.469}$$

$$2x_1 = 0, \tag{7.470}$$

$$x_2 + x_3 = 0, \tag{7.471}$$

from which

$$e_1 = \alpha \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}. \tag{7.472}$$

For $\lambda = 2$, we have

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{7.473}$$

This yields only

$$-x_2 + x_3 = 0. \tag{7.474}$$

We then see that the following eigenvector,

$$e = \begin{pmatrix} \beta \\ \gamma \\ \gamma \end{pmatrix}, \tag{7.475}$$

satisfies Eq. (7.474). Here, we have two free parameters, β and γ; we can thus extract two independent eigenvectors from this. For e_2 we arbitrarily take $\beta = 0$ and $\gamma = 1$ to get

$$e_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}. \tag{7.476}$$

For e_3 we arbitrarily take $\beta = 1$ and $\gamma = 0$ to get

$$e_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{7.477}$$

In this case e_1, e_2, e_3 are orthogonal even though e_2 and e_3 correspond to the same eigenvalue.

Example 7.46

For $y \in \mathbb{L}_2[0,1]$, find the eigenvalues and eigenvectors of $\mathbf{L} = -d^2/dt^2$, operating on functions which vanish at 0 and 1. Also find $||\mathbf{L}||_2$.

The eigenvalue problem is

$$\mathbf{L}y = -\frac{d^2y}{dt^2} = \lambda y, \qquad y(0) = y(1) = 0, \tag{7.478}$$

or

$$\frac{d^2y}{dt^2} + \lambda y = 0, \qquad y(0) = y(1) = 0. \tag{7.479}$$

The solution of this differential equation is

$$y(t) = a \sin \lambda^{1/2} t + b \cos \lambda^{1/2} t. \tag{7.480}$$

The boundary condition $y(0) = 0$ gives $b = 0$. The other condition $y(1) = 0$ gives $a \sin \lambda^{1/2} = 0$. A nontrivial solution can only be obtained if

$$\sin \lambda^{1/2} = 0. \tag{7.481}$$

There are an infinite but countable number of values of λ for which this can be satisfied. These are $\lambda_n = n^2\pi^2$, $n = 1, 2, \cdots$. The eigenvectors (also called *eigenfunctions* in this case) $y_n(t)$, $n = 1, 2, \cdots$ are

$$y_n(t) = \sin n\pi t. \tag{7.482}$$

The differential operator is self-adjoint so that the eigenvalues are real and the eigenfunctions are orthogonal.

Consider $||\mathbf{L}||_2$. Referring to the definition of Eq. (7.299), we see $||\mathbf{L}||_2 = \infty$, since by allowing y to be any eigenfunction, we have

$$\frac{||\mathbf{L}y||_2}{||y||_2} = \frac{||\lambda y||_2}{||y||_2}, \tag{7.483}$$

$$= \frac{|\lambda| \cdot ||y||_2}{||y||_2}, \tag{7.484}$$

$$= |\lambda|. \tag{7.485}$$

And since $\lambda = n^2\pi^2$, $n = 1, 2, \ldots, \infty$, the largest value that can be achieved by $||\mathbf{L}y||_2/||y||_2$ is infinite.

Example 7.47

For $x \in \mathbb{L}_2[0, 1]$, and $\mathbf{L} = d^2/ds^2 + d/ds$ with $x(0) = x(1) = 0$, find the Fourier expansion of an arbitrary function $f(s)$ in terms of the eigenfunctions of $\mathbf{L}$. Find the series representation of the "top hat" function

$$f(s) = H\left(s - \frac{1}{4}\right) - H\left(s - \frac{3}{4}\right). \tag{7.486}$$

We seek expressions for α_n in

$$f(s) = \sum_{n=1}^{N} \alpha_n x_n(s). \tag{7.487}$$

Here $x_n(s)$ is an eigenfunction of $\mathbf{L}$.

The eigenvalue problem is

$$\mathbf{L}x = \frac{d^2x}{ds^2} + \frac{dx}{ds} = \lambda x, \qquad x(0) = x(1) = 0. \tag{7.488}$$

It is easily shown that the eigenvalues of $\mathbf{L}$ are given by

$$\lambda_n = -\frac{1}{4} - n^2\pi^2, \qquad n = 1, 2, 3, \ldots \tag{7.489}$$

where n is a positive integer, and the unnormalized eigenfunctions of $\mathbf{L}$ are

$$x_n(s) = e^{-s/2} \sin(n\pi s), \qquad n = 1, 2, 3, \ldots \tag{7.490}$$

Although the eigenvalues are real, the eigenfunctions are not orthogonal. We see this, for example, by forming $<x_1, x_2>$:

$$<x_1, x_2> = \int_0^1 \underbrace{e^{-s/2} \sin(\pi s)}_{=x_1(s)} \underbrace{e^{-s/2} \sin(2\pi s)}_{=x_2(s)}\, ds, \tag{7.491}$$

$$<x_1, x_2> = \frac{4(1+e)\pi^2}{e(1+\pi^2)(1+9\pi^2)} \neq 0. \tag{7.492}$$

By using integration by parts, we calculate the adjoint operator to be

$$\mathbf{L}^* y = \frac{d^2 y}{ds^2} - \frac{dy}{ds} = \lambda^* y, \qquad y(0) = y(1) = 0. \tag{7.493}$$

We then find the eigenvalues of the adjoint operator to be the same as those of the operator (this is true because the eigenvalues are real; in general they are complex conjugates of one another).

$$\lambda_m^* = \overline{\lambda_m} = -\frac{1}{4} - m^2\pi^2, \qquad m = 1, 2, 3, \ldots \tag{7.494}$$

where m is a positive integer.

The unnormalized eigenfunctions of the adjoint are

$$y_m(s) = e^{s/2} \sin(m\pi s), \qquad m = 1, 2, 3, \ldots \tag{7.495}$$

Now, since by definition $<y_m, \mathbf{L}x_n> = <\mathbf{L}^* y_m, x_n>$, we have

$$\begin{aligned} <y_m, \mathbf{L}x_n> - <\mathbf{L}^* y_m, x_n> &= 0, & (7.496)\\ <y_m, \lambda_n x_n> - <\lambda_m^* y_m, x_n> &= 0, & (7.497)\\ \lambda_n <y_m, x_n> - \overline{\lambda_m^*} <y_m, x_n> &= 0, & (7.498)\\ (\lambda_n - \lambda_m) <y_m, x_n> &= 0. & (7.499) \end{aligned}$$

So, for $m = n$, we get $<y_n, x_n> \neq 0$, and for $m \neq n$, we get $<y_m, x_n> = 0$. Thus, we must have the so-called bi-orthogonality condition

$$<y_m, x_n> = D_{mn}, \tag{7.500}$$

$$D_{mn} = 0 \qquad \text{if} \qquad m \neq n. \tag{7.501}$$

Here D_{mn} is a diagonal matrix which can be reduced to the identity matrix with proper normalization.

Now consider the following series of operations on the original form of the expansion we seek

$$\begin{aligned} f(s) &= \sum_{n=1}^{N} \alpha_n x_n(s), & (7.502)\\ <y_j(s), f(s)> &= <y_j(s), \sum_{n=1}^{N} \alpha_n x_n(s)>, & (7.503)\\ <y_j(s), f(s)> &= \sum_{n=1}^{N} \alpha_n <y_j(s), x_n(s)>, & (7.504)\\ <y_j(s), f(s)> &= \alpha_j <y_j(s), x_j(s)>, & (7.505)\\ \alpha_j &= \frac{<y_j(s), f(s)>}{<y_j(s), x_j(s)>}, & (7.506)\\ \alpha_n &= \frac{<y_n(s), f(s)>}{<y_n(s), x_n(s)>}, \qquad n = 1, 2, 3, \ldots & (7.507) \end{aligned}$$

Now in the case at hand, it is easily shown that

$$<y_n(s), x_n(s)> = \frac{1}{2}, \qquad n = 1, 2, 3, \ldots, \tag{7.508}$$

so we have

$$\alpha_n = 2 <y_n(s), f(s)>. \tag{7.509}$$

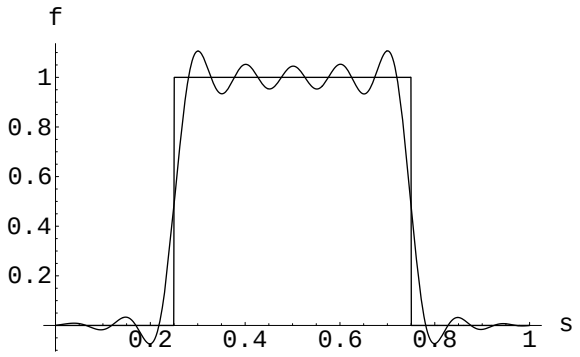

Figure 7.12: Twenty-term Fourier series approximation to a top hat function in terms of a non-orthogonal basis.

The N-term approximate representation of $f(s)$ is thus given by

$$f(s) \sim \sum_{n=1}^{N} \underbrace{\left(2\int_0^1 e^{t/2}\sin(n\pi t)\, f(t)\, dt\right)}_{=\alpha_n} \underbrace{e^{-s/2}\sin(n\pi s)}_{=x_n(s)}, \tag{7.510}$$

$$\sim 2\int_0^1 e^{(t-s)/2} f(t) \sum_{n=1}^{N} \sin(n\pi t)\sin(n\pi s)\, dt, \tag{7.511}$$

$$\sim \int_0^1 e^{(t-s)/2} f(t) \sum_{n=1}^{N} \left(\cos(n\pi(s-t)) - \cos(n\pi(s+t))\right) dt. \tag{7.512}$$

For the top hat function, a two-term expansion yields

$$f(s) \sim \underbrace{\frac{2\sqrt{2}e^{1/8}\left(-1+2\pi+e^{1/4}(1+2\pi)\right)}{1+4\pi^2}}_{=\alpha_1} \underbrace{e^{-s/2}\sin(\pi s)}_{=x_1(s)} \underbrace{-\frac{4(e^{1/8}+e^{3/8})}{1+16\pi^2}}_{=\alpha_2} \underbrace{e^{-s/2}\sin(2\pi s)}_{=x_2(s)} + \ldots. \tag{7.513}$$

A plot of a twenty-term series expansion of the top hat function is shown in Fig. 7.12.

In this exercise, the eigenfunctions of the adjoint are closely related to the reciprocal basis functions. In fact, we could have easily adjusted the constants on the eigenfunctions to obtain a true reciprocal basis. Taking

$$x_n = \sqrt{2}e^{-s/2}\sin(n\pi s), \tag{7.514}$$

$$y_m = \sqrt{2}e^{s/2}\sin(m\pi s), \tag{7.515}$$

gives $<y_m, x_n> = \delta_{mn}$, as desired for a set of reciprocal basis functions. We see that getting the Fourier coefficients for eigenfunctions of a non-self-adjoint operator requires consideration of the adjoint operator. We also note that it is often a difficult exercise in problems with practical significance to actually find the adjoint operator and its eigenfunctions.

7.5 Equations

The existence and uniqueness of the solution x of the equation

$$\mathbf{L}x = y, \tag{7.516}$$

for given linear operator $\mathbf{L}$ and y is governed by the following theorems.

Theorem

If the range of $\mathbf{L}$ is closed, $\mathbf{L}x = y$ has a solution if and only if y is orthogonal to every solution of the adjoint homogeneous equation $\mathbf{L}^* z = 0$.

Theorem

The solution of $\mathbf{L}x = y$ is non-unique if the solution of the homogeneous equation $\mathbf{L}x = 0$ is also non-unique, and conversely.

There are two basic ways in which the equation can be solved.

- Inverse: If an inverse of $\mathbf{L}$ exists then

$$x = \mathbf{L}^{-1} y. \tag{7.517}$$

- Eigenvector expansion: Assume that x, y belong to a vector space $\mathbb{S}$ and the eigenvectors $(e_1, e_2, \cdots)$ of $\mathbf{L}$ span $\mathbb{S}$. Then we can write

$$y = \sum_n \alpha_n e_n, \tag{7.518}$$

$$x = \sum_n \beta_n e_n, \tag{7.519}$$

where the α's are known and the β's are unknown. We get

$$\mathbf{L}x = y, \tag{7.520}$$

$$\mathbf{L}\underbrace{\left(\sum_n \beta_n e_n\right)}_{x} = \underbrace{\sum_n \alpha_n e_n}_{y}, \tag{7.521}$$

$$\sum_n \mathbf{L}\beta_n e_n = \sum_n \alpha_n e_n, \tag{7.522}$$

$$\sum_n \beta_n \mathbf{L} e_n = \sum_n \alpha_n e_n, \tag{7.523}$$

$$\sum_n \beta_n \lambda_n e_n = \sum_n \alpha_n e_n, \tag{7.524}$$

$$\sum_n \underbrace{(\beta_n \lambda_n - \alpha_n)}_{=0} e_n = 0, \tag{7.525}$$

where the λs are the eigenvalues of $\mathbf{L}$. Since the e_n are linearly independent, we must demand for all n that

$$\beta_n \lambda_n = \alpha_n. \tag{7.526}$$

If all $\lambda_n \neq 0$, then $\beta_n = \alpha_n/\lambda_n$ and we have the unique solution

$$x = \sum_n \frac{\alpha_n}{\lambda_n} e_n. \tag{7.527}$$

If, however, one of the λ's, λ_k say, is zero, we still have $\beta_n = \alpha_n/\lambda_n$ for $n \neq k$. For $n = k$, there are two possibilities:

- If $\alpha_k \neq 0$, no solution is possible since equation (7.526) is not satisfied for $n = k$.
- If $\alpha_k = 0$, we have the non-unique solution

$$x = \sum_{n \neq k} \frac{\alpha_n}{\lambda_n} e_n + \gamma e_k, \tag{7.528}$$

where γ is an arbitrary scalar. Equation (7.526) is satisfied $\forall n$.

Example 7.48

Solve for x in $\mathbf{L}x = y$ if $\mathbf{L} = d^2/dt^2$, with side conditions $x(0) = x(1) = 0$, and $y(t) = 2t$, via an eigenfunction expansion.

This problem of course has an exact solution via straightforward integration:

$$\frac{d^2x}{dt^2} = 2t; \qquad x(0) = x(1) = 0, \tag{7.529}$$

integrates to yield

$$x(t) = \frac{t}{3}(t^2 - 1). \tag{7.530}$$

However, let's use the series expansion technique. This can be more useful in other problems in which exact solutions do not exist. First, find the eigenvalues and eigenfunctions of the operator:

$$\frac{d^2x}{dt^2} = \lambda x; \qquad x(0) = x(1) = 0. \tag{7.531}$$

This has general solution

$$x(t) = A \sin\left(\sqrt{-\lambda} t\right) + B \cos\left(\sqrt{-\lambda} t\right). \tag{7.532}$$

To satisfy the boundary conditions, we require that $B = 0$ and $\lambda = -n^2\pi^2$, so

$$x(t) = A \sin\left(n\pi t\right). \tag{7.533}$$

This suggests that we expand $y(t) = 2t$ in a Fourier sine series. We know from Eq. (7.220) that the Fourier sine series for $y(t) = 2t$ is

$$2t = \sum_{n=1}^{\infty} \frac{4(-1)^{n+1}}{(n\pi)} \sin(n\pi t). \tag{7.534}$$

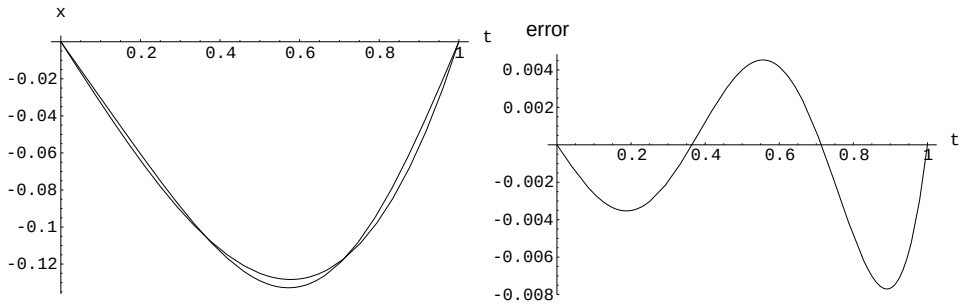

Figure 7.13: Approximate and exact solution $x(t)$; Error in solution $x_p(t) - x(t)$.

For $x(t)$ then we have

$$x(t) = \sum_{n=1}^{\infty} \frac{\alpha_n e_n}{\lambda_n} = \sum_{n=1}^{\infty} \frac{4(-1)^{n+1}}{(n\pi)\lambda_n} \sin(n\pi t). \tag{7.535}$$

Substituting in for $\lambda_n = -n^2\pi^2$, we get

$$x(t) = \sum_{n=1}^{\infty} \frac{4(-1)^{n+1}}{(-n\pi)^3} \sin(n\pi t). \tag{7.536}$$

Retaining only two terms in the expansion for $x(t)$,

$$x(t) \sim -\frac{4}{\pi^3} \sin(\pi t) + \frac{1}{2\pi^3} \sin(2\pi t), \tag{7.537}$$

gives a very good approximation for the solution, which as shown in Fig. 7.13, has a peak error of about 0.008.

Example 7.49

Solve $\mathbf{A}x = y$ using the eigenvector expansion technique when

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad y = \begin{pmatrix} 3 \\ 4 \end{pmatrix}. \tag{7.538}$$

We already know from an earlier example, p. 285, that for $\mathbf{A}$

$$\lambda_1 = 1, \qquad e_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \tag{7.539}$$

$$\lambda_2 = 3, \qquad e_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{7.540}$$

We want to express y as

$$y = \alpha_1 e_1 + \alpha_2 e_2. \tag{7.541}$$

Since the eigenvectors are orthogonal, we have from Eq. (7.186)

$$\alpha_1 = \frac{<e_1, y>}{<e_1, e_1>} = \frac{3-4}{1+1} = -\frac{1}{2}, \tag{7.542}$$

$$\alpha_2 = \frac{<e_2, y>}{<e_2, e_2>} = \frac{3+4}{1+1} = \frac{7}{2}, \tag{7.543}$$

so

$$y = -\frac{1}{2}e_1 + \frac{7}{2}e_2. \tag{7.544}$$

Then

$$\begin{aligned} x &= \frac{\alpha_1}{\lambda_1}e_1 + \frac{\alpha_2}{\lambda_2}e_2, && (7.545)\\ x &= -\frac{1}{2}\frac{1}{\lambda_1}e_1 + \frac{7}{2}\frac{1}{\lambda_2}e_2, && (7.546)\\ x &= -\frac{1}{2}\frac{1}{1}e_1 + \frac{7}{2}\frac{1}{3}e_2, && (7.547)\\ x &= -\frac{1}{2}\frac{1}{1}\begin{pmatrix} 1 \\ -1 \end{pmatrix} + \frac{7}{2}\frac{1}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, && (7.548)\\ x &= \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \end{pmatrix}. && (7.549) \end{aligned}$$

Example 7.50

Solve $\mathbf{A}x = y$ using the eigenvector expansion technique when

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}, \qquad y = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \qquad y = \begin{pmatrix} 3 \\ 6 \end{pmatrix}. \tag{7.550}$$

We first note that the two column space vectors,

$$\begin{pmatrix} 2 \\ 4 \end{pmatrix}, \qquad \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \tag{7.551}$$

are linearly dependent. They span $\mathbb{R}^1$, but not $\mathbb{R}^2$.

It is easily shown that for $\mathbf{A}$

$$\lambda_1 = 4, \qquad e_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \tag{7.552}$$

$$\lambda_2 = 0, \qquad e_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}. \tag{7.553}$$

First consider $y = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$. We want to express y as

$$y = \alpha_1 e_1 + \alpha_2 e_2. \tag{7.554}$$

For this non-symmetric matrix, the eigenvectors are linearly independent, so they form a basis. However they are not orthogonal, so there is not a direct way to compute α_1 and α_2. Matrix inversion shows that $\alpha_1 = 5/2$ and $\alpha_2 = -1/2$, so

$$y = \frac{5}{2}e_1 - \frac{1}{2}e_2. \tag{7.555}$$

Since the eigenvectors form a basis, y can be represented with an eigenvector expansion. However no solution for x exists because $\lambda_2 = 0$ and $\alpha_2 \neq 0$, hence the coefficient $\beta_2 = \alpha_2/\lambda_2$ does not exist.

However, for $y = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$, we can say that

$$y = 3e_1 + 0e_2. \tag{7.556}$$

We note that $(3, 6)^T$ is a scalar multiple of the so-called column space vector of $\mathbf{A}$, $(2, 4)^T$. Consequently,

$$\begin{aligned} x &= \frac{\alpha_1}{\lambda_1}e_1 + \frac{\alpha_2}{\lambda_2}e_2, & (7.557)\\ &= \frac{\alpha_1}{\lambda_1}e_1 + \frac{0}{0}e_2, & (7.558)\\ &= \frac{3}{4}e_1 + \gamma e_2, & (7.559)\\ &= \frac{3}{4}\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \gamma \begin{pmatrix} -1 \\ 2 \end{pmatrix}, & (7.560)\\ &= \begin{pmatrix} 3/4 - \gamma \\ 3/2 + 2\gamma \end{pmatrix}, & (7.561) \end{aligned}$$

where γ is an arbitrary constant. Note that the vector $e_2 = (-1, 2)^T$ lies in the null space of $\mathbf{A}$ since

$$\begin{aligned} \mathbf{A}e_2 &= \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}\begin{pmatrix} -1 \\ 2 \end{pmatrix}, & (7.562)\\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}. & (7.563) \end{aligned}$$

Since e_2 lies in the null space, any scalar multiple of e_2, say γe_2, also lies in the null space. We can conclude that for arbitrary y, the inverse does not exist. For vectors y which lie in the column space of $\mathbf{A}$, the inverse exists, but it is not unique; arbitrary vectors from the null space of $\mathbf{A}$ are admitted as part of the solution.

7.6 Method of weighted residuals

The method of weighted residuals is a quite general technique to solve equations. Two important methods which have widespread use in the engineering world, spectral methods and the even more pervasive finite element method, are special types of weighted residual methods.

Consider the differential equation

$$\mathbf{L}y = f(t), \qquad t \in [a, b], \tag{7.564}$$

with homogeneous boundary conditions. Here, $\mathbf{L}$ is a differential operator that is not necessarily linear. We will work with functions and inner products in $\mathbb{L}_2[a,b]$ space.

Approximate $y(t)$ by

$$y(t) \approx y_p(t) = \sum_{n=1}^{N} \alpha_n \phi_n(t), \tag{7.565}$$

where $\phi_n(t)$, $(n = 1, \cdots, N)$ are linearly independent functions (called *trial functions*) which satisfy the boundary conditions. Forcing the trial functions to satisfy the boundary conditions, in addition to having æsthetic appeal, makes it much more likely that if convergence is obtained, the convergence will be to a solution which satisfies the differential equation and boundary conditions. The trial functions can be orthogonal or non-orthogonal.[21] The constants α_n, $(n = 1, \cdots, N)$ are to be determined. Substituting into the equation, we get a residual

$$r(t) = \mathbf{L}y_p(t) - f(t). \tag{7.566}$$

Note that the residual $r(t)$ is not the error in the solution, $e(t)$, where

$$e(t) = y(t) - y_p(t). \tag{7.567}$$

The residual will almost always be non-zero for $t \in [a,b]$. However, if $r(t) = 0$, then $e(t) = 0$. We can choose α_n such that the residual, computed in a weighted average over the domain, is zero. To achieve this, we select now a set of linearly independent *weighting functions* $\psi_m(t)$, $(m = 1, \cdots, N)$ and make them orthogonal to the residual. Thus,

$$<\psi_m(t), r(t)> = 0, \qquad m = 1, \cdots, N. \tag{7.568}$$

These are N equations for the constants α_n.

There are several special ways in which the weight functions can be selected.

- Galerkin[22] : $\psi_i(t) = \phi_i(t)$.
- Collocation: $\psi_m(t) = \delta(t - t_m)$. Thus, $r(t_m) = 0$.
- Subdomain $\psi_m(t) = 1$ for $t_{m-1} \leq t < t_m$ and zero everywhere else. Note that these functions are orthogonal to each other. Also this method is easily shown to reduce to the well known finite volume method.

[21] It is occasionally advantageous, especially in the context of what is known as wavelet-based methods, to add extra functions which are linearly dependent into the set of trial functions. Such a basis is known as a *frame*. We will not consider these here; some background is given by Daubechies.

[22] Boris Gigorievich Galerkin, 1871-1945, Belarussian-born Russian-based engineer and mathematician, a participant, witness, and victim of much political turbulence, did much of his early great work in the Czar's prisons, developed a finite element method in 1915, professor of structural mechanics at what was once, and is now again, St. Petersburg (at one time known as Petrograd, and later Leningrad).

- Least squares: Minimize $||r(t)||$. This gives

$$\frac{\partial ||r||^2}{\partial \alpha_m} = \frac{\partial}{\partial \alpha_m} \int_a^b r^2 \, dt, \tag{7.569}$$

$$= 2 \int_a^b r \underbrace{\frac{\partial r}{\partial \alpha_m}}_{=\psi_m(t)} \, dt. \tag{7.570}$$

  So this method corresponds to $\psi_n = \partial r / \partial \alpha_n$.

- Moments: $\psi_m(t) = t^{m-1}$, $m = 1, 2, \ldots$.

If the trial functions are orthogonal and the method is Galerkin, we will, following Fletcher, who builds on the work of Finlayson, define the method to be a *spectral method*. Other less restrictive definitions are in common usage in the present literature, and there is no single consensus on what precisely constitutes a spectral method.[23]

Example 7.51

For $x \in \mathbb{L}_2[0,1]$, find a one-term approximate solution of the equation

$$\frac{d^2x}{dt^2} + x = t - 1, \tag{7.571}$$

with $x(0) = -1$, $x(1) = 1$.

It is easy to show that the exact solution is

$$x(t) = -1 + t + \csc(1) \sin(t). \tag{7.572}$$

[23] An important school in spectral methods, exemplified in the work of Gottlieb and Orszag, Canuto, *et al.*, and Fornberg, uses a looser nomenclature, which is not always precisely defined. In these works, spectral methods are distinguished from finite difference methods and finite element methods in that spectral methods employ basis functions which have global rather than local support; that is spectral methods' basis functions have non-zero values throughout the entire domain. While orthogonality of the basis functions within a Galerkin framework is often employed, it is not demanded that this be the distinguishing feature by those authors. Within this school, less emphasis is placed on the framework of the method of weighted residuals, and the spectral method is divided into subclasses known as Galerkin, tau, and collocation. The collocation method this school defines is identical to that defined here, and is also called by this school the "pseudospectral" method. In nearly all understandings of the word "spectral," a convergence rate which is more rapid than those exhibited by finite difference or finite element methods exists. In fact the accuracy of a spectral method should grow exponentially with the number of nodes for a spectral method, as opposed to that for a finite difference or finite element, whose accuracy grows only with the number of nodes raised to some power.

Another concern which arises with methods of this type is how many terms are necessary to properly model the desired frequency level. For example, take our equation to be $d^2u/dt^2 = 1 + u^2; u(0) = u(\pi) = 0$, and take $u = \sum_{n=1}^{N} a_n \sin(nt)$. If $N = 1$, we get $r(t) = -a_1 \sin t - 1 - a_1^2 \sin^2 t$. Expanding the square of the sin term, we see the error has higher order frequency content: $r(t) = -a_1 \sin t - 1 - a_1^2(1/2 - 1/2\cos(2t))$. The result is that if we want to get things right at a given level, we may have to reach outside that level. How far outside we have to reach will be problem dependent.

Here we will see how well the method of weighted residuals can approximate this known solution. The real value of the method is for problems in which exact solutions are not known.

Let $y = x - (2t-1)$, so that $y(0) = y(1) = 0$. The transformed differential equation is

$$\frac{d^2y}{dt^2} + y = -t. \tag{7.573}$$

Let us consider a one-term approximation,

$$y \simeq y_p(t) = \alpha\phi(t). \tag{7.574}$$

There are many choices of basis functions $\phi(t)$. Let's try finite dimensional non-trivial polynomials which match the boundary conditions. If we choose $\phi(t) = a$, a constant, we must take $a = 0$ to satisfy the boundary conditions, so this does not work. If we choose $\phi(t) = a + bt$, we must take $a = 0, b = 0$ to satisfy both boundary conditions, so this also does not work. We can find a quadratic polynomial which is non-trivial and satisfies both boundary conditions:

$$\phi(t) = t(1-t). \tag{7.575}$$

Then

$$y_p(t) = \alpha t(1-t). \tag{7.576}$$

We have to determine α. Substituting into Eq. (7.566), the residual is found to be

$$r(t) = \mathbf{L}y_p - f(t) = \frac{d^2y_p}{dt^2} + y_p - f(t), \tag{7.577}$$

$$= \underbrace{-2\alpha}_{d^2y_p/dt^2} + \underbrace{\alpha t(1-t)}_{y_p} - \underbrace{(-t)}_{f(t)} = t - \alpha(t^2 - t + 2). \tag{7.578}$$

Then, we choose α such that

$$<\psi(t), r(t)> = <\psi(t), t - \alpha(t^2 - t + 2)> = \int_0^1 \psi(t) \underbrace{\left(t - \alpha(t^2 - t + 2)\right)}_{=r(t)} \, dt = 0. \tag{7.579}$$

The form of the weighting function $\psi(t)$ is dictated by the particular method we choose:

1. Galerkin: $\psi(t) = \phi(t) = t(1-t)$. The inner product gives $\frac{1}{12} - \frac{3}{10}\alpha = 0$, so that for non-trivial solution, $\alpha = \frac{5}{18} = 0.277$.

$$y_p(t) = 0.277t(1-t). \tag{7.580}$$

$$x_p(t) = 0.277t(1-t) + 2t - 1. \tag{7.581}$$

2. Collocation: Choose $\psi(t) = \delta(t - \frac{1}{2})$ which gives $-\frac{7}{2}\alpha + 1 = 0$, from which $\alpha = \frac{2}{7} = 0.286$.

$$y_p(t) = 0.286t(1-t), \tag{7.582}$$

$$x_p(t) = 0.286t(1-t) + 2t - 1. \tag{7.583}$$

3. Subdomain: $\psi(t) = 1$, from which $-\frac{11}{6}\alpha + \frac{1}{2} = 0$, and $\alpha = \frac{3}{11} = 0.273$

$$y_p(t) = 0.273t(1-t), \tag{7.584}$$

$$x_p(t) = 0.273t(1-t) + 2t - 1. \tag{7.585}$$

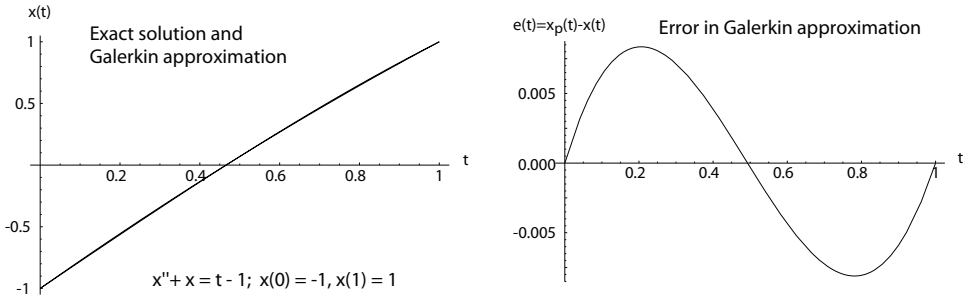

Figure 7.14: One-term estimate $x_p(t)$ and exact solution $x(t)$; Error in solution $x_p(t) - x(t)$.

4. Least squares: $\psi(t) = \frac{\partial r(t)}{\partial \alpha} = -t^2 + t - 2$. Thus, $-\frac{11}{12} + \frac{101}{30}\alpha = 0$, from which $\alpha = \frac{55}{202} = 0.273$.

$$y_p(t) = 0.273t(1-t), \tag{7.586}$$
$$x_p(t) = 0.273t(1-t) + 2t - 1. \tag{7.587}$$

5. Moments: $\psi(t) = 1$ which, for this case, is the same as the subdomain method previously reported.

$$y_p(t) = 0.273t(1-t), \tag{7.588}$$
$$x_p(t) = 0.273t(1-t) + 2t - 1. \tag{7.589}$$

The approximate solution determined by the Galerkin method is overlaid against the exact solution in Fig. 7.14. Also shown is the error in the approximation. The approximation is surprisingly accurate. Note that the error, $e(t) = x_p(t) - x(t)$, is available because in this case we have the exact solution.

Some simplification can arise through use of integration by parts. This has the result of admitting basis functions which have less stringent requirements on the continuity of their derivatives. It is also a commonly used strategy in the finite element technique.

Example 7.52

Consider a slight variant of the previous example problem, and employ integration by parts.

$$\frac{d^2y}{dt^2} + y = f(t), \qquad y(0) = 0, \quad y(1) = 0. \tag{7.590}$$

Again, take a one-term expansion

$$y_p(t) = \alpha\phi(t). \tag{7.591}$$

At this point, we will only require $\phi(t)$ to satisfy the boundary conditions, and will specify it later. The residual in the approximation is

$$r(t) = \frac{d^2y_p}{dt^2} + y_p - f(t) = \alpha\frac{d^2\phi}{dt^2} + \alpha\phi - f(t). \tag{7.592}$$

Now set a weighted residual to zero. We will also require the weighting function $\psi(t)$ to vanish at the boundaries.

$$<\psi, r> = \int_0^1 \psi(t) \underbrace{\left(\alpha \frac{d^2\phi}{dt^2} + c\phi(t) - f(t)\right)}_{=r(t)} \, dt = 0. \tag{7.593}$$

Rearranging, we get

$$\alpha \int_0^1 \left(\psi(t)\frac{d^2\phi}{dt^2} + \psi(t)\phi(t)\right) \, dt = \int_0^1 \psi(t) f(t) \, dt. \tag{7.594}$$

Now integrate by parts to get

$$c\left(\psi(t)\frac{d\phi}{dt}\bigg|_0^1 + \int_0^1 \left(\psi(t)\phi(t) - \frac{d\psi}{dt}\frac{d\phi}{dt}\right) \, dt\right) = \int_0^1 \psi(t) f(t) \, dt. \tag{7.595}$$

Since we have required $\psi(0) = \psi(1) = 0$, this simplifies to

$$\alpha \int_0^1 \left(\psi(t)\phi(t) - \frac{d\psi}{dt}\frac{d\phi}{dt}\right) \, dt = \int_0^1 \psi(t) f(t) \, dt. \tag{7.596}$$

So, the basis function ϕ only needs an integrable first derivative rather than an integrable second derivative. As an aside, we note that the term on the left hand side bears resemblance (but differs by a sign) to an inner product in the Sobolov space $\overline{\mathbb{W}}_2^1[0,1]$ in which the Sobolov inner product $<.,.>_s$ (an extension of the inner product for Hilbert space) is $<\psi(t), \phi(t)>_s = \int_0^1 \left(\overline{\psi(t)}\phi(t) + \overline{\frac{d\psi}{dt}}\frac{d\phi}{dt}\right) dt$.

Taking now, as before, $\phi = t(1-t)$ and then choosing a Galerkin method so $\psi(t) = \phi(t) = t(1-t)$, and $f(t) = -t$, we get

$$\alpha \int_0^1 \left(t^2(1-t)^2 - (1-2t)^2\right) \, dt = \int_0^1 t(1-t)(-t) \, dt, \tag{7.597}$$

which gives

$$\alpha\left(-\frac{3}{10}\right) = -\frac{1}{12}, \tag{7.598}$$

so

$$\alpha = \frac{5}{18}, \tag{7.599}$$

as was found earlier. So

$$y_p = \frac{5}{18}t(1-t), \tag{7.600}$$

with the Galerkin method.

Example 7.53

For $y \in \mathbb{L}_2[0,1]$, find a two-term spectral approximation (which by our definition of "spectral" mandates a Galerkin formulation) to the solution of

$$\frac{d^2y}{dt^2} + \sqrt{t}\, y = 1, \qquad y(0) = 0, \qquad y(1) = 0. \tag{7.601}$$

Let's try polynomial basis functions. At a minimum, these basis functions must satisfy the boundary conditions. Assumption of the first basis function to be a constant or linear gives rise to a trivial basis function when the boundary conditions are enforced. The first non-trivial basis function is a quadratic:

$$\phi_1(t) = a_0 + a_1 t + a_2 t^2. \tag{7.602}$$

We need $\phi_1(0) = 0$ and $\phi_1(1) = 0$. The first condition gives $a_0 = 0$; the second gives $a_1 = -a_2$, so we have $\phi_1 = a_1(t - t^2)$. Since the magnitude of a basis function is arbitrary, a_1 can be set to unity to give

$$\phi_1(t) = t(1 - t). \tag{7.603}$$

Alternatively, we could have chosen the magnitude in such a fashion to guarantee an orthonormal basis function, but that is a secondary concern for the purposes of this example.

We need a second linearly independent basis function for the two-term approximation. We try a third order polynomial:

$$\phi_2(t) = b_0 + b_1 t + b_2 t^2 + b_3 t^3. \tag{7.604}$$

Enforcing the boundary conditions as before gives $b_0 = 0$ and $b_1 = -(b_2 + b_3)$, so

$$\phi_2(t) = -(b_2 + b_3)t + b_2 t^2 + b_3 t^3. \tag{7.605}$$

To achieve a spectral method (which in general is not necessary to achieve an approximate solution!), we enforce $<\phi_1, \phi_2> = 0$:

$$\int_0^1 \underbrace{t(1-t)}_{=\phi_1(t)} \underbrace{\left(-(b_2+b_3)t + b_2 t^2 + b_3 t^3\right)}_{=\phi_2(t)} \, dt = 0, \tag{7.606}$$

$$-\frac{b_2}{30} - \frac{b_3}{20} = 0, \tag{7.607}$$

$$b_2 = -\frac{3}{2} b_3. \tag{7.608}$$

Substituting and factoring gives

$$\phi_2(t) = \frac{b_3}{2}\, t(1-t)(2t-1). \tag{7.609}$$

Again, because ϕ_2 is a basis function, the lead constant is arbitrary; we take for convenience $b_3 = 2$ to give

$$\phi_2 = t(1-t)(2t-1). \tag{7.610}$$

Again, b_3 could alternatively have been chosen to yield an orthonormal basis function.

Now we want to choose α_1 and α_2 so that our approximate solution

$$y_p(t) = \alpha_1 \phi_1(t) + \alpha_2 \phi_2(t), \tag{7.611}$$

has a zero weighted residual. With

$$\mathbf{L} = \left(\frac{d^2}{dt^2} + \sqrt{t} \right), \tag{7.612}$$

we have the residual as

$$r(t) = \mathbf{L} y_p(t) - f(t) = \mathbf{L}\left(\alpha_1 \phi_1(t) + \alpha_2 \phi_2(t)\right) - 1 = \alpha_1 \mathbf{L}\phi_1(t) + \alpha_2 \mathbf{L}\phi_2(t) - 1. \tag{7.613}$$

To drive the weighted residual to zero, take

$$<\psi_1, r> = \alpha_1 <\psi_1, \mathbf{L}\phi_1> + \alpha_2 <\psi_1, \mathbf{L}\phi_2> - <\psi_1, 1> = 0, \tag{7.614}$$

$$<\psi_2, r> = \alpha_1 <\psi_2, \mathbf{L}\phi_1> + \alpha_2 <\psi_2, \mathbf{L}\phi_2> - <\psi_2, 1> = 0. \tag{7.615}$$

This is easily cast in matrix form as a linear system of equations for the unknowns α_1 and α_2

$$\begin{pmatrix} <\psi_1, \mathbf{L}\phi_1> & <\psi_1, \mathbf{L}\phi_2> \\ <\psi_2, \mathbf{L}\phi_1> & <\psi_2, \mathbf{L}\phi_2> \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} <\psi_1, 1> \\ <\psi_2, 1> \end{pmatrix}. \tag{7.616}$$

We choose the Galerkin method, and thus set $\psi_1 = \phi_1$ and $\psi_2 = \phi_2$, so

$$\begin{pmatrix} <\phi_1, \mathbf{L}\phi_1> & <\phi_1, \mathbf{L}\phi_2> \\ <\phi_2, \mathbf{L}\phi_1> & <\phi_2, \mathbf{L}\phi_2> \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} <\phi_1, 1> \\ <\phi_2, 1> \end{pmatrix}. \tag{7.617}$$

Each of the inner products represents a definite integral which is easily evaluated via computer algebra. For example,

$$<\phi_1, \mathbf{L}\phi_1> = \int_0^1 \underbrace{t(1-t)}_{\phi_1} \underbrace{\left(-2 + (1-t)t^{3/2}\right)}_{\mathbf{L}\phi_1} \, dt = -\frac{215}{693}. \tag{7.618}$$

When each inner product is evaluated, the following system results

$$\begin{pmatrix} -\frac{215}{693} & \frac{16}{9009} \\ \frac{16}{9009} & -\frac{197}{1001} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix}. \tag{7.619}$$

Inverting the system, it is found that

$$\alpha_1 = -\frac{760617}{1415794} = -0.537, \qquad \alpha_2 = -\frac{3432}{707897} = -0.00485. \tag{7.620}$$

Thus, the estimate for the solution is

$$y_p(t) = -0.537\ t(1-t) - 0.00485\ t(1-t)(2t-1). \tag{7.621}$$

The two-term approximate solution determined is overlaid against a more accurate solution obtained by numerical integration of the full equation in Fig. 7.15. Also shown is the error in the approximation. The two-term solution is surprisingly accurate.

By normalizing the basis functions, we can find an orthonormal expansion. One finds that

$$||\phi_1||_2 = \sqrt{\int_0^1 \phi_1^2 \, dt}, \tag{7.622}$$

$$= \sqrt{\int_0^1 t^2(1-t)^2 \, dt}, \tag{7.623}$$

$$= \frac{1}{\sqrt{30}}, \tag{7.624}$$

$$||\phi_2||_2 = \sqrt{\int_0^1 \phi_2^2 \, dt}, \tag{7.625}$$

$$= \sqrt{\int_0^1 t^2(1-t)^2(2t-1)^2 \, dt}, \tag{7.626}$$

$$= \frac{1}{\sqrt{210}}. \tag{7.627}$$

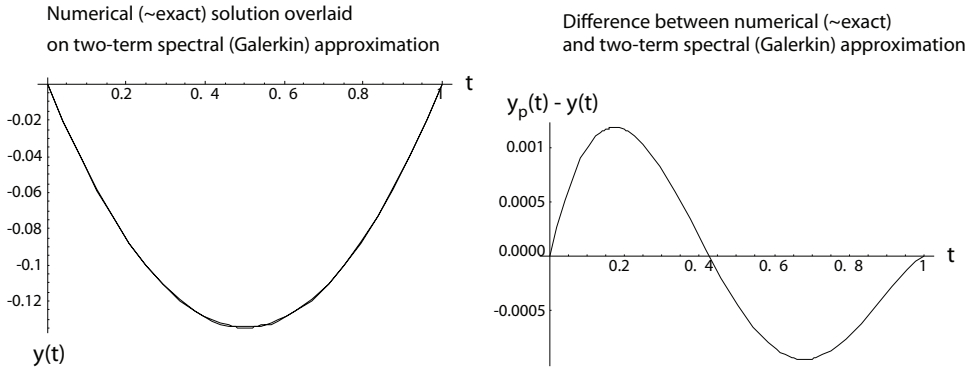

Figure 7.15: Two-term spectral (Galerkin) estimate $y_p(t)$ and highly accurate numerical solution $y(t)$; Error in approximation $y_p(t) - y(t)$.

The approximate solution can then be rewritten as an orthonormal expansion:

$$y_p(t) = -\frac{760617}{1415794\sqrt{30}}(\sqrt{30}t(1-t)) - \frac{3432}{707897\sqrt{210}}(\sqrt{210}t(1-t)(2t-1)), \tag{7.628}$$

$$= -0.981\underbrace{(\sqrt{30}t(1-t))}_{\varphi_1} - 0.000335\underbrace{(\sqrt{210}t(1-t)(2t-1))}_{\varphi_2}. \tag{7.629}$$

Because the trial functions have been normalized, one can directly compare the coefficients' magnitude. It is seen that the bulk of the solution is captured by the first term.

Example 7.54

For the equation of the previous example,

$$\frac{d^2y}{dt^2} + \sqrt{t}\, y = 1, \qquad y(0) = 0, \qquad y(1) = 0, \tag{7.630}$$

examine the convergence rates for a collocation method as the number of modes becomes large.

Let us consider a set of trial functions which do not happen to be orthogonal, but are, of course, linearly independent. Take

$$\phi_n(t) = t^n(t-1), \qquad n = 1, \ldots, N. \tag{7.631}$$

So we seek to find a vector $\boldsymbol{\alpha} = \alpha_n, n = 1, \ldots, N$, such that for a given number of collocation points N the approximation

$$y_N(t) = \alpha_1\phi_1(t) + \ldots \alpha_n\phi_n(t) + \ldots + \alpha_N\phi_N(t), \tag{7.632}$$

drives a weighted residual to zero. Obviously each these trial functions satisfies both boundary conditions, and they have the advantage of being easy to program for an arbitrary number of modes, as no Gram-Schmidt orthogonalization process is necessary. The details of the analysis are similar to those of the previous example, except we perform it many times, varying the number of nodes in each calculation. For the collocation method, we take the weighting functions to be

$$\psi_n(t) = \delta(t - t_n), \qquad n = 1, \ldots, N. \tag{7.633}$$

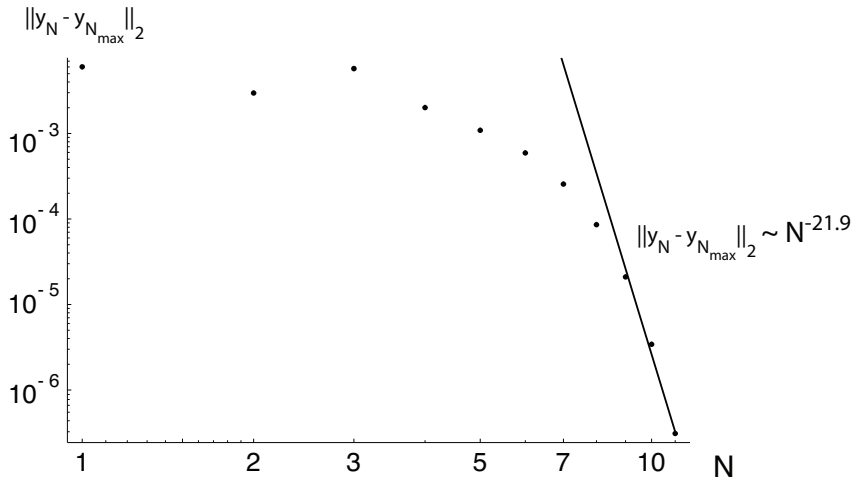

Figure 7.16: Error in solution $y_N(t) - y_{N_{max}}(t)$ as a function of number of collocation points N demonstrating exponential convergence for the spectral-type collocation method.

Here we choose $t_n = n/(N+1),\ n = 1, \ldots, N$, so that the collocation points are evenly distributed in $t \in [0, 1]$. We then form the matrix

$$\mathbf{A} = \begin{pmatrix} <\psi_1, \mathbf{L}\phi_1>, & <\psi_1, \mathbf{L}\phi_2> & \ldots & <\psi_1, \mathbf{L}\phi_N> \\ <\psi_2, \mathbf{L}\phi_1>, & <\psi_2, \mathbf{L}\phi_2> & \ldots & <\psi_2, \mathbf{L}\phi_N> \\ \vdots & \vdots & \vdots & \vdots \\ <\psi_N, \mathbf{L}\phi_1>, & <\psi_N, \mathbf{L}\phi_2> & \ldots & <\psi_N, \mathbf{L}\phi_N> \end{pmatrix}, \tag{7.634}$$

and the vector

$$\mathbf{b} = \begin{pmatrix} <\psi_1, 1> \\ \vdots \\ <\psi_N, 1> \end{pmatrix}, \tag{7.635}$$

and then solve for $\boldsymbol{\alpha}$ in

$$\mathbf{A} \cdot \boldsymbol{\alpha} = \mathbf{b}. \tag{7.636}$$

We then perform this calculation for $N = 1, \ldots, N_{max}$. We consider $N = N_{max}$ to give the most exact solution and calculate an error by finding the norm of the difference of the solution for $N < N_{max}$ and that at $N = N_{max}$:

$$e_n = ||y_N(t) - y_{N_{max}}(t)||_2 = \sqrt{\int_0^1 \left(y_N(t) - y_{N_{max}}(t)\right)^2 \, dt}. \tag{7.637}$$

A plot of the error e_N is plotted as a function of N in Fig. 7.16. We notice even on a logarithmic plot that the error reduction is accelerating as the number of nodes N increases. If the slope had relaxed to a constant, then the convergence would be a power law convergence; which is characteristic of finite difference and finite element methods. For this example of the method of weighted residuals, we see that the rate of convergence increases as the number of nodes increases, which is characteristic of *exponential convergence*. For exponential convergence, we have $e_N \sim \exp(-aN)$, where a is some positive constant; for power law convergence, we have $e_N \sim N^{-\beta}$ where β is some positive constant. At the highest value of N, $N = N_{max} = 10$, we have a local convergence rate of $O(N^{-21.9})$ which is remarkably fast. In comparison, a second order finite difference technique will converge at a rate of

$O(N^{-2})$. In general and if possible one would choose a method with the fastest convergence rate, all else being equal.

7.7 Uncertainty quantification via polynomial chaos

The methods of this chapter can be applied to account for how potential uncertainties present in model parameters affect the solutions of differential equations. To study this, we will introduce a stochastic nature into our parameters. There are many ways to deal with these so-called stochastic differential equations. One important method is known variously as "polynomial chaos," "Wiener[24]-Askey[25] chaos," as well as other names. The term "chaos" in this context was introduced by Wiener; it is in no way connected to the more modern interpretation of chaos from non-linear dynamics, as will be considered in Sec. 9.11.3.

Polynomial chaos is relevant, for example, to a differential equation of the form

$$\frac{dy}{dt} = f(y;k), \qquad y(0) = y_o, \tag{7.638}$$

where k is a parameter. For an individual calculation, k is a fixed constant. But because k is taken to possess an intrinsic uncertainty, it is allowed to take on a slightly different value for the next calculation. We expect a solution of the form $y = y(t;k)$; that is, the effect of the parameter will be realized in the solution. One way to handle the uncertainty in k is to examine a large number of solutions, each for a different value of k. The values chosen for k are driven by its uncertainty distribution, assumed to be known. We thus see how uncertain k is manifested in the solution y. This is known as the Monte Carlo method; it is an effective strategy, although potentially expensive.

For many problems, we can more easily quantify the uncertainty of the output y by propagating the known uncertainty of k via polynomial chaos, which has as its foundation notions from linear analysis. The method has the advantage of being a fully deterministic way to account for stochastic effects in differential equations. There are many variants on this method; we shall focus only on one canonical linear example which illustrates key aspects of the technique for ordinary differential equations. The method can be extended to algebraic and partial differential equations, both for scalar equations as well as for systems.

Example 7.55

Given that

$$\frac{dy}{dt} = -ky, \qquad y(0) = 1, \tag{7.639}$$

[24]Norbert Wiener, 1894-1964, American mathematician.

[25]Richard Askey, 1933-, American mathematician.

and that k has an associated uncertainty, such that

$$k = \mu + \sigma\xi, \tag{7.640}$$

where μ and σ are known constants, and $\xi \in (-\infty, \infty)$ is a random variable with a Gaussian distribution about a mean of zero with standard deviation of unity, find a two-term estimate of the behavior of $y(t)$ which accounts for the variation in k.

For our $k = \mu + \sigma\xi$, the mean value of k can be easily shown to be be μ, and the standard deviation of k is σ. The solution to to Eq. (7.639) is

$$y = e^{-kt} = e^{-(\mu+\sigma\xi)t}, \tag{7.641}$$

and will have different values, depending on the value k possess for that calculation. If there is no uncertainty in k, i.e. $\sigma = 0$, the solution to Eq. (7.639) is obviously

$$y = e^{-\mu t}. \tag{7.642}$$

Let us now try to account for the uncertainty in k in predicting the behavior of y when $\sigma \neq 0$. Let us imagine that k has an $N+1$-term Fourier expansion of

$$k(\xi) = \sum_{n=0}^{N} \alpha_n \phi_n(\xi). \tag{7.643}$$

where $\phi_n(\xi)$ are a known set of basis functions. Now as the random input ξ is varied, k will vary. And we expect the output $y(t)$ to vary, so we can imagine that we really seek $y(t,k) = y(t,k(\xi))$. Dispensing with k in favor of ξ, we can actually seek $y(t,\xi)$. Let us assume that $y(t,\xi)$ has a similar Fourier expansion,

$$y(t,\xi) = \sum_{n=0}^{N} y_n(t)\phi_n(\xi), \tag{7.644}$$

where we have also employed a separation of variables technique, with $\phi_n(\xi)$, $n = 0, \ldots, N$ as a set of basis functions and $y_n(t)$ as the time-dependent amplitude of each basis function. Let us choose the basis functions to be orthogonal:

$$<\phi_n(\xi), \phi_m(\xi)> = 0, \qquad n \neq m. \tag{7.645}$$

Since the domain of ξ is doubly infinite, a good choice for the basis functions is the Hermite polynomials; following standard practice, we choose the probabilists' form, $\phi_n(\xi) = He_n(\xi)$, Sec. 5.1.4.2, recalling $He_0(\xi) = 1$, $He_1(\xi) = \xi$, $He_2(\xi) = -1 + \xi^2$, ... Note that other non-Gaussian distributions of parametric uncertainty can render other basis functions to be better choices.

When we equip the inner product with the weighting function

$$w(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}, \tag{7.646}$$

we find our chosen basis functions are orthogonal:

$$<\phi_n(\xi), \phi_m(\xi)> = \int_{-\infty}^{\infty} \phi_n(\xi)\phi_m(\xi)w(\xi)d\xi, \tag{7.647}$$

$$= \int_{-\infty}^{\infty} He_n(\xi)He_m(\xi)\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi, \tag{7.648}$$

$$= n!\delta_{nm}. \tag{7.649}$$

Let us first find the coefficients α_n in the Fourier-Hermite expansion of $k(\xi)$:

$$k(\xi) = \sum_{n=0}^{N} \alpha_n \phi_n(\xi), \tag{7.650}$$

$$<\phi_m(\xi), k(\xi)> = <\phi_m(\xi), \sum_{n=0}^{\infty} \alpha_n \phi_n(\xi)>, \tag{7.651}$$

$$= \sum_{n=0}^{\infty} \alpha_n <\phi_m(\xi), \phi_n(\xi)>, \tag{7.652}$$

$$= \sum_{n=0}^{N} \alpha_n n! \delta_{mn}, \tag{7.653}$$

$$= m! \alpha_m, \tag{7.654}$$

$$\alpha_n = \frac{<\phi_n(\xi), k(\xi)>}{n!}, \tag{7.655}$$

$$= \frac{<\phi_n(\xi), \mu + \sigma\xi>}{n!}, \tag{7.656}$$

$$= \frac{1}{\sqrt{2\pi} n!} \int_{-\infty}^{\infty} He_n(\xi) \left(\mu + \sigma\xi\right) e^{-\xi^2/2} d\xi \tag{7.657}$$

Because of the polynomial nature of $k = \mu + \sigma\xi$ and the orthogonality of the polynomial functions, there are only two non-zero terms in the expansion: $\alpha_0 = \mu$ and $\alpha_1 = \sigma$; thus,

$$k(\xi) = \mu + \sigma\xi = \alpha_0 He_0(\xi) + \alpha_1 He_1(\xi) = \mu He_0(\xi) + \sigma He_1(\xi). \tag{7.658}$$

So for this simple distribution of $k(\xi)$, the infinite Fourier series expansion of Eq. (7.644) is a finite two-term expansion. We actually could have seen this by inspection, but it was useful to go through the formal exercise.

Now, substitute the expansions of Eqs. (7.643, 7.644) into the governing Eq. (7.639):

$$\frac{d}{dt} \left(\underbrace{\sum_{n=0}^{N} y_n(t) \phi_n(\xi)}_{y} \right) = - \left(\underbrace{\sum_{n=1}^{N} \alpha_n \phi_n(\xi)}_{k} \right) \left(\underbrace{\sum_{m=0}^{N} y_m(t) \phi_m(\xi)}_{y} \right). \tag{7.659}$$

Equation (7.659) forms $N+1$ ordinary differential equations, still with an explicit dependency on ξ. We need an initial condition for each of them. The initial condition $y(0) = 1$ can be recast as

$$y(0, \xi) = 1 = \sum_{n=0}^{N} y_n(0) \phi_n(\xi). \tag{7.660}$$

Now we could go through the same formal exercise as for k to determine the Fourier expansion of $y(0) = 1$. But since $\phi_0(\xi) = 1$, we see by inspection the set of $N+1$ initial conditions are

$$y_0(0) = 1, \qquad y_1(0) = 0, \qquad y_2(0) = 0, \qquad \ldots, \qquad y_N(0) = 0. \tag{7.661}$$

Let us now rearrange Eq. (7.659) to get

$$\sum_{n=0}^{N} \frac{dy_n}{dt} \phi_n(\xi) = - \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n y_m(t) \phi_n(\xi) \phi_m(\xi). \tag{7.662}$$

We still need to remove the explicit dependency on the random variable ξ. To achieve this, we will take the inner product of Eq. (7.662) with a set of functions, so as to simplify the system into a cleaner system of ordinary differential equations. Let us choose to invoke a Galerkin procedure by taking the inner product of Eq. (7.662) with $\phi_l(\xi)$:

$$<\phi_l(\xi), \sum_{n=0}^{N} \frac{dy_n}{dt}\phi_n(\xi)> = -<\phi_l(\xi), \sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n y_m(t)\phi_n(\xi)\phi_m(\xi)>, \tag{7.663}$$

$$\sum_{n=0}^{N} \frac{dy_n}{dt}<\phi_l(\xi), \phi_n(\xi)> = -\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n y_m(t)<\phi_l(\xi), \phi_n(\xi)\phi_m(\xi)>, \tag{7.664}$$

$$\frac{dy_l}{dt}<\phi_l(\xi), \phi_l(\xi)> = -\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n y_m(t)<\phi_l(\xi), \phi_n(\xi)\phi_m(\xi)>, \tag{7.665}$$

$$\frac{dy_l}{dt} = -\frac{1}{<\phi_l(\xi), \phi_l(\xi)>}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n y_m(t)<\phi_l(\xi), \phi_n(\xi)\phi_m(\xi)>, \tag{7.666}$$

$$\frac{dy_l}{dt} = -\frac{1}{l!}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n y_m(t)<\phi_l(\xi), \phi_n(\xi)\phi_m(\xi)>, \qquad l = 0, \ldots, N. \tag{7.667}$$

Equation (7.667) forms $N+1$ ordinary differential equations, with $N+1$ initial conditions provided by Eq. (7.661). All dependency on ξ is removed by explicit evaluation of the inner products for all l, n, and m. We could have arrived at an analogous system of ordinary differential equations had we chosen any of the other standard set of functions for the inner product. For example, Dirac delta functions would have led to a collocation method. Note the full expression of the unusual inner product which appears in Eq. (7.667) is

$$<\phi_l(\xi), \phi_n(\xi)\phi_m(\xi)> = \int_{-\infty}^{\infty} \phi_l(\xi)\phi_n(\xi)\phi_m(\xi)\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi. \tag{7.668}$$

This relation can be reduced further, but it is not straightforward.

When $N = 1$, we have a two-term series, with $l = 0, 1$. Detailed evaluation of all inner products yields two ordinary differential equations:

$$\frac{dy_0}{dt} = -\mu y_o - \sigma y_1, \qquad y_0(0) = 1, \tag{7.669}$$

$$\frac{dy_1}{dt} = -\sigma y_0 - \mu y_1, \qquad y_1(0) = 0. \tag{7.670}$$

Note when $\sigma = 0$, $y_0(t) = e^{-\mu t}$, $y_1(t) = 0$, and we recover our original non-stochastic result. For $\sigma \neq 0$, this linear system can be solved exactly using methods of the upcoming Section 9.5.1. Direct substitution reveals that the solution is in fact

$$y_0(t) = e^{-\mu t}\cosh(\sigma t), \tag{7.671}$$

$$y_1(t) = -e^{-\mu t}\sinh(\sigma t). \tag{7.672}$$

Thus, the two-term approximation is

$$y(t, \xi) \sim y_0(t)\phi_0(\xi) + y_1(t)\phi_1(\xi), \tag{7.673}$$

$$= e^{-\mu t}(\cosh(\sigma t) - \sinh(\sigma t)\,\xi). \tag{7.674}$$

The non-stochastic solution $e^{-\mu t}$ is obviously modulated by the uncertainty. Even when $\xi = 0$, there is a weak modulation by $\cosh(\sigma t) \sim 1 + \sigma^2 t^2/ + \sigma^4 t^4/24 + \ldots$.

Standard probability theory lets us estimate the mean value of $y(t,\xi)$, which we call $\overline{y}(t)$, over a range of normally distributed values of ξ:

$$\overline{y}(t) = \int_{-\infty}^{\infty} y(t,\xi)\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi, \tag{7.675}$$

$$= \int_{-\infty}^{\infty} e^{-\mu t}\left(\cosh(\sigma t) - \sinh(\sigma t)\,\xi\right)\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi, \tag{7.676}$$

$$= e^{-\mu t}\cosh(\sigma t). \tag{7.677}$$

Thus, the mean value of y is $y_0(t) = e^{-\mu t}\cosh(\sigma t)$. The standard deviation of the solution, $\sigma_s(t)$ is found by a similar process

$$\sigma_s(t) = \sqrt{\int_{-\infty}^{\infty}(y(t,\xi)-\overline{y}(t))^2\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi}, \tag{7.678}$$

$$= \sqrt{\int_{-\infty}^{\infty}\left(-e^{-\mu t}\sinh(\sigma t)\,\xi\right)^2\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi}, \tag{7.679}$$

$$= e^{-\mu t}\sinh(\sigma t). \tag{7.680}$$

Note that $\sigma_s(t) = |y_1(t)|$. Also note that σ_s is distinct from σ, the standard deviation of the parameter k.

All of this is easily verified by direct calculation. If we take $\mu = 1$ and $\sigma = 1/3$, we have $k = 1+\xi/3$, recalling that ξ is a random number, with a Gaussian distribution with unity standard deviation about zero. Let us examine various predictions at $t = 1$. Ignoring all stochastic effects, we might naïvely predict that the expected value of y should be

$$y(t=1) = e^{-\mu t} = e^{-(1)(1)} = 0.367879, \tag{7.681}$$

with no standard deviation. However, if we execute so-called Monte Carlo simulations where k is varied through its range, calculate y at $t = 1$ for each realization of k, and then take the mean value of all predictions, we find for 10^6 simulations that the mean value is

$$y_{Monte\ Carlo}(t=1) = 0.388856. \tag{7.682}$$

This number will slightly change if a different set of random values of k are tested. Remarkably though, $y_{Monte\ Carlo}$ is well predicted by our polynomial chaos estimate of

$$y_0(t=1) = e^{-\mu t}\cosh(\sigma t) = e^{-(1)(1)}\cosh\left(\frac{1}{3}\right) = 0.388507. \tag{7.683}$$

We could further improve the Monte Carlo estimate by taking more samples. We could further improve the polynomial chaos estimate by including more terms in the expansion. As the number of Monte Carlo estimates and the number terms in the polynomial chaos expansion approached infinity, the two estimates would converge. And they would converge to a number different than that of the naïve estimate. The exponential function warped the effect of the Gaussian distributed k such that the realization of y at $t = 1$ was distorted to a greater value.

For the same 10^6 simulations, the Monte Carlo method predicts a standard deviation of y at $t = 1$ of

$$\sigma_{s,\ Monte\ Carlo} = 0.133325. \tag{7.684}$$

This number is well estimated by the magnitude, $|y_1(t=1)|$:

$$|y_1(t=1)| = e^{-\mu t}\sinh(\sigma t) = e^{-(1)(1)}\sinh\left(\frac{1}{3}\right) = 0.124910. \tag{7.685}$$

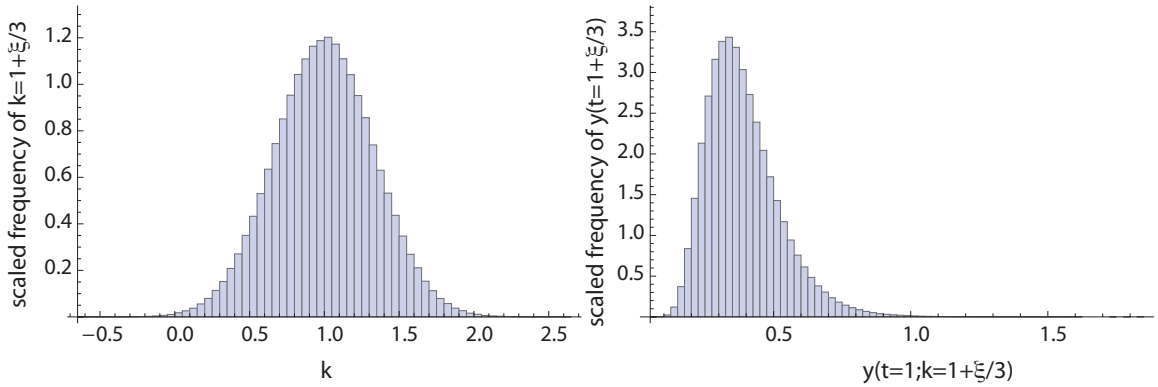

Figure 7.17: Histograms for distribution of k and $y(t=1;k=1+\xi/3)$ for 10^6 Monte Carlo simulations for various values of k.

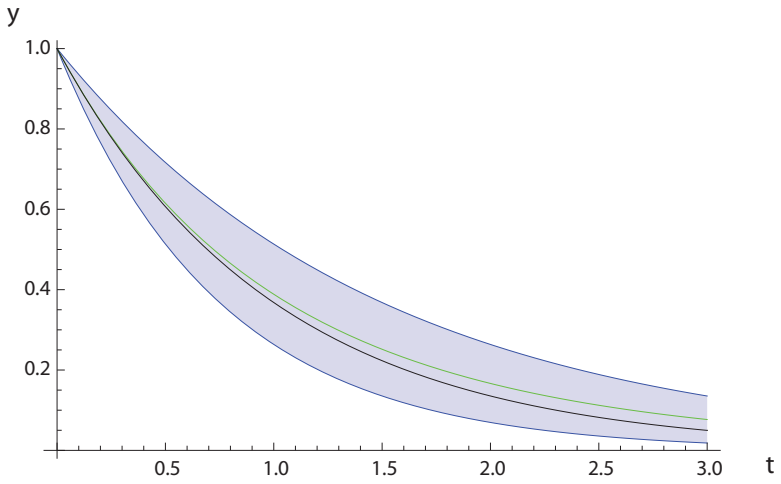

Figure 7.18: Estimates of $y(t)$ which satisfies $dy/dt = -ky$, $y(0) = 1$, for $k = 1 + \xi/3$, where ξ is a random variable, normally distributed about zero.

Again, both estimates could be improved by more samples, and more terms in the expansion, respectively.

Histograms of the scaled frequency of occurrence of k and $y(t = 1; k = 1 + \xi/3)$ within bins of specified width from the Monte Carlo method for 10^6 realizations are plotted in Fig. 7.17. We show fifty bins within which the scaled number of occurrences of k and $y(t = 1)$ are realized. The scaling factor applied to the number of occurrences was selected so that the area under the curve is unity. This is achieved by scaling the number of occurrences within a bin by the product of the total number of occurrences and the bin width; this allows the scaled number of occurrences to be thought of as a probability density. As designed, k appears symmetric about its mean value of unity, with a standard deviation of 1/3. Detailed analysis would reveal that k in fact has a Gaussian distribution. But $y(t = 1; k = 1 + \xi/3)$ does not have a Gaussian distribution about its mean, as it has been skewed by the dynamics of the differential equation for y. The time-evolution of y is plotted in Fig. 7.18. The black line gives the naïve estimate, e^{-t}. The green line gives $y_0(t)$, and the two blue lines give $y_0(t) \pm y_1(t)$, that is, the mean value of y, plus or minus one standard deviation.

Problems

1. Use a one-term collocation method with a polynomial basis function to find an approximation for

$$y'''' + (1+x)y = 1,$$

with $y(0) = y'(0) = y'(1) = y''(1) = 0$.

2. Use two-term spectral, collocation, subdomain, least squares and moments methods to solve the equation

$$y'''' + (1+x)y = 1,$$

with $y(0) = y'(0) = y(1) = y''(1) = 0$. Compare graphically with the exact solution.

3. If $x_1, x_2, \cdots, x_N$ and $y_1, y_2, \cdots, y_N$ are real numbers, show that

$$\left(\sum_{n=1}^{N} x_n y_n \right)^2 \leq \left(\sum_{n=1}^{N} x_n^2 \right) \left(\sum_{n=1}^{N} y_n^2 \right).$$

4. If $x, y \in \mathbb{X}$, an inner product space, and x is orthogonal to y, then show that $||x + \alpha y|| = ||x - \alpha y||$ where α is a scalar.

5. For an inner product space, show that

$$\begin{aligned} <x, y+z> &= <x, y> + <x, z>, \\ <\alpha x, y> &= \overline{\alpha} <x, y>, \\ <x, y> &= <y, x> \text{ in a real vector space.} \end{aligned}$$

6. The linear operator $\mathbf{A} : \mathbb{X} \rightarrow \mathbb{Y}$, where $\mathbb{X} = \mathbb{R}^2$, $\mathbb{Y} = \mathbb{R}^2$. The norms in $\mathbb{X}$ and $\mathbb{Y}$ are defined by

$$\begin{aligned} x &= (\xi_1, \xi_2)^T \in \mathbb{X}, ||x||_\infty = \max\left(|\xi_1|, |\xi_2|\right), \\ y &= (\eta_1, \eta_2)^T \in \mathbb{Y}, ||y||_1 = |\eta_1| + |\eta_2|. \end{aligned}$$

Find $||\mathbf{A}||$ if $\mathbf{A} = \begin{pmatrix} 3 & -1 \\ 5 & -2 \end{pmatrix}$.

7. Let $\mathbb{Q}$, $\mathbb{C}$ and $\mathbb{R}$ be the sets of all rational, complex and real numbers respectively. For the following determine if $\mathbb{A}$ is a vector space over the field $\mathbb{F}$. For finite-dimensional vector spaces, find also a set of basis vectors.

   (a) $\mathbb{A}$ is the set of all polynomials which are all exactly of degree n, $\mathbb{F} = \mathbb{R}$.

   (b) $\mathbb{A}$ is the set of all functions with continuous second derivatives over the interval $[0, L]$ and satisfying the differential equation $y'' + 2y' + y = 0$, $\mathbb{F} = \mathbb{R}$.

   (c) $\mathbb{A} = \mathbb{R}, \mathbb{F} = \mathbb{R}$.

   (d) $\mathbb{A} = \{(a_1, a_2, a_3)$ such that $a_1, a_2 \in \mathbb{Q}, 2a_1 + a_2 = 4a_3\}, \mathbb{F} = \mathbb{Q}$.

   (e) $\mathbb{A} = \mathbb{C}, \mathbb{F} = \mathbb{Q}$.

   (f) $\mathbb{A} = \{ae^x + be^{-2x}$ such that $a, b \in \mathbb{R}, x \in [0, 1]\}, \mathbb{F} = \mathbb{R}$.

8. Which of the following subsets of $\mathbb{R}^3$ constitute a subspace of $\mathbb{R}^3$ where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$:
   (a) All x with $x_1 = x_2$ and $x_3 = 0$.
   (b) All x with $x_1 = x_2 + 1$.
   (c) All x with positive x_1, x_2, x_3.
   (d) All x with $x_1 - x_2 + x_3 =$ constant k.
9. Given a set $\mathbb{S}$ of linearly independent vectors in a vector space $\mathbb{V}$, show that any subset of $\mathbb{S}$ is also linearly independent.
10. Do the following vectors, $(3, 1, 4, -1)^T, (1, -4, 0, 4)^T, (-1, 2, 2, 1)^T, (-1, 9, 5, -6)^T$, form a basis in $\mathbb{R}^4$?
11. Given x_1, the iterative procedure $x_{n+1} = \mathbf{L}x_n$ generates $x_2, x_3, x_4, \cdots$, where $\mathbf{L}$ is a linear operator and all the x's belong to a complete normed space. Show that $\{x_n,\ n = 1, 2, \cdots\}$ is a Cauchy sequence if $||\mathbf{L}|| < 1$. Does it converge? If so find the limit.
12. If $\{e_n, n = 1, 2, \cdots\}$ is an orthonormal set in a Hilbert space $\mathbb{H}$, show that for every $x \in \mathbb{H}$, the vector $y = \sum_{n=1}^{N} <x, e_n> e_n$ exists in $\mathbb{H}$, and that $x - y$ is orthogonal to every e_n.
13. Let the linear operator $\mathbf{A} : \mathbb{C}^2 \rightarrow \mathbb{C}^2$ be represented by the matrix $\mathbf{A} = \begin{pmatrix} 2 & -4 \\ 1 & 5 \end{pmatrix}$. Find $||\mathbf{A}||$ if all vectors in the domain and range are within a Hilbert space.
14. Let the linear operator $\mathbf{A} : \mathbb{C}^2 \rightarrow \mathbb{C}^2$ be represented by the matrix $\mathbf{A} = \begin{pmatrix} 2+i & -4 \\ 1 & 5 \end{pmatrix}$. Find $||\mathbf{A}||$ if all vectors in the domain and range are within a Hilbert space.
15. Using the inner product $(x, y) = \int_a^b w(t)x(t)y(t)\ dt$, where $w(t) > 0$ for $a \leq t \leq b$, show that the Sturm-Liouville operator
$$\mathbf{L} = \frac{1}{w(t)} \left(\frac{d}{dt} \left(p(t) \frac{d}{dt} \right) + r(t) \right),$$
with $\alpha x(a) + \beta x'(a) = 0$, and $\gamma x(b) + \delta x'(b) = 0$ is self-adjoint.
16. For elements x, y and z of an inner product space, prove the Apollonius[26] identity:
$$||z - x||_2^2 + ||z - y||_2^2 = \frac{1}{2}||x - y||_2^2 + 2 \left|\left| z - \frac{1}{2}(x + y) \right|\right|_2^2.$$
17. If $x, y \in \mathbb{X}$ an inner product space, and x is orthogonal to y, then show that $||x + ay||_2 = ||x - ay||_2$ where a is a scalar.
18. Using the Gram-Schmidt procedure, find the first three members of the orthonormal set belonging to $\mathbb{L}_2(-\infty, \infty)$, using the basis functions $\{\exp(-t^2/2),\ t\exp(-t^2/2),\ t^2\exp(-t^2/2), \cdots\}$. You may need the following definite integral
$$\int_{-\infty}^{\infty} \exp(-t^2/2)\ dt = \sqrt{2\pi}.$$
19. Let $C(0,1)$ be the space of all continuous functions in (0,1) with the norm
$$||f||_2 = \sqrt{\int_0^1 |f(t)|^2\ dt}.$$

[26] Apollonius of Perga, ca. 262 BC-ca. 190 BC, Greek astronomer and geometer.

Show that

$$f_n(t) = \begin{cases} 2^n t^{n+1} \text{ for } 0 \le t < \frac{1}{2} \\ 1 - 2^n(1-t)^{n+1} \text{ for } \frac{1}{2} \le t \le 1, \end{cases}$$

belongs to $C(0,1)$. Show also that $\{f_n, n = 1, \cdots\}$ is a Cauchy sequence, and that $C(0,1)$ is not complete.

20. Find the first three terms of the Fourier-Legendre series for $f(x) = \cos(\pi x/2)$ for $x \in [-1, 1]$. Compare graphically with exact function.

21. Find the first three terms of the Fourier-Legendre series for

$$f(x) = \begin{cases} -1, & \text{for } -1 \le x < 0, \\ 1, & \text{for } 0 \le x \le 1. \end{cases}$$

22. Consider

$$\frac{d^3y}{dt^3} + 2t^3 y = 1 - t, \qquad y(0) = 0 \qquad y(2) = 0 \qquad \frac{dy}{dt}(0) = 0.$$

Choosing polynomials as the basis functions, use a Galerkin and moments method to obtain a two-term estimate to $y(t)$. Plot your approximations and the exact solution on a single curve. Plot the residual in both methods for $t \in [0, 2]$

23. Solve

$$x'' + 2xx' + t = 0,$$

with $x(0) = 0$, $x(4) = 0$, approximately using a two-term weighted residual method where the basis functions are of the type $\sin \lambda t$. Do both a spectral (as a consequence Galerkin) and pseudospectral (as a consequence collocation) method. Plot your approximations and the exact solution on a single curve. Plot the residual in both methods for $x \in [0, 4]$.

24. Show that the set of solutions of the linear equations

$$\begin{aligned} x_1 + 3x_2 + x_3 - x_4 &= 0, \\ -2x_1 + 2x_2 - x_3 + x_4 &= 0, \end{aligned}$$

form a vector space. Find the dimension and a set of basis vectors.

25. Let

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

For $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^3$, find $||\mathbf{A}||$ if the norm of $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ is given by

$$||\mathbf{x}||_\infty = \max(|x_1|, |x_2|, |x_3|).$$

26. For any complete orthonormal set $\{\phi_i,\ i = 1, 2, \cdots\}$ in a Hilbert space $\mathbb{H}$, show that

$$\begin{aligned} u &= \sum_i <u, \phi_i> \phi_i, \\ <u, v> &= \sum_i <u, \phi_i><v, \phi_i>, \\ ||u||_2^2 &= \sum_i |<u, \phi_i>|^2, \end{aligned}$$

where u and v belong to $\mathbb{H}$.

27. Show that the set $\mathbb{P}^4[0,1]$ of all polynomials of degree 4 or less in the interval $0 < x < 1$ is a vector space. What is the dimension of this space?

28. Show that

$$(x_1^2 + x_2^2 + \ldots + x_N^2)(y_1^2 + y_2^2 + \ldots + y_N^2) \geq (x_1 y_1 + x_2 y_2 + \ldots + x_N y_N)^2,$$

where $x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_N$ are real numbers.

29. Show that the functions $e_1(t), e_2(t), \ldots, e_N(t)$ are orthogonal in $\mathbb{L}_2(0,1]$, where

$$e_n(t) = \begin{cases} 1 & \frac{n-1}{N} < t \leq \frac{n}{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Expand t^2 in terms of these functions.

30. Find one-term collocation approximations for *all* solutions of

$$\frac{d^2y}{dx^2} + y^4 = 1,$$

with $y(0) = 0$, $y(1) = 0$.

31. Show that

$$\sqrt{\int_a^b (f(x) + g(x))^2 \; dx} \leq \sqrt{\int_a^b (f(x))^2 \; dx} + \sqrt{\int_a^b (g(x))^2 \; dx},$$

where $f(x)$ and $y(x)$ belong to $\mathbb{L}_2[a,b]$.

32. Find the eigenvalues and eigenfunctions of the operator

$$\mathbf{L} = -\left(\frac{d^2}{dx^2} + 2\frac{d}{dx} + 1\right),$$

which operates on functions $y \in \mathbb{L}_2[0,5]$ that vanish at $x = 0$ and $x = 5$.

33. Find the supremum and infimum of the set $\mathbb{S} = \{1/n, \text{ where } n = 1, 2, \cdots\}$.

34. Find the $\mathbb{L}_2[0,1]$ norm of the function $f(x) = x + 1$.

35. Find the distance between the functions x and x^3 under the $\mathbb{L}_2[0,1]$ norm.

36. Find the inner product of the functions x and x^3 using the $\mathbb{L}_2$[0,1] definition.

37. Find the Green's function for the problem

$$\frac{d^2x}{dt^2} + k^2 x = f(t), \text{ with } x(0) = a, \; x(\pi) = b.$$

Write the solution of the differential equation in terms of this function.

38. Find the first three terms of the Fourier-Legendre series for

$$f(x) = \begin{cases} -2 & \text{for } -1 \leq x < 0 \\ 1 & \text{for } 0 \leq x \leq 1 \end{cases}$$

Graph $f(x)$ and its approximation.

39. Find the null space of

(a) the matrix operator

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix},$$

(b) the differential operator

$$\mathbf{L} = \frac{d^2}{dt^2} + k^2.$$

40. Test the positive definiteness of a diagonal matrix with positive real numbers on the diagonal.

41. Let $\mathbb{S}$ be a subspace of $\mathbb{L}_2[0,1]$ such that for every $x \in \mathbb{S}$, $x(0) = 0$, and $\dot{x}(0) = 1$. Find the eigenvalues and eigenfunctions of $\mathbf{L} = -d^2/dt^2$ operating on elements of $\mathbb{S}$.

42. Show that

$$\lim_{\epsilon \to 0} \int_\alpha^\beta f(x) \Delta_\epsilon (x-a) dx = f(a),$$

for $a \in (\alpha, \beta)$, where

$$\Delta_\epsilon (x-a) = \begin{cases} 0, & \text{if } x < a - \frac{\epsilon}{2}, \\ \frac{1}{\epsilon}, & \text{if } a - \frac{\epsilon}{2} \leq x \leq a + \frac{\epsilon}{2}, \\ 0, & \text{if } x > a + \frac{\epsilon}{2}. \end{cases}$$

43. Consider functions of two variables in a domain Ω with the inner product defined as

$$< u, v > = \iint_\Omega u(x,y) v(x,y) \; dx \; dy.$$

Find the space of functions such that the Laplacian operator is self-adjoint.

44. Find the eigenvalues and eigenfunctions of the operator $\mathbf{L}$ where

$$\mathbf{L}y = (1-t^2) \frac{d^2 y}{dt^2} - t \frac{dy}{dt},$$

with $t \in [-1, 1]$ and $y(-1) = y(1) = 0$. Show that there exists a weight function $r(x)$ such that the eigenfunctions are orthogonal in $[-1, 1]$ with respect to it.

45. Show that the eigenvalues of an operator and its adjoint are complex conjugates of each other.

46. Using an eigenvector expansion, find the general solution of $\mathbf{A} \cdot \mathbf{x} = \mathbf{y}$ where

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \\ \mathbf{y} &= \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}. \end{aligned}$$

47. Show graphically that the Fourier trigonometric series representation of the function

$$f(t) = \begin{cases} -1, & \text{if } -\pi \leq t < 0, \\ 1, & \text{if } 0 \leq t \leq \pi, \end{cases}$$

always has an overshoot near $x = 0$, however many terms one takes *(Gibbs phenomenon)*. Estimate the overshoot.

48. Let $\{e_1, \cdots, e_N\}$ be an orthonormal set in an inner product space $\mathbb{S}$. Approximate $x \in \mathbb{S}$ by $y = \beta_1 e_1 + \cdots + \beta_N e_N$, where the β's are to be selected. Show that $||x - y||$ is a minimum if we choose $\beta_i = <x, e_i>$.

49. (a) Starting with a vector in the direction $(1, 2, 0)^T$ use the Gram-Schmidt procedure to find a set of orthonormal vectors in $\mathbb{R}^3$. Using these vectors, construct (b) an orthogonal matrix $\mathbf{Q}$, and then find (c) the angles between $\mathbf{x}$ and $\mathbf{Q} \cdot \mathbf{x}$, where $\mathbf{x}$ is $(1, 0, 0)^T$, $(0, 1, 0)^T$ and $(0, 0, 1)^T$, respectively. The orthogonal matrix $\mathbf{Q}$ is defined as a matrix having orthonormal vectors in its columns.

50. Find the null space of the operator $\mathbf{L}$ defined by $\mathbf{L}x = (d^2/dt^2)x(t)$. Also find the eigenvalues and eigenfunctions (in terms of real functions) of $\mathbf{L}$ with $x(0) = 1, (dx/dt)(0) = 0$.

51. Find all approximate solutions of the boundary value problem

$$\frac{d^2y}{dx^2} + y + 5y^2 = -x,$$

with $y(0) = y(1) = 0$ using a two-term collocation method. Compare graphically with the exact solution determined by numerical methods.

52. Find a one-term approximation for the boundary value problem

$$y'' - y = -x^3,$$

with $y(0) = y(1) = 0$, using the collocation, Galerkin, least-squares, and moments methods. Compare graphically with the exact solution.

53. Consider the sequence $\{\frac{1+\frac{1}{N}}{2+\frac{1}{N}}\}$ in $\mathbb{R}^N$. Show that this is a Cauchy sequence. Does it converge?

54. Prove that $(\mathbf{L}_a\mathbf{L}_b)^* = \mathbf{L}_b^*\mathbf{L}_a^*$ when $\mathbf{L}_a$ and $\mathbf{L}_b$ are linear operators which operate on vectors in a Hilbert space.

55. If $\{x_i\}$ is a sequence in an inner product space such that the series $||x_1|| + ||x_2|| + \cdots$ converges, show that $\{s_N\}$ is a Cauchy sequence, where $s_N = x_1 + x_2 + \cdots + x_N$.

56. If $\mathbf{L}(x) = a_0(t)\frac{d^2x}{dt^2} + a_1(t)\frac{dx}{dt} + a_2(t)x$, find the operator that is formally adjoint to it.

57. If

$$y(t) = \mathbf{L}(x(t)) = \int_0^t x(\tau)\ d\tau,$$

where $y(t)$ and $x(t)$ are real functions in some properly defined space, find the eigenvalues and eigenfunctions of the operator $\mathbf{L}$.

58. Using a dual basis, expand the vector $(1, 3, 2)^T$ in terms of the basis vectors $(1, 1, 1)^T$, $(1, 0, -1)^T$, and $(1, 0, 1)^T$ in $\mathbb{R}^3$. The inner product is defined as usual.

59. With $f_1(x) = 1 + i + x$ and $f_2(x) = 1 + ix + ix^2$,

    a) Find the $\overline{\mathbb{L}}_2[0, 1]$ norms of $f_1(x)$ and $f_2(x)$.

    b) Find the inner product of $f_1(x)$ and $f_2(x)$ under the $\overline{\mathbb{L}}_2[0, 1]$ norm.

    c) Find the "distance" between $f_1(x)$ and $f_2(x)$ under the $\overline{\mathbb{L}}_2[0, 1]$ norm.

60. Show the vectors $u_1 = (-i, 0, 2, 1 + i)^T, u_2 = (1, 2, i, 3)^T, u_3 = (3 + i, 3 - i, 0, -2)^T, u_4 = (1, 0, 1, 3)^T$ form a basis in $\mathbb{C}^4$. Find the set of reciprocal basis vectors. For $x \in \mathbb{C}^4$, and $x = (i, 3 - i, -2, 2)^T$, express x as an expansion in the above-defined basis vectors. That is find α_i such that $x = \alpha_i u_i$.

61. The following norms can be used in $\mathbb{R}^N$, where $x = (\xi_1, \cdots, \xi_N) \in \mathbb{R}^N$.
    (a) $||x||_\infty = \max_{1 \le n \le N} |\xi_n|$,
    (b) $||x||_1 = \sum_{n=1}^N |\xi_n|$,
    (c) $||x||_2 = (\sum_{n=1}^N |\xi_n|^2)^{1/2}$,
    (d) $||x||_p = (\sum_{n=1}^N |\xi_n|^p)^{1/p},\ 1 \le p < \infty$.
    Show by examples that these are all valid norms.

62. Show that the set of all matrices $\mathbb{A} : \mathbb{R}^N \to \mathbb{R}^N$ is a vector space under the usual rules of matrix manipulation.

63. Show that if $\mathbf{A}$ is a linear operator such that
    (a) $\mathbf{A} : (\mathbb{R}^N, ||\cdot||_\infty) \to (\mathbb{R}^N, ||\cdot||_1)$, then $||\mathbf{A}|| = \sum_{i,j=1}^N A_{ij}$.
    (b) $\mathbf{A} : (\mathbb{R}^N, ||\cdot||_\infty) \to (\mathbb{R}^N, ||\cdot||_\infty)$, then $||\mathbf{A}|| = \max_{1 \le i \le N} \sum_{j=1}^N A_{ij}$.

64. If
$$\mathbf{L}u = a(x)\frac{d^2u}{dx^2} + b(x)\frac{du}{dx} + c(x)u,$$
    show
$$\mathbf{L}^* u = \frac{d^2}{dx^2}(au) - \frac{d}{dx}(bu) + cu.$$

65. Consider the function $x(t) = \sin(4t)$ for $t \in [0, 1]$. Project $x(t)$ onto the space spanned by the functions $u_m(t)$ so as to find the coefficients α_m, where $x(t) \simeq x_p(t) = \sum_{m=1}^M \alpha_m u_m(t)$ when the basis functions are
    (a) $M = 2$; $u_1(t) = t$, $u_2(t) = t^2$.
    (b) $M = 3$; $u_1(t) = 1$, $u_2(t) = t^2$, $u_3(t) = \tan t$.
    In each case plot $x(t)$ and its approximation on the same plot.

66. Project the vector $x = (1, 2, 3, 4)^T$ onto the space spanned by the vectors, u_1, u_2, so as to find the projection $x \simeq x_p = \alpha_1 u_1 + \alpha_2 u_2$.
    (a) $u_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $u_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$,
    (b) $u_1 = \begin{pmatrix} i \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $u_2 = \begin{pmatrix} i \\ 1 \\ i \\ 1 \end{pmatrix}$.

67. Show the vectors $u_1 = (-i, 3, 2-i, 1+i)^T, u_2 = (i+1, 2, i, 3)^T, u_3 = (3+i, 3-i, 0, -2)^T, u_4 = (1, 0, 2, 3)^T$ form a basis in $\mathbb{C}^4$. Find the set of reciprocal basis vectors. For $x \in \mathbb{C}^4$, and $x = (i, 3-i, -5, 2+i)^T$,
    (a) express x as an expansion in the above-defined basis vectors. That is find α_i such that $x = \sum_{i=1}^4 \alpha_i u_i$,
    (b) project onto the space spanned by u_1, u_2, and u_3. That is find the best set of α_i such that $x \simeq x_p = \sum_{i=1}^3 \alpha_i u_i$.

68. Consider $d^2y/dt^2 = -ky$, $y(0) = 1$, $dy/dt(0) = 0$. With $\xi \in (-\infty, \infty)$ a random normally distributed variable with mean of zero and standard deviation of unity, consider $k = \mu + \sigma\xi$. Use the method of polynomial chaos to get a two-term estimate for y when $\mu = 1$, $\sigma = 1/10$. Compare the expected value of $y(t = 10)$ with that of a Monte Carlo simulation and that when $\sigma = 0$.

Chapter 8

Linear algebra

see Kaplan, Chapter 1,
see Lopez, Chapters 33, 34,
see Riley, Hobson, and Bence, Chapter 7,
see Michel and Herget,
see Golub and Van Loan,
see Strang, Linear Algebra and its Applications,
see Strang, Introduction to Applied Mathematics.

The key problem in linear algebra is addressing the equation

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}, \tag{8.1}$$

where $\mathbf{A}$ is a known constant rectangular matrix, $\mathbf{b}$ is a known column vector, and $\mathbf{x}$ is an unknown column vector. In this chapter, we will more often consider $\mathbf{A}$ to be an alibi transformation in which the coordinate axes remain fixed, though occasionally we revert to alias transformations. To explicitly indicate the dimension of the matrices and vectors, we sometimes write this in expanded form:

$$\mathbf{A}_{N\times M} \cdot \mathbf{x}_{M\times 1} = \mathbf{b}_{N\times 1}, \tag{8.2}$$

where $N, M \in \mathbb{N}$ are the positive integers which give the dimensions. If $N = M$, the matrix is square, and solution techniques are usually straightforward. For $N \neq M$, which arises often in physical problems, the issues are not as straightforward. In some cases we find an infinite number of solutions; in others we find none. Relaxing our equality constraint, we can, however, always find a vector $\mathbf{x}_p$

$$\mathbf{x}_p = \mathbf{x} \text{ such that } ||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2 \to \min. \tag{8.3}$$

This vector $\mathbf{x}_p$ is the best solution to the equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, for cases in which there is no exact solution. Depending on the problem, it may turn out that $\mathbf{x}_p$ is not unique. It will

always be the case, however, that of all the vectors $\mathbf{x}_p$ which minimize $||\mathbf{A}\cdot\mathbf{x}-\mathbf{b}||_2$, that one of them, $\hat{\mathbf{x}}$, will itself have a minimum norm. We will define here the residual $\mathbf{r}$ as

$$\mathbf{r} = \mathbf{A}\cdot\mathbf{x} - \mathbf{b}. \tag{8.4}$$

In general, we will seek an $\mathbf{x}$ that minimizes $||\mathbf{r}||_2$.

8.1 Determinants and rank

We can take the determinant of a square matrix $\mathbf{A}$, written $\det \mathbf{A}$. Details of computation of determinants are found in any standard reference and will not be repeated here. Properties of the determinant include

- $\det \mathbf{A}_{N\times N}$ is equal to the volume of a parallelepiped in N-dimensional space whose edges are formed by the rows of $\mathbf{A}$.
- If all elements of a row (or column) are multiplied by a scalar, the determinant is also similarly multiplied.
- The elementary operation of subtracting a multiple of one row from another leaves the determinant unchanged.
- If two rows (or columns) of a matrix are interchanged the sign of the determinant changes.

A *singular* matrix is one whose determinant is zero. The *rank* of a matrix is the size r of the largest square non-singular matrix that can be formed by deleting rows and columns.

While the determinant is useful to some ends in linear algebra, most of the common problems are better solved without using the determinant at all; in fact it is probably a fair generalization to say that the determinant is less, rather than more, useful than imagined by many. It is useful in solving linear systems of equations of small dimension, but becomes much too cumbersome relative to other methods for commonly encountered large systems of linear algebraic equations. While it can be used to find the rank, there are also other more efficient means to calculate this. Further, while a zero value for the determinant almost always has significance, other values do not. Some matrices which are particularly ill-conditioned for certain problems often have a determinant which gives no clue as to difficulties which may arise.

8.2 Matrix algebra

We will denote a matrix of size $N \times M$ as

$$\mathbf{A}_{N\times M} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2m} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} & \cdots & a_{nM} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{Nm} & \cdots & a_{NM} \end{pmatrix}. \tag{8.5}$$

Addition of matrices can be defined as

$$\mathbf{A}_{N\times M} + \mathbf{B}_{N\times M} = \mathbf{C}_{N\times M}, \tag{8.6}$$

where the elements of $\mathbf{C}$ are obtained by adding the corresponding elements of $\mathbf{A}$ and $\mathbf{B}$. Multiplication of a matrix by a scalar α can be defined as

$$\alpha\mathbf{A}_{N\times M} = \mathbf{B}_{N\times M}, \tag{8.7}$$

where the elements of $\mathbf{B}$ are the corresponding elements of $\mathbf{A}$ multiplied by α.

It can be shown that the set of all $N \times M$ matrices is a vector space. We will also refer to an $N \times 1$ matrix as an N-dimensional column vector. Likewise a $1 \times M$ matrix will be called an M-dimensional row vector. Unless otherwise stated vectors are assumed to be column vectors. In this sense the inner product of two vectors $\mathbf{x}_{N\times 1}$ and $\mathbf{y}_{N\times 1}$ is $<\mathbf{x}, \mathbf{y}> = \bar{\mathbf{x}}^T \cdot \mathbf{y}$. In this chapter matrices will be represented by upper-case bold-faced letters, such as $\mathbf{A}$, and vectors by lower-case bold-faced letters, such as $\mathbf{x}$.

8.2.1 Column, row, left and right null spaces

The M *column vectors* $\mathbf{c}_m \in \mathbb{C}^N$, $m = 1, 2, \ldots, M$, of the matrix $\mathbf{A}_{N\times M}$ are each one of the columns of $\mathbf{A}$. The *column space* is the subspace of $\mathbb{C}^M$ spanned by the column vectors. The N *row vectors* $\mathbf{r}_n \in \mathbb{C}^M$, $n = 1, 2, \ldots, N$, of the same matrix are each one of the rows. The *row space* is the subspace of $\mathbb{C}^N$ spanned by the row vectors. The column space vectors and the row space vectors span spaces of the same dimension. Consequently, the column space and row space have the same dimension. The right null space is the set of all vectors $\mathbf{x}_{M\times 1} \in \mathbb{C}^M$ for which $\mathbf{A}_{N\times M} \cdot \mathbf{x}_{M\times 1} = \mathbf{0}_{N\times 1}$. The left null space is the set of all vectors $\mathbf{y}_{N\times 1} \in \mathbb{C}^N$ for which $\overline{\mathbf{y}}^T_{N\times 1} \cdot \mathbf{A}_{N\times M} = \overline{\mathbf{y}}_{1\times N} \cdot \mathbf{A}_{N\times M} = \mathbf{0}_{1\times M}$.

If we have $\mathbf{A}_{N\times M} : \mathbb{C}^M \to \mathbb{C}^N$, and recall that the rank of $\mathbf{A}$ is r, then we have the following important results:

- The column space of $\mathbf{A}_{N\times M}$ has dimension r, $(r \leq M)$.
- The left null space of $\mathbf{A}_{N\times M}$ has dimension $N - r$.

- The row space of $\mathbf{A}_{N\times M}$ has dimension r, $(r \leq N)$.
- The right null space of $\mathbf{A}_{N\times M}$ has dimension $M - r$.

We also can show

$$\mathbb{C}^N = \text{column space} \oplus \text{left null space}, \tag{8.8}$$
$$\mathbb{C}^M = \text{row space} \oplus \text{right null space}. \tag{8.9}$$

Also

- Any vector $\mathbf{x} \in \mathbb{C}^M$ can be written as a linear combination of vectors in the row space and the right null space.
- Any M-dimensional vector $\mathbf{x}$ which is in the right null space of $\mathbf{A}$ is orthogonal to any M-dimensional vector in the row space. This comes directly from the definition of the right null space $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$.
- Any vector $\mathbf{y} \in \mathbb{C}^N$ can be written as the sum of vectors in the column space and the left null space.
- Any N-dimensional vector $\mathbf{y}$ which is in the left null space of $\mathbf{A}$ is orthogonal to any N-dimensional vector in the column space. This comes directly from the definition of the left null space $\overline{\mathbf{y}}^T \cdot \mathbf{A} = \mathbf{0}^T$.

Example 8.1

Find the column and row spaces of

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \tag{8.10}$$

and their dimensions.

Restricting ourselves to real vectors, we note first that in the equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, $\mathbf{A}$ is an operator which maps three-dimensional real vectors $\mathbf{x}$ into vectors $\mathbf{b}$ which are elements of a two-dimensional real space, i.e.

$$\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^2. \tag{8.11}$$

The column vectors are

$$\mathbf{c}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{8.12}$$
$$\mathbf{c}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{8.13}$$
$$\mathbf{c}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{8.14}$$

The column space consists of the vectors $\alpha_1\mathbf{c}_1 + \alpha_2\mathbf{c}_2 + \alpha_3\mathbf{c}_3$, where the α's are any scalars. Since only two of the $\mathbf{c}_i$'s are linearly independent, the dimension of the column space is also two. We can see this by looking at the sub-determinant

$$\det\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1, \tag{8.15}$$

which indicates the rank, $r = 2$. Note that

- $\mathbf{c}_1 + 2\mathbf{c}_2 = \mathbf{c}_3$.
- The three column vectors thus lie in a single two-dimensional plane.
- The three column vectors are thus said to span a two-dimensional subspace of $\mathbb{R}^3$.

The two row vectors are

$$\mathbf{r}_1 = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}, \tag{8.16}$$

$$\mathbf{r}_2 = \begin{pmatrix} 0 & 1 & 2 \end{pmatrix}. \tag{8.17}$$

The row space consists of the vectors $\beta_1\mathbf{r}_1 + \beta_2\mathbf{r}_2$, where the β's are any scalars. Since the two $\mathbf{r}_i$'s are linearly independent, the dimension of the row space is also two. That is the two row vectors are both three dimensional, but span a two-dimensional subspace.

We note for instance, if $\mathbf{x} = (1, 2, 1)^T$, that $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ gives

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}. \tag{8.18}$$

So

$$\mathbf{b} = 1\mathbf{c}_1 + 2\mathbf{c}_2 + 1\mathbf{c}_3. \tag{8.19}$$

That is $\mathbf{b}$ is a linear combination of the column space vectors and thus lies in the column space of $\mathbf{A}$. We note for this problem that since an arbitrary $\mathbf{b}$ is two-dimensional and the dimension of the column space is two, that we can represent an arbitrary $\mathbf{b}$ as some linear combination of the column space vectors. For example, we can also say that $\mathbf{b} = 2\mathbf{c}_1 + 4\mathbf{c}_2$. We also note that $\mathbf{x}$ in general *does not* lie in the row space of $\mathbf{A}$, since $\mathbf{x}$ is an arbitrary three-dimensional vector, and we only have enough row vectors to span a two-dimensional subspace (i.e. a plane embedded in a three-dimensional space). However, as will be seen, $\mathbf{x}$ does lie in the space defined by the combination of the row space of $\mathbf{A}$, and the *right null space* of $\mathbf{A}$ (the set of vectors $\mathbf{x}$ for which $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$). In special cases, $\mathbf{x}$ will in fact lie in the row space of $\mathbf{A}$.

8.2.2 Matrix multiplication

Multiplication of matrices $\mathbf{A}$ and $\mathbf{B}$ can be defined if they are of the proper sizes. Thus

$$\mathbf{A}_{N\times L} \cdot \mathbf{B}_{L\times M} = \mathbf{C}_{N\times M}. \tag{8.20}$$

It may be better to say here that $\mathbf{A}$ is a linear operator which operates on elements which are in a space of dimension $L \times M$ so as to generate elements which are in a space of dimension $N \times M$; that is, $\mathbf{A} : \mathbb{R}^L \times \mathbb{R}^M \to \mathbb{R}^N \times \mathbb{R}^M$.

Example 8.2

Consider the matrix operator

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ -3 & 3 & 1 \end{pmatrix}, \tag{8.21}$$

which operates on 3×4 matrices, i.e.

$$\mathbf{A} : \mathbb{R}^3 \times \mathbb{R}^4 \rightarrow \mathbb{R}^2 \times \mathbb{R}^4, \tag{8.22}$$

and show how it acts on another matrix.

We can use $\mathbf{A}$ to operate on a 3×4 matrix as follows:

$$\begin{pmatrix} 1 & 2 & 1 \\ -3 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 & -2 \\ 2 & -4 & 1 & 3 \\ -1 & 4 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 4 & -4 & 5 & 6 \\ 2 & -8 & -6 & 17 \end{pmatrix}. \tag{8.23}$$

Note the operation does not exist if the order is reversed.

A vector operating on a vector can yield a scalar or a matrix, depending on the order of operation.

Example 8.3

Consider the vector operations $\mathbf{A}_{1\times 3} \cdot \mathbf{B}_{3\times 1}$ and $\mathbf{B}_{3\times 1} \cdot \mathbf{A}_{1\times 3}$ where

$$\mathbf{A}_{1\times 3} = \mathbf{a}^T = \begin{pmatrix} 2 & 3 & 1 \end{pmatrix}, \tag{8.24}$$

$$\mathbf{B}_{3\times 1} = \mathbf{b} = \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix}. \tag{8.25}$$

Then

$$\mathbf{A}_{1\times 3} \cdot \mathbf{B}_{3\times 1} = \mathbf{a}^T \cdot \mathbf{b} = \begin{pmatrix} 2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix} = (2)(3) + (3)(-2) + (1)(5) = 5. \tag{8.26}$$

This is the ordinary inner product $<\mathbf{a}, \mathbf{b}>$. The commutation of this operation however yields a matrix:

$$\mathbf{B}_{3\times 1} \cdot \mathbf{A}_{1\times 3} = \mathbf{b}\mathbf{a}^T = \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} (3)(2) & (3)(3) & (3)(1) \\ (-2)(2) & (-2)(3) & (-2)(1) \\ (5)(2) & (5)(3) & (5)(1) \end{pmatrix}, \tag{8.27}$$

$$= \begin{pmatrix} 6 & 9 & 3 \\ -4 & -6 & -2 \\ 10 & 15 & 5 \end{pmatrix}. \tag{8.28}$$

This is the *dyadic product* of the two vectors. Note that for vector (lower case notation) the dyadic product usually is not characterized by the "dot" operator that we use for the vector inner product.

A special case is that of a *square* matrix $\mathbf{A}_{N\times N}$ of size N. For square matrices of the same size both $\mathbf{A}\cdot\mathbf{B}$ and $\mathbf{B}\cdot\mathbf{A}$ exist. While $\mathbf{A}\cdot\mathbf{B}$ and $\mathbf{B}\cdot\mathbf{A}$ both yield $N\times N$ matrices, the actual value of the two products is different. In what follows, we will often assume that we are dealing with square matrices.
Properties of matrices include

1. $(\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C} = \mathbf{A}\cdot(\mathbf{B}\cdot\mathbf{C})$ (associative),
2. $\mathbf{A}\cdot(\mathbf{B}+\mathbf{C}) = \mathbf{A}\cdot\mathbf{B}+\mathbf{A}\cdot\mathbf{C}$ (distributive),
3. $(\mathbf{A}+\mathbf{B})\cdot\mathbf{C} = \mathbf{A}\cdot\mathbf{C}+\mathbf{B}\cdot\mathbf{C}$ (distributive),
4. $\mathbf{A}\cdot\mathbf{B} \neq \mathbf{B}\cdot\mathbf{A}$ in general (not commutative),
5. $\det \mathbf{A}\cdot\mathbf{B} = (\det\mathbf{A})(\det\mathbf{B})$.

8.2.3 Definitions and properties

8.2.3.1 Identity

The *identity* matrix $\mathbf{I}$ is a square diagonal matrix with 1 on the main diagonal. With this definition, we get

$$\mathbf{A}_{N\times M}\cdot\mathbf{I}_{M\times M} = \mathbf{A}_{N\times M}, \tag{8.29}$$
$$\mathbf{I}_{N\times N}\cdot\mathbf{A}_{N\times M} = \mathbf{A}_{N\times M}, \quad \text{or, more compactly,} \tag{8.30}$$
$$\mathbf{A}\cdot\mathbf{I} = \mathbf{I}\cdot\mathbf{A} = \mathbf{A}, \tag{8.31}$$

where the unsubscripted identity matrix is understood to be square with the correct dimension for matrix multiplication.

8.2.3.2 Nilpotent

A square matrix $\mathbf{A}$ is called *nilpotent* if there exists a positive integer n for which $\mathbf{A}^n = 0$.

8.2.3.3 Idempotent

A square matrix $\mathbf{A}$ is called *idempotent* if $\mathbf{A}\cdot\mathbf{A} = \mathbf{A}$. The identity matrix $\mathbf{I}$ is idempotent. Projection matrices $\mathbf{P}$, see Eq. (7.160), are idempotent. All idempotent matrices which are not the identity matrix are singular. The trace of an idempotent matrix gives its rank. More generally, a function f is idempotent if $f(f(x)) = f(x)$. As an example, the absolute value function is idempotent since $\text{abs}(\text{abs}(x)) = \text{abs}(x)$.

8.2.3.4 Diagonal

A *diagonal* matrix $\mathbf{D}$ has nonzero terms only along its main diagonal. The sum and product of diagonal matrices are also diagonal. The determinant of a diagonal matrix is the product of all diagonal elements.

8.2.3.5 Transpose

Here we expand on the earlier discussion of Sec. 6.2.3. The *transpose* $\mathbf{A}^T$ of a matrix $\mathbf{A}$ is an operation in which the terms above and below the diagonal are interchanged. For any matrix $\mathbf{A}_{N\times M}$, we find that $\mathbf{A}\cdot\mathbf{A}^T$ and $\mathbf{A}^T\cdot\mathbf{A}$ are square matrices of size N and M, respectively. Properties of the transpose include

1. $\det\mathbf{A} = \det\mathbf{A}^T$,
2. $(\mathbf{A}_{N\times M}\cdot\mathbf{B}_{M\times N})^T = \mathbf{B}^T\cdot\mathbf{A}^T$,
3. $(\mathbf{A}_{N\times N}\cdot\mathbf{x}_{N\times 1})^T\cdot\mathbf{y}_{N\times 1} = \mathbf{x}^T\cdot\mathbf{A}^T\cdot\mathbf{y} = \mathbf{x}^T\cdot(\mathbf{A}^T\cdot\mathbf{y})$.

8.2.3.6 Symmetry, anti-symmetry, and asymmetry

To reiterate the earlier discussion of Sec. 6.2.3, a *symmetric* matrix is one for which $\mathbf{A}^T = \mathbf{A}$. An *anti-symmetric* or *skew-symmetric* matrix is one for which $\mathbf{A}^T = -\mathbf{A}$. Any matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \frac{1}{2}(\mathbf{A}+\mathbf{A}^T) + \frac{1}{2}(\mathbf{A}-\mathbf{A}^T), \tag{8.32}$$

where $(1/2)(\mathbf{A}+\mathbf{A}^T)$ is symmetric and $(1/2)(\mathbf{A}-\mathbf{A}^T)$ is anti-symmetric. An *asymmetric* matrix is neither symmetric nor anti-symmetric.

8.2.3.7 Triangular

A lower (or upper) triangular matrix is one in which all entries above (or below) the main diagonal are zero. Lower triangular matrices are often denoted by $\mathbf{L}$, and upper triangular matrices by either $\mathbf{U}$ or $\mathbf{R}$.

8.2.3.8 Positive definite

A *positive definite* matrix $\mathbf{A}$ is a matrix for which $\mathbf{x}^T\cdot\mathbf{A}\cdot\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x}$. A positive definite matrix has real, positive eigenvalues. Every positive definite matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{U}^T\cdot\mathbf{U}$, where $\mathbf{U}$ is an upper triangular matrix (Cholesky[1] decomposition).

[1] after André-Louis Cholesky, 1875-1918, French mathematician and military officer.

8.2.3.9 Permutation

A *permutation* matrix $\mathbf{P}$ is a square matrix composed of zeroes and a single one in each column. None of the ones occur in the same row. It effects a row exchange when it operates on a general matrix $\mathbf{A}$. It is never singular, and is in fact its own inverse, $\mathbf{P} = \mathbf{P}^{-1}$, so $\mathbf{P} \cdot \mathbf{P} = \mathbf{I}$. Also $||\mathbf{P}||_2 = 1$, and $|\det \mathbf{P}| = 1$. However, we can have $\det \mathbf{P} = \pm 1$, so it can be either a rotation or a reflection.

The permutation matrix $\mathbf{P}$ is not to be confused with a projection matrix $\mathbf{P}$, which is usually denoted in the same way. The context should be clear as to which matrix is intended.

Example 8.4

Find the permutation matrix $\mathbf{P}$ which effects the exchange of the first and second rows of $\mathbf{A}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 5 & 7 \\ 2 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \end{pmatrix}. \tag{8.33}$$

To construct $\mathbf{P}$, we begin with at 3×3 identity matrix $\mathbf{I}$. For a first and second row exchange, we replace the ones in the $(1,1)$ and $(2,2)$ slot with zero, then replace the zeroes in the $(1,2)$ and $(2,1)$ slot with ones. Thus

$$\mathbf{P} \cdot \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 5 & 7 \\ 2 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 1 & 2 \\ 1 & 3 & 5 & 7 \\ 3 & 1 & 3 & 2 \end{pmatrix}. \tag{8.34}$$

Example 8.5

Find the rank and right null space of

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 5 & 4 & 9 \\ 2 & 4 & 6 \end{pmatrix}. \tag{8.35}$$

The rank of $\mathbf{A}$ is not three since

$$\det \mathbf{A} = 0. \tag{8.36}$$

Since

$$\begin{vmatrix} 1 & 0 \\ 5 & 4 \end{vmatrix} \neq 0, \tag{8.37}$$

the rank of $\mathbf{A}$ is 2.

Let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \tag{8.38}$$

belong to the right null space of $\mathbf{A}$. Then

$$\begin{aligned} x_1 + x_3 &= 0, & (8.39)\\ 5x_1 + 4x_2 + 9x_3 &= 0, & (8.40)\\ 2x_1 + 4x_2 + 6x_3 &= 0. & (8.41) \end{aligned}$$

One strategy to solve singular systems is to take one of the variables to be a known parameter, and see if the resulting system can be solved. If the resulting system remains singular, take a second variable to be a second parameter. This *ad hoc* method will later be made systematic.

So here take $x_1 = t$, and consider the first two equations, which gives

$$\begin{pmatrix} 0 & 1 \\ 4 & 9 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -t \\ -5t \end{pmatrix}. \tag{8.42}$$

Solving, we find $x_2 = t$, $x_3 = -t$. So,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} t \\ t \\ -t \end{pmatrix} = t \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \qquad t \in \mathbb{R}^1. \tag{8.43}$$

Therefore, the right null space is the straight line in $\mathbb{R}^3$ which passes through (0,0,0) and (1,1,-1).

8.2.3.10 Inverse

Definition: A matrix $\mathbf{A}$ has an inverse $\mathbf{A}^{-1}$ if $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}$.

Theorem

A unique inverse exists if the matrix is non-singular.

Properties of the inverse include

1. $(\mathbf{A} \cdot \mathbf{B})^{-1} = \mathbf{B}^{-1} \cdot \mathbf{A}^{-1}$,
2. $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$,
3. $\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1}$.

If a_{ij} and a_{ij}^{-1} are the elements of $\mathbf{A}$ and $\mathbf{A}^{-1}$, and we define the *cofactor* as

$$c_{ij} = (-1)^{i+j} m_{ij}, \tag{8.44}$$

where the *minor*, m_{ij} is the determinant of the matrix obtained by canceling out the j-th row and i-th column, then the inverse is

$$a_{ij}^{-1} = \frac{c_{ij}}{\det \mathbf{A}}. \tag{8.45}$$

The inverse of a diagonal matrix is also diagonal, but with the reciprocals of the original diagonal elements.

Example 8.6

Find the inverse of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{8.46}$$

The inverse is

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{8.47}$$

We can confirm that $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}$.

8.2.3.11 Similar matrices

Matrices $\mathbf{A}$ and $\mathbf{B}$ are *similar* if there exists a non-singular matrix $\mathbf{S}$ such that $\mathbf{B} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$. Similar matrices have the same determinant, eigenvalues, multiplicities and eigenvectors.

8.2.4 Equations

In general, for matrices that are not necessarily square, the equation $\mathbf{A}_{N\times M} \cdot \mathbf{x}_{M\times 1} = \mathbf{b}_{N\times 1}$ is solvable iff $\mathbf{b}$ can be expressed as combinations of the columns of $\mathbf{A}$. Problems in which $M < N$ are *over-constrained*; in special cases, those in which $\mathbf{b}$ is in the column space of $\mathbf{A}$, a unique solution $\mathbf{x}$ exists. However in general no solution $\mathbf{x}$ exists; nevertheless, one can find an $\mathbf{x}$ which will minimize $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$. This is closely related to what is known as the method of least squares. Problems in which $M > N$ are generally *under-constrained*, and have an infinite number of solutions $\mathbf{x}$ which will satisfy the original equation. Problems for which $M = N$ (square matrices) have a unique solution $\mathbf{x}$ when the rank r of $\mathbf{A}$ is equal to N. If $r < N$, then the problem is under-constrained.

8.2.4.1 Over-constrained systems

Example 8.7

For $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{b} \in \mathbb{R}^3$, $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^3$, consider

$$\begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}. \tag{8.48}$$

Here it turns out that $\mathbf{b} = (0, 1, 3)^T$ is *not* in the column space of $\mathbf{A}$, and there is no solution $\mathbf{x}$ for which $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$! The column space is a plane defined by two vectors; the vector $\mathbf{b}$ does not happen to lie in the plane defined by the column space. However, we can find a solution $\mathbf{x} = \mathbf{x}_p$, where $\mathbf{x}_p$ can be shown to minimize the Euclidean norm of the residual $||\mathbf{A} \cdot \mathbf{x}_p - \mathbf{b}||_2$. This is achieved by the following

procedure, the same employed earlier in Sec. 7.3.2.6, in which we operate on both vectors $\mathbf{A}\cdot\mathbf{x}_p$ and $\mathbf{b}$ by the operator $\mathbf{A}^T$ so as to map both vectors *into the same space*, namely the row space of $\mathbf{A}$. Once the vectors are in the same space, a unique inversion is possible.

$$\begin{aligned} \mathbf{A}\cdot\mathbf{x}_p &\simeq \mathbf{b}, & (8.49)\\ \mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x}_p &= \mathbf{A}^T\cdot\mathbf{b}, & (8.50)\\ \mathbf{x}_p &= (\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T\cdot\mathbf{b}. & (8.51)\end{aligned}$$

These operations are, numerically,

$$\begin{aligned} \begin{pmatrix}1&1&1\\2&0&1\end{pmatrix}\begin{pmatrix}1&2\\1&0\\1&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} &= \begin{pmatrix}1&1&1\\2&0&1\end{pmatrix}\begin{pmatrix}0\\1\\3\end{pmatrix}, & (8.52)\\ \begin{pmatrix}3&3\\3&5\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} &= \begin{pmatrix}4\\3\end{pmatrix}, & (8.53)\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} &= \begin{pmatrix}\frac{11}{6}\\-\frac{1}{2}\end{pmatrix}. & (8.54)\end{aligned}$$

Note the resulting $\mathbf{x}_p$ *will not satisfy* $\mathbf{A}\cdot\mathbf{x}_p=\mathbf{b}$. We can define the difference of $\mathbf{A}\cdot\mathbf{x}_p$ and $\mathbf{b}$ as the residual vector, see Eq. (8.4), $\mathbf{r}=\mathbf{A}\cdot\mathbf{x}_p-\mathbf{b}$. In fact, $||\mathbf{r}||_2=||\mathbf{A}\cdot\mathbf{x}_p-\mathbf{b}||_2=2.0412$. If we tried any nearby $\mathbf{x}$, say $\mathbf{x}=(2,-3/5)^T$, $||\mathbf{A}\cdot\mathbf{x}-\mathbf{b}||_2=2.0494>2.0412$. Since the problem is linear, this minimum is global; if we take $\mathbf{x}=(10,-24)^T$, then $||\mathbf{A}\cdot\mathbf{x}-\mathbf{b}||_2=42.5911>2.0412$. Though we have not proved it, our $\mathbf{x}_p$ is the unique vector which minimizes the Euclidean norm of the residual.

Further manipulation shows that we can write our solution as a combination of vectors in the row space of $\mathbf{A}$. As the dimension of the right null space of $\mathbf{A}$ is zero, there is no possible contribution from the right null space vectors.

$$\begin{aligned} \begin{pmatrix}\frac{11}{6}\\-\frac{1}{2}\end{pmatrix} &= \alpha_1\begin{pmatrix}1\\2\end{pmatrix}+\alpha_2\begin{pmatrix}1\\0\end{pmatrix}, & (8.55)\\ \begin{pmatrix}\frac{11}{6}\\-\frac{1}{2}\end{pmatrix} &= \begin{pmatrix}1&1\\2&0\end{pmatrix}\begin{pmatrix}\alpha_1\\\alpha_2\end{pmatrix}, & (8.56)\\ \begin{pmatrix}\alpha_1\\\alpha_2\end{pmatrix} &= \begin{pmatrix}-\frac{1}{4}\\\frac{25}{12}\end{pmatrix}. & (8.57)\end{aligned}$$

So

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} = \underbrace{-\frac{1}{4}\begin{pmatrix}1\\2\end{pmatrix}+\frac{25}{12}\begin{pmatrix}1\\0\end{pmatrix}}_{\text{linear combination of row space vectors}}. \quad (8.58)$$

We could also have chosen to expand in terms of the other row space vector $(1,1)^T$, since any two of the three row space vectors span the space $\mathbb{R}^2$.

The vector $\mathbf{A}\cdot\mathbf{x}_p$ actually represents the *projection* of $\mathbf{b}$ onto the subspace spanned by the column vectors (i.e. the column space). Call the projected vector $\mathbf{b}_p$:

$$\mathbf{b}_p=\mathbf{A}\cdot\mathbf{x}_p=\underbrace{\mathbf{A}\cdot(\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T}_{\text{projection matrix, }\mathbf{P}}\cdot\mathbf{b}. \quad (8.59)$$

For this example $\mathbf{b}_p=(5/6,11/6,4/3)^T$. We can think of $\mathbf{b}_p$ as the shadow cast by $\mathbf{b}$ onto the column space. Here, following Eq. (7.160), we have the projection matrix $\mathbf{P}$ as

$$\mathbf{P}=\mathbf{A}\cdot(\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T. \quad (8.60)$$

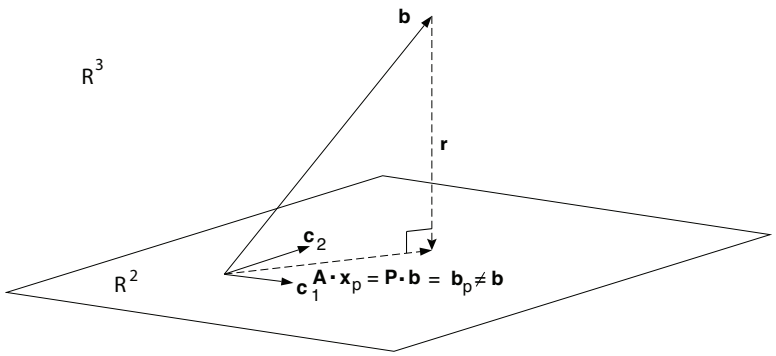

Figure 8.1: Plot for $\mathbf{b}$ which lies outside of column space (space spanned by $\mathbf{c}_1$ and $\mathbf{c}_2$) of $\mathbf{A}$.

A sketch of this system is shown in Fig. 8.1. Here we sketch what might represent this example in which the column space of $\mathbf{A}$ does not span the entire space $\mathbb{R}^3$, and for which $\mathbf{b}$ lies outside of the column space of $\mathbf{A}$. In such a case $||\mathbf{A} \cdot \mathbf{x}_p - \mathbf{b}||_2 > 0$. We have $\mathbf{A}$ as a matrix which maps two-dimensional vectors $\mathbf{x}$ into three-dimensional vectors $\mathbf{b}$. Our space is $\mathbb{R}^3$, and embedded within that space are two column vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ which span a column space $\mathbb{R}^2$, which is represented by a plane within a three-dimensional volume. Since $\mathbf{b}$ lies outside the column space, there exists no unique vector $\mathbf{x}$ for which $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$.

Example 8.8

For $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{b} \in \mathbb{R}^3$, consider $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^3$,

$$\begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix}. \tag{8.61}$$

The column space of $\mathbf{A}$ is spanned by the two column vectors

$$\mathbf{c}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{c}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}. \tag{8.62}$$

Our equation can also be cast in the form which makes the contribution of the column vectors obvious:

$$x_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix}. \tag{8.63}$$

Here we have the unusual case that $\mathbf{b} = (5, 1, 3)^T$ is in the column space of $\mathbf{A}$ (in fact $\mathbf{b} = \mathbf{c}_1 + 2\mathbf{c}_2$), and we have a unique solution of

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{8.64}$$

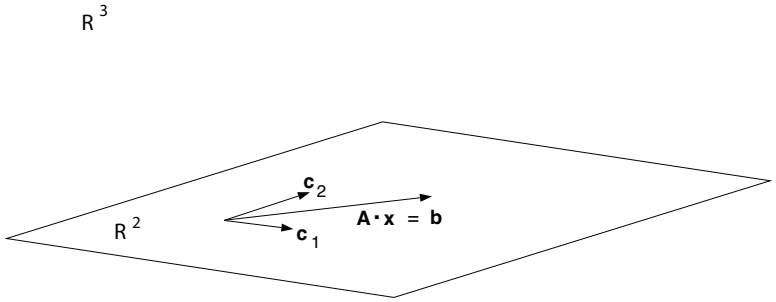

Figure 8.2: Plot for $\mathbf{b}$ which lies in column space (space spanned by $\mathbf{c}_1$ and $\mathbf{c}_2$) of $\mathbf{A}$.

In most cases, however, it is not obvious that $\mathbf{b}$ lies in the column space. We can still operate on both sides by the transpose and solve, which will reveal the correct result:

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix}, \tag{8.65}$$

$$\begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 13 \end{pmatrix}, \tag{8.66}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{8.67}$$

A quick check of the residual shows that in fact $\mathbf{r} = \mathbf{A} \cdot \mathbf{x}_p - \mathbf{b} = \mathbf{0}$. So, we have an exact solution for which $\mathbf{x} = \mathbf{x}_p$.

Note that the solution vector $\mathbf{x}$ lies entirely in the row space of $\mathbf{A}$; here, it is identically the first row vector $\mathbf{r}_1 = (1, 2)^T$. Note also that here the column space is a two-dimensional subspace, in this case a plane defined by the two column vectors, embedded within a three-dimensional space. The operator $\mathbf{A}$ maps arbitrary two-dimensional vectors $\mathbf{x}$ into the three-dimensional $\mathbf{b}$; however, these $\mathbf{b}$ vectors are confined to a two-dimensional subspace within the greater three-dimensional space. Consequently, we cannot always expect to find a vector $\mathbf{x}$ for arbitrary $\mathbf{b}$!

A sketch of this system is shown in Fig. 8.2. Here we sketch what might represent this example in which the column space of $\mathbf{A}$ does not span the entire space $\mathbb{R}^3$, but for which $\mathbf{b}$ lies in the column space of $\mathbf{A}$. In such a case $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2 = 0$. We have $\mathbf{A}$ as a matrix which maps two-dimensional vectors $\mathbf{x}$ into three-dimensional vectors $\mathbf{b}$. Our space is $\mathbb{R}^3$ and embedded within that space are two column vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ which span a column space $\mathbb{R}^2$, which is represented by a plane within a three-dimensional volume. Since $\mathbf{b}$ in this example happens to lie in the column space, there exists a unique vector $\mathbf{x}$ for which $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$.

8.2.4.2 Under-constrained systems

Example 8.9

Consider now $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^2$ such that

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \tag{8.68}$$

In this case operating on both sides by the transpose is not useful because $(\mathbf{A}^T \cdot \mathbf{A})^{-1}$ does not exist. We take an alternate strategy.

Certainly $\mathbf{b} = (1,3)^T$ lies in the column space of $\mathbf{A}$, since for example, $\mathbf{b} = 0(1,2)^T - 2(1,0)^T + 3(1,1)^T$. Setting $x_1 = t$, where t is an arbitrary number, lets us solve for x_2, x_3:

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} t \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \tag{8.69}$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1-t \\ 3-2t \end{pmatrix}. \tag{8.70}$$

Inversion gives

$$\begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -2+t \\ 3-2t \end{pmatrix}, \tag{8.71}$$

so

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} t \\ -2+t \\ 3-2t \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \\ 3 \end{pmatrix} + \underbrace{t \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}}_{\text{right null space}}, \qquad t \in \mathbb{R}^1. \tag{8.72}$$

A useful way to think of problems such as this which are undetermined is that *the matrix* $\mathbf{A}$ *maps the additive combination of a unique vector from the row space of* $\mathbf{A}$ *plus an arbitrary vector from the right null space of* $\mathbf{A}$ *into the vector* $\mathbf{b}$. Here the vector $(1,1,-2)^T$ is in the right null space; however, the vector $(0,-2,3)^T$ has components in both the right null space and the row space. Let us extract the parts of $(0,-2,3)^T$ which are in each space. Since the row space and right null space are linearly independent, they form a basis, and we can say

$$\begin{pmatrix} 0 \\ -2 \\ 3 \end{pmatrix} = \underbrace{a_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}}_{\text{row space}} + \underbrace{a_3 \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}}_{\text{right null space}}. \tag{8.73}$$

In matrix form, we then get

$$\begin{pmatrix} 0 \\ -2 \\ 3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -2 \end{pmatrix}}_{\text{invertible}} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}. \tag{8.74}$$

The coefficient matrix is non-singular and thus invertible. Solving, we get

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ 1 \\ -\frac{4}{3} \end{pmatrix}. \tag{8.75}$$

So $\mathbf{x}$ can be rewritten as

$$\mathbf{x} = \underbrace{-\frac{2}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}}_{\text{row space}} + \underbrace{\left(t - \frac{4}{3}\right) \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}}_{\text{right null space}}, \qquad t \in \mathbb{R}^1. \tag{8.76}$$

The first two terms in the right-hand side of Eq. (8.76) are the unique linear combination of the row space vectors, while the third term is from the right null space. As by definition, $\mathbf{A}$ maps any vector

from the right null space into the zero element, it makes no contribution to forming $\mathbf{b}$; hence, one can allow for an arbitrary constant. Note the analogy here with solutions to inhomogeneous differential equations. The right null space vector can be thought of as a solution to the homogeneous equation, and the terms with the row space vectors can be thought of as particular solutions.

We can also write the solution $\mathbf{x}$ in matrix form. The matrix is composed of three column vectors, which are the original two row space vectors and the right null space vector, which together form a basis in $\mathbb{R}^3$:

$$\mathbf{x} = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -2 \end{pmatrix} \begin{pmatrix} -\frac{2}{3} \\ 1 \\ t - \frac{4}{3} \end{pmatrix}, \qquad t \in \mathbb{R}^1. \tag{8.77}$$

While the right null space vector is orthogonal to both row space vectors, the row space vectors are not orthogonal to themselves, so this basis is not orthogonal. Leaving out the calculational details, we can use the Gram-Schmidt procedure to cast the solution on an orthonormal basis:

$$\mathbf{x} = \underbrace{\frac{1}{\sqrt{3}} \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} + \sqrt{2} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}}_{\text{row space}} + \underbrace{\sqrt{6}\left(t - \frac{4}{3}\right) \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\sqrt{\frac{2}{3}} \end{pmatrix}}_{\text{right null space}}, \qquad t \in \mathbb{R}^1. \tag{8.78}$$

The first two terms are in the row space, now represented on an orthonormal basis, the third is in the right null space. In matrix form, we can say that

$$\mathbf{x} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \sqrt{2} \\ \sqrt{6}\left(t - \frac{4}{3}\right) \end{pmatrix}, \qquad t \in \mathbb{R}^1. \tag{8.79}$$

Of course, there are other orthonormal bases on which the system can be cast.

We see that the minimum length of the vector $\mathbf{x}$ occurs when $t = 4/3$, that is when $\mathbf{x}$ is entirely in the row space. In such a case we have

$$\min ||\mathbf{x}||_2 = \sqrt{\left(\frac{1}{\sqrt{3}}\right)^2 + \left(\sqrt{2}\right)^2} = \sqrt{\frac{7}{3}}. \tag{8.80}$$

Lastly note that here, we achieved a reasonable answer by setting $x_1 = t$ at the outset. We could have achieved an equivalent result by starting with $x_2 = t$, or $x_3 = t$. This will not work in all problems, as will be discussed in Sec. 8.8.3 on *row echelon form*.

8.2.4.3 Simultaneously over- and under-constrained systems

Some systems of equations are both over- and under-constrained simultaneously. This often happens when the rank r of the matrix is less than both N and M, the matrix dimensions. Such matrices are known as less than full rank matrices.

Example 8.10

Consider $\mathbf{A} : \mathbb{R}^4 \rightarrow \mathbb{R}^3$ such that

$$\begin{pmatrix} 1 & 2 & 0 & 4 \\ 3 & 2 & -1 & 3 \\ -1 & 2 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}. \tag{8.81}$$

Using elementary row operations to perform Gaussian elimination gives rise to the equivalent system:

$$\begin{pmatrix} 1 & 0 & -1/2 & -1/2 \\ 0 & 1 & 1/4 & 9/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{8.82}$$

We immediately see that there is a problem in the last equation, which purports $0 = 1$! What is actually happening is that $\mathbf{A}$ is not full rank $r = 3$, but actually has $r = 2$, so vectors $\mathbf{x} \in \mathbb{R}^4$ are mapped into a two-dimensional subspace. So, we do not expect to find any solution to this problem, since our vector $\mathbf{b}$ is an arbitrary three-dimensional vector which most likely does not lie in the two-dimensional subspace. We can, however, find an $\mathbf{x}$ which minimizes the Euclidean norm of the residual. We return to the original equation and operate on a both sides with $\mathbf{A}^T$ to form $\mathbf{A}^T \cdot \mathbf{A} \cdot \mathbf{x} = \mathbf{A}^T \cdot \mathbf{b}$. It can be easily verified that if we chose to operate on the system which was reduced by Gaussian elimination that we would *not* recover a solution which minimized $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||$!

$$\begin{pmatrix} 1 & 3 & -1 \\ 2 & 2 & 2 \\ 0 & -1 & 1 \\ 4 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 4 \\ 3 & 2 & -1 & 3 \\ -1 & 2 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & 3 & -1 \\ 2 & 2 & 2 \\ 0 & -1 & 1 \\ 4 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \tag{8.83}$$

$$\begin{pmatrix} 11 & 6 & -4 & 8 \\ 6 & 12 & 0 & 24 \\ -4 & 0 & 2 & 2 \\ 8 & 24 & 2 & 50 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 12 \\ -1 \\ 23 \end{pmatrix}. \tag{8.84}$$

This operation has mapped both sides of the equation into the same space, namely, the column space of $\mathbf{A}^T$, which is also the row space of $\mathbf{A}$. Since the rank of $\mathbf{A}$ is $r = 2$, the dimension of the row space is also two, and now the vectors on both sides of the equation have been mapped into the same plane. Again using row operations to perform Gaussian elimination gives rise to

$$\begin{pmatrix} 1 & 0 & -1/2 & -1/2 \\ 0 & 1 & 1/4 & 9/4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1/4 \\ 7/8 \\ 0 \\ 0 \end{pmatrix}. \tag{8.85}$$

This equation suggests that here x_3 and x_4 are arbitrary, so we set $x_3 = s$, $x_4 = t$ and, treating s and t as known quantities, reduce the system to the following

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/4 + s/2 + t/2 \\ 7/8 - s/4 - 9t/4 \end{pmatrix}, \tag{8.86}$$

so

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1/4 \\ 7/8 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1/2 \\ -1/4 \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 1/2 \\ -9/4 \\ 0 \\ 1 \end{pmatrix}. \tag{8.87}$$

The vectors which are multiplied by s and t are in the right null space of $\mathbf{A}$. The vector $(1/4, 7/8, 0, 0)^T$ is not entirely in the row space of $\mathbf{A}$; it has components in both the row space and right null space. We can, thus, decompose this vector into a linear combination of row space vectors and right null space vectors using the procedure in the previous section, solving the following equation for the coefficients $a_1, \ldots, a_4$, which are the coefficients of the row and right null space vectors:

$$\begin{pmatrix} 1/4 \\ 7/8 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 1/2 & 1/2 \\ 2 & 2 & -1/4 & -9/4 \\ 0 & -1 & 1 & 0 \\ 4 & 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}. \tag{8.88}$$

Solving, we get

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} -3/244 \\ 29/244 \\ 29/244 \\ -75/244 \end{pmatrix}. \tag{8.89}$$

So we can recast the solution as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \underbrace{-\frac{3}{244}\begin{pmatrix} 1 \\ 2 \\ 0 \\ 4 \end{pmatrix} + \frac{29}{244}\begin{pmatrix} 3 \\ 2 \\ -1 \\ 3 \end{pmatrix}}_{\text{row space}} + \underbrace{\left(s + \frac{29}{244}\right)\begin{pmatrix} 1/2 \\ -1/4 \\ 1 \\ 0 \end{pmatrix} + \left(t - \frac{75}{244}\right)\begin{pmatrix} 1/2 \\ -9/4 \\ 0 \\ 1 \end{pmatrix}}_{\text{right null space}}. \tag{8.90}$$

This choice of $\mathbf{x}$ guarantees that we minimize $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$, which in this case is 1.22474. So there are no vectors $\mathbf{x}$ which satisfy the original equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, but there are a doubly infinite number of vectors $\mathbf{x}$ which can minimize the Euclidean norm of the residual.

We can choose special values of s and t such that we minimize $||\mathbf{x}||_2$ while maintaining $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ at its global minimum. This is done simply by forcing the magnitude of the right null space vectors to zero, so we choose $s = -29/244$, $t = 75/244$, giving

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \underbrace{-\frac{3}{244}\begin{pmatrix} 1 \\ 2 \\ 0 \\ 4 \end{pmatrix} + \frac{29}{244}\begin{pmatrix} 3 \\ 2 \\ -1 \\ 3 \end{pmatrix}}_{\text{row space}} = \begin{pmatrix} 21/61 \\ 13/61 \\ -29/244 \\ 75/244 \end{pmatrix}. \tag{8.91}$$

This vector has $||\mathbf{x}||_2 = 0.522055$.

8.2.4.4 Square systems

A set of N linear algebraic equations in N unknowns can be represented as

$$\mathbf{A}_{N\times N} \cdot \mathbf{x}_{N\times 1} = \mathbf{b}_{N\times 1}. \tag{8.92}$$

There is a unique solution if $\det \mathbf{A} \neq 0$ and either no solution or an infinite number of solutions otherwise. In the case where there are no solutions, one can still find an $\mathbf{x}$ which minimizes the normed residual $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$.

Theorem

(Cramer's rule) The solution of the equation is

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}}, \tag{8.93}$$

where $\mathbf{A}_i$ is the matrix obtained by replacing the i-th column of $\mathbf{A}$ by y. While generally valid, Cramer's rule is most useful for low dimension systems. For large systems, Gaussian elimination is a more efficient technique.

Example 8.11

For $\mathbf{A}$: $\mathbb{R}^2 \to \mathbb{R}^2$, Solve for $\mathbf{x}$ in $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$:

$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}. \tag{8.94}$$

By Cramer's rule

$$x_1 = \frac{\begin{vmatrix} 4 & 2 \\ 5 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix}} = \frac{-2}{-4} = \frac{1}{2}, \tag{8.95}$$

$$x_2 = \frac{\begin{vmatrix} 1 & 4 \\ 3 & 5 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix}} = \frac{-7}{-4} = \frac{7}{4}. \tag{8.96}$$

So

$$\mathbf{x} = \begin{pmatrix} \frac{1}{2} \\ \frac{7}{4} \end{pmatrix}. \tag{8.97}$$

We get the same result by Gaussian elimination. Subtracting three times the first row from the second yields

$$\begin{pmatrix} 1 & 2 \\ 0 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ -7 \end{pmatrix}. \tag{8.98}$$

Thus, $x_2 = 7/4$. Back substitution into the first equation then gives $x_1 = 1/2$.

Example 8.12

With $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$, find the most general $\mathbf{x}$ which best satisfies $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ for

$$\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}. \tag{8.99}$$

Obviously, there is no unique solution to this system since the determinant of the coefficient matrix is zero. The rank of $\mathbf{A}$ is 1, so in actuality, $\mathbf{A}$ maps vectors from $\mathbb{R}^2$ into a one-dimensional subspace,

$\mathbb{R}^1$. For a general $\mathbf{b}$, which does not lie in the one-dimensional subspace, we can find the best solution $\mathbf{x}$ by first multiplying both sides by $\mathbf{A}^T$:

$$\begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \tag{8.100}$$

$$\begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}. \tag{8.101}$$

This operation maps both sides of the equation into the column space of $\mathbf{A}^T$, which is the row space of $\mathbf{A}$, which has dimension 1. Since the vectors are now in the same space, a solution can be found. Using row reductions to perform Gaussian elimination, we get

$$\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/5 \\ 0 \end{pmatrix}. \tag{8.102}$$

We set $x_2 = t$, where t is any arbitrary real number and solve to get

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/5 \\ 0 \end{pmatrix} + t \begin{pmatrix} -2 \\ 1 \end{pmatrix}. \tag{8.103}$$

The vector which t multiplies, $(-2, 1)^T$, is in the right null space of $\mathbf{A}$. We can recast the vector $(1/5, 0)^T$ in terms of a linear combination of the row space vector $(1, 2)^T$ and the right null space vector to get the final form of the solution:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\frac{1}{25} \begin{pmatrix} 1 \\ 2 \end{pmatrix}}_{\text{row space}} + \underbrace{\left(t - \frac{2}{25} \right) \begin{pmatrix} -2 \\ 1 \end{pmatrix}}_{\text{right null space}}. \tag{8.104}$$

This choice of $\mathbf{x}$ guarantees that the Euclidean norm of the residual $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ is minimized. In this case the Euclidean norm of the residual is 1.89737. The vector $\mathbf{x}$ with the smallest norm that minimizes $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ is found by setting the magnitude of the right null space contribution to zero, so we can take $t = 2/25$ giving

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\frac{1}{25} \begin{pmatrix} 1 \\ 2 \end{pmatrix}}_{\text{row space}}. \tag{8.105}$$

This gives rise to $||\mathbf{x}||_2 = 0.0894427$.

8.3 Eigenvalues and eigenvectors

8.3.1 Ordinary eigenvalues and eigenvectors

Much of the general discussion of eigenvectors and eigenvalues has been covered in Chap. 7, see especially Sec. 7.4.4, and will not be repeated here. A few new concepts are introduced, and some old ones reinforced.

First, we recall that when one refers to eigenvectors, one typically is referring to the *right* eigenvectors which arise from $\mathbf{A} \cdot \mathbf{e} = \lambda \mathbf{I} \cdot \mathbf{e}$; if no distinction is made, it can be assumed

that it is the right set that is being discussed. Though it does not arise as often, there are occasions when one requires the *left* eigenvectors which arise from $\overline{\mathbf{e}}^T \cdot \mathbf{A} = \overline{\mathbf{e}}^T \cdot \mathbf{I}\lambda$. Some important properties and definitions involving eigenvalues are listed next:

- If the matrix $\mathbf{A}$ is self-adjoint, it can be shown that it has the same left and right eigenvectors.
- If $\mathbf{A}$ is not self-adjoint, it has different left and right eigenvectors. The eigenvalues are the same for both left and right eigenvectors of the same operator, whether or not the system is self-adjoint.
- The polynomial equation that arises in the eigenvalue problem is the *characteristic equation* of the matrix.
- The *Cayley-Hamilton*[2] theorem states that a matrix satisfies its own characteristic equation.
- If a matrix is triangular, then its eigenvalues are its diagonal terms.
- Eigenvalues of $\mathbf{A} \cdot \mathbf{A} = \mathbf{A}^2$ are the square of the eigenvalues of $\mathbf{A}$.
- Every eigenvector of $\mathbf{A}$ is also an eigenvector of $\mathbf{A}^2$.
- A matrix $\mathbf{A}$ has *spectral radius*, $\rho(\mathbf{A})$, defined as the largest of the absolute values of its eigenvalues:
$$\rho(\mathbf{A}) \equiv \max_n(|\lambda_n|). \tag{8.106}$$
- Recall from Eq. (7.301) that a matrix $\mathbf{A}$ has a *spectral norm*, $||\mathbf{A}||_2$ where
$$||\mathbf{A}||_2 = \sqrt{\max_i(\kappa_i)}, \tag{8.107}$$
where for real valued $\mathbf{A}$, κ_i is an eigenvalue of $\mathbf{A}^T \cdot \mathbf{A}$. Note in general $\rho(\mathbf{A}) \neq ||\mathbf{A}||_2$.
- If $\mathbf{A}$ is self-adjoint, $\rho(\mathbf{A}) = ||\mathbf{A}||_2$.
- In general, Gelfand's[3] formula holds
$$\rho(\mathbf{A}) = \lim_{k\to\infty} ||\mathbf{A}^k||^{1/k}. \tag{8.108}$$
The norm here holds for any matrix norm, including our spectral norm.
- The *trace* of a matrix is the sum of the terms on the leading diagonal.

[2]after Arthur Cayley, 1821-1895, English mathematician, and William Rowan Hamilton, 1805-1865, Anglo-Irish mathematician.

[3]Israel Gelfand, 1913-2009, Soviet mathematician.

- The trace of a $N \times N$ matrix is the sum of its N eigenvalues.
- The product of the N eigenvalues is the determinant of the matrix.

Example 8.13

Demonstrate the theorems and definitions just described for

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix}. \tag{8.109}$$

The characteristic equation is

$$\lambda^3 - 6\lambda^2 + 11\lambda - 6 = 0. \tag{8.110}$$

The Cayley-Hamilton theorem is easily verified by direct substitution:

$$\mathbf{A}^3 - 6\mathbf{A}^2 + 11\mathbf{A} - 6\mathbf{I} = 0, \tag{8.111}$$

$$\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix}\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix}\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} - 6\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix}\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} + 11\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} - 6\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{8.112}$$

$$\begin{pmatrix} -30 & 19 & -38 \\ -10 & 13 & -24 \\ 52 & -26 & 53 \end{pmatrix} + \begin{pmatrix} 36 & -30 & 60 \\ -12 & -18 & 24 \\ -96 & 48 & -102 \end{pmatrix} + \begin{pmatrix} 0 & 11 & -22 \\ 22 & 11 & 0 \\ 44 & -22 & 55 \end{pmatrix} + \begin{pmatrix} -6 & 0 & 0 \\ 0 & -6 & 0 \\ 0 & 0 & -6 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{8.113}$$

Considering the traditional right eigenvalue problem, $\mathbf{A}\cdot\mathbf{e} = \lambda\mathbf{I}\cdot\mathbf{e}$, it is easily shown that the eigenvalues and (right) eigenvectors for this system are

$$\lambda_1 = 1, \qquad \mathbf{e}_1 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}, \tag{8.114}$$

$$\lambda_2 = 2, \qquad \mathbf{e}_2 = \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \end{pmatrix}, \tag{8.115}$$

$$\lambda_3 = 3, \qquad \mathbf{e}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}. \tag{8.116}$$

One notes that while the eigenvectors do form a basis in $\mathbb{R}^3$, that they are not orthogonal; this is a consequence of the matrix not being self-adjoint (or more specifically asymmetric). The spectral radius is $\rho(\mathbf{A}) = 3$. Now

$$\mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A} = \begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} \begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} = \begin{pmatrix} -6 & 5 & -10 \\ 2 & 3 & -4 \\ 16 & -8 & 17 \end{pmatrix}. \tag{8.117}$$

It is easily shown that the eigenvalues for $\mathbf{A}^2$ are $1, 4, 9$, precisely the squares of the eigenvalues of $\mathbf{A}$.

The trace is

$$\text{tr}(\mathbf{A}) = 0 + 1 + 5 = 6. \tag{8.118}$$

Note this is the equal to the sum of the eigenvalues

$$\sum_{i=1}^{3} \lambda_i = 1 + 2 + 3 = 6. \tag{8.119}$$

Note also that

$$\det \mathbf{A} = 6 = \lambda_1 \lambda_2 \lambda_3 = (1)(2)(3) = 6. \tag{8.120}$$

Note that since all the eigenvalues are positive, $\mathbf{A}$ is a positive matrix. It is not positive definite. Note for instance if $\mathbf{x} = (-1, 1, 1)^T$, that $\mathbf{x}^T \cdot \mathbf{A} \cdot \mathbf{x} = -1$. We might ask about the positive definiteness of the symmetric part of $\mathbf{A}$, $\mathbf{A}_s = (\mathbf{A} + \mathbf{A}^T)/2$:

$$\mathbf{A}_s = \begin{pmatrix} 0 & \frac{3}{2} & 1 \\ \frac{3}{2} & 1 & -1 \\ 1 & -1 & 5 \end{pmatrix}. \tag{8.121}$$

In this case $\mathbf{A}_s$ has real eigenvalues, both positive and negative, $\lambda_1 = 5.32, \lambda_2 = -1.39, \lambda_3 = 2.07$. Because of the presence of a negative eigenvalue in the symmetric part of $\mathbf{A}$, we can conclude that both $\mathbf{A}$ and $\mathbf{A}_s$ are not positive definite.

We also note that for real-valued problems $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{A} \in \mathbb{R}^{N \times N}$, the antisymmetric part of a matrix can never be positive definite by the following argument. We can say $\mathbf{x}^T \cdot \mathbf{A} \cdot \mathbf{x} = \mathbf{x}^T \cdot (\mathbf{A}_s + \mathbf{A}_a) \cdot \mathbf{x}$. Then one has $\mathbf{x}^T \cdot \mathbf{A}_a \cdot \mathbf{x} = 0$ for all $\mathbf{x}$ because the tensor inner product of the real antisymmetric $\mathbf{A}_a$ with the symmetric $\mathbf{x}^T$ and $\mathbf{x}$ is identically zero. So to test the positive definiteness of a real $\mathbf{A}$, it suffices to consider the positive definiteness of its symmetric part: $\mathbf{x}^T \cdot \mathbf{A}_s \cdot \mathbf{x} \geq 0$.

For complex-valued problems, $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{A} \in \mathbb{C}^{N \times N}$, it is not quite as simple. Recalling that the eigenvalues of an antisymmetric matrix $\mathbf{A}_a$ are purely imaginary, we have, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}_a$, that $\overline{\mathbf{x}}^T \cdot \mathbf{A}_a \cdot \mathbf{x} = \overline{\mathbf{x}}^T \cdot (\lambda)\mathbf{x} = \overline{\mathbf{x}}^T \cdot (i\lambda_I)\mathbf{x} = i\lambda_I \overline{\mathbf{x}}^T \cdot \mathbf{x} = i\lambda_I ||\mathbf{x}||_2^2$, where $\lambda_I \in \mathbb{R}^1$. Hence whenever the vector $\mathbf{x}$ is an eigenvector of $\mathbf{A}_a$, the quantity $\overline{\mathbf{x}}^T \cdot \mathbf{A}_a \cdot \mathbf{x}$ is a pure imaginary number.

We can also easily solve the left eigenvalue problem, $\overline{\mathbf{e}}_L^T \cdot \mathbf{A} = \lambda \overline{\mathbf{e}}_L^T \cdot \mathbf{I}$:

$$\lambda_1 = 1, \qquad \mathbf{e}_{(L1)} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \tag{8.122}$$

$$\lambda_2 = 2, \qquad \mathbf{e}_{L2} = \begin{pmatrix} -3 \\ 1 \\ -2 \end{pmatrix}, \tag{8.123}$$

$$\lambda_3 = 3, \qquad \mathbf{e}_{L3} = \begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix}. \tag{8.124}$$

We see eigenvalues are the same, but the left and right eigenvectors are different.

We find $||\mathbf{A}||_2$ by considering eigenvalues of $\mathbf{A}^T \cdot \mathbf{A}$, the real variable version of that described in Eq. (7.301):

$$\mathbf{A}^T \cdot \mathbf{A} = \begin{pmatrix} 0 & 2 & 4 \\ 1 & 1 & -2 \\ -2 & 0 & 5 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix}, \tag{8.125}$$

$$= \begin{pmatrix} 20 & -6 & 20 \\ -6 & 6 & -12 \\ 20 & -12 & 29 \end{pmatrix}. \tag{8.126}$$

This matrix has eigenvalues $\kappa = 49.017, 5.858, 0.125$. The spectral norm is the square root of the largest, giving

$$||\mathbf{A}||_2 = \sqrt{49.017} = 7.00122. \tag{8.127}$$

The eigenvector of $\mathbf{A}^T \cdot \mathbf{A}$ corresponding to $\kappa = 49.017$ is $\hat{\mathbf{e}}_1 = (0.5829, -0.2927, 0.7579)^T$. When we compute the quantity associated with the norm of an operator, we find this vector maps to the norm:

$$\frac{||\mathbf{A} \cdot \hat{\mathbf{e}}_1||_2}{||\hat{\mathbf{e}}_1||_2} = \frac{\left|\left|\begin{pmatrix} 0 & 1 & -2 \\ 2 & 1 & 0 \\ 4 & -2 & 5 \end{pmatrix} \cdot \begin{pmatrix} 0.5829 \\ -0.2927 \\ 0.7579 \end{pmatrix}\right|\right|_2}{\left|\left|\begin{pmatrix} 0.582944 \\ -0.292744 \\ 0.757943 \end{pmatrix}\right|\right|_2} = \frac{\left|\left|\begin{pmatrix} -1.80863 \\ 0.873144 \\ 6.70698 \end{pmatrix}\right|\right|_2}{1} = 7.00122. \tag{8.128}$$

Had we chosen the eigenvector associated with the eigenvalue of largest magnitude, $\mathbf{e}_3 = (-1, -1, 1)^T$, we would have found $||\mathbf{A} \cdot \mathbf{e}_3||_2 / ||\mathbf{e}_3||_2 = 3$, the spectral radius. Obviously, this is not the maximum of this operation and thus cannot be a norm.

We can easily verify Gelfand's theorem by direct calculation of $||\mathbf{A}^k||_2^{1/k}$ for various k. We find the following.

k	$\|\|\mathbf{A}^k\|\|_2^{1/k}$
1	7.00122
2	5.27011
3	4.61257
4	4.26334
5	4.03796
10	3.52993
100	3.04984
1000	3.00494
∞	3

As $k \to \infty$, $||\mathbf{A}^k||_2^{1/k}$ approaches the spectral radius $\rho(\mathbf{A}) = 3$.

8.3.2 Generalized eigenvalues and eigenvectors in the second sense

On p. 288, we studied generalized eigenvectors in the first sense. Here we consider a distinct problem which leads to *generalized eigenvalues and eigenvectors in the second sense.*

Consider the problem

$$\mathbf{A}\cdot\mathbf{e} = \lambda\mathbf{B}\cdot\mathbf{e}, \tag{8.129}$$

where $\mathbf{A}$ and $\mathbf{B}$ are square matrices, possibly singular, of dimension $N\times N$, $\mathbf{e}$ is a generalized eigenvector in the second sense, and λ is a generalized eigenvalue. If $\mathbf{B}$ were not singular, we could form $(\mathbf{B}^{-1}\cdot\mathbf{A})\cdot\mathbf{e} = \lambda\mathbf{I}\cdot\mathbf{e}$, which amounts to an ordinary eigenvalue problem. But let us assume that the inverses do not exist. Then Eq. (8.129) can be re-cast as

$$(\mathbf{A}-\lambda\mathbf{B})\cdot\mathbf{e} = \mathbf{0}. \tag{8.130}$$

For non-trivial solutions, we simply require

$$\det(\mathbf{A}-\lambda\mathbf{B}) = 0, \tag{8.131}$$

and analyze in a similar manner.

Example 8.14

Find the generalized eigenvalues and eigenvectors in the second sense for

$$\underbrace{\begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix}}_{\mathbf{A}}\cdot\mathbf{e} = \lambda\underbrace{\begin{pmatrix}1 & 0\\ 1 & 0\end{pmatrix}}_{\mathbf{B}}\cdot\mathbf{e}. \tag{8.132}$$

Here $\mathbf{B}$ is obviously singular. We rewrite as

$$\begin{pmatrix}1-\lambda & 2\\ 2-\lambda & 1\end{pmatrix}\cdot\begin{pmatrix}e_1\\ e_2\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}. \tag{8.133}$$

For a non-trivial solution, we require

$$\begin{vmatrix}1-\lambda & 2\\ 2-\lambda & 1\end{vmatrix} = 0, \tag{8.134}$$

which gives a generalized eigenvalue of

$$\begin{aligned} 1-\lambda-2(2-\lambda) &= 0, && (8.135)\\ 1-\lambda-4+2\lambda &= 0, && (8.136)\\ \lambda &= 3. && (8.137)\end{aligned}$$

For $\mathbf{e}$, we require

$$\begin{aligned}\begin{pmatrix}1-3 & 2\\ 2-3 & 1\end{pmatrix}\cdot\begin{pmatrix}e_1\\ e_2\end{pmatrix} &= \begin{pmatrix}0\\ 0\end{pmatrix}, && (8.138)\\ \begin{pmatrix}-2 & 2\\ -1 & 1\end{pmatrix}\cdot\begin{pmatrix}e_1\\ e_2\end{pmatrix} &= \begin{pmatrix}0\\ 0\end{pmatrix}. && (8.139)\end{aligned}$$

By inspection, the generalized eigenvector in the second sense

$$\mathbf{e} = \begin{pmatrix}e_1\\ e_2\end{pmatrix} = \alpha\begin{pmatrix}1\\ 1\end{pmatrix}, \tag{8.140}$$

satisfies Eq. (8.132) when $\lambda = 3$, and α is any scalar.

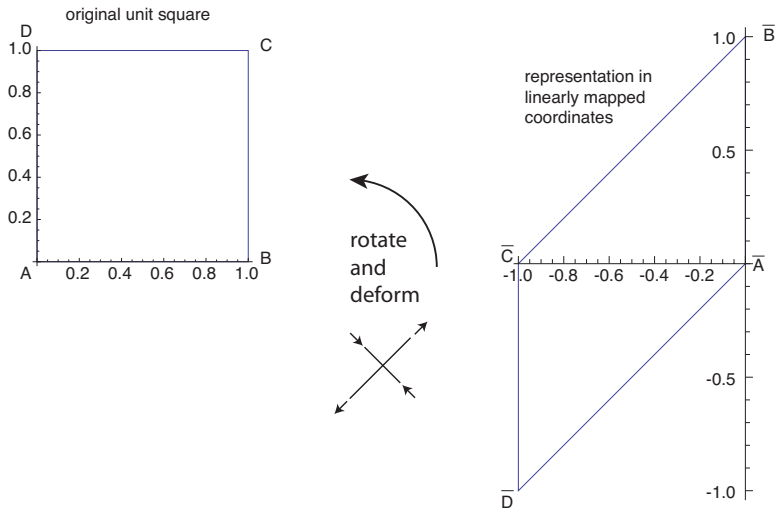

Figure 8.3: Unit square transforming via stretching and rotation under a linear area- and orientation-preserving alibi mapping.

8.4 Matrices as linear mappings

By considering a matrix as an operator which effects a linear mapping and applying it to a specific geometry, one can better envision the characteristics of the matrix. This is demonstrated in the following example.

Example 8.15

Consider how the matrix

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \tag{8.141}$$

acts on vectors $\mathbf{x}$, including those that form a unit square with vertices as $A : (0,0), B : (1,0), C : (1,1), D : (0,1)$.

The original square has area of $\mathcal{A} = 1$. Each of the vertices map under the linear homogeneous transformation to

$$\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{\overline{A}}, \quad \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{B} = \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\overline{B}}, \tag{8.142}$$

$$\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{C} = \underbrace{\begin{pmatrix} -1 \\ 0 \end{pmatrix}}_{\overline{C}}, \quad \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{D} = \underbrace{\begin{pmatrix} -1 \\ -1 \end{pmatrix}}_{\overline{D}}. \tag{8.143}$$

In the mapped space, the square has transformed to a parallelogram. This is plotted in Fig. 8.3. Here, the alibi approach to the mapping is clearly evident. We keep the coordinate axes fixed in Fig. 8.3 and rotate and stretch the vectors, instead of keeping the vectors fixed and rotating the axes, as would have been done in an alias transformation. Now we have

$$\det \mathbf{A} = (0)(-1) - (1)(-1) = 1. \tag{8.144}$$

Thus, the mapping is both orientation- and area-preserving. The orientation-preserving feature is obvious by inspecting the locations of the points A, B, C, and D in both configurations shown in Fig. 8.3. We easily calculate the area in the mapped space by combining the areas of two triangles which form the parallelogram:

$$\mathcal{A} = \frac{1}{2}(1)(1) + \frac{1}{2}(1)(1) = 1. \tag{8.145}$$

The eigenvalues of $\mathbf{A}$ are $-(1/2) \pm \sqrt{3/2}$, both of which have magnitude of unity. Thus, the spectral radius $\rho(\mathbf{A}) = 1$. However, the spectral norm of $\mathbf{A}$ is non-unity, because

$$\mathbf{A}^T \cdot \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \tag{8.146}$$

which has eigenvalues

$$\kappa = \frac{1}{2}(3 \pm \sqrt{5}). \tag{8.147}$$

The spectral norm is the square root of the maximum eigenvalue of $\mathbf{A}^T \cdot \mathbf{A}$, which is

$$||\mathbf{A}||_2 = \sqrt{\frac{1}{2}(3 + \sqrt{5})} = 1.61803. \tag{8.148}$$

It will later be shown, Sec. 8.8.4, that the action of $\mathbf{A}$ on the unit square can be decomposed into a deformation and a rotation. Both are evident in Fig. 8.3.

8.5 Complex matrices

If $\mathbf{x}$ and $\mathbf{y}$ are complex vectors, we know that their inner product involves the conjugate transpose. The conjugate transpose operation occurs so often we give it a name, the Hermitian transpose, and denote it by a superscript H. Thus, we define the inner product as

$$<\mathbf{x}, \mathbf{y}> = \overline{\mathbf{x}}^T \cdot \mathbf{y} = \mathbf{x}^H \cdot \mathbf{y}. \tag{8.149}$$

Then the norm is given by

$$||\mathbf{x}||_2 = +\sqrt{\mathbf{x}^H \cdot \mathbf{x}}. \tag{8.150}$$

Example 8.16

If

$$\mathbf{x} = \begin{pmatrix} 1+i \\ 3-2i \\ 2 \\ -3i \end{pmatrix}, \tag{8.151}$$

find $||\mathbf{x}||_2$.

$$||\mathbf{x}||_2 = +\sqrt{\mathbf{x}^H \cdot \mathbf{x}} = +\sqrt{(1-i, 3+2i, 2, +3i)\begin{pmatrix} 1+i \\ 3-2i \\ 2 \\ -3i \end{pmatrix}} = +\sqrt{2+13+4+9} = 2\sqrt{7}. \tag{8.152}$$

Example 8.17

If

$$\mathbf{x} = \begin{pmatrix} 1+i \\ -2+3i \\ 2-i \end{pmatrix}, \tag{8.153}$$

$$\mathbf{y} = \begin{pmatrix} 3 \\ 4-2i \\ 3+3i \end{pmatrix}, \tag{8.154}$$

find $<\mathbf{x}, \mathbf{y}>$.

$$\begin{aligned} <\mathbf{x}, \mathbf{y}> &= \mathbf{x}^H \cdot \mathbf{y}, & (8.155)\\ &= (1-i, -2-3i, 2+i)\begin{pmatrix} 3 \\ 4-2i \\ 3+3i \end{pmatrix}, & (8.156)\\ &= (3-3i) + (-14-8i) + (3+9i), & (8.157)\\ &= -8-2i. & (8.158) \end{aligned}$$

Likewise, the conjugate or Hermitian transpose of a matrix $\mathbf{A}$ is $\mathbf{A}^H$, given by the transpose of the matrix with each element being replaced by its conjugate:

$$\mathbf{A}^H = \bar{\mathbf{A}}^T. \tag{8.159}$$

As the Hermitian transpose is the adjoint operator corresponding to a given complex matrix, we can apply an earlier proved theorem for linear operators, Sec. 7.4.4, to deduce that the eigenvalues of a complex matrix are the complex conjugates of the Hermitian transpose of that matrix.

The Hermitian transpose is distinguished from a matrix which is Hermitian as follows. A Hermitian matrix is one which is equal to its conjugate transpose. So a matrix which equals its Hermitian transpose is Hermitian. A matrix which does not equal its Hermitian transpose is non-Hermitian. A skew-Hermitian matrix is the negative of its Hermitian transpose. A Hermitian matrix is self-adjoint.

Properties:

- $\mathbf{x}^H \cdot \mathbf{A} \cdot \mathbf{x}$ is real if $\mathbf{A}$ is Hermitian.
- The eigenvalues of a Hermitian matrix are real.
- The eigenvectors of a Hermitian matrix that correspond to different eigenvalues, are orthogonal to each other.
- The determinant of a Hermitian matrix is real.
- The spectral radius of a Hermitian matrix is equal to its spectral norm, $\rho(\mathbf{A}) = ||\mathbf{A}||_2$.
- If $\mathbf{A}$ is skew-Hermitian, then $i\mathbf{A}$ is Hermitian, and vice-versa.

Note the diagonal elements of a Hermitian matrix must be real as they must be unchanged by conjugation.

Example 8.18

Consider $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, where $\mathbf{A} : \mathbb{C}^3 \rightarrow \mathbb{C}^3$ with $\mathbf{A}$ the Hermitian matrix and $\mathbf{x}$ the complex vector:

$$\mathbf{A} = \begin{pmatrix} 1 & 2-i & 3 \\ 2+i & -3 & 2i \\ 3 & -2i & 4 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} 3+2i \\ -1 \\ 2-i \end{pmatrix}. \tag{8.160}$$

First, we have

$$\mathbf{b} = \mathbf{A} \cdot \mathbf{x} = \begin{pmatrix} 1 & 2-i & 3 \\ 2+i & -3 & 2i \\ 3 & -2i & 4 \end{pmatrix} \begin{pmatrix} 3+2i \\ -1 \\ 2-i \end{pmatrix} = \begin{pmatrix} 7 \\ 9+11i \\ 17+4i \end{pmatrix}. \tag{8.161}$$

Now, demonstrate that the properties of Hermitian matrices hold for this case. First

$$\mathbf{x}^H \cdot \mathbf{A} \cdot \mathbf{x} = \begin{pmatrix} 3-2i & -1 & 2+i \end{pmatrix} \begin{pmatrix} 1 & 2-i & 3 \\ 2+i & -3 & 2i \\ 3 & -2i & 4 \end{pmatrix} \begin{pmatrix} 3+2i \\ -1 \\ 2-i \end{pmatrix} = 42 \in \mathbb{R}^1. \tag{8.162}$$

The eigenvalues and (right, same as left here) eigenvectors are

$$\lambda_1 = 6.51907, \qquad \mathbf{e}_1 = \begin{pmatrix} 0.525248 \\ 0.132451 + 0.223964i \\ 0.803339 - 0.105159i \end{pmatrix}, \tag{8.163}$$

$$\lambda_2 = -0.104237, \qquad \mathbf{e}_2 = \begin{pmatrix} -0.745909 \\ -0.385446 + 0.0890195i \\ 0.501844 - 0.187828i \end{pmatrix}, \tag{8.164}$$

$$\lambda_3 = -4.41484, \qquad \mathbf{e}_3 = \begin{pmatrix} 0.409554 \\ -0.871868 - 0.125103i \\ -0.116278 - 0.207222i \end{pmatrix}. \tag{8.165}$$

By inspection $\rho(\mathbf{A}) = 6.51907$. Because $\mathbf{A}$ is Hermitian, we also have $||\mathbf{A}||_2 = \rho(\mathbf{A}) = 6.51907$. We find this by first finding the eigenvalues of $\mathbf{A}^H \cdot \mathbf{A}$, which are 42.4983, 19.4908, and 0.010865. The square roots of these are 6.51907, 4.41484, and 0.104237; the spectral norm is the maximum, 6.51907.

Check for orthogonality between two of the eigenvectors, e.g. $\mathbf{e}_1, \mathbf{e}_2$:

$$\begin{aligned} <\mathbf{e}_1, \mathbf{e}_2> &= \mathbf{e}_1^H \cdot \mathbf{e}_2, && (8.166)\\ &= \begin{pmatrix} 0.525248 & 0.132451 - 0.223964i & 0.803339 + 0.105159i \end{pmatrix} \cdot \begin{pmatrix} -0.745909 \\ -0.385446 + 0.0890195i \\ 0.501844 - 0.187828i \end{pmatrix}, && (8.167)\\ &= 0 + 0i. && (8.168) \end{aligned}$$

The same holds for other eigenvectors. It can then be shown that

$$\det \mathbf{A} = 3, \tag{8.169}$$

which is also equal to the product of the eigenvalues. This also tells us that $\mathbf{A}$ is not volume-preserving, but it is orientation-preserving.

Lastly,

$$i\mathbf{A} = \begin{pmatrix} i & 1+2i & 3i \\ -1+2i & -3i & -2 \\ 3i & 2 & 4i \end{pmatrix}, \tag{8.170}$$

is skew-symmetric. It is easily shown the eigenvalues of $i\mathbf{A}$ are

$$\lambda_1 = 6.51907i, \qquad \lambda_2 = -0.104237i, \qquad \lambda_3 = -4.41484i. \tag{8.171}$$

Note the eigenvalues of this matrix are just those of the previous multiplied by i.

8.6 Orthogonal and unitary matrices

8.6.1 Orthogonal matrices

Expanding on a topic introduced on p. 24, discussed on p. 183 and briefly discussed on p. 287, a set of N N-dimensional real orthonormal vectors $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\}$ can be formed into an *orthogonal* matrix

$$\mathbf{Q} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_N \\ \vdots & \vdots & & \vdots \end{pmatrix}. \tag{8.172}$$

Properties of orthogonal matrices include

1. $\mathbf{Q}^T = \mathbf{Q}^{-1}$, and both are orthogonal.
2. $\mathbf{Q}^T \cdot \mathbf{Q} = \mathbf{Q} \cdot \mathbf{Q}^T = \mathbf{I}$.
3. $||\mathbf{Q}||_2 = 1$, when the domain and range of $\mathbf{Q}$ are in Hilbert spaces.
4. $||\mathbf{Q} \cdot \mathbf{x}||_2 = ||\mathbf{x}||_2$, where $\mathbf{x}$ is a vector.

5. $(\mathbf{Q}\cdot\mathbf{x})^T\cdot(\mathbf{Q}\cdot\mathbf{y}) = \mathbf{x}^T\cdot\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are real vectors.

6. Eigenvalues of $\mathbf{Q}$ have $|\lambda_i| = 1, \lambda_i \in \mathbb{C}^1$, thus, $\rho(\mathbf{Q}) = 1$.

7. $|\det \mathbf{Q}| = 1$.

Geometrically, an orthogonal matrix is an operator which transforms but does not stretch a vector. For an orthogonal matrix to be a rotation, which is orientation-preserving, we must have $\det \mathbf{Q} = 1$. Rotation matrices, reflection matrices, and permutation matrices $\mathbf{P}$ are all orthogonal matrices. Recall that permutation matrices can also be reflection or rotation matrices.

Example 8.19

Find the orthogonal matrix corresponding to

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{8.173}$$

The normalized eigenvectors are $(1/\sqrt{2}, 1/\sqrt{2})^T$ and $(-1/\sqrt{2}, 1/\sqrt{2})^T$. The orthogonal matrix is thus

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \tag{8.174}$$

In the sense of Eq. (6.54), we can say

$$\mathbf{Q} = \begin{pmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix}, \tag{8.175}$$

and the angle of rotation of the coordinate axes is $\alpha = \pi/4$. We calculate the eigenvalues of $\mathbf{Q}$ to be $\lambda = (1 \pm i)/\sqrt{2}$, which in exponential form becomes $\lambda = \exp(\pm i\pi/4)$, and so we see the rotation angle is embedded within the argument of the polar representation of the eigenvalues. We also see $|\lambda| = 1$. Note that $\mathbf{Q}$ is *not* symmetric. Also note that $\det \mathbf{Q} = 1$, so this orthogonal matrix is also a rotation matrix.

If $\boldsymbol{\xi}$ is an unrotated Cartesian vector, and our transformation to a rotated frame is $\boldsymbol{\xi} = \mathbf{Q}\cdot\mathbf{x}$, so that $\mathbf{x} = \mathbf{Q}^T\cdot\boldsymbol{\xi}$, we see that the Cartesian unit vector $\boldsymbol{\xi} = (1,0)^T$ is represented in the rotated coordinate system by

$$\mathbf{x} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}}_{\mathbf{Q}^T} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}. \tag{8.176}$$

Thus, the counterclockwise rotation of the axes through angle $\alpha = \pi/4$ gives the Cartesian unit vector $(1,0)^T$ a new representation of $(1/\sqrt{2}, -1/\sqrt{2})^T$. We see that the other Cartesian unit vector $\boldsymbol{\xi} = (0,1)^T$ is represented in the rotated coordinate system by

$$\mathbf{x} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}}_{\mathbf{Q}^T} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{8.177}$$

Had $\det \mathbf{Q} = -1$, the transformation would have been non-orientation preserving.

Example 8.20

Analyze the three-dimensional orthogonal matrix

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \end{pmatrix}. \tag{8.178}$$

Direct calculation reveals $||\mathbf{Q}||_2 = 1$, $\det \mathbf{Q} = 1$, and $\mathbf{Q}^T = \mathbf{Q}^{-1}$, so clearly the matrix is a volume- and orientation-preserving rotation matrix. It can also be shown to have a set of eigenvalues and eigenvectors of

$$\lambda_1 = 1, \quad \mathbf{e}_1 = \begin{pmatrix} 0.886452 \\ 0.36718 \\ 0.281747 \end{pmatrix}, \tag{8.179}$$

$$\lambda_2 = \exp(2.9092i), \quad \mathbf{e}_2 = \begin{pmatrix} -0.18406 + 0.27060i \\ -0.076240 - 0.653281i \\ 0.678461 \end{pmatrix}, \tag{8.180}$$

$$\lambda_3 = \exp(-2.9092i), \quad \mathbf{e}_3 = \begin{pmatrix} -0.18406 - 0.27060i \\ -0.076240 + 0.653281i \\ 0.678461 \end{pmatrix}. \tag{8.181}$$

As expected, each eigenvalue has $|\lambda| = 1$. It can be shown that the eigenvector $\mathbf{e}_1$ which is associated with real eigenvalue, $\lambda_1 = 1$, is aligned with the so-called Euler axis, i.e. the axis in three-space about which the rotation occurs. The remaining two eigenvalues are of the form $\exp(\pm\alpha i)$, where α is the angle of rotation about the Euler axis. For this example, we have $\alpha = 2.9092$.

Example 8.21

Consider the composite action of three rotations on a vector $\mathbf{x}$:

$$\mathbf{Q}_1 \cdot \mathbf{Q}_2 \cdot \mathbf{Q}_3 \cdot \mathbf{x} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_1 & -\sin\alpha_1 \\ 0 & \sin\alpha_1 & \cos\alpha_1 \end{pmatrix}}_{\mathbf{Q}_1} \cdot \underbrace{\begin{pmatrix} \cos\alpha_2 & 0 & \sin\alpha_2 \\ 0 & 1 & 0 \\ -\sin\alpha_2 & 0 & \cos\alpha_2 \end{pmatrix}}_{\mathbf{Q}_2} \cdot \underbrace{\begin{pmatrix} \cos\alpha_3 & -\sin\alpha_3 & 0 \\ \sin\alpha_3 & \cos\alpha_3 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{Q}_3} \cdot \mathbf{x}. \tag{8.182}$$

It is easy to verify that $||\mathbf{Q}_1||_2 = ||\mathbf{Q}_2||_2 = ||\mathbf{Q}_3||_2 = 1$, $\det \mathbf{Q}_1 = \det \mathbf{Q}_2 = \det \mathbf{Q}_3 = 1$, so each is a rotation. For $\mathbf{Q}_3$, we find eigenvalues of $\lambda = 1, \cos\alpha_3 \pm i \sin\alpha_3$. These can be rewritten as $\lambda = 1, e^{\pm\alpha_3 i}$. The eigenvector associated with the eigenvalue of 1 is $(0, 0, 1)$. Thus, we can consider $\mathbf{Q}_3$ to effect a rotation of α_3 about the 3-axis. Similarly, $\mathbf{Q}_2$ effects a rotation of α_2 about the 2-axis, and $\mathbf{Q}_1$ effects a rotation of α_1 about the 1-axis.

So the action of the combination of rotations on a vector $\mathbf{x}$ is an initial rotation of α_3 about the 3-axis: $\mathbf{Q}_3 \cdot \mathbf{x}$. This vector is then rotated through α_2 about the 2-axis: $\mathbf{Q}_2 \cdot (\mathbf{Q}_3 \cdot \mathbf{x})$. Finally, there is a rotation through α_1 about the 1-axis: $\mathbf{Q}_1 \cdot (\mathbf{Q}_2 \cdot (\mathbf{Q}_3 \cdot \mathbf{x}))$. This is called a 3-2-1 rotation through

the so-called *Euler angles* of α_3, α_2, and α_1. Note because in general matrix multiplication does not commute, that the result will depend on the order of application of the rotations, e.g.

$$\mathbf{Q}_1 \cdot \mathbf{Q}_2 \cdot \mathbf{Q}_3 \cdot \mathbf{x} \neq \mathbf{Q}_2 \cdot \mathbf{Q}_1 \cdot \mathbf{Q}_3 \cdot \mathbf{x}, \qquad \forall \mathbf{x} \in \mathbb{R}^3, \quad \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3 \in \mathbb{R}^3 \times \mathbb{R}^3. \tag{8.183}$$

In contrast, it is not difficult to show that rotations in two dimensions do commute

$$\mathbf{Q}_1 \cdot \mathbf{Q}_2 \cdot \mathbf{Q}_3 \cdot \mathbf{x} = \mathbf{Q}_2 \cdot \mathbf{Q}_1 \cdot \mathbf{Q}_3 \cdot \mathbf{x}, \qquad \forall \mathbf{x} \in \mathbb{R}^2, \quad \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3 \in \mathbb{R}^2 \times \mathbb{R}^2. \tag{8.184}$$

8.6.2 Unitary matrices

A unitary matrix $\mathbf{U}$ is a complex matrix with orthonormal columns. It is the complex analog of an orthogonal matrix.

Properties of unitary matrices include

- $\mathbf{U}^H = \mathbf{U}^{-1}$, and both are unitary.
- $\mathbf{U}^H \cdot \mathbf{U} = \mathbf{U} \cdot \mathbf{U}^H = \mathbf{I}$.
- $||\mathbf{U}||_2 = 1$, when the domain and range of $\mathbf{U}$ are in Hilbert spaces.
- $||\mathbf{U} \cdot \mathbf{x}||_2 = ||\mathbf{x}||_2$, where $\mathbf{x}$ is a vector.
- $(\mathbf{U} \cdot \mathbf{x})^H \cdot (\mathbf{U} \cdot \mathbf{y}) = \mathbf{x}^H \cdot \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are vectors.
- Eigenvalues of $\mathbf{U}$ have $|\lambda_i| = 1, \lambda_i \in \mathbb{C}^1$, thus, $\rho(\mathbf{U}) = 1$.
- Eigenvectors of $\mathbf{U}$ corresponding to different eigenvalues are orthogonal.
- $|\det \mathbf{U}| = 1$.

If $\det \mathbf{U} = 1$, one is tempted to say the unitary matrix operating on a vector induces a pure rotation in a complex space; however, the notion of an angle of rotation is elusive.

Example 8.22

Consider the unitary matrix

$$\mathbf{U} = \begin{pmatrix} \frac{1+i}{\sqrt{3}} & \frac{1-2i}{\sqrt{15}} \\ \frac{1}{\sqrt{3}} & \frac{1+3i}{\sqrt{15}} \end{pmatrix}. \tag{8.185}$$

The column vectors are easily seen to be normal. They are also orthogonal:

$$\left(\frac{1-i}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right) \begin{pmatrix} \frac{1-2i}{\sqrt{15}} \\ \frac{1+3i}{\sqrt{15}} \end{pmatrix} = 0 + 0i. \tag{8.186}$$

The matrix itself is not Hermitian. Still, its Hermitian transpose exists:

$$\mathbf{U}^H = \begin{pmatrix} \frac{1-i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1+2i}{\sqrt{15}} & \frac{1-3i}{\sqrt{15}} \end{pmatrix}. \tag{8.187}$$

It is then easily verified that

$$\mathbf{U}^{-1} = \mathbf{U}^H, \tag{8.188}$$

$$\mathbf{U} \cdot \mathbf{U}^H = \mathbf{U}^H \cdot \mathbf{U} = \mathbf{I}. \tag{8.189}$$

The eigensystem is

$$\lambda_1 = -0.0986232 + 0.995125i, \qquad \mathbf{e}_1 = \begin{pmatrix} 0.688191 - 0.425325i \\ 0.587785 \end{pmatrix}, \tag{8.190}$$

$$\lambda_2 = 0.934172 + 0.356822i, \qquad \mathbf{e}_2 = \begin{pmatrix} -0.306358 + 0.501633i \\ -0.721676 - 0.36564i \end{pmatrix}. \tag{8.191}$$

It is easily verified that the eigenvectors are orthogonal and the eigenvalues have magnitude of one. We find $\det \mathbf{U} = (1+2i)/\sqrt{5}$, which yields $|\det \mathbf{U}| = 1$. Also, $||\mathbf{U}||_2 = 1$.

8.7 Discrete Fourier transforms

It is a common practice in experimental and theoretical science and engineering to decompose a function or a signal into its Fourier modes. The amplitudes of these modes is often a useful description of the function. A Fourier transform is a linear integral operator which operates on continuous functions and yields results from which amplitudes of each frequency component can be determined. Its discrete analog is the Discrete Fourier transform (DFT). The DFT is a matrix which operates on a vector of data to yield a vector of transformed data. There exists a popular, albeit complicated, algorithm to compute the DFT, known as the Fast Fourier Transform (FFT). This will not be studied here; instead, a simpler and slower method is presented, which will be informally known as a Slow Fourier Transform (SFT). This discussion will simply present the algorithm for the SFT and demonstrate its use by example.

The Fourier transform (FT) $Y(\kappa)$ of a function $y(x)$ is defined as

$$Y(\kappa) = \mathcal{Y}[y(x)] = \int_{-\infty}^{\infty} y(x) e^{-(2\pi i)\kappa x} \, dx, \tag{8.192}$$

and the inverse FT is defined as

$$y(x) = \mathcal{Y}^{-1}[Y(\kappa)] = \int_{-\infty}^{\infty} Y(\kappa) e^{(2\pi i)\kappa x} \, d\kappa. \tag{8.193}$$

Here κ is the wavenumber, and is the reciprocal of the wavelength. The FT has a discrete analog. The connection between the two is often not transparent in the literature. With some

effort a connection *can* be made at the expense of diverging from one school's notation to the other's. Here, we will be satisfied with a form which demonstrates the analogs between the continuous and discrete transform, but will not be *completely* linked. To make the connection, one can construct a discrete approximation to the integral of the FT, and with some effort, arrive at an equivalent result.

For the DFT, consider a function $y(x),\ x \in [x_{min}, x_{max}]$, $x \in \mathbb{R}^1, y \in \mathbb{R}^1$. Now discretize the domain into N uniformly distributed points so that every x_j is mapped to a y_j for $j = 0, \ldots, N-1$. Here we comply with the traditional, yet idiosyncratic, limits on j which are found in many texts on DFT. This offsets standard vector and matrix numbering schemes by one, and so care must be exercised in implementing these algorithms with common software. We seek a discrete analog of the continuous Fourier transformation of the form

$$y_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} c_k \exp\left((2\pi i)k\left(\frac{N-1}{N}\right)\left(\frac{x_j - x_{min}}{x_{max} - x_{min}}\right)\right), \qquad j = 0, \ldots, N-1. \quad (8.194)$$

Here k plays the role of κ, and c_k plays the role of $Y(\kappa)$. For uniformly spaced x_j, one has

$$j = (N-1)\left(\frac{x_j - x_{min}}{x_{max} - x_{min}}\right), \quad (8.195)$$

so that we then seek

$$y_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} c_k \exp\left((2\pi i)\frac{kj}{N}\right), \qquad j = 0, \ldots, N-1. \quad (8.196)$$

Now consider the equation

$$z^N = 1, \qquad z \in \mathbb{C}^1. \quad (8.197)$$

This equation has N distinct roots

$$z = e^{2\pi i \frac{j}{N}}, \qquad j = 0, \ldots, N-1, \quad (8.198)$$

Taking for convenience

$$w \equiv e^{2\pi i/N}, \quad (8.199)$$

one sees that the N roots are also described by $w^0, w^1, w^2, \ldots, w^{N-1}$. Now define the following matrix

$$\mathbf{F} = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}. \quad (8.200)$$

It is easy to demonstrate for arbitrary N that $\mathbf{F}$ is unitary, that is

$$\mathbf{F}^H \cdot \mathbf{F} = \mathbf{I}. \quad (8.201)$$

Since $\mathbf{F}$ is unitary, it is immediately known that $\mathbf{F}^H = \mathbf{F}^{-1}$, that $||\mathbf{F}||_2 = 1$, that the eigenvalues of $\mathbf{F}$ have magnitude of unity, and that the column vectors of $\mathbf{F}$ are orthonormal. Note that $\mathbf{F}$ is not Hermitian. Also note that many texts omit the factor $1/\sqrt{N}$ in the definition of $\mathbf{F}$; this is not a major problem, but does render $\mathbf{F}$ to be non-unitary.

Now given a vector $\mathbf{y} = y_j,\ j = 0, \ldots, N-1$, the DFT is defined as the following mapping

$$\mathbf{c} = \mathbf{F}^H \cdot \mathbf{y}. \tag{8.202}$$

The inverse transform is trivial due to the unitary nature of $\mathbf{F}$:

$$\begin{aligned} \mathbf{F} \cdot \mathbf{c} &= \mathbf{F} \cdot \mathbf{F}^H \cdot \mathbf{y}, & (8.203)\\ \mathbf{F} \cdot \mathbf{c} &= \mathbf{F} \cdot \mathbf{F}^{-1} \cdot \mathbf{y}, & (8.204)\\ \mathbf{F} \cdot \mathbf{c} &= \mathbf{I} \cdot \mathbf{y}, & (8.205)\\ \mathbf{y} &= \mathbf{F} \cdot \mathbf{c}. & (8.206) \end{aligned}$$

Because our $\mathbf{F}$ is unitary, it preserves the length of vectors. Thus, it induces a Parseval's equation

$$||\mathbf{y}||_2 = ||\mathbf{c}||_2. \tag{8.207}$$

Example 8.23

Consider a five term DFT of the function

$$y = x^2, \qquad x \in [0, 4]. \tag{8.208}$$

Take then for $N = 5$, a set of uniformly distributed points in the domain and their image in the range:

$$\begin{aligned} x_0 = 0, \quad x_1 = 1, \quad x_2 = 2, \quad x_3 = 3, \quad x_4 = 4, & \qquad (8.209)\\ y_0 = 0, \quad y_1 = 1, \quad y_2 = 4, \quad y_3 = 9, \quad y_4 = 16. & \qquad (8.210) \end{aligned}$$

Now for $N = 5$, one has

$$w = e^{2\pi i/5} = \underbrace{\left(\frac{1}{4}(-1+\sqrt{5})\right)}_{=\Re(w)} + \underbrace{\left(\frac{1}{2}\sqrt{\frac{1}{2}(5+\sqrt{5})}\right)}_{=\Im(w)} i = 0.309 + 0.951i. \tag{8.211}$$

The five distinct roots of $z^5 = 1$ are

$$\begin{aligned} z^{(0)} &= w^0 = 1, & (8.212)\\ z^{(1)} &= w^1 = 0.309 + 0.951i, & (8.213)\\ z^{(2)} &= w^2 = -0.809 + 0.588i, & (8.214)\\ z^{(3)} &= w^3 = -0.809 - 0.588i, & (8.215)\\ z^{(4)} &= w^4 = 0.309 - 0.951i. & (8.216) \end{aligned}$$

The matrix $\mathbf{F}$ is then

$$\mathbf{F} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 \\ 1 & w^2 & w^4 & w^6 & w^8 \\ 1 & w^3 & w^6 & w^9 & w^{12} \\ 1 & w^4 & w^8 & w^{12} & w^{16} \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 \\ 1 & w^2 & w^4 & w^1 & w^3 \\ 1 & w^3 & w^1 & w^4 & w^2 \\ 1 & w^4 & w^3 & w^2 & w^1 \end{pmatrix}, \tag{8.217}$$

$$= \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0.309+0.951i & -0.809+0.588i & -0.809-0.588i & 0.309-0.951i \\ 1 & -0.809+0.588i & 0.309-0.951i & 0.309+0.951i & -0.809-0.588i \\ 1 & -0.809-0.588i & 0.309+0.951i & 0.309-0.951i & -0.809+0.588i \\ 1 & 0.309-0.951i & -0.809-0.588i & -0.809+0.588i & 0.309+0.951i \end{pmatrix}. \tag{8.218}$$

Now $\mathbf{c} = \mathbf{F}^H \cdot \mathbf{y}$, so

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0.309-0.951i & -0.809-0.588i & -0.809+0.588i & 0.309+0.951i \\ 1 & -0.809-0.588i & 0.309+0.951i & 0.309-0.951i & -0.809+0.588i \\ 1 & -0.809+0.588i & 0.309-0.951i & 0.309+0.951i & -0.809-0.588i \\ 1 & 0.309+0.951i & -0.809+0.588i & -0.809-0.588i & 0.309-0.951i \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \end{pmatrix},$$

$$= \begin{pmatrix} 13.416 \\ -2.354+7.694i \\ -4.354+1.816i \\ -4.354-1.816i \\ -2.354-7.694i \end{pmatrix}. \tag{8.219}$$

Now one is often interested in the magnitude of the components of $\mathbf{c}$, which gives a measure of the so-called energy associated with each Fourier mode. So one calculates a vector of the magnitude of each component as

$$\begin{pmatrix} \sqrt{c_0\overline{c_0}} \\ \sqrt{c_1\overline{c_1}} \\ \sqrt{c_2\overline{c_2}} \\ \sqrt{c_3\overline{c_3}} \\ \sqrt{c_4\overline{c_4}} \end{pmatrix} = \begin{pmatrix} |c_0| \\ |c_1| \\ |c_2| \\ |c_3| \\ |c_4| \end{pmatrix} = \begin{pmatrix} 13.4164 \\ 8.0463 \\ 4.7178 \\ 4.7178 \\ 8.0463 \end{pmatrix}. \tag{8.220}$$

Now due to a phenomena known as *aliasing*, explained in detail in standard texts, the values of c_k which have the most significance are the first half $c_k, k = 0, \ldots, N/2$.

Here

$$||\mathbf{y}||_2 = ||\mathbf{c}||_2 = \sqrt{354} = 18.8149. \tag{8.221}$$

Note that by construction

$$y_0 = \frac{1}{\sqrt{5}}\left(c_0 + c_1 + c_2 + c_3 + c_4\right), \tag{8.222}$$

$$y_1 = \frac{1}{\sqrt{5}}\left(c_0 + c_1 e^{2\pi i/5} + c_2 e^{4\pi i/5} + c_3 e^{6\pi i/5} + c_4 e^{8\pi i/5}\right), \tag{8.223}$$

$$y_2 = \frac{1}{\sqrt{5}}\left(c_0 + c_1 e^{4\pi i/5} + c_2 e^{8\pi i/5} + c_3 e^{12\pi i/5} + c_4 e^{16\pi i/5}\right), \tag{8.224}$$

$$y_3 = \frac{1}{\sqrt{5}}\left(c_0 + c_1 e^{6\pi i/5} + c_2 e^{12\pi i/5} + c_3 e^{18\pi i/5} + c_4 e^{24\pi i/5}\right), \tag{8.225}$$

$$y_4 = \frac{1}{\sqrt{5}}\left(c_0 + c_1 e^{8\pi i/5} + c_2 e^{16\pi i/5} + c_3 e^{24\pi i/5} + c_4 e^{32\pi i/5}\right). \tag{8.226}$$

In general, it is seen that y_j can be described by

$$y_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} c_k \exp\left((2\pi i)\frac{kj}{N}\right), \qquad j = 0, \ldots, N-1. \tag{8.227}$$

Realizing now that for a uniform discretization, such as done here, that

$$\Delta x = \frac{x_{max} - x_{min}}{N-1}, \tag{8.228}$$

and that

$$x_j = j\Delta x + x_{min}, \qquad j = 0, \ldots, N-1, \tag{8.229}$$

one has

$$x_j = j\left(\frac{x_{max} - x_{min}}{N-1}\right) + x_{min}, \qquad j = 0, \ldots, N-1. \tag{8.230}$$

Solving for j, one gets

$$j = (N-1)\left(\frac{x_j - x_{min}}{x_{max} - x_{min}}\right), \tag{8.231}$$

so that y_j can be expressed as a Fourier-type expansion in terms of x_j as

$$y_j = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} c_k \exp\left((2\pi i)k\left(\frac{N-1}{N}\right)\left(\frac{x_j - x_{min}}{x_{max} - x_{min}}\right)\right), \qquad j = 0, \ldots, N-1. \tag{8.232}$$

Here, the wavenumber of mode k, κ_k, is seen to be

$$\kappa_k = k\frac{N-1}{N}. \tag{8.233}$$

And as $N \to \infty$, one has

$$\kappa_k \sim k. \tag{8.234}$$

Example 8.24

The real power of the DFT is seen in its ability to select amplitudes of modes of signals at certain frequencies. Consider the signal for $x \in [0, 3]$

$$y(x) = 10\sin\left((2\pi)\frac{2x}{3}\right) + 2\sin\left((2\pi)\frac{10x}{3}\right) + \sin\left((2\pi)\frac{100x}{3}\right). \tag{8.235}$$

Rescaling the domain so as to take $x \in [0, 3]$ into $\tilde{x} \in [0, 1]$ via the transformation $\tilde{x} = x/3$, one has

$$y(\tilde{x}) = 10\sin\left((2\pi)2\tilde{x}\right) + 2\sin\left((2\pi)10\tilde{x}\right) + \sin\left((2\pi)100\tilde{x}\right). \tag{8.236}$$

To capture the high wavenumber components of the signal, one must have a sufficiently large value of N. Note in the transformed domain that the smallest wavelength is $\lambda = 1/100 = 0.01$. So for a domain length of unity, one needs *at least* $N = 100$ sampling points. In fact, let us choose to take more points, $N = 523$. There is no problem in choosing an unusual number of points for this so-called slow Fourier

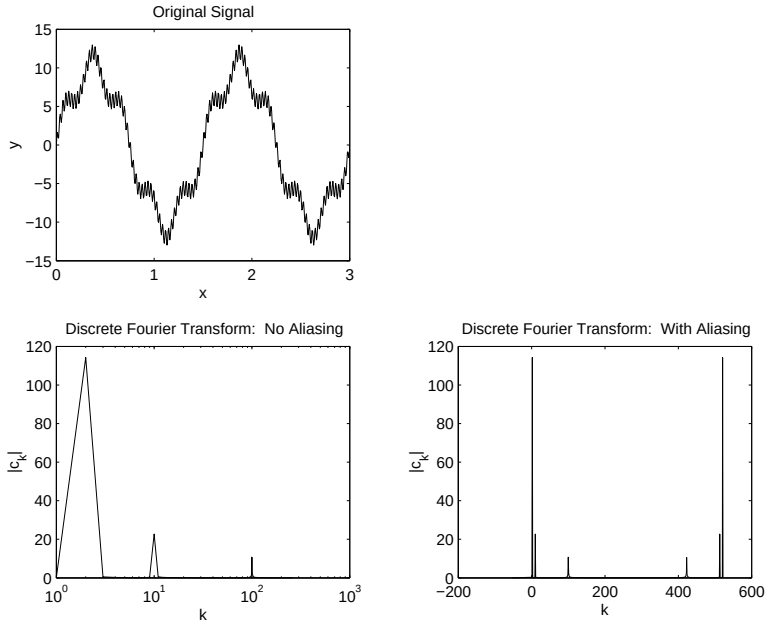

Figure 8.4: Plot of a three term sinusoid $y(x)$ and its discrete Fourier transform for $N = 523$. The first DFT is plotted from $k = 0, \ldots, N/2$ and thus represents the original signal well. The second DFT is plotted from $k = 0, \ldots, N-1$ and exhibits aliasing effects at high k.

transform. If an FFT were attempted, one would have to choose integral powers of 2 as the number of points.

A plot of the function $y(x)$ and two versions of its DFT, $|c_k|$ vs. k, is given in in Fig. 8.4 Note that $|c_k|$ has its peaks at $k = 2$, $k = 10$, and $k = 100$, equal to the wave numbers of the generating sine functions, $\kappa_1 = 2$, $\kappa_2 = 10$, and $\kappa_3 = 100$. To avoid the confusing, and non-physical, aliasing effect, only half the $|c_k|$ values have been plotted the first DFT of Fig. 8.4. The second DFT here plots all values of $|c_k|$ and thus exhibits aliasing for large k.

Example 8.25

Now take the DFT of a signal which is corrupted by so-called white, or random, noise. The signal here is given in $x \in [0, 1]$ by

$$y(x) = \sin((2\pi)10x) + \sin((2\pi)100x) + f_{rand}[-1, 1](x). \tag{8.237}$$

Here $f_{rand}[-1, 1](x)$ returns a random number between -1 and 1 for any value of x. A plot of the function $y(x)$ and two versions of its 607 point DFT, $|c_k|$ vs. k, is given in in Fig. 8.5 In the raw data plotted in Fig. 8.5, it is difficult to distinguish the signal from the random noise. But on examination of the accompanying DFT plot, it is clear that there are unambiguous components of the signal which peak at $k = 10$ and $k = 100$, which indicates there is a strong component of the signal with $\kappa = 10$ and $\kappa = 100$. Once again, to avoid the confusing, and non-physical, aliasing effect, only half the $|c_k|$ values have been plotted in the first DFT of Fig. 8.5. The second DFT gives all values of $|c_k|$ and exhibits aliasing.

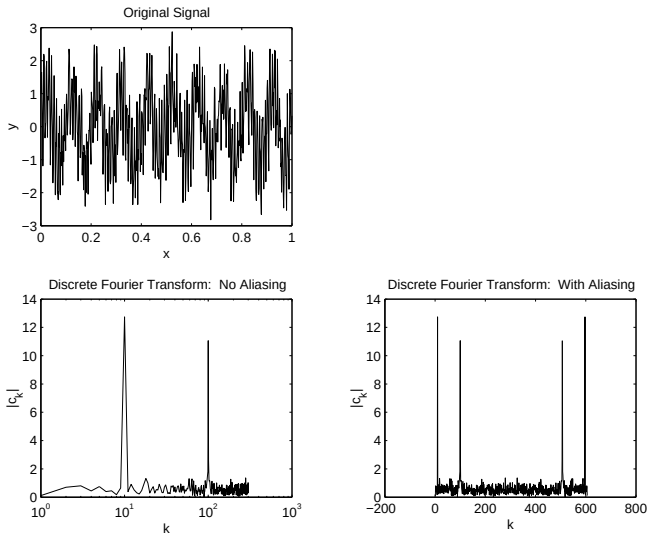

Figure 8.5: Plot of a two-term sinusoid accompanied by random noise $y(x)$ and its discrete Fourier transform for $N = 607$ points. The first DFT is plotted from $k = 0, \ldots, N/2$ and thus represents the original signal well. The second DFT is plotted from $k = 0, \ldots, N-1$ and exhibits aliasing effects at high k.

8.8 Matrix decompositions

One of the most important tasks, especially in the numerical solution of algebraic and differential equations, is decomposing general matrices into simpler components. A brief discussion will be given here of some of the more important decompositions. Full discussions can be found in Strang's text. It is noted that many popular software programs, such as `MATLAB`, `Mathematica`, `LAPACK` libraries, etc. have routines which routinely calculate these decompositions.

8.8.1 $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition

Probably the most important technique in solving linear systems of algebraic equations of the form $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, uses the decomposition

$$\mathbf{A} = \mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}, \tag{8.238}$$

where $\mathbf{A}$ is a square matrix,[4] $\mathbf{P}$ is a never-singular permutation matrix, $\mathbf{L}$ is a lower triangular matrix, $\mathbf{D}$ is a diagonal matrix, and $\mathbf{U}$ is an upper triangular matrix. The notation of $\mathbf{U}$ for the upper triangular matrix is common, and should not be confused with the identical notation for a unitary matrix. In other contexts $\mathbf{R}$ is sometimes used for an upper triangular matrix, and $\mathbf{P}$ is sometimes used for a projection matrix. All terms can be found by ordinary Gaussian elimination. The permutation matrix is necessary in case row exchanges are necessary in the Gaussian elimination.

A common numerical algorithm to solve for $\mathbf{x}$ in $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ is as follows

- Factor $\mathbf{A}$ into $\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ so that $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ becomes

$$\underbrace{\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}}_{\mathbf{A}} \cdot \mathbf{x} = \mathbf{b}. \tag{8.239}$$

- Operate on both sides of Eq. (8.239) with $\left(\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D}\right)^{-1}$ to get

$$\mathbf{U} \cdot \mathbf{x} = \left(\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D}\right)^{-1} \cdot \mathbf{b}. \tag{8.240}$$

- Solve next for the new variable $\mathbf{c}$ in the new equation

$$\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{c} = \mathbf{b}, \tag{8.241}$$

  so

$$\mathbf{c} = \left(\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D}\right)^{-1} \cdot \mathbf{b}. \tag{8.242}$$

  The triangular form of $\mathbf{L} \cdot \mathbf{D}$ renders the inversion of $(\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D})$ to be much more computationally efficient than inversion of an arbitrary square matrix.

- Substitute $\mathbf{c}$ from Eq. (8.242) into Eq. (8.240), the modified version of the original equation, to get

$$\mathbf{U} \cdot \mathbf{x} = \mathbf{c}, \tag{8.243}$$

  so

$$\mathbf{x} = \mathbf{U}^{-1} \cdot \mathbf{c}. \tag{8.244}$$

  Again since $\mathbf{U}$ is triangular, the inversion is computationally efficient.

Example 8.26

Find the $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition of the matrix:

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}. \tag{8.245}$$

[4] If $\mathbf{A}$ is not square, there is an equivalent decomposition, known as *row echelon form*, to be discussed in Sec. 8.8.3.

The process is essentially a series of row operations, which is the essence of Gaussian elimination. First we operate to transform the -22 and 16 in the first column into zeroes. Crucial in this step is the necessity of the term in the 1,1 slot, known as the *pivot*, to be non-zero. If it is zero, a row exchange will be necessary, mandating a permutation matrix which is not the identity matrix. In this case there are no such problems. We multiply the first row by $22/5$ and subtract from the second row, then multiply the first row by $-16/5$ and subtract from the third row. The factors $22/5$ and $-16/5$ will go in the 2,1 and 3,1 slots of the matrix $\mathbf{L}$. The diagonal of $\mathbf{L}$ always is filled with ones. This row operation yields

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 22/5 & 1 & 0 \\ -16/5 & 0 & 1 \end{pmatrix} \begin{pmatrix} -5 & 4 & 9 \\ 0 & -18/5 & -108/5 \\ 0 & 24/5 & 114/5 \end{pmatrix}. \tag{8.246}$$

Now multiplying the new second row by $-4/3$, subtracting this from the third row, and depositing the factor $-4/3$ into 3,2 slot of the matrix $\mathbf{L}$, we get

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 22/5 & 1 & 0 \\ -16/5 & -4/3 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} -5 & 4 & 9 \\ 0 & -18/5 & -108/5 \\ 0 & 0 & -6 \end{pmatrix}}_{\mathbf{U}}. \tag{8.247}$$

The form given in Eq. (8.247) is often described as the $\mathbf{L} \cdot \mathbf{U}$ decomposition of $\mathbf{A}$. We can force the diagonal terms of the upper triangular matrix to unity by extracting a diagonal matrix $\mathbf{D}$ to form the $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition:

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 22/5 & 1 & 0 \\ -16/5 & -4/3 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} -5 & 0 & 0 \\ 0 & -18/5 & 0 \\ 0 & 0 & -6 \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} 1 & -4/5 & -9/5 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{U}}. \tag{8.248}$$

Note that $\mathbf{D}$ *does not* contain the eigenvalues of $\mathbf{A}$. Also since there were no row exchanges necessary $\mathbf{P} = \mathbf{P}^{-1} = \mathbf{I}$, and it has not been included.

Example 8.27

Find the $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition of the matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{8.249}$$

There is a zero in the pivot, so a row exchange is necessary:

$$\mathbf{P} \cdot \mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}. \tag{8.250}$$

Performing Gaussian elimination by subtracting 1 times the first row from the second and depositing the 1 in the 2,1 slot of $\mathbf{L}$, we get

$$\mathbf{P}\cdot\mathbf{A}=\mathbf{L}\cdot\mathbf{U}\rightarrow\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}0&1&2\\1&1&1\\1&0&0\end{pmatrix}=\begin{pmatrix}1&0&0\\1&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&1&2\end{pmatrix}. \tag{8.251}$$

Now subtracting 1 times the second row, and depositing the 1 in the 3,2 slot of $\mathbf{L}$

$$\mathbf{P}\cdot\mathbf{A}=\mathbf{L}\cdot\mathbf{U}\rightarrow\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}0&1&2\\1&1&1\\1&0&0\end{pmatrix}=\begin{pmatrix}1&0&0\\1&1&0\\0&1&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}. \tag{8.252}$$

Now $\mathbf{U}$ already has ones on the diagonal, so the diagonal matrix $\mathbf{D}$ is simply the identity matrix. Using this and inverting $\mathbf{P}$, which is $\mathbf{P}$ itself(!), we get the final decomposition

$$\mathbf{A}=\mathbf{P}^{-1}\cdot\mathbf{L}\cdot\mathbf{D}\cdot\mathbf{U}\rightarrow\begin{pmatrix}0&1&2\\1&1&1\\1&0&0\end{pmatrix}=\underbrace{\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}}_{\mathbf{P}^{-1}}\underbrace{\begin{pmatrix}1&0&0\\1&1&0\\0&1&1\end{pmatrix}}_{\mathbf{L}}\underbrace{\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}}_{\mathbf{D}}\underbrace{\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}}_{\mathbf{U}}. \tag{8.253}$$

8.8.2 Cholesky decomposition

If $\mathbf{A}$ is a Hermitian positive definite matrix, we can define a Cholesky decomposition. Because $\mathbf{A}$ must be positive definite, it must be square. The Cholesky decomposition is as follows:

$$\mathbf{A}=\mathbf{U}^H\cdot\mathbf{U}. \tag{8.254}$$

Here $\mathbf{U}$ is an upper triangular matrix. One might think of $\mathbf{U}$ as the rough equivalent of the square root of the positive definite $\mathbf{A}$. We also have the related decomposition

$$\mathbf{A}=\hat{\mathbf{U}}^H\cdot\mathbf{D}\cdot\hat{\mathbf{U}}, \tag{8.255}$$

where $\hat{\mathbf{U}}$ is upper triangular with a value of unity on its diagonal, and $\mathbf{D}$ is diagonal.

If we define a lower triangular $\mathbf{L}$ as $\mathbf{L}=\mathbf{U}^H$, the Cholesky decomposition can be rewritten as

$$\mathbf{A}=\mathbf{L}\cdot\mathbf{L}^H. \tag{8.256}$$

There also exists an analogous decomposition

$$\mathbf{A}=\hat{\mathbf{L}}\cdot\mathbf{D}\cdot\hat{\mathbf{L}}^H, \tag{8.257}$$

Note also that these definitions hold as well for real $\mathbf{A}$; in such cases, we can simply replace the Hermitian transpose by the ordinary transpose.

Example 8.28

The Cholesky decomposition of a Hermitian matrix $\mathbf{A}$ is as follows

$$\mathbf{A} = \begin{pmatrix} 5 & 4i \\ -4i & 5 \end{pmatrix} = \mathbf{U}^H \cdot \mathbf{U} = \underbrace{\begin{pmatrix} \sqrt{5} & 0 \\ -\frac{4i}{\sqrt{5}} & \frac{3}{\sqrt{5}} \end{pmatrix}}_{\mathbf{U}^H} \cdot \underbrace{\begin{pmatrix} \sqrt{5} & \frac{4i}{\sqrt{5}} \\ 0 & \frac{3}{\sqrt{5}} \end{pmatrix}}_{\mathbf{U}}. \tag{8.258}$$

Note the eigenvalues of $\mathbf{A}$ are $\lambda = 1$, $\lambda = 9$, so the matrix is indeed positive definite.

We can also write in alternative form

$$\mathbf{A} = \begin{pmatrix} 5 & 4i \\ -4i & 5 \end{pmatrix} = \hat{\mathbf{U}}^H \cdot \mathbf{D} \cdot \hat{\mathbf{U}} = \underbrace{\begin{pmatrix} 1 & 0 \\ -\frac{4i}{5} & 1 \end{pmatrix}}_{\hat{\mathbf{U}}^H} \cdot \underbrace{\begin{pmatrix} 5 & 0 \\ 0 & \frac{9}{5} \end{pmatrix}}_{\mathbf{D}} \cdot \underbrace{\begin{pmatrix} 1 & \frac{4i}{5} \\ 0 & 1 \end{pmatrix}}_{\hat{\mathbf{U}}}. \tag{8.259}$$

8.8.3 Row echelon form

When $\mathbf{A}$ is not square, we can still use Gaussian elimination to cast the matrix in *row echelon form*:

$$\mathbf{A} = \mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}. \tag{8.260}$$

Again $\mathbf{P}$ is a never-singular permutation matrix, $\mathbf{L}$ is lower triangular and square, $\mathbf{D}$ is diagonal and square, $\mathbf{U}$ is upper triangular and rectangular and of the same dimension as $\mathbf{A}$. The strategy is to use row operations in such a fashion that ones or zeroes appear on the diagonal.

Example 8.29

Determine the row-echelon form of the non-square matrix,

$$\mathbf{A} = \begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix}. \tag{8.261}$$

We take 2 times the first row and subtract the result from the second row. The scalar 2 is deposited in the 2,1 slot in the $\mathbf{L}$ matrix. So

$$\mathbf{A} = \begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} 1 & -3 & 2 \\ 0 & 6 & -1 \end{pmatrix}}_{\mathbf{U}}. \tag{8.262}$$

Again, Eq. (8.262) is also known as an $\mathbf{L} \cdot \mathbf{U}$ decomposition, and is often as useful as the $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition. There is no row exchange so the permutation matrix and its inverse are the identity matrix. We extract a 1 and 6 to form the diagonal matrix $\mathbf{D}$, so the final form is

$$\mathbf{A} = \mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{P}^{-1}} \underbrace{\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} 1 & -3 & 2 \\ 0 & 1 & -\frac{1}{6} \end{pmatrix}}_{\mathbf{U}}. \tag{8.263}$$

Row echelon form is an especially useful form for under-constrained systems as illustrated in the following example.

Example 8.30

Consider solutions for the unknown $\mathbf{x}$ in the equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ where $\mathbf{A}$ is known $\mathbf{A} : \mathbb{R}^5 \to \mathbb{R}^3$, and $\mathbf{b}$ is left general, but considered to be known:

$$\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 4 & 2 & -2 & 1 & 0 \\ -2 & -1 & 1 & -2 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \tag{8.264}$$

We perform Gaussian elimination row operations on the second and third rows to get zeros in the first column:

$$\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 0 & 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & -1 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ -2b_1 + b_2 \\ b_1 + b_3 \end{pmatrix}. \tag{8.265}$$

The next round of Gaussian elimination works on the third row and yields

$$\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 0 & 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ -2b_1 + b_2 \\ 3b_1 - b_2 + b_3 \end{pmatrix}. \tag{8.266}$$

Note that the reduced third equation gives

$$0 = 3b_1 - b_2 + b_3. \tag{8.267}$$

This is the equation of a plane in $\mathbb{R}^3$. Thus, arbitrary $\mathbf{b} \in \mathbb{R}^3$ *will not* satisfy the original equation. Said another way, the operator $\mathbf{A}$ maps arbitrary five-dimensional vectors $\mathbf{x}$ into a two-dimensional subspace of a three-dimensional vector space. The rank of $\mathbf{A}$ is 2. Thus, the dimension of both the row space and the column space is 2; the dimension of the right null space is 3, and the dimension of the left null space is 1.

We also note there are two non-trivial equations remaining. The first non-zero elements from the left of each row are known as the pivots. The number of pivots is equal to the rank of the matrix. Variables which correspond to each pivot are known as *basic variables*. Variables with no pivot are known as *free variables*. Here the basic variables are x_1 and x_4, while the free variables are x_2, x_3, and x_5.

Now enforcing the constraint $3b_1 - b_2 + b_3 = 0$, without which there will be *no* solution, we can set each free variable to an arbitrary value, and then solve the resulting square system. Take $x_2 = r$, $x_3 = s$, $x_5 = t$, where here r, s, and t are arbitrary real scalar constants. So

$$\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 0 & 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ r \\ s \\ x_4 \\ t \end{pmatrix} = \begin{pmatrix} b_1 \\ -2b_1 + b_2 \\ 0 \end{pmatrix}, \tag{8.268}$$

which gives

$$\begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 - r + s - 2t \\ -2b_1 + b_2 + 4t \end{pmatrix}, \tag{8.269}$$

which yields

$$x_4 = 2b_1 - b_2 - 4t, \tag{8.270}$$

$$x_1 = \frac{1}{2}(-b_1 + b_2 - r + s + 2t). \tag{8.271}$$

Thus

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(-b_1 + b_2 - r + s + 2t) \\ r \\ s \\ 2b_1 - b_2 - 4t \\ t \end{pmatrix}, \tag{8.272}$$

$$= \begin{pmatrix} \frac{1}{2}(-b_1 + b_2) \\ 0 \\ 0 \\ 2b_1 - b_2 \\ 0 \end{pmatrix} + r \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} \frac{1}{2} \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \\ 0 \\ -4 \\ 1 \end{pmatrix}, r, s, t \in \mathbb{R}^1. \tag{8.273}$$

The coefficients r, s, and t multiply the three right null space vectors. These in combination with two independent row space vectors, form a basis for any vector $\mathbf{x}$. Thus, we can again cast the solution as a particular solution which is a unique combination of independent row space vectors and a non-unique combination of the right null space vectors (the homogeneous solution):

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \underbrace{\frac{25b_1 - 13b_2}{106} \begin{pmatrix} 2 \\ 1 \\ -1 \\ 1 \\ 2 \end{pmatrix} + \frac{-13b_1 + 11b_2}{106} \begin{pmatrix} 4 \\ 2 \\ -2 \\ 1 \\ 0 \end{pmatrix}}_{\text{row space}} + \underbrace{\hat{r} \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \hat{s} \begin{pmatrix} \frac{1}{2} \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \hat{t} \begin{pmatrix} 1 \\ 0 \\ 0 \\ -4 \\ 1 \end{pmatrix}}_{\text{right null space}}. \tag{8.274}$$

In matrix form, we can say that

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 2 & 4 & -\frac{1}{2} & \frac{1}{2} & 1 \\ 1 & 2 & 1 & 0 & 0 \\ -1 & -2 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & -4 \\ 2 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{25b_1 - 13b_2}{106} \\ \frac{-13b_1 + 11b_2}{106} \\ \hat{r} \\ \hat{s} \\ \hat{t} \end{pmatrix}. \tag{8.275}$$

Here we have taken $\hat{r} = r + (b_1 - 9b_2)/106$, $\hat{s} = s + (-b_1 + 9b_2)/106$, and $\hat{t} = (-30b_1 + 26b_2)/106$; as they are arbitrary constants multiplying vectors in the right null space, the relationship to b_1 and b_2 is actually unimportant. As before, while the null space basis vectors are orthogonal to the row space basis vectors, the entire system is not orthogonal. The Gram-Schmidt procedure could be used to cast the solution on either an orthogonal or orthonormal basis.

It is also noted that we have effectively found the $\mathbf{L} \cdot \mathbf{U}$ decomposition of $\mathbf{A}$. The terms in $\mathbf{L}$ are from the Gaussian elimination, and we have already $\mathbf{U}$:

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \rightarrow \underbrace{\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 4 & 2 & -2 & 1 & 0 \\ -2 & -1 & 1 & -2 & -6 \end{pmatrix}}_{\mathbf{A}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 0 & 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{\mathbf{U}}. \tag{8.276}$$

The $\mathbf{L}\cdot\mathbf{D}\cdot\mathbf{U}$ decomposition is

$$\underbrace{\begin{pmatrix} 2 & 1 & -1 & 1 & 2 \\ 4 & 2 & -2 & 1 & 0 \\ -2 & -1 & 1 & -2 & -6 \end{pmatrix}}_{\mathbf{A}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}}_{\mathbf{L}} \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{\mathbf{U}}. \tag{8.277}$$

There were no row exchanges, so in effect the permutation matrix $\mathbf{P}$ is the identity matrix, and there is no need to include it.

Lastly, we note that a more robust alternative to the method shown here would be to *first* apply the $\mathbf{A}^T$ operator to both sides of the equation so to map both sides into the column space of $\mathbf{A}$. Then there would be no need to restrict $\mathbf{b}$ so that it lies in the column space. Our results are then interpreted as giving us only a projection of $\mathbf{x}$. Taking $\mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x} = \mathbf{A}^T\cdot\mathbf{b}$ and then casting the result into row echelon form gives

$$\begin{pmatrix} 1 & 1/2 & -1/2 & 0 & -1 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} (1/22)(b_1 + 7b_2 + 4b_3) \\ (1/11)(b_1 - 4b_2 - 7b_3) \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{8.278}$$

This suggests we take $x_2 = r$, $x_3 = s$, and $x_5 = t$ and solve so to get

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} (1/22)(b_1 + 7b_2 + 4b_3) \\ 0 \\ 0 \\ (1/11)(b_1 - 4b_2 - 7b_3) \\ 0 \end{pmatrix} + r \begin{pmatrix} -1/2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1/2 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \\ 0 \\ -4 \\ 1 \end{pmatrix}. \tag{8.279}$$

We could go on to cast this in terms of combinations of row vectors and right null space vectors, but will not do so here. It is reiterated that this result is valid for arbitrary $\mathbf{b}$, but that it only represents a solution which minimizes the residual in $||\mathbf{A}\cdot\mathbf{x} - \mathbf{b}||_2$.

8.8.4 $\mathbf{Q}\cdot\mathbf{R}$ decomposition

The $\mathbf{Q}\cdot\mathbf{R}$ decomposition allows us to formulate a matrix as the product of an orthogonal (unitary if complex) matrix $\mathbf{Q}$ and an upper triangular matrix $\mathbf{R}$, of the same dimension as $\mathbf{A}$. That is we seek $\mathbf{Q}$ and $\mathbf{R}$ such that

$$\mathbf{A} = \mathbf{Q}\cdot\mathbf{R}. \tag{8.280}$$

The matrix $\mathbf{A}$ can be square or rectangular. See Strang for details of the algorithm. It can be thought of as a deformation due to $\mathbf{R}$ followed by a volume-preserving rotation or reflection due to $\mathbf{Q}$.

Example 8.31

The $\mathbf{Q}\cdot\mathbf{R}$ decomposition of the matrix we considered in a previous example, p. 364, is as follows:

$$\mathbf{A} = \underbrace{\begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}}_{\mathbf{A}} = \mathbf{Q}\cdot\mathbf{R} = \underbrace{\begin{pmatrix} -0.1808 & -0.4982 & 0.8480 \\ -0.7954 & -0.4331 & -0.4240 \\ 0.5785 & -0.7512 & -0.3180 \end{pmatrix}}_{\mathbf{Q}} \underbrace{\begin{pmatrix} 27.6586 & -16.4867 & -19.4153 \\ 0 & -2.0465 & -7.7722 \\ 0 & 0 & 1.9080 \end{pmatrix}}_{\mathbf{R}}. \tag{8.281}$$

Note that $\det\mathbf{Q} = 1$, so it is volume- and orientation-preserving. Noting further that $||\mathbf{Q}||_2 = 1$, we deduce that $||\mathbf{R}||_2 = ||\mathbf{A}||_2$. And it is easy to show that $||\mathbf{R}||_2 = ||\mathbf{A}||_2 = 37.9423$. Also recalling how matrices can be thought of as transformations, we see how to think of $\mathbf{A}$ as a stretching ($\mathbf{R}$) followed by rotation ($\mathbf{Q}$).

Example 8.32

Find the $\mathbf{Q}\cdot\mathbf{R}$ decomposition for our non-square matrix from p. 366,

$$\mathbf{A} = \begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix}. \tag{8.282}$$

The decomposition is

$$\mathbf{A} = \underbrace{\begin{pmatrix} 0.4472 & -0.8944 \\ 0.8944 & 0.4472 \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} 2.2361 & -1.3416 & 3.577 \\ 0 & 2.6833 & -0.4472 \end{pmatrix}}_{\mathbf{R}}. \tag{8.283}$$

Once again $\det\mathbf{Q} = 1$, so it is volume- and orientation-preserving. It is easy to show $||\mathbf{A}||_2 = ||\mathbf{R}||_2 = 4.63849$.

Example 8.33

Give a geometric interpretation of the $\mathbf{Q}\cdot\mathbf{R}$ decomposition in the context of the discussion surrounding the transformation of a unit square by the matrix $\mathbf{A}$ considered earlier on p. 348.

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}. \tag{8.284}$$

The decomposition is

$$\mathbf{A} = \underbrace{\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}}_{\mathbf{R}}. \tag{8.285}$$

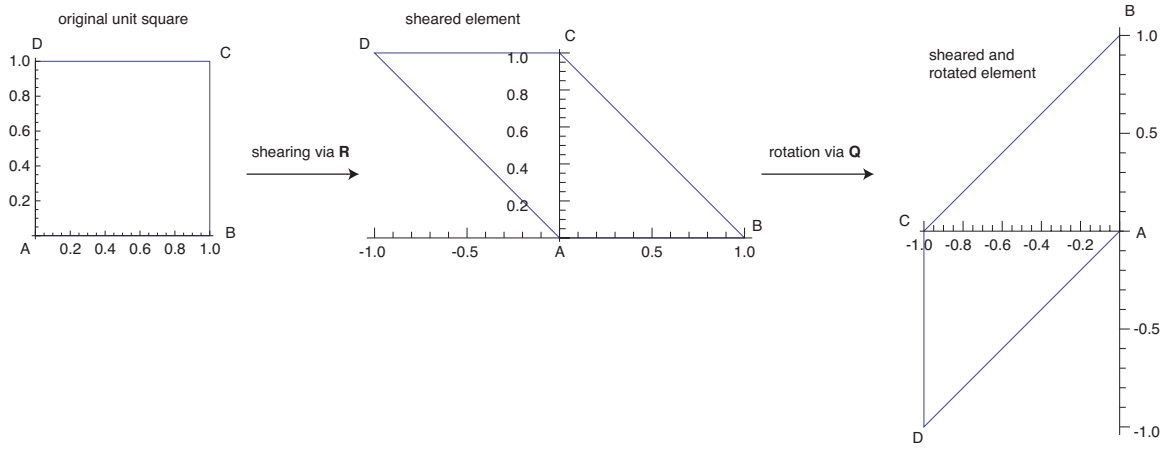

Figure 8.6: Unit square transforming via explicit stretching ($\mathbf{R}$), and rotation ($\mathbf{Q}$) under a linear area- and orientation-preserving alibi mapping.

Now $\det \mathbf{A} = 1$. Moreover, $\det \mathbf{Q} = 1$ and $\det \mathbf{R} = 1$, so both of these matrices preserve area and orientation. As usual $||\mathbf{Q}||_2 = 1$, so its operation preserves the lengths of vectors. The deformation is embodied in $\mathbf{R}$ which has $||\mathbf{R}||_2 = ||\mathbf{A}||_2 = 1.61803$. Decomposing the transformation of the unit square depicted in Fig. 8.3 by first applying $\mathbf{R}$ to each of the vertices, and then applying $\mathbf{Q}$ to each of the stretched vertices, we see that $\mathbf{R}$ effects an area- and orientation-preserving shear deformation, and $\mathbf{Q}$ effects a counter-clockwise rotation of $\pi/2$. This is depicted in Fig. 8.6.

The $\mathbf{Q}\cdot\mathbf{R}$ decomposition can be shown to be closely related to the Gram-Schmidt orthogonalization process. It is also useful in increasing the efficiency of estimating $\mathbf{x}$ for $\mathbf{A}\cdot\mathbf{x} \simeq \mathbf{b}$ when the system is over-constrained; that is $\mathbf{b}$ is not in the column space of $\mathbf{A}$, $\mathbb{R}(\mathbf{A})$. If we, as usual operate on both sides as follows,

$$\begin{aligned}
\mathbf{A}\cdot\mathbf{x} &\simeq \mathbf{b}, \qquad \mathbf{b} \notin \mathbb{R}(\mathbf{A}), && (8.286)\\
\mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x} &= \mathbf{A}^T\cdot\mathbf{b}, \qquad \mathbf{A} = \mathbf{Q}\cdot\mathbf{R}, && (8.287)\\
(\mathbf{Q}\cdot\mathbf{R})^T\cdot\mathbf{Q}\cdot\mathbf{R}\cdot\mathbf{x} &= (\mathbf{Q}\cdot\mathbf{R})^T\cdot\mathbf{b}, && (8.288)\\
\mathbf{R}^T\cdot\mathbf{Q}^T\cdot\mathbf{Q}\cdot\mathbf{R}\cdot\mathbf{x} &= \mathbf{R}^T\cdot\mathbf{Q}^T\cdot\mathbf{b}, && (8.289)\\
\mathbf{R}^T\cdot\mathbf{Q}^{-1}\cdot\mathbf{Q}\cdot\mathbf{R}\cdot\mathbf{x} &= \mathbf{R}^T\cdot\mathbf{Q}^T\cdot\mathbf{b}, && (8.290)\\
\mathbf{R}^T\cdot\mathbf{R}\cdot\mathbf{x} &= \mathbf{R}^T\cdot\mathbf{Q}^T\cdot\mathbf{b}, && (8.291)\\
\mathbf{x} &= \left(\mathbf{R}^T\cdot\mathbf{R}\right)^{-1}\cdot\mathbf{R}^T\cdot\mathbf{Q}^T\cdot\mathbf{b}, && (8.292)\\
\mathbf{Q}\cdot\mathbf{R}\cdot\mathbf{x} &= \mathbf{Q}\cdot\left(\mathbf{R}\cdot\left(\mathbf{R}^T\cdot\mathbf{R}\right)^{-1}\cdot\mathbf{R}^T\right)\cdot\mathbf{Q}^T\cdot\mathbf{b}, && (8.293)\\
\mathbf{A}\cdot\mathbf{x} &= \underbrace{\mathbf{Q}\cdot\left(\mathbf{R}\cdot\left(\mathbf{R}^T\cdot\mathbf{R}\right)^{-1}\cdot\mathbf{R}^T\right)\cdot\mathbf{Q}^T}_{\mathbf{P}}\cdot\mathbf{b}. && (8.294)
\end{aligned}$$

When rectangular $\mathbf{R}$ has no zeros on its diagonal, $\mathbf{R} \cdot \left(\mathbf{R}^T \cdot \mathbf{R}\right)^{-1} \cdot \mathbf{R}^T$ has all zeroes, except for r ones on the diagonal, where r is the rank of $\mathbf{R}$. This makes solution of over-constrained problems particularly simple. We note lastly that $\mathbf{Q} \cdot \mathbf{R} \cdot \left(\mathbf{R}^T \cdot \mathbf{R}\right)^{-1} \cdot \mathbf{R}^T \cdot \mathbf{Q}^T = \mathbf{P}$, a projection matrix, defined first in Eq. (7.160), and to be discussed in Sec. 8.9.

8.8.5 Diagonalization

Casting a matrix into a form in which all (or sometimes most) of its off-diagonal elements have zero value has its most important application in solving systems of differential equations but also in other scenarios. For many cases, we can decompose a square matrix $\mathbf{A}$ into the form

$$\mathbf{A} = \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{S}^{-1}, \tag{8.295}$$

where $\mathbf{S}$ is non-singular matrix and $\mathbf{\Lambda}$ is a diagonal matrix. To diagonalize a square matrix $\mathbf{A}$, we must find $\mathbf{S}$, a diagonalizing matrix, such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$ is diagonal. Not all matrices are diagonalizable. Note that by inversion, we can also say

$$\mathbf{\Lambda} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}. \tag{8.296}$$

Considering $\mathbf{A}$ to be the original matrix, we have subjected it to a general linear transformation, which in general stretches and rotates, to arrive at $\mathbf{\Lambda}$; this transformation has the same form as that previously considered in Eq. (7.278).

Theorem

A matrix with distinct eigenvalues can be diagonalized, but the diagonalizing matrix is not unique.

Definition: The algebraic multiplicity of an eigenvalue is the number of times it occurs. The geometric multiplicity of an eigenvalue is the number of eigenvectors it has.

Theorem

Nonzero eigenvectors corresponding to different eigenvalues are linearly independent.

Theorem

If $\mathbf{A}$ is an $N \times N$ matrix with N linearly independent right eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n, \cdots, \mathbf{e}_N\}$ corresponding to eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n, \cdots, \lambda_N\}$ (not necessarily distinct), then the $N \times N$ matrix $\mathbf{S}$ whose columns are populated by the eigenvectors of $\mathbf{A}$

$$\mathbf{S} = \begin{pmatrix} \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_n & \dots & \mathbf{e}_N \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \end{pmatrix} \tag{8.297}$$

makes

$$\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \mathbf{\Lambda}, \tag{8.298}$$

where

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \lambda_n & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda_N \end{pmatrix}, \tag{8.299}$$

is a diagonal matrix of eigenvalues. The matrices $\mathbf{A}$ and $\mathbf{\Lambda}$ are similar.

Let's see if this recipe works when we fill the columns of $\mathbf{S}$ with the eigenvectors. First operate on Eq. (8.298) with $\mathbf{S}$ to arrive at a more general version of the eigenvalue problem:

$$\mathbf{A}\cdot\mathbf{S} = \mathbf{S}\cdot\mathbf{\Lambda}, \tag{8.300}$$

$$\underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}}_{=\mathbf{A}} \underbrace{\begin{pmatrix} \vdots & \dots & \vdots \\ \mathbf{e}_1 & \dots & \mathbf{e}_N \\ \vdots & \dots & \vdots \end{pmatrix}}_{=\mathbf{S}} = \underbrace{\begin{pmatrix} \vdots & \dots & \vdots \\ \mathbf{e}_1 & \dots & \mathbf{e}_N \\ \vdots & \dots & \vdots \end{pmatrix}}_{=\mathbf{S}} \underbrace{\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix}}_{=\mathbf{\Lambda}}, \tag{8.301}$$

$$= \underbrace{\begin{pmatrix} \vdots & \dots & \vdots \\ \lambda_1\mathbf{e}_1 & \dots & \lambda_N\mathbf{e}_N \\ \vdots & \dots & \vdots \end{pmatrix}}_{=\mathbf{S}\cdot\mathbf{\Lambda}}, \tag{8.302}$$

$$\mathbf{A}\cdot\mathbf{e}_1 + \cdots + \mathbf{A}\cdot\mathbf{e}_N = \lambda_1\mathbf{e}_1 + \cdots + \lambda_N\mathbf{e}_N, \tag{8.303}$$

$$\mathbf{A}\cdot\mathbf{e}_1 + \cdots + \mathbf{A}\cdot\mathbf{e}_N = \lambda_1\mathbf{I}\cdot\mathbf{e}_1 + \cdots + \lambda_N\mathbf{I}\cdot\mathbf{e}_N. \tag{8.304}$$

Rearranging, we get

$$\underbrace{(\mathbf{A}-\lambda_1\mathbf{I})\cdot\mathbf{e}_1}_{=0} + \ldots + \underbrace{(\mathbf{A}-\lambda_N\mathbf{I})\cdot\mathbf{e}_N}_{=0} = \mathbf{0}. \tag{8.305}$$

Now $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ are linearly independent. Thus, this induces N eigenvalue problems for each eigenvector:

$$\mathbf{A}\cdot\mathbf{e}_1 = \lambda_1\mathbf{I}\cdot\mathbf{e}_1, \tag{8.306}$$

$$\mathbf{A}\cdot\mathbf{e}_2 = \lambda_2\mathbf{I}\cdot\mathbf{e}_2, \tag{8.307}$$

$$\vdots \quad \vdots$$

$$\mathbf{A}\cdot\mathbf{e}_N = \lambda_N\mathbf{I}\cdot\mathbf{e}_N. \tag{8.308}$$

Note also the effect of post-multiplication of both sides of Eq. (8.300) by $\mathbf{S}^{-1}$:

$$\mathbf{A}\cdot\mathbf{S}\cdot\mathbf{S}^{-1} = \mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}, \tag{8.309}$$

$$\mathbf{A} = \mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}. \tag{8.310}$$

Example 8.34

Diagonalize the matrix considered in a previous example, p. 364:

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}, \tag{8.311}$$

and check. See the example around Eq. (8.245).

The eigenvalue-eigenvector pairs are

$$\lambda_1 = -6, \qquad \mathbf{e}_1 = \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}, \tag{8.312}$$

$$\lambda_2 = 3, \qquad \mathbf{e}_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \tag{8.313}$$

$$\lambda_3 = 6, \qquad \mathbf{e}_3 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}. \tag{8.314}$$

$$\tag{8.315}$$

Then

$$\mathbf{S} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} -1 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix}. \tag{8.316}$$

The inverse is

$$\mathbf{S}^{-1} = \begin{pmatrix} -\frac{4}{3} & \frac{2}{3} & 1 \\ -\frac{5}{3} & \frac{4}{3} & 1 \\ \frac{2}{3} & -\frac{1}{3} & 0 \end{pmatrix}. \tag{8.317}$$

Thus,

$$\mathbf{A} \cdot \mathbf{S} = \begin{pmatrix} 6 & 3 & 12 \\ 12 & 6 & 6 \\ -6 & 0 & 12 \end{pmatrix}, \tag{8.318}$$

and

$$\mathbf{\Lambda} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \begin{pmatrix} -6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}. \tag{8.319}$$

Let us also note the complementary decomposition of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{S}^{-1} = \underbrace{\begin{pmatrix} -1 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix}}_{\mathbf{S}} \underbrace{\begin{pmatrix} -6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{pmatrix} -\frac{4}{3} & \frac{2}{3} & 1 \\ -\frac{5}{3} & \frac{4}{3} & 1 \\ \frac{2}{3} & -\frac{1}{3} & 0 \end{pmatrix}}_{\mathbf{S}^{-1}} = \underbrace{\begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}}_{\mathbf{A}}. \tag{8.320}$$

Note that because the matrix is not symmetric, the eigenvectors are not orthogonal, e.g. $\mathbf{e}_1^T \cdot \mathbf{e}_2 = -5$.

Note that if $\mathbf{A}$ is symmetric (Hermitian), then its eigenvectors must be orthogonal; thus, it is possible to normalize the eigenvectors so that the matrix $\mathbf{S}$ is in fact orthogonal (unitary if complex). Thus, for symmetric $\mathbf{A}$ we have

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{-1}. \tag{8.321}$$

Since $\mathbf{Q}^{-1} = \mathbf{Q}^T$, we have

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^T. \tag{8.322}$$

Geometrically, the action of a symmetric $\mathbf{A}$ on a geometric entity can be considered as volume-preserving rotation or reflection via $\mathbf{Q}^T$, followed by a stretching due to $\mathbf{\Lambda}$, completed by another volume-preserving rotation or reflection via $\mathbf{Q}$, which acts opposite to the effect of $\mathbf{Q}^T$. Note also that with $\mathbf{A} \cdot \mathbf{S} = \mathbf{S} \cdot \mathbf{\Lambda}$, the column vectors of $\mathbf{S}$ (which are the right eigenvectors of $\mathbf{A}$) form a basis in $\mathbb{C}^N$.

Example 8.35

Consider the action of the matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \tag{8.323}$$

on a unit square in terms of the diagonal decomposition of $\mathbf{A}$.

We first note that $\det \mathbf{A} = 1$, so it preserves volumes and orientations. We easily calculate that $||\mathbf{A}||_2 = 3/2 + \sqrt{5}/2 = 2.61803$, so it has the potential to stretch a vector. It is symmetric, so it has real eigenvalues, which are $\lambda = 3/2 \pm \sqrt{5}/2$. Its spectral radius is thus $\rho(\mathbf{A}) = 3/2 + \sqrt{5}/2$, which is equal to its spectral norm. Its eigenvectors are orthogonal, so they can be orthonormalized to form an orthogonal matrix. After detailed calculation, one finds the diagonal decomposition to be

$$\mathbf{A} = \underbrace{\begin{pmatrix} \sqrt{\frac{5+\sqrt{5}}{10}} & -\sqrt{\frac{2}{5+\sqrt{5}}} \\ \sqrt{\frac{2}{5+\sqrt{5}}} & \sqrt{\frac{5+\sqrt{5}}{10}} \end{pmatrix}}_{\mathbf{Q}} \underbrace{\begin{pmatrix} \frac{3+\sqrt{5}}{2} & 0 \\ 0 & \frac{3-\sqrt{5}}{2} \end{pmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{pmatrix} \sqrt{\frac{5+\sqrt{5}}{10}} & \sqrt{\frac{2}{5+\sqrt{5}}} \\ -\sqrt{\frac{2}{5+\sqrt{5}}} & \sqrt{\frac{5+\sqrt{5}}{10}} \end{pmatrix}}_{\mathbf{Q}^T} \tag{8.324}$$

The action of this composition of matrix operations on a unit square is depicted in Fig. 8.7. The first rotation is induced by $\mathbf{Q}^T$ and is clockwise through an angle of $\pi/5.67511 = 31.717°$. This is followed by an eigen-stretching of $\mathbf{\Lambda}$. The action is completed by a rotation induced by $\mathbf{Q}$. The second rotation reverses the angle of the first in a counterclockwise rotation of $\pi/5.67511 = 31.717°$.

Consider now the right eigensystem of the adjoint of $\mathbf{A}$, denoted by $\mathbf{A}^*$:

$$\mathbf{A}^* \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{\Lambda}^*, \tag{8.325}$$

where $\mathbf{\Lambda}^*$ is the diagonal matrix containing the eigenvalues of $\mathbf{A}^*$, and $\mathbf{V}$ is the matrix whose columns are populated by the (right) eigenvectors of $\mathbf{A}^*$. Now we know from an earlier proof, Sec. 7.4.4, that the eigenvalues of the adjoint are the complex conjugates of

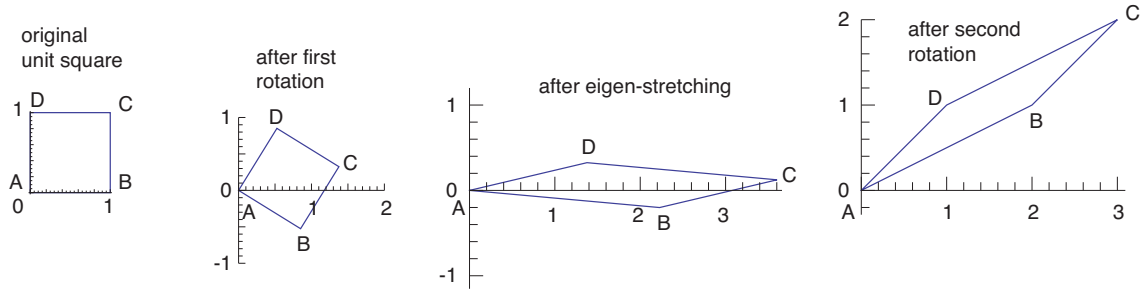

Figure 8.7: Unit square transforming via rotation, stretching, and rotation of the diagonalization decomposition under a linear area- and orientation-preserving alibi mapping.

those of the original operator, thus $\mathbf{\Lambda}^* = \mathbf{\Lambda}^H$. Also the adjoint operator for matrices is the Hermitian transpose. So, we find that

$$\mathbf{A}^H \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{\Lambda}^H. \tag{8.326}$$

Taking the Hermitian transpose of both sides, we recover

$$\mathbf{V}^H \cdot \mathbf{A} = \mathbf{\Lambda} \cdot \mathbf{V}^H. \tag{8.327}$$

So we see clearly that the left eigenvectors of a linear operator are the right eigenvectors of the adjoint of that operator.

It is also possible to show that, remarkably, when we take the product of the matrix of right eigenvectors of the operator with the matrix of right eigenvectors of its adjoint, that we obtain a diagonal matrix, which we denote as $\mathbf{D}$:

$$\mathbf{S}^H \cdot \mathbf{V} = \mathbf{D}. \tag{8.328}$$

Equivalently, this states that the inner product of the left eigenvector matrix with the right eigenvector matrix is diagonal. Let us see how this comes about. Let $\mathbf{s}_i$ be a right eigenvector of $\mathbf{A}$ with eigenvalue λ_i and $\mathbf{v}_j$ be a left eigenvector of $\mathbf{A}$ with eigenvalue λ_j. Then

$$\mathbf{A} \cdot \mathbf{s}_i = \lambda_i \mathbf{s}_i, \tag{8.329}$$

and

$$\mathbf{v}_j^H \cdot \mathbf{A} = \lambda_j \mathbf{v}_j^H. \tag{8.330}$$

If we premultiply the first eigen-relation, Eq. (8.329), by $\mathbf{v}_j^H$, we obtain

$$\underbrace{\mathbf{v}_j^H \cdot \mathbf{A}}_{=\lambda_j \mathbf{v}_j^H} \cdot \mathbf{s}_i = \mathbf{v}_j^H \cdot (\lambda_i \mathbf{s}_i)\,. \tag{8.331}$$

Substituting from the second eigen-relation, Eq. (8.330) and rearranging, Eq. (8.331) becomes

$$\lambda_j \mathbf{v}_j^H \cdot \mathbf{s}_i = \lambda_i \mathbf{v}_j^H \cdot \mathbf{s}_i. \tag{8.332}$$

Rearranging

$$(\lambda_j - \lambda_i)\left(\mathbf{v}_j^H \cdot \mathbf{s}_i\right) = 0. \tag{8.333}$$

Now if $i \neq j$ and $\lambda_i \neq \lambda_j$, we must have

$$\mathbf{v}_j^H \cdot \mathbf{s}_i = 0, \tag{8.334}$$

or, taking the Hermitian transpose,

$$\mathbf{s}_i^H \cdot \mathbf{v}_j = 0. \tag{8.335}$$

If $i = j$, then all we can say is $\mathbf{s}_i^H \cdot \mathbf{v}_j$ is some arbitrary scalar. Hence we have shown the desired relation that $\mathbf{S}^H \cdot \mathbf{V} = \mathbf{D}$.

Since eigenvectors have an arbitrary magnitude, it is a straightforward process to scale either $\mathbf{V}$ or $\mathbf{S}$ such that the diagonal matrix is actually the identity matrix. Here we choose to scale $\mathbf{V}$, given that our task was to find the reciprocal basis vectors of $\mathbf{S}$. We take then

$$\mathbf{S}^H \cdot \hat{\mathbf{V}} = \mathbf{I}. \tag{8.336}$$

Here $\hat{\mathbf{V}}$ denotes the matrix in which each eigenvector (column) of the original $\mathbf{V}$ has been scaled such that Eq. (8.336) is achieved. Hence $\hat{\mathbf{V}}$ is seen to give the set of reciprocal basis vectors for the basis defined by $\mathbf{S}$:

$$\mathbf{S}^R = \hat{\mathbf{V}}. \tag{8.337}$$

It is also easy to see then that the inverse of the matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \hat{\mathbf{V}}^H. \tag{8.338}$$

Example 8.36

For a matrix $\mathbf{A}$ considered in an earlier example, p. 363, consider the basis formed by its matrix of eigenvectors $\mathbf{S}$, and use the properly scaled matrix of eigenvectors of $\mathbf{A}^* = \mathbf{A}^H$ to determine the reciprocal basis $\mathbf{S}^R$.

We will take

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}. \tag{8.339}$$

As found before, the eigenvalue- (right) eigenvector pairs are

$$\lambda_1 = -6, \qquad \mathbf{e}_{1R} = \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}, \tag{8.340}$$

$$\lambda_2 = 3, \qquad \mathbf{e}_{2R} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \tag{8.341}$$

$$\lambda_3 = 6, \qquad \mathbf{e}_{3R} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}. \tag{8.342}$$

(8.343)

Then we take the matrix of basis vectors to be

$$\mathbf{S} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} -1 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix}. \tag{8.344}$$

The adjoint of $\mathbf{A}$ is

$$\mathbf{A}^H = \begin{pmatrix} -5 & -22 & 16 \\ 4 & 14 & -8 \\ 9 & 18 & -6 \end{pmatrix}. \tag{8.345}$$

The eigenvalues-(right) eigenvectors of $\mathbf{A}^H$, which are the left eigenvectors of $\mathbf{A}$, are found to be

$$\lambda_1 = -6, \qquad \mathbf{e}_{1L} = \begin{pmatrix} -4 \\ 2 \\ 3 \end{pmatrix}, \tag{8.346}$$

$$\lambda_2 = 3, \qquad \mathbf{e}_{2L} = \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix}, \tag{8.347}$$

$$\lambda_3 = 6, \qquad \mathbf{e}_{3L} = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}. \tag{8.348}$$

$$\tag{8.349}$$

So the matrix of right eigenvectors of the adjoint, which contains the left eigenvectors of the original matrix, is

$$\mathbf{V} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{e}_{1L} & \mathbf{e}_{2L} & \mathbf{e}_{3L} \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} -4 & -5 & -2 \\ 2 & 4 & 1 \\ 3 & 3 & 0 \end{pmatrix}. \tag{8.350}$$

We indeed find that the inner product of $\mathbf{S}$ and $\mathbf{V}$ is a diagonal matrix $\mathbf{D}$:

$$\mathbf{S}^H \cdot \mathbf{V} = \begin{pmatrix} -1 & -2 & 1 \\ 1 & 2 & 0 \\ 2 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} -4 & -5 & -2 \\ 2 & 4 & 1 \\ 3 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -3 \end{pmatrix}. \tag{8.351}$$

Using our knowledge of $\mathbf{D}$, we individually scale each column of $\mathbf{V}$ to form the desired reciprocal basis

$$\hat{\mathbf{V}} = \begin{pmatrix} -4/3 & -5/3 & 2/3 \\ 2/3 & 4/3 & -1/3 \\ 1 & 1 & 0 \end{pmatrix} = \mathbf{S}^R. \tag{8.352}$$

Then we see that the inner product of $\mathbf{S}$ and the reciprocal basis $\hat{\mathbf{V}} = \mathbf{S}^R$ is indeed the identity matrix:

$$\mathbf{S}^H \cdot \hat{\mathbf{V}} = \begin{pmatrix} -1 & -2 & 1 \\ 1 & 2 & 0 \\ 2 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} -4/3 & -5/3 & 2/3 \\ 2/3 & 4/3 & -1/3 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.353}$$

8.8.6 Jordan canonical form

A square matrix $\mathbf{A}$ without a sufficient number of linearly independent eigenvectors can still be decomposed into a near-diagonal form:

$$\mathbf{A} = \mathbf{S} \cdot \mathbf{J} \cdot \mathbf{S}^{-1}, \tag{8.354}$$

This form is known as the Jordan[5] (upper) canonical form in which the near-diagonal matrix $\mathbf{J}$:

$$\mathbf{J} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}, \tag{8.355}$$

has zeros everywhere except for eigenvalues along the principal diagonal and unity above the missing eigenvectors. The form is sometimes called a Jordan normal form.

Consider the eigenvalue λ of algebraic multiplicity $N - L + 1$ of the matrix $\mathbf{A}_{N\times N}$. Then

$$(\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{e} = \mathbf{0}, \tag{8.356}$$

gives some linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L$. If $L = N$, the algebraic multiplicity is unity, and the matrix can be diagonalized. If, however, $L < N$ we need $N - L$ more linearly independent vectors. These are the *generalized eigenvectors*. One can be obtained from

$$(\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g}_1 = \mathbf{e}, \tag{8.357}$$

and others from

$$(\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g}_{j+1} = \mathbf{g}_j \quad \text{for} \quad j = 1, 2, \ldots, N - L - 1. \tag{8.358}$$

This procedure is continued until N linearly independent eigenvectors and generalized eigenvectors are obtained, which is the most that we can have in $\mathbb{R}^N$. Then

$$\mathbf{S} = \begin{pmatrix} \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ \mathbf{e}_1 & \ldots & \mathbf{e}_L & \mathbf{g}_1 & \ldots & \mathbf{g}_{N-L} \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \end{pmatrix} \tag{8.359}$$

gives $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \mathbf{J}$, where $\mathbf{J}$ is of the Jordan canonical form.

Notice that $\mathbf{g}_n$ also satisfies $(\mathbf{A} - \lambda\mathbf{I})^n \cdot \mathbf{g}_n = \mathbf{0}$. For example, if

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g} &= \mathbf{e}, & (8.360)\\ (\mathbf{A} - \lambda\mathbf{I}) \cdot (\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g} &= (\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{e}, & (8.361)\\ (\mathbf{A} - \lambda\mathbf{I}) \cdot (\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g} &= \mathbf{0}, & (8.362)\\ (\mathbf{A} - \lambda\mathbf{I})^2 \cdot \mathbf{g} &= \mathbf{0}. & (8.363) \end{aligned}$$

However any solution of Eq. (8.363) is not necessarily a generalized eigenvector.

[5]Marie Ennemond Camille Jordan, 1838-1922, French mathematician.

Example 8.37

Find the Jordan canonical form of

$$\mathbf{A} = \begin{pmatrix} 4 & 1 & 3 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \tag{8.364}$$

The eigenvalues are $\lambda = 4$ with multiplicity three. For this value

$$(\mathbf{A} - \lambda \mathbf{I}) = \begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{8.365}$$

The eigenvectors are obtained from $(\mathbf{A} - \lambda \mathbf{I}) \cdot \mathbf{e}_1 = \mathbf{0}$, which gives $x_2 + 3x_3 = 0$, $x_3 = 0$. The most general form of the eigenvector is

$$\mathbf{e}_1 = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix}. \tag{8.366}$$

Only one eigenvector can be obtained from this eigenvalue. To get a generalized eigenvector, we take $(\mathbf{A} - \lambda \mathbf{I}) \cdot \mathbf{g}_1 = \mathbf{e}_1$, which gives $x_2 + 3x_3 = a$, $x_3 = 0$, so that

$$\mathbf{g}_1 = \begin{pmatrix} b \\ a \\ 0 \end{pmatrix}. \tag{8.367}$$

Another generalized eigenvector can be similarly obtained from $(\mathbf{A} - \lambda \mathbf{I}) \cdot \mathbf{g}_2 = \mathbf{g}_1$, so that $x_2 + 3x_3 = b$, $x_3 = a$. Thus, we get

$$\mathbf{g}_2 = \begin{pmatrix} c \\ b - 3a \\ a \end{pmatrix}. \tag{8.368}$$

From the eigenvector and generalized eigenvectors

$$\mathbf{S} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{e}_1 & \mathbf{g}_1 & \mathbf{g}_2 \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} a & b & c \\ 0 & a & b - 3a \\ 0 & 0 & a \end{pmatrix}, \tag{8.369}$$

and

$$\mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{a} & -\frac{b}{a^2} & \frac{-b^2+3ba+ca}{a^3} \\ 0 & \frac{1}{a} & \frac{-b+3a}{a^2} \\ 0 & 0 & \frac{1}{a} \end{pmatrix}. \tag{8.370}$$

The Jordan canonical form is

$$\mathbf{J} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \tag{8.371}$$

Note that in Eq. (8.370), a, b, and c are any constants. Choosing $a = 1$, $b = c = 0$, for example, simplifies the algebra giving

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}, \tag{8.372}$$

and

$$\mathbf{S}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.373}$$

8.8.7 Schur decomposition

The Schur[6] decomposition is as follows:

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{R} \cdot \mathbf{Q}^T. \tag{8.374}$$

Here $\mathbf{Q}$ is an orthogonal (unitary if complex) matrix, and $\mathbf{R}$ is upper triangular, with the eigenvalues this time along the diagonal. The matrix $\mathbf{A}$ must be square.

Example 8.38

The Schur decomposition of the matrix we diagonalized in a previous example, p. 364, is as follows:

$$\mathbf{A} = \underbrace{\begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix}}_{\mathbf{A}} = \mathbf{Q} \cdot \mathbf{R} \cdot \mathbf{Q}^T = \tag{8.375}$$

$$\underbrace{\begin{pmatrix} -0.4082 & 0.1826 & 0.8944 \\ -0.8165 & 0.3651 & -0.4472 \\ 0.4082 & 0.9129 & 0 \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} -6 & -20.1246 & 31.0376 \\ 0 & 3 & 5.7155 \\ 0 & 0 & 6 \end{pmatrix}}_{\mathbf{R}} \cdot \underbrace{\begin{pmatrix} -0.4082 & -0.8165 & 0.4082 \\ 0.1826 & 0.3651 & 0.9129 \\ 0.8944 & -0.4472 & 0 \end{pmatrix}}_{\mathbf{Q}^T}. \tag{8.376}$$

This decomposition was achieved with numerical software. This particular $\mathbf{Q}$ has $\det \mathbf{Q} = -1$, so if it were used in a coordinate transformation it would be volume-preserving but not orientation-preserving. Since the Schur decomposition is non-unique, it could be easily re-calculated if one also wanted to preserve orientation.

Example 8.39

The Schur decomposition of another matrix considered in earlier examples, see p. 348, is as follows:

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \tag{8.377}$$

$$= \underbrace{\begin{pmatrix} \frac{1+\sqrt{3}i}{2\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1-\sqrt{3}i}{2\sqrt{2}} \end{pmatrix}}_{\mathbf{U}} \cdot \underbrace{\begin{pmatrix} \frac{-1+\sqrt{3}i}{2} & \frac{-1+\sqrt{3}i}{2} \\ 0 & \frac{-1-\sqrt{3}i}{2} \end{pmatrix}}_{\mathbf{R}} \cdot \underbrace{\begin{pmatrix} \frac{1-\sqrt{3}i}{2\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1+\sqrt{3}i}{2\sqrt{2}} \end{pmatrix}}_{\mathbf{U}^H}. \tag{8.378}$$

[6] Issai Schur, 1875-1941, Belrussian-born German-based mathematician.

This is a non-unique decomposition. Unusually, the form given here is exact; most require numerical approximation. Note that because $\mathbf{R}$ has the eigenvalues of $\mathbf{A}$ on its diagonal, that we must consider complex unitary matrices. When this is recomposed, we recover the original real $\mathbf{A}$. Once again, we have $||\mathbf{R}||_2 = ||\mathbf{A}||_2 = 1.61803$. Here $\det \mathbf{U} = \det \mathbf{U}^H = 1$, so both are area- and orientation-preserving.

We can imagine the operation of $\mathbf{A}$ on a real vector $\mathbf{x}$ as an initial rotation into the complex plane effected by application of $\mathbf{U}^H$: $\mathbf{x}' = \mathbf{U}^H \cdot \mathbf{x}$. This is followed by an eigen-stretching effected by $\mathbf{R}$: $\mathbf{x}'' = \mathbf{R} \cdot \mathbf{x}'$. Application of $\mathbf{U}$ rotates back into the real plane: $\mathbf{x}''' = \mathbf{U} \cdot \mathbf{x}''$. The composite effect is $\mathbf{x}''' = \mathbf{U} \cdot \mathbf{R} \cdot \mathbf{U}^H \cdot \mathbf{x} = \mathbf{A} \cdot \mathbf{x}$.

If $\mathbf{A}$ is symmetric, then the upper triangular matrix $\mathbf{R}$ reduces to the diagonal matrix with eigenvalues on the diagonal, $\mathbf{\Lambda}$; the Schur decomposition is in this case simply $\mathbf{A} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^T$.

8.8.8 Singular value decomposition

The singular value decomposition (SVD) is used for non-square matrices and is the most general form of diagonalization. Any complex matrix $\mathbf{A}_{N\times M}$ can be factored into the form

$$\mathbf{A}_{N\times M} = \mathbf{Q}_{N\times N} \cdot \mathbf{B}_{N\times M} \cdot \mathbf{Q}^H_{M\times M}, \tag{8.379}$$

where $\mathbf{Q}_{N\times N}$ and $\mathbf{Q}^H_{M\times M}$ are orthogonal (unitary, if complex) matrices, and $\mathbf{B}$ has positive numbers μ_i, $(i = 1, 2, \ldots, r)$ in the first r positions on the main diagonal, and zero everywhere else. It turns out that r is the rank of $\mathbf{A}_{N\times M}$. The columns of $\mathbf{Q}_{N\times N}$ are the eigenvectors of $\mathbf{A}_{N\times M} \cdot \mathbf{A}^H_{N\times M}$. The columns of $\mathbf{Q}_{M\times M}$ are the eigenvectors of $\mathbf{A}^H_{N\times M} \cdot \mathbf{A}_{N\times M}$. The values μ_i, $(i = 1, 2, \ldots, r) \in \mathbb{R}^1$ are called the singular values of $\mathbf{A}$. They are analogous to eigenvalues and are in fact the positive square roots of the eigenvalues of $\mathbf{A}_{N\times M} \cdot \mathbf{A}^H_{N\times M}$ or $\mathbf{A}^H_{N\times M} \cdot \mathbf{A}_{N\times M}$. Note that since the matrix from which the eigenvalues are drawn is Hermitian, that the eigenvalues, and thus the singular values, are guaranteed real. Note also that if $\mathbf{A}$ itself is square and Hermitian, that the absolute value of the eigenvalues of $\mathbf{A}$ will equal its singular values. If $\mathbf{A}$ is square and non-Hermitian, there is no simple relation between its eigenvalues and singular values. The factorization $\mathbf{Q}_{N\times N} \cdot \mathbf{B}_{N\times M} \cdot \mathbf{Q}^H_{M\times M}$ is called the singular value decomposition.

As discussed by Strang, the column vectors of $\mathbf{Q}_{N\times N}$ and $\mathbf{Q}_{M\times M}$ are even more than orthonormal. They also must be chosen in such a way that $\mathbf{A}_{N\times M} \cdot \mathbf{Q}_{M\times M}$ is a scalar multiple of $\mathbf{Q}_{N\times N}$. This comes directly from post-multiplying the general form of the singular value decomposition by $\mathbf{Q}_{M\times M}$: $\mathbf{A}_{N\times M} \cdot \mathbf{Q}_{M\times M} = \mathbf{Q}_{N\times N} \cdot \mathbf{B}_{N\times M}$. So in fact a more robust way of computing the singular value decomposition is to first compute one of the orthogonal matrices, and then compute the other orthogonal matrix with which the first one is consistent.

Example 8.40

Find the singular value decomposition of the matrix from p. 366,

$$\mathbf{A}_{2\times 3} = \begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix}. \tag{8.380}$$

The matrix is real so we do not need to consider the conjugate transpose; we will retain the notation for generality though here the ordinary transpose would suffice. First consider $\mathbf{A} \cdot \mathbf{A}^H$:

$$\mathbf{A} \cdot \mathbf{A}^H = \underbrace{\begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & 2 \\ -3 & 0 \\ 2 & 3 \end{pmatrix}}_{\mathbf{A}^H} = \begin{pmatrix} 14 & 8 \\ 8 & 13 \end{pmatrix}. \tag{8.381}$$

The diagonal eigenvalue matrix and corresponding orthogonal matrix composed of the normalized eigenvectors in the columns are

$$\mathbf{\Lambda}_{2\times 2} = \begin{pmatrix} 21.5156 & 0 \\ 0 & 5.48439 \end{pmatrix}, \qquad \mathbf{Q}_{2\times 2} = \begin{pmatrix} 0.728827 & -0.684698 \\ 0.684698 & 0.728827 \end{pmatrix}. \tag{8.382}$$

Next we consider $\mathbf{A}^H \cdot \mathbf{A}$:

$$\mathbf{A}^H \cdot \mathbf{A} = \underbrace{\begin{pmatrix} 1 & 2 \\ -3 & 0 \\ 2 & 3 \end{pmatrix}}_{\mathbf{A}^H} \underbrace{\begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix}}_{\mathbf{A}} = \begin{pmatrix} 5 & -3 & 8 \\ -3 & 9 & -6 \\ 8 & -6 & 13 \end{pmatrix}. \tag{8.383}$$

The diagonal eigenvalue matrix and corresponding orthogonal matrix composed of the normalized eigenvectors in the columns are

$$\mathbf{\Lambda}_{3\times 3} = \begin{pmatrix} 21.52 & 0 & 0 \\ 0 & 5.484 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{Q}_{3\times 3} = \begin{pmatrix} 0.4524 & 0.3301 & -0.8285 \\ -0.4714 & 0.8771 & 0.09206 \\ 0.7571 & 0.3489 & 0.5523 \end{pmatrix}. \tag{8.384}$$

We take

$$\mathbf{B}_{2\times 3} = \begin{pmatrix} \sqrt{21.52} & 0 & 0 \\ 0 & \sqrt{5.484} & 0 \end{pmatrix} = \begin{pmatrix} 4.639 & 0 & 0 \\ 0 & 2.342 & 0 \end{pmatrix}, \tag{8.385}$$

and can easily verify that

$$\mathbf{Q}_{2\times 2} \cdot \mathbf{B}_{2\times 3} \cdot \mathbf{Q}_{3\times 3}^H = \underbrace{\begin{pmatrix} 0.7288 & -0.6847 \\ 0.6847 & 0.7288 \end{pmatrix}}_{\mathbf{Q}_{2\times 2}} \underbrace{\begin{pmatrix} 4.639 & 0 & 0 \\ 0 & 2.342 & 0 \end{pmatrix}}_{\mathbf{B}_{2\times 3}} \underbrace{\begin{pmatrix} 0.4524 & -0.4714 & 0.7571 \\ 0.3301 & 0.8771 & 0.3489 \\ -0.8285 & 0.09206 & 0.5523 \end{pmatrix}}_{\mathbf{Q}_{3\times 3}^H}, \tag{8.386}$$

$$= \begin{pmatrix} 1 & -3 & 2 \\ 2 & 0 & 3 \end{pmatrix} = \mathbf{A}_{2\times 3}. \tag{8.387}$$

The singular values here are $\mu_1 = 4.639$, $\mu_2 = 2.342$. As an aside, both $\det \mathbf{Q}_{2\times 2} = 1$ and $\det \mathbf{Q}_{3\times 3} = 1$, so they are orientation-preserving.

Let's see how we can get another singular value decomposition of the same matrix. Here we will employ the more robust technique of computing the decomposition. The orthogonal matrices $\mathbf{Q}_{3\times 3}$ and $\mathbf{Q}_{2\times 2}$ are not unique as one can multiply any row or column by -1 and still maintain orthonormality. For example, instead of the value found earlier, let us presume that we found

$$\mathbf{Q}_{3\times 3} = \begin{pmatrix} -0.4524 & 0.3301 & -0.8285 \\ 0.4714 & 0.8771 & 0.09206 \\ -0.7571 & 0.3489 & 0.5523 \end{pmatrix}. \tag{8.388}$$

Here, the first column of the original $\mathbf{Q}_{3\times3}$ has been multiplied by -1. If we used this new $\mathbf{Q}_{3\times3}$ in conjunction with the previously found matrices to form $\mathbf{Q}_{2\times2}\cdot\mathbf{A}_{2\times3}\cdot\mathbf{Q}_{3\times3}^H$, we would not recover $\mathbf{A}_{2\times3}$! The more robust way is to take

$$\mathbf{A}_{2\times3} = \mathbf{Q}_{2\times2}\cdot\mathbf{B}_{2\times3}\cdot\mathbf{Q}_{3\times3}^H, \tag{8.389}$$

$$\mathbf{A}_{2\times3}\cdot\mathbf{Q}_{3\times3} = \mathbf{Q}_{2\times2}\cdot\mathbf{B}_{2\times3}, \tag{8.390}$$

$$\underbrace{\begin{pmatrix}1 & -3 & 2\\ 2 & 0 & 3\end{pmatrix}}_{\mathbf{A}_{2\times3}}\underbrace{\begin{pmatrix}-0.4524 & 0.3301 & -0.8285\\ 0.4714 & 0.8771 & 0.09206\\ -0.7571 & 0.3489 & 0.5523\end{pmatrix}}_{\mathbf{Q}_{3\times3}} = \underbrace{\begin{pmatrix}q_{11} & q_{12}\\ q_{21} & q_{22}\end{pmatrix}}_{\mathbf{Q}_{2\times2}}\underbrace{\begin{pmatrix}4.639 & 0 & 0\\ 0 & 2.342 & 0\end{pmatrix}}_{\mathbf{B}_{2\times3}}, \tag{8.391}$$

$$\begin{pmatrix}-3.381 & -1.603 & 0\\ -3.176 & 1.707 & 0\end{pmatrix} = \begin{pmatrix}4.639q_{11} & 2.342q_{12} & 0\\ 4.639q_{21} & 2.342q_{22} & 0\end{pmatrix}. \tag{8.392}$$

Solving for q_{ij}, we find that

$$\mathbf{Q}_{2\times2} = \begin{pmatrix}-0.7288 & -0.6847\\ -0.6847 & 0.7288\end{pmatrix}. \tag{8.393}$$

It is easily seen that this version of $\mathbf{Q}_{2\times2}$ differs from the first version by a sign change in the first column. Direct substitution shows that the new decomposition also recovers $\mathbf{A}_{2\times3}$:

$$\mathbf{Q}_{2\times2}\cdot\mathbf{B}_{2\times3}\cdot\mathbf{Q}_{3\times3}^H = \underbrace{\begin{pmatrix}-0.7288 & -0.6847\\ -0.6847 & 0.7288\end{pmatrix}}_{\mathbf{Q}_{2\times2}}\underbrace{\begin{pmatrix}4.639 & 0 & 0\\ 0 & 2.342 & 0\end{pmatrix}}_{\mathbf{B}_{2\times3}}\underbrace{\begin{pmatrix}-0.4524 & 0.4714 & -0.7571\\ 0.3301 & 0.8771 & 0.3489\\ -0.8285 & 0.09206 & 0.5523\end{pmatrix}}_{\mathbf{Q}_{3\times3}^H}, \tag{8.394}$$

$$= \begin{pmatrix}1 & -3 & 2\\ 2 & 0 & 3\end{pmatrix} = \mathbf{A}_{2\times3}. \tag{8.395}$$

Both of the orthogonal matrices $\mathbf{Q}$ used in this section have determinant of -1, so they do not preserve orientation.

Example 8.41

The singular value decomposition of another matrix considered in earlier examples, p. 348, is as follows:

$$\mathbf{A} = \begin{pmatrix}0 & -1\\ 1 & -1\end{pmatrix}, \tag{8.396}$$

$$= \underbrace{\begin{pmatrix}\sqrt{\frac{2}{5+\sqrt{5}}} & -\sqrt{\frac{2}{5-\sqrt{5}}}\\ \sqrt{\frac{2}{5-\sqrt{5}}} & \sqrt{\frac{2}{5+\sqrt{5}}}\end{pmatrix}}_{\mathbf{Q}_2}\cdot\underbrace{\begin{pmatrix}\sqrt{\frac{1}{2}(3+\sqrt{5})} & 0\\ 0 & \sqrt{\frac{1}{2}(3-\sqrt{5})}\end{pmatrix}}_{\mathbf{B}}\cdot\underbrace{\begin{pmatrix}\sqrt{\frac{2}{5+\sqrt{5}}} & -\sqrt{\frac{2}{5-\sqrt{5}}}\\ \sqrt{\frac{2}{5-\sqrt{5}}} & \sqrt{\frac{2}{5+\sqrt{5}}}\end{pmatrix}}_{\mathbf{Q}_1^T}. \tag{8.397}$$

The singular value decomposition here is $\mathbf{A} = \mathbf{Q}_2\cdot\mathbf{B}\cdot\mathbf{Q}_1^T$. All matrices are 2×2, since $\mathbf{A}$ is square of dimension 2×2. Interestingly $\mathbf{Q}_2 = \mathbf{Q}_1^T$. Both induce a counterclockwise rotation of

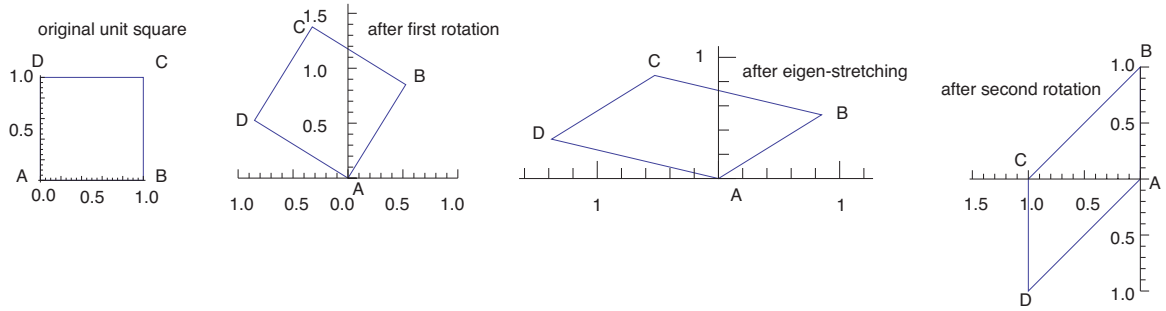

Figure 8.8: Unit square transforming via rotation, stretching, and rotation of the singular value decomposition under a linear area- and orientation-preserving alibi mapping.

$\alpha = \arcsin\sqrt{2/(5-\sqrt{5})} = \pi/3.0884 = 58.2°$. We also have $\det \mathbf{Q}_1 = \det \mathbf{Q}_2 = ||\mathbf{Q}_2||_2 = ||\mathbf{Q}_1||_2 = 1$. Thus, both are pure rotations. By inspection $||\mathbf{B}||_2 = ||\mathbf{A}||_2 = \sqrt{(3+\sqrt{5})/2} = 1.61803$.

The action of this composition of matrix operations on a unit square is depicted in Fig. 8.8. The first rotation is induced by $\mathbf{Q}_1^T$. This is followed by an eigen-stretching of $\mathbf{B}$. The action is completed by a rotation induced by $\mathbf{Q}_2$.

It is also easily shown that the singular values of a square Hermitian matrix are identical to the eigenvalues of that matrix. The singular values of a square non-Hermitian matrix are not, in general, the eigenvalues of that matrix.

8.8.9 Hessenberg form

A square matrix $\mathbf{A}$ can be decomposed into Hessenberg[7] form

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{H} \cdot \mathbf{Q}^T, \tag{8.398}$$

where $\mathbf{Q}$ is an orthogonal (or unitary) matrix and $\mathbf{H}$ has zeros below the first sub-diagonal. When $\mathbf{A}$ is Hermitian, $\mathbf{Q}$ is tridiagonal, which is very easy to invert numerically. Also $\mathbf{H}$ has the same eigenvalues as $\mathbf{A}$. Here the $\mathbf{H}$ of the Hessenberg form is not to be confused with the Hessian matrix, which often is denoted by the same symbol; see Eq. (1.283).

Example 8.42

The Hessenberg form of our example square matrix $\mathbf{A}$ from p. 364 is

$$\mathbf{A} = \begin{pmatrix} -5 & 4 & 9 \\ -22 & 14 & 18 \\ 16 & -8 & -6 \end{pmatrix} = \mathbf{Q} \cdot \mathbf{H} \cdot \mathbf{Q}^T = \tag{8.399}$$

[7]Karl Hessenberg, 1904-1959, German mathematician and engineer.

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.8087 & 0.5882 \\ 0 & 0.5882 & 0.8087 \end{pmatrix}}_{\mathbf{Q}} \underbrace{\begin{pmatrix} -5 & 2.0586 & 9.6313 \\ 27.2029 & 2.3243 & -24.0451 \\ 0 & 1.9459 & 5.6757 \end{pmatrix}}_{\mathbf{H}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.8087 & 0.5882 \\ 0 & 0.5882 & 0.8087 \end{pmatrix}}_{\mathbf{Q}^T}. \tag{8.400}$$

The matrix $\mathbf{Q}$ found here has determinant of -1; it could be easily recalculated to arrive at an orientation-preserving value of $+1$.

8.9 Projection matrix

Here we consider a topic discussed earlier in a broader context, the projection matrix defined in Eq. (7.160). The vector $\mathbf{A}\cdot\mathbf{x}$ belongs to the column space of $\mathbf{A}$. Here $\mathbf{A}$ is not necessarily square. Consider the equation $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are given. If the given vector $\mathbf{b}$ does not lie in the column space of $\mathbf{A}$, the equation cannot be solved for $\mathbf{x}$. Still, we would like to find $\mathbf{x}_p$ such that

$$\mathbf{A}\cdot\mathbf{x}_p=\mathbf{b}_p, \tag{8.401}$$

which does lie in the column space of $\mathbf{A}$, such that $\mathbf{b}_p$ is the projection of $\mathbf{b}$ onto the column space. The residual vector from Eq. (8.4) is also expressed as

$$\mathbf{r}=\mathbf{b}_p-\mathbf{b}. \tag{8.402}$$

For a projection, this residual should be orthogonal to all vectors $\mathbf{A}\cdot\mathbf{z}$ which belong to the column space, where the components of $\mathbf{z}$ are arbitrary. Enforcing this condition, we get

$$\begin{aligned} \mathbf{0} &= (\mathbf{A}\cdot\mathbf{z})^T\cdot\mathbf{r}, && (8.403)\\ &= (\mathbf{A}\cdot\mathbf{z})^T\cdot\underbrace{(\mathbf{b}_p-\mathbf{b})}_{\mathbf{r}}, && (8.404)\\ &= \mathbf{z}^T\cdot\mathbf{A}^T\cdot(\underbrace{\mathbf{A}\cdot\mathbf{x}_p}_{\mathbf{b}_p}-\mathbf{b}), && (8.405)\\ &= \mathbf{z}^T\cdot(\mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x}_p-\mathbf{A}^T\cdot\mathbf{b}). && (8.406) \end{aligned}$$

Since $\mathbf{z}$ is an arbitrary vector,

$$\mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x}_p-\mathbf{A}^T\cdot\mathbf{b}=\mathbf{0}. \tag{8.407}$$

from which

$$\begin{aligned} \mathbf{A}^T\cdot\mathbf{A}\cdot\mathbf{x}_p &= \mathbf{A}^T\cdot\mathbf{b}, && (8.408)\\ \mathbf{x}_p &= (\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T\cdot\mathbf{b}, && (8.409)\\ \mathbf{A}\cdot\mathbf{x}_p &= \mathbf{A}\cdot(\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T\cdot\mathbf{b}, && (8.410)\\ \mathbf{b}_p &= \underbrace{\mathbf{A}\cdot(\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T}_{\equiv\mathbf{P}}\cdot\mathbf{b}. && (8.411) \end{aligned}$$

This is equivalent to that given in Eq. (7.160). The projection matrix $\mathbf{P}$ defined by $\mathbf{b}_p = \mathbf{P}\cdot\mathbf{b}$ is

$$\mathbf{P} = \mathbf{A}\cdot(\mathbf{A}^T\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^T. \tag{8.412}$$

The projection matrix for an operator $\mathbf{A}$, when operating on an arbitrary vector $\mathbf{b}$ yields the projection of $\mathbf{b}$ onto the column space of $\mathbf{A}$. Note that many vectors $\mathbf{b}$ could have the same projection onto the column space of $\mathbf{A}$. It can be shown that an $N\times N$ matrix $\mathbf{P}$ is a projection matrix iff $\mathbf{P}\cdot\mathbf{P} = \mathbf{P}$. Because of this, the projection matrix is idempotent: $\mathbf{P}\cdot\mathbf{x} = \mathbf{P}\cdot\mathbf{P}\cdot\mathbf{x} = \ldots = \mathbf{P}^n\cdot\mathbf{x}$. Moreover, the rank of $\mathbf{P}$ is its trace.

Example 8.43

Determine and analyze the projection matrix associated with projecting a vector $\mathbf{b}\in\mathbb{R}^3$ onto the two-dimensional space spanned by the basis vectors $(1,2,3)^T$ and $(1,1,1)^T$.

We form the matrix $\mathbf{A}$ by populating its columns with the basis vectors. So

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}. \tag{8.413}$$

The we find the projection matrix $\mathbf{P}$ via Eq. (8.412):

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}\cdot\left(\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}\cdot\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}\right)^{-1}\cdot\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{pmatrix}. \tag{8.414}$$

By inspection $\mathbf{P}$ is self-adjoint, thus it is guaranteed to possess real eigenvalues, which are $\lambda = 1, 1, 0$. There is one non-zero eigenvalue for each of the two linearly independent basis vectors which form $\mathbf{A}$. It is easily shown that $||\mathbf{P}||_2 = 1$, $\rho(\mathbf{P}) = 1$, and $\det\mathbf{P} = 0$. Thus, $\mathbf{P}$ is singular. This is because it maps vectors in three space to two space. The rank of $\mathbf{P}$ is 2 as is its trace. Note that, as required of all projection matrices, $\mathbf{P}\cdot\mathbf{P} = \mathbf{P}$:

$$\begin{pmatrix} \frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{pmatrix}\cdot\begin{pmatrix} \frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{pmatrix} = \begin{pmatrix} \frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{pmatrix}. \tag{8.415}$$

That is to say, $\mathbf{P}$ is idempotent.

It is easily shown the singular value decomposition of $\mathbf{P}$ is equivalent to a diagonalization, giving

$$\mathbf{P} = \mathbf{Q}\cdot\mathbf{\Lambda}\cdot\mathbf{Q}^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & -\sqrt{\frac{2}{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \end{pmatrix}\cdot\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\cdot\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{pmatrix}. \tag{8.416}$$

The matrix $\mathbf{Q}$ has $||\mathbf{Q}||_2 = 1$ and $\det\mathbf{Q} = 1$, so it is a true rotation. Thus, when $\mathbf{P}$ is applied to a vector $\mathbf{b}$ to obtain $\mathbf{b}_p$, we can consider $\mathbf{b}$ to be first rotated into the configuration aligned with the two basis vectors via application of $\mathbf{Q}^T$. Then in this configuration, one of the modes of $\mathbf{b}$ is suppressed via application of $\mathbf{\Lambda}$, while the other two modes are preserved. The result is returned to its original configuration via application of $\mathbf{Q}$, which precisely provides a counter-rotation to $\mathbf{Q}^T$. Note also that the decomposition is equivalent to that previously discussed on p. 372; here, $\mathbf{\Lambda} = \mathbf{R}\cdot\left(\mathbf{R}^T\cdot\mathbf{R}\right)^{-1}\cdot\mathbf{R}^T$, where $\mathbf{R}$ is as was defined on p. 372. Note specifically that $\mathbf{P}$ has rank $r = 2$ and that $\mathbf{\Lambda}$ has $r = 2$ values of unity on its diagonal.

8.10 Method of least squares

One important application of projection matrices is the method of least squares. This method is often used to fit data to a given functional form. The form is most often in terms of polynomials, but there is absolutely no restriction; trigonometric functions, logarithmic functions, Bessel functions can all serve as well. Now if one has say, ten data points, one can in principle, find a ninth order polynomial which will pass through all the data points. Often times, especially when there is much experimental error in the data, such a function may be subject to wild oscillations, which are unwarranted by the underlying physics, and thus is not useful as a predictive tool. In such cases, it may be more useful to choose a lower order curve which does not exactly pass through all experimental points, but which does minimize the residual.

In this method, one

- examines the data,
- makes a non-unique judgment of what the functional form might be,
- substitutes each data point into the assumed form so as to form an over-constrained system of linear equations, and
- uses the technique associated with projection matrices to solve for the coefficients which best represent the given data.

8.10.1 Unweighted least squares

This is the most common method used when one has equal confidence in all the data.

Example 8.44

Find the best straight line to approximate the measured data relating x to t.

t	x
0	5
1	7
2	10
3	12
6	15

(8.417)

A straight line fit will have the form

$$x = a_0 + a_1 t, \tag{8.418}$$

where a_0 and a_1 are the terms to be determined. Substituting each data point to the assumed form, we get five equations in two unknowns:

$$5 = a_0 + 0a_1, \tag{8.419}$$

$$\begin{aligned} 7 &= a_0 + 1a_1, && (8.420)\\ 10 &= a_0 + 2a_1, && (8.421)\\ 12 &= a_0 + 3a_1, && (8.422)\\ 15 &= a_0 + 6a_1. && (8.423) \end{aligned}$$

Rearranging, we get

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \\ 10 \\ 12 \\ 15 \end{pmatrix}. \tag{8.424}$$

This is of the form $\mathbf{A} \cdot \mathbf{a} = \mathbf{b}$. We then find that

$$\mathbf{a} = \left(\mathbf{A}^T \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}^T \cdot \mathbf{b}. \tag{8.425}$$

Substituting, we find that

$$\underbrace{\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}}_{\mathbf{a}} = \left[\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \end{pmatrix}}_{\mathbf{A}^T} \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \end{pmatrix}}_{\mathbf{A}} \right]^{-1} \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \end{pmatrix}}_{\mathbf{A}^T} \underbrace{\begin{pmatrix} 5 \\ 7 \\ 10 \\ 12 \\ 15 \end{pmatrix}}_{\mathbf{b}} = \begin{pmatrix} 5.7925 \\ 1.6698 \end{pmatrix}. \tag{8.426}$$

So the best fit estimate is

$$x = 5.7925 + 1.6698\ t. \tag{8.427}$$

The Euclidean norm of the residual is $||\mathbf{A} \cdot \mathbf{a} - \mathbf{b}||_2 = 1.9206$. This represents the ℓ_2 residual of the prediction. A plot of the raw data and the best fit straight line is shown in Fig. 8.9.

8.10.2 Weighted least squares

If one has more confidence in some data points than others, one can define a weighting function to give more priority to those particular data points.

Example 8.45

Find the best straight line fit for the data in the previous example. Now however, assume that we have five times the confidence in the accuracy of the final two data points, relative to the other points. Define a square weighting matrix $\mathbf{W}$:

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}. \tag{8.428}$$

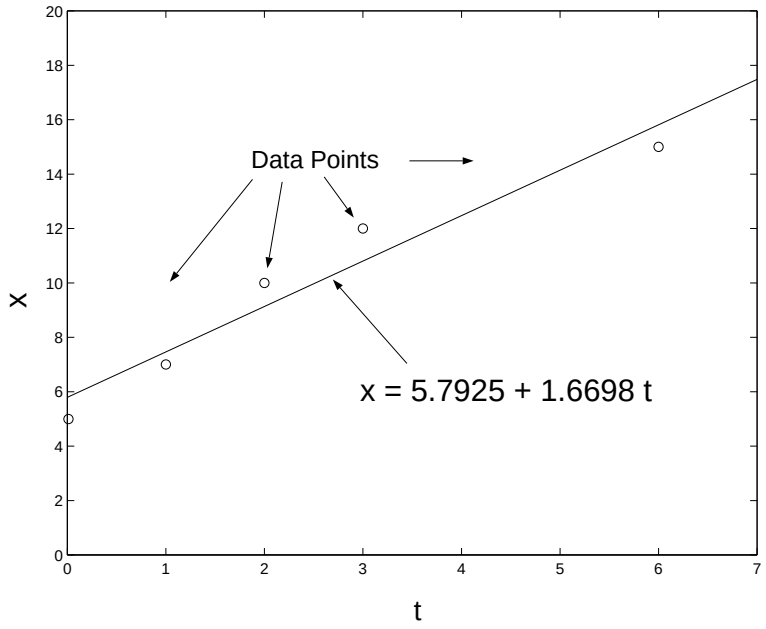

Figure 8.9: Plot of $x - t$ data and best least squares straight line fit.

Now we perform the following operations:

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{b}, \tag{8.429}$$

$$\mathbf{W} \cdot \mathbf{A} \cdot \mathbf{a} = \mathbf{W} \cdot \mathbf{b}, \tag{8.430}$$

$$(\mathbf{W} \cdot \mathbf{A})^T \cdot \mathbf{W} \cdot \mathbf{A} \cdot \mathbf{a} = (\mathbf{W} \cdot \mathbf{A})^T \cdot \mathbf{W} \cdot \mathbf{b}, \tag{8.431}$$

$$\mathbf{a} = \left((\mathbf{W} \cdot \mathbf{A})^T \cdot \mathbf{W} \cdot \mathbf{A} \right)^{-1} (\mathbf{W} \cdot \mathbf{A})^T \cdot \mathbf{W} \cdot \mathbf{b}. \tag{8.432}$$

With values of $\mathbf{W}$ from Eq. (8.428), direct substitution leads to

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 8.0008 \\ 1.1972 \end{pmatrix}. \tag{8.433}$$

So the best weighted least squares fit is

$$x = 8.0008 + 1.1972\ t. \tag{8.434}$$

A plot of the raw data and the best fit straight line is shown in Fig. 8.10.

When the measurements are independent and equally reliable, $\mathbf{W}$ is the identity matrix. If the measurements are independent but not equally reliable, $\mathbf{W}$ is at most diagonal. If the measurements are not independent, then non-zero terms can appear off the diagonal in $\mathbf{W}$. It is often advantageous, for instance in problems in which one wants to control a process in real time, to give priority to recent data estimates over old data estimates and to continually employ a least squares technique to estimate future system behavior. The previous example does just that. A famous fast algorithm for such problems is known as a *Kalman*[8] *Filter*.

[8]Rudolf Emil Kálmán, 1930-, Hungarian/American electrical engineer.

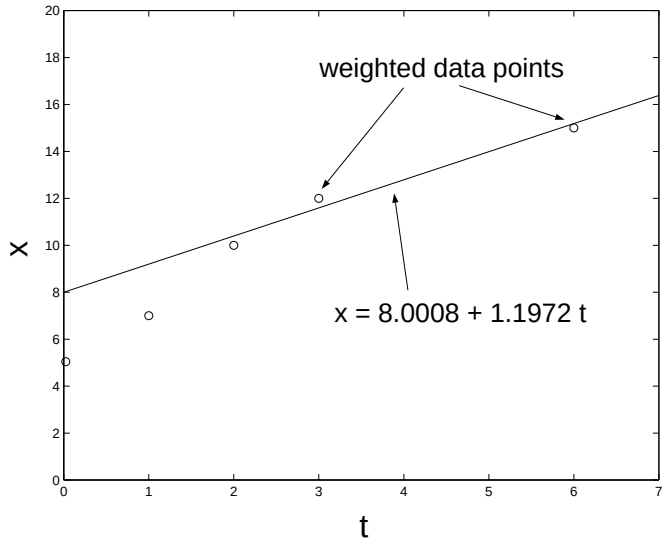

Figure 8.10: Plot of $x - t$ data and best weighted least squares straight line fit.

8.11 Matrix exponential

Definition: The exponential matrix is defined as

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \cdots \tag{8.435}$$

Thus

$$\begin{aligned} e^{\mathbf{A}t} &= \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^3 + \cdots, & (8.436)\\ \frac{d}{dt}\left(e^{\mathbf{A}t}\right) &= \mathbf{A} + \mathbf{A}^2t + \frac{1}{2!}\mathbf{A}^3t^2 + \cdots, & (8.437)\\ &= \mathbf{A}\cdot\underbrace{\left(\mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^3 + \cdots\right)}_{=e^{\mathbf{A}t}}, & (8.438)\\ &= \mathbf{A}\cdot e^{\mathbf{A}t}. & (8.439) \end{aligned}$$

Properties of the matrix exponential include

$$\begin{aligned} e^{a\mathbf{I}} &= e^a\mathbf{I}, & (8.440)\\ (e^{\mathbf{A}})^{-1} &= e^{-\mathbf{A}}, & (8.441)\\ e^{\mathbf{A}(t+s)} &= e^{\mathbf{A}t}e^{\mathbf{A}s}. & (8.442) \end{aligned}$$

But $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ only if $\mathbf{A}\cdot\mathbf{B} = \mathbf{B}\cdot\mathbf{A}$. Thus, $e^{t\mathbf{I}+s\mathbf{A}} = e^te^{s\mathbf{A}}$.

Example 8.46

Find $e^{\mathbf{A}t}$ if

$$\mathbf{A} = \begin{pmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{pmatrix}. \tag{8.443}$$

We have

$$\mathbf{A} = a\mathbf{I} + \mathbf{B}, \tag{8.444}$$

where

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{8.445}$$

Thus

$$\mathbf{B}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{8.446}$$

$$\mathbf{B}^3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{8.447}$$

$$\vdots \tag{8.448}$$

$$\mathbf{B}^n = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ for } n \geq 4. \tag{8.449}$$

$$\tag{8.450}$$

Furthermore

$$\mathbf{I} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{I} = \mathbf{B}. \tag{8.451}$$

Thus

$$\begin{aligned} e^{\mathbf{A}t} &= e^{(a\mathbf{I}+\mathbf{B})t}, && (8.452)\\ &= e^{at\mathbf{I}} \cdot e^{\mathbf{B}t}, && (8.453)\\ &= \left(\underbrace{\mathbf{I} + at\mathbf{I} + \frac{1}{2!}a^2t^2\mathbf{I}^2 + \frac{1}{3!}a^3t^3\mathbf{I}^3 + \cdots}_{=e^{at\mathbf{I}}=e^{at}\mathbf{I}} \right) \cdot \left(\underbrace{\mathbf{I} + \mathbf{B}t + \frac{1}{2!}\mathbf{B}^2t^2 + \overbrace{\frac{1}{3!}\mathbf{B}^3t^3 + \cdots}^{=0}}_{=e^{\mathbf{B}t}} \right), && (8.454)\\ &= e^{at}\mathbf{I} \cdot \left(\mathbf{I} + \mathbf{B}t + \mathbf{B}^2\frac{t^2}{2} \right), && (8.455)\\ &= e^{at} \begin{pmatrix} 1 & t & \frac{t^2}{2}, \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}. && (8.456) \end{aligned}$$

If $\mathbf{A}$ can be diagonalized, the calculation is simplified. Then

$$e^{\mathbf{A}t} = e^{\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}t} = \mathbf{I} + \mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}t + \ldots + \frac{1}{N!}\left(\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}t\right)^N. \tag{8.457}$$

Noting that

$$\left(\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}\right)^2 = \mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}\cdot\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1} = \mathbf{S}\cdot\mathbf{\Lambda}^2\cdot\mathbf{S}^{-1}, \tag{8.458}$$

$$\left(\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}\right)^N = \mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1}\cdot\ldots\cdot\mathbf{S}\cdot\mathbf{\Lambda}\cdot\mathbf{S}^{-1} = \mathbf{S}\cdot\mathbf{\Lambda}^N\cdot\mathbf{S}^{-1}, \tag{8.459}$$

the original expression reduces to

$$e^{\mathbf{A}t} = \mathbf{S}\cdot\left(\mathbf{I} + \mathbf{\Lambda}t + \ldots + \frac{1}{N!}\left(\mathbf{\Lambda}^N t^N\right)\right)\cdot\mathbf{S}^{-1}, \tag{8.460}$$

$$= \mathbf{S}\cdot e^{\mathbf{\Lambda}t}\cdot\mathbf{S}^{-1}. \tag{8.461}$$

8.12 Quadratic form

At times one may be given a polynomial equation for which one wants to determine conditions under which the expression is positive. For example if we have

$$f(\xi_1,\xi_2,\xi_3) = 18\xi_1^2 - 16\xi_1\xi_2 + 5\xi_2^2 + 12\xi_1\xi_3 - 4\xi_2\xi_3 + 6\xi_3^2, \tag{8.462}$$

it is not obvious whether or not there exist (ξ_1,ξ_2,ξ_3) which will give positive or negative values of f. However, it is easily verified that f can be rewritten as

$$f(\xi_1,\xi_2,\xi_3) = 2(\xi_1-\xi_2+\xi_3)^2 + 3(2\xi_1-\xi_2)^2 + 4(\xi_1+\xi_3)^2. \tag{8.463}$$

So in this case $f \geq 0$ for all (ξ_1,ξ_2,ξ_3). How to demonstrate positivity (or non-positivity) of such expressions is the topic of this section. A *quadratic form* is an expression

$$f(\xi_1,\cdots,\xi_N) = \sum_{j=1}^{N}\sum_{i=1}^{N} a_{ij}\xi_i\xi_j, \tag{8.464}$$

where $\{a_{ij}\}$ is a real, symmetric matrix which we will also call $\mathbf{A}$. The surface represented by the equation $\sum_{j=1}^{N}\sum_{i=1}^{N} a_{ij}\xi_i\xi_j =$ constant is a *quadric* surface. With the coefficient matrix defined, we can represent f as

$$f = \boldsymbol{\xi}^T\cdot\mathbf{A}\cdot\boldsymbol{\xi}. \tag{8.465}$$

Now, by Eq. (8.321), $\mathbf{A}$ can be decomposed as $\mathbf{Q}\cdot\mathbf{\Lambda}\cdot\mathbf{Q}^{-1}$, where $\mathbf{Q}$ is the orthogonal matrix populated by the orthonormalized eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is the corresponding diagonal matrix of eigenvalues. Thus, Eq. (8.465) becomes

$$f = \boldsymbol{\xi}^T\cdot\underbrace{\mathbf{Q}\cdot\mathbf{\Lambda}\cdot\mathbf{Q}^{-1}}_{\mathbf{A}}\cdot\boldsymbol{\xi}. \tag{8.466}$$

Since $\mathbf{Q}$ is orthogonal, $\mathbf{Q}^T = \mathbf{Q}^{-1}$, and we find

$$f = \boldsymbol{\xi}^T \cdot \mathbf{Q} \cdot \boldsymbol{\Lambda} \cdot \mathbf{Q}^T \cdot \boldsymbol{\xi}. \tag{8.467}$$

Now, define $\mathbf{x}$ so that $\mathbf{x} = \mathbf{Q}^T \cdot \boldsymbol{\xi} = \mathbf{Q}^{-1} \cdot \boldsymbol{\xi}$. Consequently, $\boldsymbol{\xi} = \mathbf{Q} \cdot \mathbf{x}$. Thus, Eq. (8.467) becomes

$$\begin{aligned} f &= (\mathbf{Q} \cdot \mathbf{x})^T \cdot \mathbf{Q} \cdot \boldsymbol{\Lambda} \cdot \mathbf{x}, && (8.468)\\ &= \mathbf{x}^T \cdot \mathbf{Q}^T \cdot \mathbf{Q} \cdot \boldsymbol{\Lambda} \cdot \mathbf{x}, && (8.469)\\ &= \mathbf{x}^T \cdot \mathbf{Q}^{-1} \cdot \mathbf{Q} \cdot \boldsymbol{\Lambda} \cdot \mathbf{x}, && (8.470)\\ &= \mathbf{x}^T \cdot \boldsymbol{\Lambda} \cdot \mathbf{x}. && (8.471) \end{aligned}$$

This *standard form* of a quadratic form is one in which the cross-product terms (i.e. $\xi_i \xi_j$, $i \neq j$) do not appear.

Theorem

(Principal axis theorem) If $\mathbf{Q}$ is the orthogonal matrix and $\lambda_1, \cdots, \lambda_N$ the eigenvalues corresponding to $\{a_{ij}\}$, a change in coordinates

$$\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_N \end{pmatrix} = \mathbf{Q} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \tag{8.472}$$

will reduce the quadratic form, Eq. (8.464), to its standard quadratic form

$$f(x_1, \ldots, x_N) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_N x_N^2. \tag{8.473}$$

It is perhaps better to consider this as an alias rather than an alibi transformation.

Example 8.47

Change

$$f(\xi_1, \xi_2) = 2\xi_1^2 + 2\xi_1\xi_2 + 2\xi_2^2, \tag{8.474}$$

to a standard quadratic form.

For $N = 2$, Eq. (8.464) becomes

$$f(\xi_1, \xi_2) = a_{11}\xi_1^2 + (a_{12} + a_{21})\xi_1\xi_2 + a_{22}\xi_2^2. \tag{8.475}$$

We choose $\{a_{ij}\}$ such that the matrix is symmetric. This gives us

$$\begin{aligned} a_{11} &= 2, && (8.476)\\ a_{12} &= 1, && (8.477)\\ a_{21} &= 1, && (8.478)\\ a_{22} &= 2. && (8.479) \end{aligned}$$

So we get

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{8.480}$$

The eigenvalues of $\mathbf{A}$ are $\lambda = 1$, $\lambda = 3$. The orthogonal matrix corresponding to $\mathbf{A}$ is

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \mathbf{Q}^{-1} = \mathbf{Q}^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{8.481}$$

The transformation $\boldsymbol{\xi} = \mathbf{Q} \cdot \mathbf{x}$ is

$$\xi_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \tag{8.482}$$

$$\xi_2 = \frac{1}{\sqrt{2}}(-x_1 + x_2). \tag{8.483}$$

We have $\det \mathbf{Q} = 1$, so the transformation is orientation-preserving. The inverse transformation $\mathbf{x} = \mathbf{Q}^{-1} \cdot \boldsymbol{\xi} = \mathbf{Q}^T \cdot \boldsymbol{\xi}$ is

$$x_1 = \frac{1}{\sqrt{2}}(\xi_1 - \xi_2), \tag{8.484}$$

$$x_2 = \frac{1}{\sqrt{2}}(\xi_1 + \xi_2). \tag{8.485}$$

Using Eqs. (8.482,8.483) to eliminate ξ_1 and ξ_2 in Eq. (8.474), we get a result in the form of Eq. (8.473):

$$f(x_1, x_2) = x_1^2 + 3x_2^2. \tag{8.486}$$

In terms of the original variables, we get

$$f(\xi_1, \xi_2) = \frac{1}{2}(\xi_1 - \xi_2)^2 + \frac{3}{2}(\xi_1 + \xi_2)^2. \tag{8.487}$$

Example 8.48

Change

$$f(\xi_1, \xi_2, \xi_3) = 18\xi_1^2 - 16\xi_1\xi_2 + 5\xi_2^2 + 12\xi_1\xi_3 - 4\xi_2\xi_3 + 6\xi_3^2, \tag{8.488}$$

to a standard quadratic form.

For $N = 3$, Eq. (8.464) becomes

$$f(\xi_1, \xi_2, \xi_3) = \begin{pmatrix} \xi_1 & \xi_2 & \xi_3 \end{pmatrix} \begin{pmatrix} 18 & -8 & 6 \\ -8 & 5 & -2 \\ 6 & -2 & 6 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \boldsymbol{\xi}^T \cdot \mathbf{A} \cdot \boldsymbol{\xi}. \tag{8.489}$$

The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1, \lambda_2 = 4, \lambda_3 = 24$. The orthogonal matrix corresponding to $\mathbf{A}$ is

$$\mathbf{Q} = \begin{pmatrix} -\frac{4}{\sqrt{69}} & -\frac{1}{\sqrt{30}} & \frac{13}{\sqrt{230}} \\ -\frac{7}{\sqrt{69}} & \sqrt{\frac{2}{15}} & -3\sqrt{\frac{2}{115}} \\ \frac{2}{\sqrt{69}} & \sqrt{\frac{5}{6}} & \sqrt{\frac{5}{46}} \end{pmatrix}, \qquad \mathbf{Q}^{-1} = \mathbf{Q}^T = \begin{pmatrix} -\frac{4}{\sqrt{69}} & -\frac{7}{\sqrt{69}} & \frac{2}{\sqrt{69}} \\ -\frac{1}{\sqrt{30}} & \sqrt{\frac{2}{15}} & \sqrt{\frac{5}{6}} \\ \frac{13}{\sqrt{230}} & -3\sqrt{\frac{2}{115}} & \sqrt{\frac{5}{46}} \end{pmatrix}. \tag{8.490}$$

For this non-unique choice of $\mathbf{Q}$, we note that $\det \mathbf{Q} = -1$, so it fails to satisfy the requirements of a right-handed coordinate system. For the purposes of this particular problem, this fact has no consequence. The inverse transformation $\mathbf{x} = \mathbf{Q}^{-1} \cdot \boldsymbol{\xi} = \mathbf{Q}^T \cdot \boldsymbol{\xi}$ is

$$x_1 = \frac{-4}{\sqrt{69}}\xi_1 - \frac{7}{\sqrt{69}}\xi_2 + \frac{2}{\sqrt{69}}\xi_3, \tag{8.491}$$

$$x_2 = -\frac{1}{\sqrt{30}}\xi_1 + \sqrt{\frac{2}{15}}\xi_2 + \sqrt{\frac{5}{6}}\xi_3, \tag{8.492}$$

$$x_3 = \frac{13}{\sqrt{230}}\xi_1 - 3\sqrt{\frac{2}{115}}\xi_2 + \sqrt{\frac{5}{46}}\xi_3. \tag{8.493}$$

Directly imposing then the standard quadratic form of Eq. (8.473) onto Eq. (8.488), we get

$$f(x_1, x_2, x_3) = x_1^2 + 4x_2^2 + 24x_3^2. \tag{8.494}$$

In terms of the original variables, we get

$$\begin{aligned} f(\xi_1, \xi_2, \xi_3) = & \left(\frac{-4}{\sqrt{69}}\xi_1 - \frac{7}{\sqrt{69}}\xi_2 + \frac{2}{\sqrt{69}}\xi_3\right)^2 \\ & +4\left(-\frac{1}{\sqrt{30}}\xi_1 + \sqrt{\frac{2}{15}}\xi_2 + \sqrt{\frac{5}{6}}\xi_3\right)^2 \\ & +24\left(\frac{13}{\sqrt{230}}\xi_1 - 3\sqrt{\frac{2}{115}}\xi_2 + \sqrt{\frac{5}{46}}\xi_3\right)^2. \end{aligned} \tag{8.495}$$

It is clear that $f(\xi_1, \xi_2, \xi_3)$ is positive definite. Moreover, by performing the multiplications, it is easily seen that the original form is recovered. Further manipulation would also show that

$$f(\xi_1, \xi_2, \xi_3) = 2(\xi_1 - \xi_2 + \xi_3)^2 + 3(2\xi_1 - \xi_2)^2 + 4(\xi_1 + \xi_3)^2, \tag{8.496}$$

so we see the particular quadratic form is not unique.

8.13 Moore-Penrose inverse

We seek the Moore-Penrose[9] inverse: $\mathbf{A}^+_{M\times N}$ such that the following four conditions are satisfied

$$\mathbf{A}_{N\times M} \cdot \mathbf{A}^+_{M\times N} \cdot \mathbf{A}_{N\times M} = \mathbf{A}_{N\times M}, \tag{8.497}$$

$$\mathbf{A}^+_{M\times N} \cdot \mathbf{A}_{N\times M} \cdot \mathbf{A}^+_{M\times N} = \mathbf{A}^+_{M\times N}, \tag{8.498}$$

$$\left(\mathbf{A}_{N\times M} \cdot \mathbf{A}^+_{M\times N}\right)^H = \mathbf{A}_{N\times M} \cdot \mathbf{A}^+_{M\times N}, \tag{8.499}$$

$$\left(\mathbf{A}^+_{M\times N} \cdot \mathbf{A}_{N\times M}\right)^H = \mathbf{A}^+_{M\times N} \cdot \mathbf{A}_{N\times M}. \tag{8.500}$$

[9] after Eliakim Hastings Moore, 1862-1932, American mathematician, and Sir Roger Penrose, 1931-, English mathematician. It is also credited to Arne Bjerhammar, 1917-2011, Swedish geodesist.

This will be achieved if we define

$$\mathbf{A}^+_{M\times N} = \mathbf{Q}_{M\times M}\cdot\mathbf{B}^+_{M\times N}\cdot\mathbf{Q}^H_{N\times N}. \tag{8.501}$$

The matrix $\mathbf{B}^+$ is $M\times N$ with μ_i^{-1}, $(i=1,2,\ldots)$ in the first r positions on the main diagonal. This is closely related to the $N\times M$ matrix $\mathbf{B}$, defined in Sec. 8.8.8, having μ_i on the same diagonal positions. The Moore-Penrose inverse, $\mathbf{A}^+_{M\times N}$, is also known as the *pseudo-inverse*. This is because in the special case in which $N\leq M$ and $N=r$ that it can be shown that

$$\mathbf{A}_{N\times M}\cdot\mathbf{A}^+_{M\times N} = \mathbf{I}_{N\times N}. \tag{8.502}$$

Let's check this with our definitions for the case when $N\leq M$, $N=r$.

$$\begin{aligned}
\mathbf{A}_{N\times M}\cdot\mathbf{A}^+_{M\times N} &= \left(\mathbf{Q}_{N\times N}\cdot\mathbf{B}_{N\times M}\cdot\mathbf{Q}^H_{M\times M}\right)\cdot\left(\mathbf{Q}_{M\times M}\cdot\mathbf{B}^+_{M\times N}\cdot\mathbf{Q}^H_{N\times N}\right), && (8.503)\\
&= \mathbf{Q}_{N\times N}\cdot\mathbf{B}_{N\times M}\cdot\mathbf{Q}^{-1}_{M\times M}\cdot\mathbf{Q}_{M\times M}\cdot\mathbf{B}^+_{M\times N}\cdot\mathbf{Q}^H_{N\times N}, && (8.504)\\
&= \mathbf{Q}_{N\times N}\cdot\mathbf{B}_{N\times M}\cdot\mathbf{B}^+_{M\times N}\cdot\mathbf{Q}^H_{N\times N}, && (8.505)\\
&= \mathbf{Q}_{N\times N}\cdot\mathbf{I}_{N\times N}\cdot\mathbf{Q}^H_{N\times N}, && (8.506)\\
&= \mathbf{Q}_{N\times N}\cdot\mathbf{Q}^H_{N\times N}, && (8.507)\\
&= \mathbf{Q}_{N\times N}\cdot\mathbf{Q}^{-1}_{N\times N}, && (8.508)\\
&= \mathbf{I}_{N\times N}. && (8.509)
\end{aligned}$$

We note for this special case that precisely because of the way we defined $\mathbf{B}^+$ that $\mathbf{B}_{N\times M}\cdot\mathbf{B}^+_{M\times N}=\mathbf{I}_{N\times N}$. When $N>M$, $\mathbf{B}_{N\times M}\cdot\mathbf{B}^+_{M\times N}$ yields a matrix with r ones on the diagonal and zeros elsewhere.

Example 8.49

Find the Moore-Penrose inverse, $\mathbf{A}^+_{3\times 2}$, of $\mathbf{A}_{2\times 3}$ from the matrix of a previous example, p. 366:

$$\mathbf{A}_{2\times 3} = \begin{pmatrix} 1 & -3 & 2\\ 2 & 0 & 3\end{pmatrix}. \tag{8.510}$$

$$\mathbf{A}^+_{3\times 2} = \mathbf{Q}_{3\times 3}\cdot\mathbf{B}^+_{3\times 2}\cdot\mathbf{Q}^H_{2\times 2}, \tag{8.511}$$

$$\mathbf{A}^+_{3\times 2} = \begin{pmatrix} 0.452350 & 0.330059 & -0.828517\\ -0.471378 & 0.877114 & 0.0920575\\ 0.757088 & 0.348902 & 0.552345\end{pmatrix}\cdot\begin{pmatrix}\frac{1}{4.6385} & 0\\ 0 & \frac{1}{2.3419}\\ 0 & 0\end{pmatrix}\cdot\begin{pmatrix} 0.728827 & 0.684698\\ -0.684698 & 0.728827\end{pmatrix}, \tag{8.512}$$

$$\mathbf{A}^+_{3\times 2} = \begin{pmatrix} -0.0254237 & 0.169492\\ -0.330508 & 0.20339\\ 0.0169492 & 0.220339\end{pmatrix}. \tag{8.513}$$

Note that

$$\mathbf{A}_{2\times 3}\cdot\mathbf{A}^+_{3\times 2} = \begin{pmatrix} 1 & -3 & 2\\ 2 & 0 & 3\end{pmatrix}\cdot\begin{pmatrix} -0.0254237 & 0.169492\\ -0.330508 & 0.20339\\ 0.0169492 & 0.220339\end{pmatrix} = \begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}. \tag{8.514}$$

Both $\mathbf{Q}$ matrices have a determinant of $+1$ and are thus volume- and orientation-preserving.

Example 8.50

Use the Moore-Penrose inverse to solve the problem $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$ studied in an earlier example, p. 341:

$$\begin{pmatrix}1 & 2\\ 3 & 6\end{pmatrix}\cdot\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}2\\ 0\end{pmatrix}. \tag{8.515}$$

We first seek the singular value decomposition of $\mathbf{A}$, $\mathbf{A}=\mathbf{Q}_2\cdot\mathbf{B}\cdot\mathbf{Q}_1^H$. Now

$$\mathbf{A}^H\cdot\mathbf{A}=\begin{pmatrix}1 & 3\\ 2 & 6\end{pmatrix}\cdot\begin{pmatrix}1 & 2\\ 3 & 6\end{pmatrix}=\begin{pmatrix}10 & 20\\ 20 & 40\end{pmatrix}. \tag{8.516}$$

The eigensystem with normalized eigenvectors corresponding to $\mathbf{A}^H\cdot\mathbf{A}$ is

$$\lambda_1 = 50, \qquad \mathbf{e}_1=\begin{pmatrix}\frac{1}{\sqrt{5}}\\ \frac{2}{\sqrt{5}}\end{pmatrix}, \tag{8.517}$$

$$\lambda_2 = 0, \qquad \mathbf{e}_2=\begin{pmatrix}-\frac{2}{\sqrt{5}}\\ \frac{1}{\sqrt{5}}\end{pmatrix}, \tag{8.518}$$

so

$$\mathbf{Q}_1 = \begin{pmatrix}\frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}}\\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}}\end{pmatrix}, \tag{8.519}$$

$$\mathbf{B} = \begin{pmatrix}\sqrt{50} & 0\\ 0 & 0\end{pmatrix}=\begin{pmatrix}5\sqrt{2} & 0\\ 0 & 0\end{pmatrix}, \tag{8.520}$$

so taking $\mathbf{A}\cdot\mathbf{Q}_1=\mathbf{Q}_2\cdot\mathbf{B}$, gives

$$\underbrace{\begin{pmatrix}1 & 2\\ 3 & 6\end{pmatrix}}_{\mathbf{A}}\cdot\underbrace{\begin{pmatrix}\frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}}\\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}}\end{pmatrix}}_{=\mathbf{Q}_1}=\underbrace{\begin{pmatrix}q_{11} & q_{12}\\ q_{21} & q_{22}\end{pmatrix}}_{\mathbf{Q}_2}\cdot\underbrace{\begin{pmatrix}5\sqrt{2} & 0\\ 0 & 0\end{pmatrix}}_{\mathbf{B}}, \tag{8.521}$$

$$\sqrt{5}\begin{pmatrix}1 & 0\\ 3 & 0\end{pmatrix}=\begin{pmatrix}5\sqrt{2}q_{11} & 0\\ 5\sqrt{2}q_{21} & 0\end{pmatrix}. \tag{8.522}$$

Solving, we get

$$\begin{pmatrix}q_{11}\\ q_{21}\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{10}}\\ \frac{3}{\sqrt{10}}\end{pmatrix}. \tag{8.523}$$

Imposing orthonormality to find q_{12} and q_{22}, we get

$$\begin{pmatrix}q_{12}\\ q_{22}\end{pmatrix}=\begin{pmatrix}\frac{3}{\sqrt{10}}\\ -\frac{1}{\sqrt{10}}\end{pmatrix}, \tag{8.524}$$

so

$$\mathbf{Q}_2=\begin{pmatrix}\frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}}\\ \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}}\end{pmatrix}, \tag{8.525}$$

and

$$\mathbf{A} = \mathbf{Q}_2 \cdot \mathbf{B} \cdot \mathbf{Q}_1^H = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{pmatrix}}_{\mathbf{Q}_2} \cdot \underbrace{\begin{pmatrix} 5\sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}}_{\mathbf{B}} \cdot \underbrace{\begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix}}_{\mathbf{Q}_1^H} = \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}. \tag{8.526}$$

As an aside, note that $\mathbf{Q}_1$ is orientation-preserving, while $\mathbf{Q}_2$ is not, though this property is not important for this analysis.

We will need $\mathbf{B}^+$, which is easily calculated by taking the inverse of each diagonal term of $\mathbf{B}$:

$$\mathbf{B}^+ = \begin{pmatrix} \frac{1}{5\sqrt{2}} & 0 \\ 0 & 0 \end{pmatrix}. \tag{8.527}$$

Now the Moore-Penrose inverse is

$$\mathbf{A}^+ = \mathbf{Q}_1 \cdot \mathbf{B}^+ \cdot \mathbf{Q}_2^H = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix}}_{\mathbf{Q}_1} \cdot \underbrace{\begin{pmatrix} \frac{1}{5\sqrt{2}} & 0 \\ 0 & 0 \end{pmatrix}}_{\mathbf{B}^+} \cdot \underbrace{\begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{pmatrix}}_{\mathbf{Q}_2^H} = \begin{pmatrix} \frac{1}{50} & \frac{3}{50} \\ \frac{2}{50} & \frac{6}{50} \end{pmatrix}. \tag{8.528}$$

Direct multiplication shows that $\mathbf{A} \cdot \mathbf{A}^+ \neq \mathbf{I}$. This is a consequence of $\mathbf{A}$ not being a full rank matrix. However, the four Moore-Penrose conditions are satisfied: $\mathbf{A} \cdot \mathbf{A}^+ \cdot \mathbf{A} = \mathbf{A}$, $\mathbf{A}^+ \cdot \mathbf{A} \cdot \mathbf{A}^+ = \mathbf{A}^+$, $(\mathbf{A} \cdot \mathbf{A}^+)^H = \mathbf{A} \cdot \mathbf{A}^+$, and $(\mathbf{A}^+ \cdot \mathbf{A})^H = \mathbf{A}^+ \cdot \mathbf{A}$.

Lastly, applying the Moore-Penrose inverse operator to the vector $\mathbf{b}$ to form $\mathbf{x} = \mathbf{A}^+ \cdot \mathbf{b}$, we get

$$\mathbf{x} = \mathbf{A}^+ \cdot \mathbf{b} = \underbrace{\begin{pmatrix} \frac{1}{50} & \frac{3}{50} \\ \frac{2}{50} & \frac{6}{50} \end{pmatrix}}_{\mathbf{A}^+} \cdot \underbrace{\begin{pmatrix} 2 \\ 0 \end{pmatrix}}_{\mathbf{b}} = \begin{pmatrix} \frac{1}{25} \\ \frac{2}{25} \end{pmatrix}. \tag{8.529}$$

We see that the Moore-Penrose operator acting on $\mathbf{b}$ has yielded an $\mathbf{x}$ vector which is in the row space of $\mathbf{A}$. As there is no right null space component, it is the minimum length vector that minimizes the residual $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$. It is fully consistent with the solution we found using Gaussian elimination in an earlier example, p. 341.

Problems

1. Find the $\mathbf{x}$ with smallest $||\mathbf{x}||_2$ which minimizes $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ for

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & -1 & 3 \\ 3 & -1 & 5 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

2. Find the most general $\mathbf{x}$ which minimizes $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ for

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 3 & -2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

3. Find $\mathbf{x}$ with the smallest $||\mathbf{x}||_2$ which minimizes $||\mathbf{A}\cdot\mathbf{x}-\mathbf{b}||_2$ for

$$\mathbf{A}=\begin{pmatrix}1&0&1&4\\1&0&2&-1\\2&1&3&-2\end{pmatrix},\qquad \mathbf{b}=\begin{pmatrix}2\\1\\-3\end{pmatrix}.$$

4. Find $e^{\mathbf{A}}$ if

$$\mathbf{A}=\begin{pmatrix}1&1&1\\0&3&2\\0&0&5\end{pmatrix}.$$

5. Diagonalize or reduce to Jordan canonical form

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

6. Find the eigenvectors and generalized eigenvectors of

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

7. Decompose $\mathbf{A}$ into Jordan form $\mathbf{S}\cdot\mathbf{J}\cdot\mathbf{S}^{-1}$, $\mathbf{P}^{-1}\cdot\mathbf{L}\cdot\mathbf{D}\cdot\mathbf{U}$, $\mathbf{Q}\cdot\mathbf{R}$, Schur form, and Hessenberg form

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

8. Find the matrix $\mathbf{S}$ that will convert the following to the Jordan canonical form

(a) $\begin{pmatrix}6&-1&-3&1\\-1&6&1&-3\\-3&1&6&-1\\1&-3&-1&6\end{pmatrix}$,

(b) $\begin{pmatrix}8&-2&-2&0\\0&6&2&-4\\-2&0&8&-2\\2&-4&0&6\end{pmatrix}$,

and show the Jordan canonical form.

9. Show that the eigenvectors and generalized eigenvectors of $\begin{pmatrix}1&1&2&0\\0&1&3&0\\0&0&2&2\\0&0&0&1\end{pmatrix}$ span the space.

10. Find the projection matrix onto the space spanned by $(1,2,3)^T$ and $(2,3,5)^T$. Find the projection of the vector $(7,8,9)^T$ onto this space.

11. Reduce $4x^2+4y^2+2z^2-4xy+4yz+4zx$ to standard quadratic form.

12. Find the inverse of
$$\begin{pmatrix} 1/4 & 1/2 & 3/4 \\ 3/4 & 1/2 & 1/4 \\ 1/4 & 1/2 & 1/2 \end{pmatrix}.$$

13. Find $\exp \begin{pmatrix} 0 & 0 & i \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

14. Find the nth power of $\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$.

15. If
$$\mathbf{A} = \begin{pmatrix} 5 & 4 \\ 1 & 2 \end{pmatrix},$$
find a matrix $\mathbf{S}$ such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$ is a diagonal matrix. Show by multiplication that it is indeed diagonal.

16. Determine if $\mathbf{A} = \begin{pmatrix} 6 & 2 \\ -2 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 8 & 6 \\ -3 & -1 \end{pmatrix}$ are similar.

17. Find the eigenvalues, eigenvectors, and the matrix $\mathbf{S}$ such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$ is diagonal or of Jordan form, where $\mathbf{A}$ is

    (a) $\begin{pmatrix} 5 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & -2 \end{pmatrix}$,

    (b) $\begin{pmatrix} -2 & 0 & 2 \\ 2 & 1 & 0 \\ 0 & 0 & -2i \end{pmatrix}$,

    (c) $\begin{pmatrix} 3 & 0 & -1 \\ -1 & 2 & 2i \\ 1 & 0 & 1+i \end{pmatrix}$.

18. Put each of the matrices above in $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ form.

19. Put each of the matrices above in $\mathbf{Q} \cdot \mathbf{R}$ form.

20. Put each of the matrices above in Schur form.

21. Let
$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$
Find $\mathbf{S}$ such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \mathbf{J}$, where $\mathbf{J}$ is of the Jordan form. Show by multiplication that $\mathbf{A} \cdot \mathbf{S} = \mathbf{S} \cdot \mathbf{J}$.

22. Show that
$$e^{\mathbf{A}} = \begin{pmatrix} \cos(1) & \sin(1) \\ -\sin(1) & \cos(1) \end{pmatrix},$$
if
$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

23. Write $\mathbf{A}$ in row echelon form
$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & 0 \\ 1 & 0 & 1 & 2 \end{pmatrix}.$$

24. Show that the function
$$f(x, y, z) = x^2 + y^2 + z^2 + yz - zx - xy,$$
is always non-negative.

25. If $\mathbf{A} : \ell_2^2 \to \ell_2^2$, find $||\mathbf{A}||$ when
$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$
Also find its inverse and adjoint.

26. Is the quadratic form
$$f(x_1, x_2, x_3) = 4x_1^2 + 2x_1x_2 + 4x_1x_3,$$
positive definite?

27. Find the Schur decomposition and Cholesky decompositions of $\mathbf{A}$:
$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

28. Find the $\mathbf{x}$ with minimum $||\mathbf{x}||_2$ which minimizes $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||_2$ in the following problems:

    (a)
    $$\mathbf{A} = \begin{pmatrix} -4 & 1 & 0 \\ 2 & 0 & 0 \\ -2 & 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix},$$

    (b)
    $$\mathbf{A} = \begin{pmatrix} 1 & 3 & 2 & 5 & 6 \\ 7 & 2 & 1 & -4 & 5 \\ 1 & 4 & 2 & 13 & 7 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix}.$$

29. In each part of the previous problem, find the right null space and show the most general solution vector can be represented as a linear combination of a unique vector in the row space of $\mathbf{A}$ plus an arbitrary scalar multiple of the right null space of $\mathbf{A}$.

30. An experiment yields the following data:

t	x
0.00	1.001
0.10	1.089
0.23	1.240
0.70	1.654
0.90	1.738
1.50	2.120
2.65	1.412
3.00	1.301

    We have fifteen times as much confidence in the first four data points than we do in all the others. Find the least squares best fit coefficients a, b, and c if the assumed functional form is

(a) $x = a + bt + ct^2$,

(b) $x = a + b\sin t + c\sin 2t$.

Plot on a single graph the data points and the two best fit estimates. Which best fit estimate has the smallest least squares residual?

31. For
$$\mathbf{A} = \begin{pmatrix} 8 & 5 & -2 & -1 \\ 6 & 8 & -2 & 8 \\ -1 & 2 & 0 & 1 \end{pmatrix},$$
a) find the $\mathbf{P}^{-1} \cdot \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{U}$ decomposition, and

b) find the singular values and the singular value decomposition.

32. For the complex matrices $\mathbf{A}$ find eigenvectors, eigenvalues, demonstrate whether or not the eigenvectors are orthogonal, find (if possible) the matrix $\mathbf{S}$ such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$ is of Jordan form, and find the singular value decomposition if
$$\mathbf{A} = \begin{pmatrix} 2+i & 2 \\ 2 & 1 \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} 2 & 4i & 2+i \\ -4i & 1 & 3 \\ 2-i & 3 & -2 \end{pmatrix}.$$

33. Consider the action of the matrix
$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix},$$
on the unit square with vertices at $(0,0)$, $(1,0)$, $(1,1)$, and $(0,1)$. Give a plot of the original unit square and its image following the alibi mapping. Also decompose $\mathbf{A}$ under a) $\mathbf{Q} \cdot \mathbf{R}$ decomposition, and b) singular value decomposition, and for each decomposition plot the series of mappings under the action of each component of the decomposition.

Chapter 9

Dynamical systems

see Kaplan, Chapter 9,
see Drazin,
see Lopez, Chapter 12,
see Hirsch and Smale,
see Guckenheimer and Holmes,
see Wiggins,
see Strogatz.

In this chapter we consider the evolution of systems, often called *dynamic systems*. Generally, we will be concerned with systems which can be described by sets of ordinary differential equations, both linear and non-linear. Some other classes of systems will also be studied.

9.1 Paradigm problems

We first consider some paradigm problems which will illustrate the techniques used to solve non-linear systems of ordinary differential equations. Systems of equations are typically more complicated than scalar differential equations. The fundamental procedure for analyzing systems of non-linear ordinary differential equations is to

- Cast the system into a standard form.
- Identify the equilibria of the system.
- If possible, linearize the system about its equilibria.
- If linearizable, ascertain the stability of the linearized system to small disturbances.
- If not linearizable, attempt to ascertain the stability of the non-linear system near its equilibria.
- Solve the full non-linear system.

9.1.1 Autonomous example

First consider a simple example of what is known as an *autonomous* system. An autonomous system of ordinary differential equations can be written in the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}). \tag{9.1}$$

Notice that the independent variable t does not appear explicitly.

Example 9.1

For $\mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^1, f : \mathbb{R}^2 \to \mathbb{R}^2$, consider

$$\frac{dx_1}{dt} = x_2 - x_1^2 = f_1(x_1, x_2), \tag{9.2}$$

$$\frac{dx_2}{dt} = x_2 - x_1 = f_2(x_1, x_2). \tag{9.3}$$

The curves defined in the (x_1, x_2) plane by $f_1 = 0$ and $f_2 = 0$ are very useful in determining both the fixed points (found at the intersection) and in the behavior of the system of differential equations. In fact one can sketch trajectories of paths in this phase space by inspection in many cases. The loci of points where $f_1 = 0$ and $f_2 = 0$ are plotted in Fig. 9.1. The zeroes are found at $(x_1, x_2)^T = (0, 0)^T,\ (1, 1)^T$. Linearize about both points by neglecting quadratic and higher powers of deviations from the critical points to find the local behavior of the solution near these points. Near (0,0), the linearization is

$$\frac{dx_1}{dt} = x_2, \tag{9.4}$$

$$\frac{dx_2}{dt} = x_2 - x_1, \tag{9.5}$$

or

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{9.6}$$

This is of the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}. \tag{9.7}$$

And with

$$\mathbf{S} \cdot \mathbf{z} \equiv \mathbf{x}, \tag{9.8}$$

where $\mathbf{S}$ is a *constant* matrix, we get

$$\frac{d}{dt}(\mathbf{S} \cdot \mathbf{z}) = \mathbf{S} \cdot \frac{d\mathbf{z}}{dt} = \mathbf{A} \cdot \mathbf{S} \cdot \mathbf{z}, \tag{9.9}$$

$$\frac{d\mathbf{z}}{dt} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} \cdot \mathbf{z}. \tag{9.10}$$

At this point we assume that $\mathbf{A}$ has distinct eigenvalues and linearly independent eigenvectors; other cases are easily handled. If we choose $\mathbf{S}$ such that its columns contain the eigenvectors of $\mathbf{A}$, we will get a diagonal matrix, which will lead to a set of uncoupled differential equations; each of these can be solved individually. So for our $\mathbf{A}$, standard linear algebra gives

$$\mathbf{S} = \begin{pmatrix} \frac{1}{2} + \frac{\sqrt{3}}{2}i & \frac{1}{2} - \frac{\sqrt{3}}{2}i \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{S}^{-1} = \begin{pmatrix} \frac{i}{\sqrt{3}} & \frac{1}{2} + \frac{\sqrt{3}}{6}i \\ -\frac{i}{\sqrt{3}} & \frac{1}{2} - \frac{\sqrt{3}}{6}i \end{pmatrix}. \tag{9.11}$$

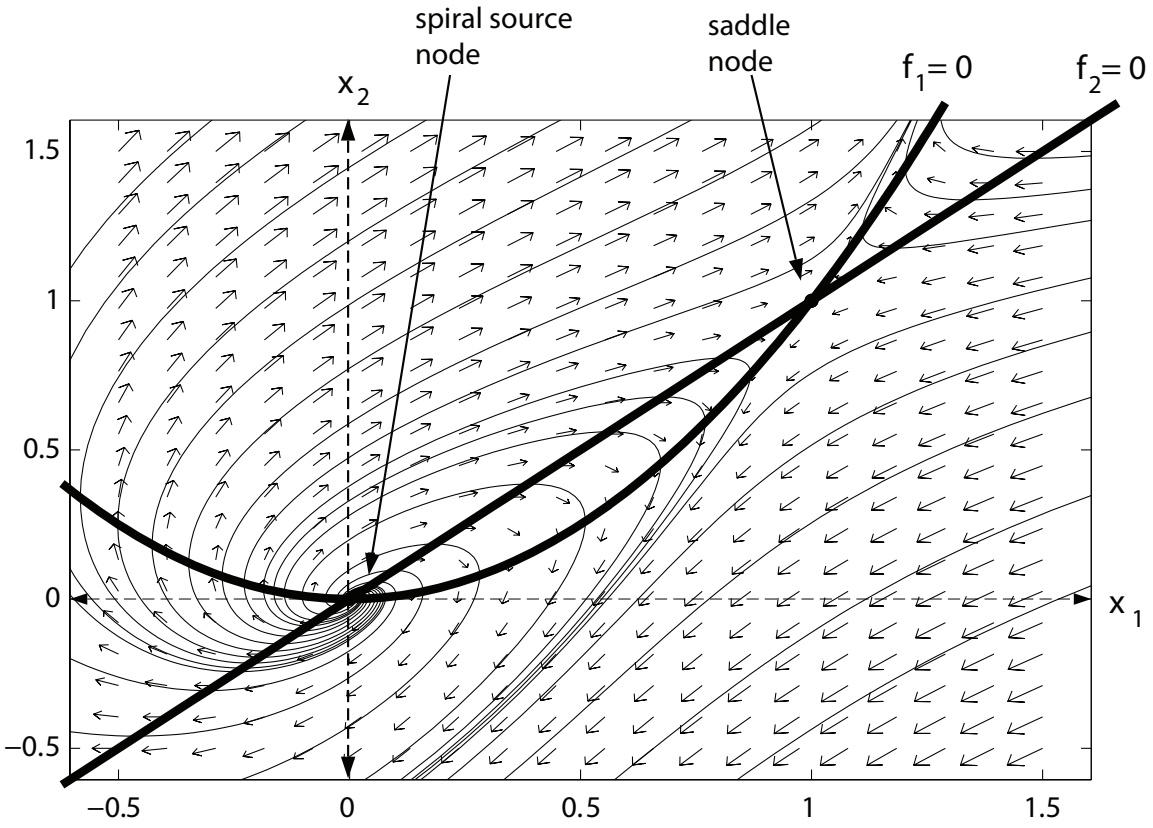

Figure 9.1: Phase plane for $dx_1/dt = x_2 - x_1^2$, $dx_2/dt = x_2 - x_1$, along with equilibrium points $(0,0)$ and $(1,1)$, separatrices $x_2 - x_1^2 = 0$, $x_2 - x_1 = 0$, solution trajectories, and corresponding vector field.

With this choice we get the eigenvalue matrix

$$\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \begin{pmatrix} \frac{1}{2} - \frac{\sqrt{3}}{2} i & 0 \\ 0 & \frac{1}{2} + \frac{\sqrt{3}}{2} i \end{pmatrix}. \tag{9.12}$$

So we get two uncoupled equations for z:

$$\frac{dz_1}{dt} = \underbrace{\left(\frac{1}{2} - \frac{\sqrt{3}}{2} i \right)}_{=\lambda_1} z_1, \tag{9.13}$$

$$\frac{dz_2}{dt} = \underbrace{\left(\frac{1}{2} + \frac{\sqrt{3}}{2} i \right)}_{=\lambda_2} z_2, \tag{9.14}$$

which have solutions

$$z_1 = c_1 \exp\left(\left(\frac{1}{2} - \frac{\sqrt{3}}{2} i \right) t \right), \tag{9.15}$$

$$z_2 = c_2 \exp\left(\left(\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)t\right). \tag{9.16}$$

Then we form $\mathbf{x}$ by taking $\mathbf{x} = \mathbf{S}\cdot\mathbf{z}$ so that

$$x_1 = \left(\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)\underbrace{c_1 \exp\left(\left(\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)t\right)}_{=z_1} + \left(\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)\underbrace{c_2 \exp\left(\left(\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)t\right)}_{=z_2}, \tag{9.17}$$

$$x_2 = \underbrace{c_1 \exp\left(\left(\frac{1}{2}-\frac{\sqrt{3}}{2}i\right)t\right)}_{=z_1} + \underbrace{c_2 \exp\left(\left(\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)t\right)}_{=z_2}. \tag{9.18}$$

Since there is a positive real coefficient in the exponential terms, both x_1 and x_2 grow exponentially. The imaginary component indicates that this is an oscillatory growth. Hence, there is no tendency for a solution which is initially close to $(0,0)$, to remain there. So the fixed point is *unstable*.

Consider the next fixed point near $(1,1)$. First define a new set of local variables:

$$\tilde{x}_1 = x_1 - 1, \tag{9.19}$$

$$\tilde{x}_2 = x_2 - 1. \tag{9.20}$$

Then

$$\frac{dx_1}{dt} = \frac{d\tilde{x}_1}{dt} = (\tilde{x}_2+1) - (\tilde{x}_1+1)^2, \tag{9.21}$$

$$\frac{dx_2}{dt} = \frac{d\tilde{x}_2}{dt} = (\tilde{x}_2+1) - (\tilde{x}_1+1). \tag{9.22}$$

Expanding, we get

$$\frac{d\tilde{x}_1}{dt} = (\tilde{x}_2+1) - \tilde{x}_1^2 - 2\tilde{x}_1 - 1, \tag{9.23}$$

$$\frac{d\tilde{x}_2}{dt} = (\tilde{x}_2+1) - (\tilde{x}_1+1). \tag{9.24}$$

Linearizing about $(\tilde{x}_1, \tilde{x}_2) = (0,0)$, we find

$$\frac{d\tilde{x}_1}{dt} = \tilde{x}_2 - 2\tilde{x}_1, \tag{9.25}$$

$$\frac{d\tilde{x}_2}{dt} = \tilde{x}_2 - \tilde{x}_1, \tag{9.26}$$

or

$$\frac{d}{dt}\begin{pmatrix}\tilde{x}_1\\ \tilde{x}_2\end{pmatrix} = \begin{pmatrix}-2 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}\tilde{x}_1\\ \tilde{x}_2\end{pmatrix}. \tag{9.27}$$

Going through an essentially identical exercise gives the eigenvalues to be

$$\lambda_1 = -\frac{1}{2}+\frac{\sqrt{5}}{2} > 0, \tag{9.28}$$

$$\lambda_2 = -\frac{1}{2}-\frac{\sqrt{5}}{2} < 0, \tag{9.29}$$

which in itself shows the solution to be essentially unstable since there is a positive eigenvalue. After the usual linear algebra and back transformations, one obtains the local solution:

$$x_1 = 1 + c_1 \left(\frac{3-\sqrt{5}}{2}\right) \exp\left(\left(-\frac{1}{2}+\frac{\sqrt{5}}{2}\right)t\right) + c_2 \left(\frac{3+\sqrt{5}}{2}\right) \exp\left(\left(-\frac{1}{2}-\frac{\sqrt{5}}{2}\right)t\right), \quad (9.30)$$

$$x_2 = 1 + c_1 \exp\left(\left(-\frac{1}{2}+\frac{\sqrt{5}}{2}\right)t\right) + c_2 \exp\left(\left(-\frac{1}{2}-\frac{\sqrt{5}}{2}\right)t\right). \quad (9.31)$$

Note that while this solution is generally unstable, if one has the special case in which $c_1 = 0$, that the fixed point in fact is stable. Such is characteristic of a *saddle node*.

As an interesting aside, we can use Eq. (6.371) to calculate the curvature field for this system. With the notation of the present section, the curvature field is given by

$$\kappa = \frac{\sqrt{(\mathbf{f}^T \cdot \mathbf{F} \cdot \mathbf{F}^T \cdot \mathbf{f})(\mathbf{f}^T \cdot \mathbf{f}) - (\mathbf{f}^T \cdot \mathbf{F}^T \cdot \mathbf{f})^2}}{(\mathbf{f}^T \cdot \mathbf{f})^{3/2}}, \quad (9.32)$$

where $\mathbf{F}$, the gradient of the vector field $\mathbf{f}$, is given by the analog of Eq. (6.370)

$$\mathbf{F} = \nabla \mathbf{f}^T. \quad (9.33)$$

So with

$$\mathbf{f} = \begin{pmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{pmatrix} = \begin{pmatrix} x_2 - x_1^2 \\ x_2 - x_1 \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -2x_1 & -1 \\ 1 & 1 \end{pmatrix}, \quad (9.34)$$

detailed calculation reveals that

$$\kappa = \frac{\sqrt{\left(-x_1^2 + x_1^3 + x_1^4 + x_1 x_2 - x_1^2 x_2 - 2x_1^3 x_2 - x_2^2 + 2x_1 x_2^2\right)^2}}{\left(x_1^2 + x_1^4 - 2x_1 x_2 - 2x_1^2 x_2 + 2x_2^2\right)^{3/2}} \quad (9.35)$$

A plot of the curvature field is shown in Fig. 9.2. Because κ varies over orders of magnitude, the contours are for $\ln \kappa$ to more easily visualize the variation. Regions of high curvature are noted near both critical points and in the regions between the curves $x_2 = x_1$ and $x_2 = x_1^2$ for $x_1 \in [0,1]$. Comparison with Fig. 9.1 reveals consistency.

9.1.2 Non-autonomous example

Next, consider a more complicated example. Among other things, the system as originally cast is *non-autonomous* in that the independent variable t appears explicitly. Additionally, it is coupled and contains hidden singularities. Some operations are necessary in order to cast the system in standard form.

Example 9.2

For $\mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^1, f : \mathbb{R}^2 \times \mathbb{R}^1 \to \mathbb{R}^2$, analyze

$$t\frac{dx_1}{dt} + x_2 x_1 \frac{dx_2}{dt} = x_1 + t \qquad = f_1(x_1, x_2, t), \quad (9.36)$$

$$x_1\frac{dx_1}{dt} + x_2^2 \frac{dx_2}{dt} = x_1 t \qquad = f_2(x_1, x_2, t), \quad (9.37)$$

$$x_1(0) = x_{10}, \qquad x_2(0) = x_{20}. \quad (9.38)$$

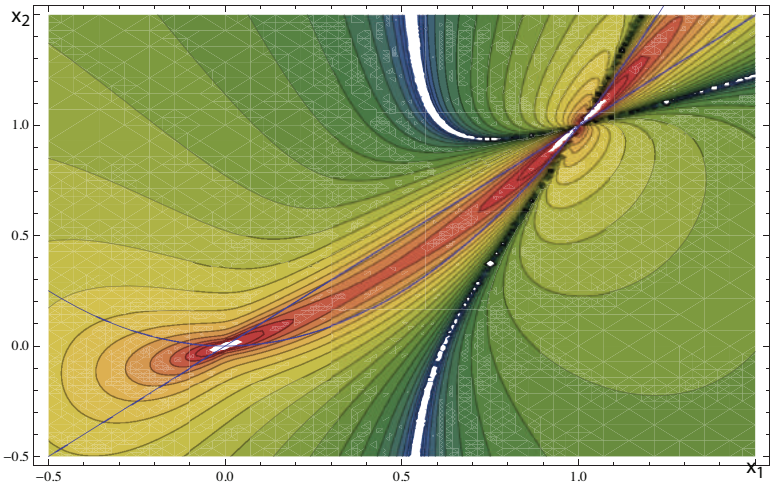

Figure 9.2: Contours of $\ln \kappa$, where κ is trajectory curvature for trajectories of solutions to $dx_1/dt = x_2 - x_1^2$, $dx_2/dt = x_2 - x_1$. Separatrices $x_2 - x_1^2 = 0$ and $x_2 - x_1 = 0$ are also plotted. Red shading corresponds to large trajectory curvature; blue shading corresponds to small trajectory curvature.

Let

$$\frac{dt}{ds} = 1, \qquad t(0) = 0, \tag{9.39}$$

and further $y_1 = x_1, y_2 = x_2, y_3 = t$. Then with $s \in \mathbb{R}^1, y \in \mathbb{R}^3, g : \mathbb{R}^3 \to \mathbb{R}^3$,

$$\begin{aligned} y_3 \frac{dy_1}{ds} + y_2 y_1 \frac{dy_2}{ds} &= y_1 + y_3 \qquad = g_1(y_1, y_2, y_3), & (9.40)\\ y_1 \frac{dy_1}{ds} + y_2^2 \frac{dy_2}{ds} &= y_1 y_3 \qquad = g_2(y_1, y_2, y_3), & (9.41)\\ \frac{dy_3}{ds} &= 1 \qquad = g_3(y_1, y_2, y_3), & (9.42)\\ y_1(0) = y_{10}, \qquad y_2(0) &= y_{20}, \qquad y_3(0) = 0. & (9.43) \end{aligned}$$

In matrix form, we have

$$\begin{pmatrix} y_3 & y_2 y_1 & 0 \\ y_1 & y_2^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{dy_1}{ds} \\ \frac{dy_2}{ds} \\ \frac{dy_3}{ds} \end{pmatrix} = \begin{pmatrix} y_1 + y_3 \\ y_1 y_3 \\ 1 \end{pmatrix}. \tag{9.44}$$

Inverting the coefficient matrix, we obtain the following equation which is in autonomous form:

$$\frac{d}{ds} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{y_1 y_2 - y_1^2 y_3 + y_2 y_3}{y_2 y_3 - y_1^2} \\ \frac{y_1 \left(y_3^2 - y_1 - y_3\right)}{y_2 \left(y_2 y_3 - y_1^2\right)} \\ 1 \end{pmatrix} = \begin{pmatrix} h_1(y_1, y_2, y_3) \\ h_2(y_1, y_2, y_3) \\ h_3(y_1, y_2, y_3) \end{pmatrix}. \tag{9.45}$$

There are potential singularities at $y_2 = 0$ and $y_2 y_3 = y_1^2$. Under such conditions, the determinant of the coefficient matrix is zero, and dy_i/ds is not uniquely determined. One way to address the potential singularities is by defining a new independent variable $u \in \mathbb{R}^1$ via the equation

$$\frac{ds}{du} = y_2 \left(y_2 y_3 - y_1^2\right). \tag{9.46}$$

The system, Eq. (9.45), then transforms to

$$\frac{d}{du}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_2\left(y_1 y_2 - y_1^2 y_3 + y_2 y_3\right) \\ y_1\left(y_3^2 - y_1 - y_3\right) \\ y_2\left(y_2 y_3 - y_1^2\right) \end{pmatrix} = \begin{pmatrix} p_1(y_1, y_2, y_3) \\ p_2(y_1, y_2, y_3) \\ p_3(y_1, y_2, y_3) \end{pmatrix}. \tag{9.47}$$

This equation actually has an infinite number of fixed points, all of which lie on a line in the three-dimensional phase volume. The line is given parametrically by $(y_1, y_2, y_3)^T = (0, 0, v)^T,\ v \in \mathbb{R}^1$ Here v is just a parameter used in describing the line of fixed points. However, it turns out in this case that the Taylor series expansions yield no linear contribution near any of the fixed points, so we don't get to use the standard linear analysis technique! The problem has an essential non-linear essence, even near fixed points. More potent methods would need to be employed, but the example demonstrates the principle. Figure 9.3 gives a numerically obtained solution for $y_1(u), y_2(u), y_3(u)$ along with a trajectory in (y_1, y_2, y_3) space when $y_1(0) = 1, y_2(0) = -1, y_3(0) = 0$. This corresponds to $x_1(t = 0) = 1, x_2(t = 0) = -1$.

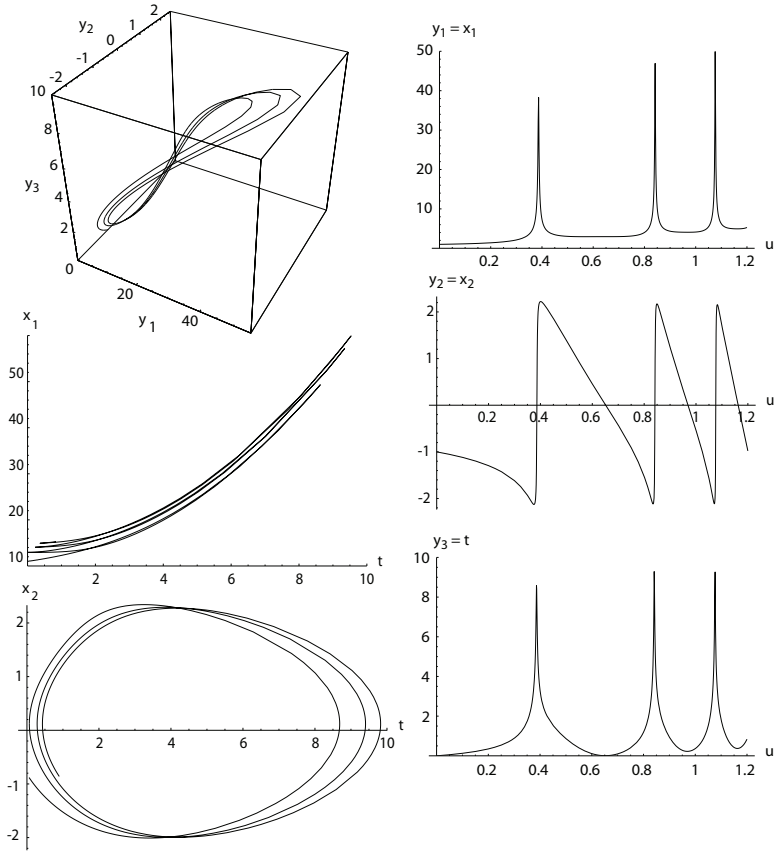

Figure 9.3: Solutions for one set of initial conditions, $y_1(0) = 1,\ y_2(0) = -1,\ y_3(0) = 0$, for second paradigm example: trajectory in phase volume (y_1, y_2, y_3); also $y_1(u), y_2(u), y_3(u)$ and $x_1(t),\ x_2(t)$. Here $y_1 = x_1,\ y_2 = x_2,\ y_3 = t$.

We note that while the solutions are monotonic in the variable u, that they are not monotonic in t, after the transformation back to $x_1(t), x_2(t)$ is effected. Also, while it appears there are points $(u = 0.38, u = 0.84, u = 1.07)$ where the derivatives $dy_1/du, dy_2/du, dy_3/du$ become unbounded, closer

inspection reveals that they are simply points of steep, but bounded, derivatives. However at points where the slope $dy_3/du = dt/du$ changes sign, the derivatives dx_1/dt and dx_2/dt formally are infinite, as is reflected in the cyclic behavior exhibited in the plots of x_1 versus t or x_2 versus t.

9.2 General theory

Consider $\mathbf{x} \in \mathbb{R}^N, t \in \mathbb{R}^1, \mathbf{g} : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^1 \to \mathbb{R}^N$. A general non-linear system of differential-algebraic equations takes on the form

$$\mathbf{g}\left(\frac{d\mathbf{x}}{dt}, \mathbf{x}, t\right) = \mathbf{0}. \tag{9.48}$$

Such general problems can be challenging. Let us here restrict to a form which is quasi-linear in the time-derivatives. Thus, consider $\mathbf{x} \in \mathbb{R}^N, t \in \mathbb{R}^1, \mathbf{A} : \mathbb{R}^N \times \mathbb{R}^1 \to \mathbb{R}^N \times \mathbb{R}^N, \mathbf{f} : \mathbb{R}^N \times \mathbb{R}^1 \to \mathbb{R}^N$. Then the quasi-linear problem of the form

$$\mathbf{A}(\mathbf{x}, t) \cdot \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t), \qquad \mathbf{x}(0) = \mathbf{x}_o, \tag{9.49}$$

can be reduced to autonomous form in the following manner. With

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \ \mathbf{A}(\mathbf{x}, t) = \begin{pmatrix} a_{11}(\mathbf{x}, t) & \dots & a_{1N}(\mathbf{x}, t) \\ \vdots & \ddots & \vdots \\ a_{N1}(\mathbf{x}, t) & \dots & a_{NN}(\mathbf{x}, t) \end{pmatrix}, \ \mathbf{f}(\mathbf{x}, t) = \begin{pmatrix} f_1(x_1, \dots, x_N, t) \\ \vdots \\ f_N(x_1, \dots, x_N, t) \end{pmatrix}, \tag{9.50}$$

define $s \in \mathbb{R}^1$ such that

$$\frac{dt}{ds} = 1, \qquad t(0) = 0. \tag{9.51}$$

Then define $\mathbf{y} \in \mathbb{R}^{N+1}, \mathbf{B} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1} \times \mathbb{R}^{N+1}, \mathbf{g} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$, such that along with $s \in \mathbb{R}^1$ that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \\ y_{N+1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \\ t \end{pmatrix}, \tag{9.52}$$

$$\mathbf{B}(\mathbf{y}) = \begin{pmatrix} a_{11}(\mathbf{y}) & \dots & a_{1N}(\mathbf{y}) & 0 \\ \vdots & \ddots & \vdots & \vdots \\ a_{N1}(\mathbf{y}) & \dots & a_{NN}(\mathbf{y}) & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \tag{9.53}$$

$$\mathbf{g}(\mathbf{y}) = \begin{pmatrix} g_1(y_1, \dots, y_{N+1}) \\ \vdots \\ g_N(y_1, \dots, y_{N+1}) \\ g_{N+1}(y_1, \dots, y_{N+1}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \dots, x_N, t) \\ \vdots \\ f_N(x_1, \dots, x_N, t) \\ 1 \end{pmatrix}. \tag{9.54}$$

Equation (9.49) then transforms to

$$\mathbf{B}(\mathbf{y}) \cdot \frac{d\mathbf{y}}{ds} = \mathbf{g}(\mathbf{y}). \tag{9.55}$$

By forming $\mathbf{B}^{-1}$, assuming $\mathbf{B}$ is non-singular, Eq. (9.55) can be written as

$$\frac{d\mathbf{y}}{ds} = \mathbf{B}^{-1}(\mathbf{y}) \cdot \mathbf{g}(\mathbf{y}), \tag{9.56}$$

or by taking

$$\mathbf{B}^{-1}(\mathbf{y}) \cdot \mathbf{g}(\mathbf{y}) \equiv \mathbf{h}(\mathbf{y}), \tag{9.57}$$

we get the form, commonly called *autonomous form*, with $s \in \mathbb{R}^1, \mathbf{y} \in \mathbb{R}^{N+1}, \mathbf{h} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$:

$$\frac{d\mathbf{y}}{ds} = \mathbf{h}(\mathbf{y}). \tag{9.58}$$

If $\mathbf{B}(\mathbf{y})$ is singular, then $\mathbf{h}$ has singularities. At such singular points, we cannot form a linearly independent set of $d\mathbf{y}/ds$, and the system is better considered as a set of differential-algebraic equations. If the source of the singularity can be identified, a singularity-free autonomous set of equations can often be written. For example, suppose $\mathbf{h}$ can be rewritten as

$$\mathbf{h}(\mathbf{y}) = \frac{\mathbf{p}(\mathbf{y})}{q(\mathbf{y})}, \tag{9.59}$$

where $\mathbf{p}$ and q have no singularities. Then we can remove the singularity by introducing the new independent variable $u \in \mathbb{R}^1$ such that

$$\frac{ds}{du} = q(\mathbf{y}). \tag{9.60}$$

Using the chain rule, the system then becomes

$$\frac{d\mathbf{y}}{ds} = \frac{\mathbf{p}(\mathbf{y})}{q(\mathbf{y})}, \tag{9.61}$$

$$\frac{ds}{du}\frac{d\mathbf{y}}{ds} = q(\mathbf{y})\frac{\mathbf{p}(\mathbf{y})}{q(\mathbf{y})}, \tag{9.62}$$

$$\frac{d\mathbf{y}}{du} = \mathbf{p}(\mathbf{y}), \tag{9.63}$$

which has no singularities.

Casting ordinary differential equations systems in autonomous form is the starting point for most problems and most theoretical development. The task from here generally proceeds as follows:

- Find all the zeroes of $\mathbf{h}$. This is an algebra problem, which can be topologically difficult for non-linear problems.

- If **h** has any singularities, redefine variables in the manner demonstrated to remove the singularity
- If possible, linearize **h** (or its equivalent) about each of its zeroes
- Perform a local analysis of the system of differential equations near zeroes.
- If the system is linear, an eigenvalue analysis is sufficient to reveal stability; for nonlinear systems, the situation is not always straightforward.

9.3 Iterated maps

A map $f : \mathbb{R}^N \to \mathbb{R}^N$ can be iterated to give a dynamical system of the form

$$x_n^{k+1} = f_n(x_1^k, x_2^k, \cdots, x_N^k), \qquad n = 1, \cdots, N. \tag{9.64}$$

Given an initial point $x_n^0, (n = 1, \ldots, N)$ in $\mathbb{R}^N$, a series of images $x_n^1, x_n^2, x_n^3, \ldots$ can be found as $k = 0, 1, 2, \ldots$. The map is dissipative or conservative according to whether the diameter of a set is larger than that of its image or the same, respectively, i.e. if the determinant of the Jacobian matrix, $\det \partial f_n/\partial x_j \leq 1$.

The point $x_i = \overline{x}_i$ is a *fixed point* of the map if it maps to itself, i.e. if

$$\overline{x}_n = f_n(\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_N), \qquad n = 1, \cdots, N. \tag{9.65}$$

The fixed point $\overline{x}_n = 0$ is linearly unstable if a small perturbation from it leads the images farther and farther away. Otherwise it is stable. A special case of this is asymptotic stability wherein the image returns arbitrarily close to the fixed point.

A linear map can be written as $x_i^{k+1} = \sum_{j=1}^N A_{ij} x_j^k$, $(i = 1, 2, \ldots)$ or $\mathbf{x}^{k+1} = \mathbf{A} \cdot \mathbf{x}^k$. The origin $\mathbf{x} = 0$ is a fixed point of this map. If $||\mathbf{A}|| > 1$, then $||\mathbf{x}^{k+1}|| > ||\mathbf{x}^k||$, and the map is unstable. Otherwise it is stable.

Example 9.3

Examine the linear stability of the fixed points of the logistics map, popularized by May.[1]

$$x^{k+1} = r x^k (1 - x^k), \tag{9.66}$$

We take $r \in [0, 4]$ so that $x^k \in [0, 1]$ maps onto $x^{k+1} \in [0, 1]$. That is, the mapping is onto itself. The fixed points are solutions of

$$\overline{x} = r\overline{x}(1 - \overline{x}), \tag{9.67}$$

which are

$$\overline{x} = 0, \qquad \overline{x} = 1 - \frac{1}{r}. \tag{9.68}$$

[1] Robert McCredie May, 1936-, Australian-Anglo ecologist.

Consider the mapping itself. For an initial seed x^0, we generate a series of x^k. For example if we take $r = 0.4$ and $x^0 = 0.3$, we get

$$\begin{aligned} x^0 &= 0.3, && (9.69)\\ x^1 = 0.4(0.3)(1-0.3) &= 0.084, && (9.70)\\ x^2 = 0.4(0.084)(1-0.084) &= 0.0307776, && (9.71)\\ x^3 = 0.4(0.0307776)(1-0.0307776) &= 0.0119321, && (9.72)\\ x^4 = 0.4(0.0119321)(1-0.0119321) &= 0.0047159, && (9.73)\\ x^5 = 0.4(0.0047159)(1-0.0047159) &= 0.00187747, && (9.74)\\ &\vdots \\ x^\infty &= 0. && (9.75) \end{aligned}$$

For this value of r, the solution approaches the fixed point of 0. Consider $r = 4/3$ and $x^0 = 0.3$

$$\begin{aligned} x^0 &= 0.3, && (9.76)\\ x^1 = (4/3)(0.3)(1-0.3) &= 0.28, && (9.77)\\ x^2 = (4/3)(0.28)(1-0.28) &= 0.2688, && (9.78)\\ x^3 = (4/3)(0.2688)(1-0.2688) &= 0.262062, && (9.79)\\ x^4 = (4/3)(0.262062)(1-0.262062) &= 0.257847, && (9.80)\\ x^5 = (4/3)(0.257847)(1-0.257847) &= 0.255149, && (9.81)\\ &\vdots \\ x^\infty &= 0.250 = 1 - \frac{1}{r}. && (9.82) \end{aligned}$$

In this case, the solution was attracted to the alternate fixed point.

To analyze the stability of each fixed point, we give it a small perturbation $\tilde{x}$. Thus, $\overline{x} + \tilde{x}$ is mapped to $\overline{x} + \tilde{\tilde{x}}$, where

$$\overline{x} + \tilde{\tilde{x}} = r(\overline{x} + \tilde{x})(1 - \overline{x} - \tilde{x}) = r(\overline{x} - \overline{x}^2 + \tilde{x} - 2\overline{x}\tilde{x} + \tilde{x}^2). \tag{9.83}$$

Neglecting small terms, we get

$$\overline{x} + \tilde{\tilde{x}} = r(\overline{x} - \overline{x}^2 + \tilde{x} - 2\overline{x}\tilde{x}) = r\overline{x}(1 - \overline{x}) + r\tilde{x}(1 - 2\overline{x}). \tag{9.84}$$

Simplifying, we get

$$\tilde{\tilde{x}} = r\tilde{x}(1 - 2\overline{x}). \tag{9.85}$$

A fixed point is stable if $|\tilde{\tilde{x}}/\tilde{x}| \leq 1$. This indicates that the perturbation is decaying. Now consider each fixed point in turn.

$\overline{x} = 0$:

$$\begin{aligned} \tilde{\tilde{x}} &= r\tilde{x}(1 - 2(0)), && (9.86)\\ \tilde{\tilde{x}} &= r\tilde{x}, && (9.87)\\ \left|\frac{\tilde{\tilde{x}}}{\tilde{x}}\right| &= r. && (9.88) \end{aligned}$$

This is stable if $r < 1$.

$\overline{x} = 1 - 1/r$:

$$\tilde{\tilde{x}} = r\tilde{x}\left(1 - 2\left(1 - \frac{1}{r}\right)\right), \tag{9.89}$$

$$\tilde{\tilde{x}} = (2 - r)\tilde{x}, \tag{9.90}$$

$$\left|\frac{\tilde{\tilde{x}}}{\tilde{x}}\right| = |2 - r|. \tag{9.91}$$

This is unstable for $r < 1$, stable for $1 \leq r \leq 3$, unstable for $r > 3$.

What happens to the map for $r > 3$. Consider $r = 3.2$ and $x^0 = 0.3$

$$x^0 = 0.3, \tag{9.92}$$

$$x^1 = 3.2(0.3)(1 - 0.3) = 0.672, \tag{9.93}$$

$$x^2 = 3.2(0.672)(1 - 0.672) = 0.705331, \tag{9.94}$$

$$x^3 = 3.2(0.705331)(1 - 0.705331) = 0.665085, \tag{9.95}$$

$$x^4 = 3.2(0.665085)(1 - 0.665085) = 0.71279, \tag{9.96}$$

$$x^5 = 3.2(0.71279)(1 - 0.71279) = 0.655105, \tag{9.97}$$

$$x^6 = 3.2(0.655105)(1 - 0.655105) = 0.723016, \tag{9.98}$$

$$x^7 = 3.2(0.723016)(1 - 0.723016) = 0.640845, \tag{9.99}$$

$$x^8 = 3.2(0.640845)(1 - 0.640845) = 0.736521, \tag{9.100}$$

$$\vdots$$

$$x^{\infty-1} = 0.799455, \tag{9.101}$$

$$x^{\infty} = 0.513045. \tag{9.102}$$

This system has bifurcated. It oscillates between two points, never going to the fixed point. The two points about which it oscillates are quite constant for this value of r. For greater values of r, the system moves between $4, 8, 16, ...$ points. Such is the essence of bifurcation phenomena. A plot, known as a bifurcation diagram, of the equilibrium values of x as a function of r is given in Fig. 9.4.

Other maps that have been studied are:

- Hénon[2] map:

$$x_{k+1} = y_k + 1 - ax_k^2, \tag{9.103}$$

$$y_{k+1} = bx_k. \tag{9.104}$$

  For $a = 1.3, b = 0.34$, the attractor is periodic, while for $a = 1.4, b = 0.34$, the map has a strange attractor.

- Dissipative standard map:

$$x_{k+1} = x_k + y_{k+1} \bmod 2\pi, \tag{9.105}$$

$$y_{k+1} = \lambda y_k + k \sin x_k. \tag{9.106}$$

  If $\lambda = 1$, the map is area preserving.

[2]Michel Hénon, 1931-, French mathematician and astronomer.

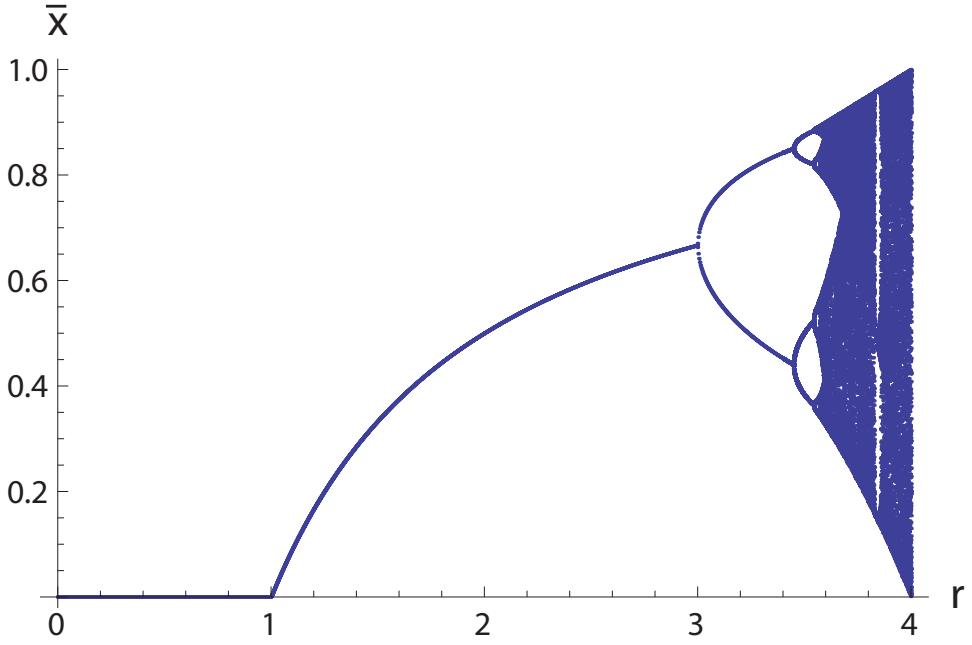

Figure 9.4: Bifurcation diagram of $\overline{x} = \lim_{k\to\infty} x^k$ as a function of r for the logistics map, $x^{k+1} = rx^k(1-x^k)$ for $r \in [0,4]$.

9.4 High order scalar differential equations

An equation with $x \in \mathbb{R}^1, t \in \mathbb{R}^1, a : \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^N, f : \mathbb{R}^1 \to \mathbb{R}^1$ of the form

$$\frac{d^N x}{dt^N} + a_N(x,t)\frac{d^{N-1}x}{dt^{N-1}} + \cdots + a_2(x,t)\frac{dx}{dt} + a_1(x,t)x = f(t), \tag{9.107}$$

can be expressed as a system of $n+1$ first order autonomous equations. Let $x = y_1$, $dx/dt = y_2, \cdots, d^{N-1}x/dt^{N-1} = y_N, t = y_{N+1}$. Then with $y \in \mathbb{R}^{N+1}, s = t \in \mathbb{R}^1, a : \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^N, f : \mathbb{R}^1 \to \mathbb{R}^1$,

$$\frac{dy_1}{ds} = y_2, \tag{9.108}$$

$$\frac{dy_2}{ds} = y_3, \tag{9.109}$$

$$\vdots$$

$$\frac{dy_{N-1}}{ds} = y_N, \tag{9.110}$$

$$\frac{dy_N}{ds} = -a_N(y_1, y_{N+1})y_N - a_{N-1}(y_1, y_{N+1})y_{N-1} - \cdots - a_1(y_1, y_{N+1})y_1 + f(y_{N+1}), \tag{9.111}$$

$$\frac{dy_{N+1}}{ds} = 1. \tag{9.112}$$

Example 9.4

For $x \in \mathbb{R}^1, t \in \mathbb{R}^1$, consider the forced Duffing equation:

$$\frac{d^2x}{dt^2} + x + x^3 = \sin(2t), \qquad x(0) = 0, \left.\frac{dx}{dt}\right|_{t=0} = 0. \tag{9.113}$$

Here $a_2(x,t) = 0, a_1(x,t) = 1 + x^2, f(t) = \sin(2t)$. Now this non-linear differential equation with homogeneous boundary conditions and forcing has no analytic solution. It can be solved numerically; most solution techniques require a recasting as a system of first order equations. To recast this as an autonomous set of equations, with $y \in \mathbb{R}^3, s \in \mathbb{R}^1$, consider

$$x = y_1, \qquad \frac{dx}{dt} = y_2, \qquad t = s = y_3. \tag{9.114}$$

Then $d/dt = d/ds$, and the equations transform to

$$\frac{d}{ds}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_2 \\ -y_1 - y_1^3 + \sin(2y_3) \\ 1 \end{pmatrix} = \begin{pmatrix} h_1(y_1,y_2,y_3) \\ h_2(y_1,y_2,y_3) \\ h_3(y_1,y_2,y_3) \end{pmatrix}, \qquad \begin{pmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{9.115}$$

Note that this system has no equilibrium point as there exists no y for which $h = 0$. Once the numerical solution is obtained, one transforms back to (x,t) space. Fig. 9.5 gives the trajectory in the (y_1, y_2, y_3) phase space and a plot of the corresponding solution $x(t)$ for $t \in [0, 50]$.

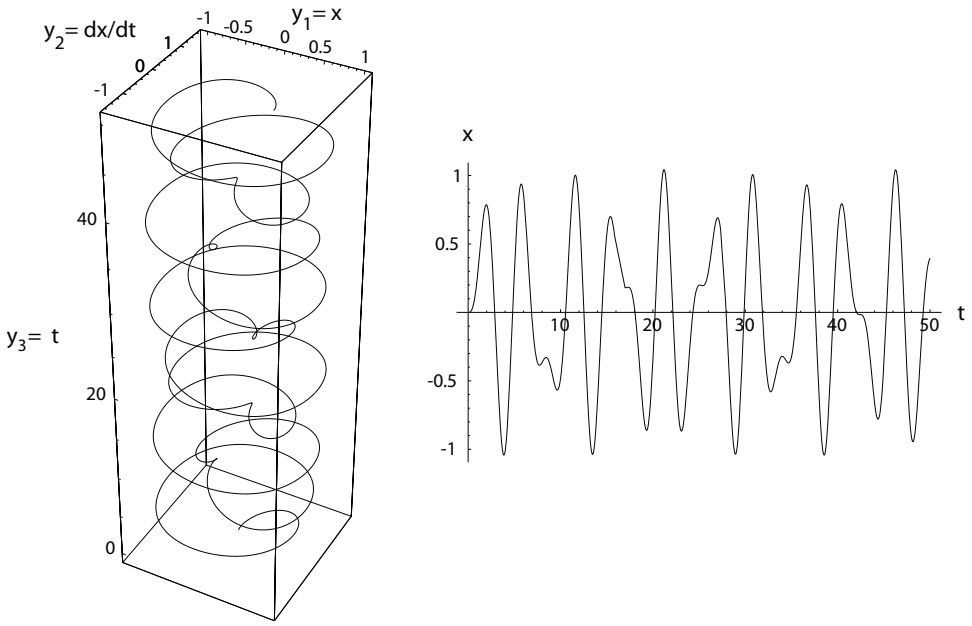

Figure 9.5: Phase space trajectory and solution $x(t)$ for forced Duffing equation.

9.5 Linear systems

For a linear system the coefficients $a_N, \ldots, a_2, a_1$ in equation (9.107) are independent of x. In general, for $\mathbf{x} \in \mathbb{R}^N, t \in \mathbb{R}^1, \mathbf{A} : \mathbb{R}^1 \to \mathbb{R}^N \times \mathbb{R}^N, \mathbf{f} : \mathbb{R}^1 \to \mathbb{R}^N$, any linear system may be written in matrix form as

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}(t) \cdot \mathbf{x} + \mathbf{f}(t), \tag{9.116}$$

where

$$\mathbf{x} = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_N(t) \end{pmatrix}, \tag{9.117}$$

$$\mathbf{A} = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1N}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2N}(t) \\ \vdots & \vdots & \vdots & \vdots \\ a_{N1}(t) & a_{N2}(t) & \cdots & a_{NN}(t) \end{pmatrix}, \tag{9.118}$$

$$\mathbf{f} = \begin{pmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_N(t) \end{pmatrix}. \tag{9.119}$$

Here $\mathbf{A}$ and $\mathbf{f}$ are known. The solution can be written as $\mathbf{x} = \mathbf{x}_H + \mathbf{x}_P$, where $\mathbf{x}_H$ is the solution to the homogeneous equation, and $\mathbf{x}_P$ is the particular solution.

9.5.1 Homogeneous equations with constant A

For $\mathbf{x} \in \mathbb{R}^N, t \in \mathbb{R}^1, \mathbf{A} \in \mathbb{R}^N \times \mathbb{R}^N$, the solution of the homogeneous equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}, \tag{9.120}$$

where $\mathbf{A}$ is a matrix of constants is obtained by setting

$$\mathbf{x} = \mathbf{e} e^{\lambda t}, \tag{9.121}$$

with a constant vector $\mathbf{e} \in \mathbb{R}^N$. Substituting into Eq. (9.120), we get

$$\lambda \mathbf{e} e^{\lambda t} = \mathbf{A} \cdot \mathbf{e} e^{\lambda t}, \tag{9.122}$$

$$\lambda \mathbf{e} = \mathbf{A} \cdot \mathbf{e}. \tag{9.123}$$

This is an eigenvalue problem where λ is an eigenvalue and $\mathbf{e}$ is an eigenvector.

In this case there is only one fixed point, namely the null vector:

$$\mathbf{x} = \mathbf{0}. \tag{9.124}$$

9.5.1.1 N eigenvectors

We will assume that there is a full set of eigenvectors even though not all the eigenvalues are distinct. If $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ are the eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, then

$$\mathbf{x} = \sum_{n=1}^{N} c_n \mathbf{e}_n e^{\lambda_n t}, \tag{9.125}$$

is the general solution, where $c_1, c_2, \ldots, c_N$ are arbitrary constants.

Example 9.5

For $\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}^1, \mathbf{A} \in \mathbb{R}^3 \times \mathbb{R}^3$, solve $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix}. \tag{9.126}$$

The eigenvalues and eigenvectors are

$$\lambda_1 = 1, \qquad \mathbf{e}_1 = \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}, \tag{9.127}$$

$$\lambda_2 = 3, \qquad \mathbf{e}_2 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \tag{9.128}$$

$$\lambda_3 = -2, \qquad \mathbf{e}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}. \tag{9.129}$$

Thus, the solution is

$$\mathbf{x} = c_1 \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} e^t + c_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} e^{3t} + c_3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} e^{-2t}. \tag{9.130}$$

Expanding, we get

$$x_1(t) = -c_1 e^t + c_2 e^{3t} - c_3 e^{-2t}, \tag{9.131}$$

$$x_2(t) = 4c_1 e^t + 2c_2 e^{3t} + c_3 e^{-2t}, \tag{9.132}$$

$$x_3(t) = c_1 e^t + c_2 e^{3t} + c_3 e^{-2t}. \tag{9.133}$$

Example 9.6

For $\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}^1, \mathbf{A} \in \mathbb{R}^3 \times \mathbb{R}^3$, solve $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & -1 \\ 2 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}. \tag{9.134}$$

The eigenvalues and eigenvectors are

$$\lambda_1 = 2, \qquad \mathbf{e}_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \tag{9.135}$$

$$\lambda_2 = 1+i, \qquad \mathbf{e}_2 = \begin{pmatrix} 1 \\ -i \\ 1 \end{pmatrix}, \tag{9.136}$$

$$\lambda_3 = 1-i, \qquad \mathbf{e}_3 = \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix}. \tag{9.137}$$

Thus, the solution is

$$\mathbf{x} = c_1 \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} e^{2t} + c_2 \begin{pmatrix} 1 \\ -i \\ 1 \end{pmatrix} e^{(1+i)t} + c_3 \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix} e^{(1-i)t}, \tag{9.138}$$

$$= c_1 \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} e^{2t} + c_2' \begin{pmatrix} \cos t \\ \sin t \\ \cos t \end{pmatrix} e^{t} + c_3' \begin{pmatrix} \sin t \\ -\cos t \\ \sin t \end{pmatrix} e^{t}, \tag{9.139}$$

$$\tag{9.140}$$

where $c_2' = c_2 + c_3$, $c_3' = i(c_2 - c_3)$.

9.5.1.2 $< N$ eigenvectors

One solution of $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ is $\mathbf{x} = e^{\mathbf{A}t} \cdot \mathbf{e}$, where $\mathbf{e}$ is a constant vector. If $\mathbf{e}_1$, $\mathbf{e}_2$,$\cdots$, $\mathbf{e}_N$ are linearly independent vectors, then $\mathbf{x}_n = e^{\mathbf{A}t} \cdot \mathbf{e}_n$, $n = 1, \cdots, N$, are linearly independent solutions. We would like to choose $\mathbf{e}_n$, $n = 1, 2, \cdots, N$, such that each $e^{\mathbf{A}t} \cdot \mathbf{e}_n$ is a series with a finite number of terms. This can be done in the following manner. Since

$$e^{\mathbf{A}t} \cdot \mathbf{e} = e^{\lambda \mathbf{I} t} \cdot e^{(\mathbf{A} - \lambda \mathbf{I})t} \cdot \mathbf{e}, \tag{9.141}$$

$$= e^{\lambda t} \mathbf{I} \cdot e^{(\mathbf{A} - \lambda \mathbf{I})t} \cdot \mathbf{e}, \tag{9.142}$$

$$= e^{\lambda t} e^{(\mathbf{A} - \lambda \mathbf{I})t} \cdot \mathbf{e}, \tag{9.143}$$

$$= e^{\lambda t} \left(\mathbf{I} + (\mathbf{A} - \lambda \mathbf{I})t + \left(\frac{1}{2!} \right) (\mathbf{A} - \lambda \mathbf{I})^2 t^2 + \cdots \right) \cdot \mathbf{e}. \tag{9.144}$$

the series will be finite if

$$(\mathbf{A} - \lambda \mathbf{I})^k \cdot \mathbf{e} = \mathbf{0}, \tag{9.145}$$

for some positive integer k.

9.5.1.3 Summary of method

The procedure to find $\mathbf{x}_n$, $(n = 1, 2, \ldots, N)$, the N linearly independent solutions of

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}, \tag{9.146}$$

where $\mathbf{A}$ is a constant, is the following. First find all eigenvalues λ_n, $n = 1, \cdots, N$, and as many eigenvectors $\mathbf{e}_k$, $i = 1, 2, \cdots, K$ as possible.

1. If $K = N$, the N linearly independent solutions are $\mathbf{x}_n = e^{\lambda_n t}\mathbf{e}_n$.

2. If $K < N$, there are only K linearly independent solutions of the type $\mathbf{x}_k = e^{\lambda_k t}\mathbf{e}_k$. To find additional solutions corresponding to a multiple eigenvalue λ, find all linearly independent $\mathbf{g}$ such that $(\mathbf{A}-\lambda\mathbf{I})^2 \cdot \mathbf{g} = 0$, but $(\mathbf{A}-\lambda\mathbf{I}) \cdot \mathbf{g} \neq 0$. Notice that generalized eigenvectors will satisfy the requirement, though it has other solutions as well. For each such $\mathbf{g}$, we have
$$e^{\mathbf{A}t} \cdot \mathbf{g} = e^{\lambda t} \left(\mathbf{g} + t(\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g} \right), \tag{9.147}$$
which is a solution.

3. If more solutions are needed, then find all linearly independent $\mathbf{g}$ for which $(\mathbf{A} - \lambda\mathbf{I})^3 \cdot \mathbf{g} = 0$, but $(\mathbf{A} - \lambda\mathbf{I})^2 \cdot \mathbf{g} \neq 0$. The corresponding solution is
$$e^{\mathbf{A}t} \cdot \mathbf{g} = e^{\lambda t} \left(\mathbf{g} + t(\mathbf{A} - \lambda\mathbf{I}) \cdot \mathbf{g} + \frac{t^2}{2}(\mathbf{A} - \lambda\mathbf{I})^2 \cdot \mathbf{g} \right). \tag{9.148}$$

4. Continue until N linearly independent solutions have been found.

A linear combination of the N linearly independent solutions

$$\mathbf{x} = \sum_{n=1}^{N} c_n \mathbf{x}_n, \tag{9.149}$$

is the general solution, where $c_1, c_2, \ldots, c_N$ are arbitrary constants.

9.5.1.4 Alternative method

As an alternative to the method just described, which is easily seen to be equivalent, we can use the Jordan canonical form in a straightforward way to arrive at the solution. Recall that the Jordan form exists for all matrices. We begin with

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}. \tag{9.150}$$

Then we use the Jordan decomposition, Eq. (8.354), $\mathbf{A} = \mathbf{S} \cdot \mathbf{J} \cdot \mathbf{S}^{-1}$ to write

$$\frac{d\mathbf{x}}{dt} = \underbrace{\mathbf{S} \cdot \mathbf{J} \cdot \mathbf{S}^{-1}}_{=\mathbf{A}} \cdot \mathbf{x}. \tag{9.151}$$

If we apply the matrix operator $\mathbf{S}^{-1}$, which is a constant, to both sides, we get

$$\frac{d}{dt}\left(\underbrace{\mathbf{S}^{-1}\cdot\mathbf{x}}_{\equiv\mathbf{z}}\right) = \mathbf{J}\cdot\underbrace{\mathbf{S}^{-1}\cdot\mathbf{x}}_{\equiv\mathbf{z}}. \tag{9.152}$$

Now taking $\mathbf{z}\equiv\mathbf{S}^{-1}\cdot\mathbf{x}$, we get

$$\frac{d\mathbf{z}}{dt} = \mathbf{J}\cdot\mathbf{z}. \tag{9.153}$$

We then solve each equation one by one, starting with the last equation $dz_N/dt = \lambda_N z_N$, and proceeding to the first. In the process of solving these equations sequentially, there will be feedback for each off-diagonal term which will give rise to a secular term in the solution. Once $\mathbf{z}$ is determined, we solve for $\mathbf{x}$ by taking $\mathbf{x} = \mathbf{S}\cdot\mathbf{z}$.

It is also noted that this method works in the common case in which the matrix $\mathbf{J}$ is diagonal; that is, it applies for cases in which there are n differential equations and n ordinary eigenvectors.

Example 9.7

For $\mathbf{x}\in\mathbb{R}^3, t\in\mathbb{R}^1, \mathbf{A}\in\mathbb{R}^3\times\mathbb{R}^3$, find the general solution of

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\cdot\mathbf{x}, \tag{9.154}$$

where

$$\mathbf{A} = \begin{pmatrix} 4 & 1 & 3 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \tag{9.155}$$

$\mathbf{A}$ has an eigenvalue $\lambda = 4$ with multiplicity three. The eigenvector is

$$\mathbf{e} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{9.156}$$

which gives a solution

$$e^{4t}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{9.157}$$

A generalized eigenvector is

$$\mathbf{g}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{9.158}$$

which leads to the solution

$$e^{4t}\left(\mathbf{g}_1 + t(\mathbf{A}-\lambda\mathbf{I})\cdot\mathbf{g}_1\right) = e^{4t}\left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + t\begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right), \tag{9.159}$$

$$= e^{4t}\begin{pmatrix} t \\ 1 \\ 0 \end{pmatrix}. \tag{9.160}$$

Another generalized eigenvector

$$\mathbf{g}_2 = \begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix}, \tag{9.161}$$

gives the solution

$$e^{4t}\left(\mathbf{g}_2 + t(\mathbf{A}-\lambda\mathbf{I})\cdot\mathbf{g}_2 + \frac{t^2}{2}(\mathbf{A}-\lambda\mathbf{I})^2\cdot\mathbf{g}_2\right) =$$

$$e^{4t}\left(\begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix} + t\begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix} + \frac{t^2}{2}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix}\right), \tag{9.162}$$

$$= e^{4t}\begin{pmatrix} \frac{t^2}{2} \\ -3+t \\ 1 \end{pmatrix}. \tag{9.163}$$

The general solution is

$$\mathbf{x} = c_1 e^{4t}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2 e^{4t}\begin{pmatrix} t \\ 1 \\ 0 \end{pmatrix} + c_3 e^{4t}\begin{pmatrix} \frac{t^2}{2} \\ -3+t \\ 1 \end{pmatrix}, \tag{9.164}$$

where c_1, c_2, c_3 are arbitrary constants.

Alternative method

Alternatively, we can simply use the Jordan decomposition to form the solution. When we form the matrix $\mathbf{S}$ from the eigenvectors and generalized eigenvectors, we have

$$\mathbf{S} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{e} & \mathbf{g}_1 & \mathbf{g}_2 \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9.165}$$

We then get

$$\mathbf{S}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}, \tag{9.166}$$

$$\mathbf{J} = \mathbf{S}^{-1}\cdot\mathbf{A}\cdot\mathbf{S} = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \tag{9.167}$$

Now with $\mathbf{z} = \mathbf{S}^{-1}\cdot\mathbf{x}$, we solve $d\mathbf{z}/dt = \mathbf{J}\cdot\mathbf{z}$,

$$\frac{d}{dt}\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}. \tag{9.168}$$

The final equation is totally uncoupled; solving $dz_3/dt = 4z_3$, we get

$$z_3(t) = c_3 e^{4t}. \tag{9.169}$$

Now consider the second equation,

$$\frac{dz_2}{dt} = 4z_2 + z_3. \tag{9.170}$$

Using our solution for z_3, we get

$$\frac{dz_2}{dt} = 4z_2 + c_3 e^{4t}. \tag{9.171}$$

Solving, we get

$$z_2(t) = c_2 e^{4t} + c_3 t e^{4t}. \tag{9.172}$$

Now consider the first equation,

$$\frac{dz_1}{dt} = 4z_1 + z_2. \tag{9.173}$$

Using our solution for z_2, we get

$$\frac{dz_1}{dt} = 4z_1 + c_2 e^{4t} + c_3 t e^{4t}. \tag{9.174}$$

Solving, we get

$$z_1(t) = c_1 e^{4t} + \frac{1}{2} t e^{4t} \left(2c_2 + t c_3\right). \tag{9.175}$$

so we have

$$\mathbf{z}(t) = \begin{pmatrix} c_1 e^{4t} + \frac{1}{2} t e^{4t} \left(2c_2 + t c_3\right) \\ c_2 e^{4t} + c_3 t e^{4t} \\ c_3 e^{4t} \end{pmatrix} \tag{9.176}$$

Then for $\mathbf{x} = \mathbf{S} \cdot \mathbf{z}$, we recover

$$\mathbf{x} = c_1 e^{4t} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + c_2 e^{4t} \begin{pmatrix} t \\ 1 \\ 0 \end{pmatrix} + c_3 e^{4t} \begin{pmatrix} \frac{t^2}{2} \\ -3 + t \\ 1 \end{pmatrix}, \tag{9.177}$$

which is identical to our earlier result.

Example 9.8

Examine the linear homogeneous system $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ in terms of an explicit finite difference approximation and give a geometric interpretation of the of the combined action of the differential and matrix operator on $\mathbf{x}$.

A first order *explicit* finite difference approximation to the differential equation takes the form

$$\begin{aligned} \frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\Delta t} &= \mathbf{A} \cdot \mathbf{x}^k, & (9.178) \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + \Delta t \mathbf{A} \cdot \mathbf{x}^k, & (9.179) \\ &= (\mathbf{I} + \Delta t \mathbf{A}) \cdot \mathbf{x}^k. & (9.180) \end{aligned}$$

Let us decompose $\mathbf{A}$ into a symmetric and anti-symmetric part:

$$\begin{aligned} \mathbf{A}_s &= \frac{\mathbf{A} + \mathbf{A}^T}{2}, & (9.181) \\ \mathbf{A}_a &= \frac{\mathbf{A} - \mathbf{A}^T}{2}, & (9.182) \end{aligned}$$

so that $\mathbf{A} = \mathbf{A}_s + \mathbf{A}_a$. Then Eq. (9.180) becomes

$$\mathbf{x}^{k+1} = (\mathbf{I} + \Delta t \mathbf{A}_s + \Delta t \mathbf{A}_a) \cdot \mathbf{x}^k. \tag{9.183}$$

Now since $\mathbf{A}_s$ is symmetric, it can be diagonally decomposed as

$$\mathbf{A}_s = \mathbf{Q} \cdot \mathbf{\Lambda}_s \cdot \mathbf{Q}^T, \tag{9.184}$$

where $\mathbf{Q}$ is an orthogonal matrix, which we will restrict to be a rotation matrix, and $\mathbf{\Lambda}_s$ is a diagonal matrix with the guaranteed real eigenvalues of $\mathbf{A}_s$ on its diagonal. It can also be shown that the anti-symmetric $\mathbf{A}_a$ has a related decomposition,

$$\mathbf{A}_a = \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H, \tag{9.185}$$

where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Lambda}_a$ is a diagonal matrix with the purely imaginary eigenvalues of $\mathbf{A}_a$ on its diagonal. Substituting Eqs. (9.184,9.185) into Eq. (9.183), we get

$$\mathbf{x}^{k+1} = \left(\mathbf{I} + \Delta t \underbrace{\mathbf{Q} \cdot \mathbf{\Lambda}_s \cdot \mathbf{Q}^T}_{\mathbf{A}_s} + \Delta t \underbrace{\mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H}_{\mathbf{A}_a} \right) \cdot \mathbf{x}^k. \tag{9.186}$$

Now since $\mathbf{Q} \cdot \mathbf{Q}^T = \mathbf{I} = \mathbf{Q} \cdot \mathbf{I} \cdot \mathbf{Q}^T$, we can operate on the first and third terms of Eq. (9.186) to get

$$\mathbf{x}^{k+1} = \left(\mathbf{Q} \cdot \mathbf{I} \cdot \mathbf{Q}^T + \Delta t \mathbf{Q} \cdot \mathbf{\Lambda}_s \cdot \mathbf{Q}^T + \Delta t \mathbf{Q} \cdot \mathbf{Q}^T \cdot \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H \cdot \mathbf{Q} \cdot \mathbf{Q}^T \right) \cdot \mathbf{x}^k, \tag{9.187}$$

$$\mathbf{x}^{k+1} = \mathbf{Q} \cdot \left(\mathbf{I} + \Delta t \mathbf{\Lambda}_s + \Delta t \mathbf{Q}^T \cdot \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H \cdot \mathbf{Q} \right) \cdot \mathbf{Q}^T \cdot \mathbf{x}^k, \tag{9.188}$$

$$\mathbf{Q}^T \cdot \mathbf{x}^{k+1} = \underbrace{\mathbf{Q}^T \cdot \mathbf{Q}}_{=\mathbf{I}} \cdot \left(\mathbf{I} + \Delta t \mathbf{\Lambda}_s + \Delta t \mathbf{Q}^T \cdot \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H \cdot \mathbf{Q} \right) \cdot \mathbf{Q}^T \cdot \mathbf{x}^k. \tag{9.189}$$

Now, let us define a rotated coordinate system as $\hat{\mathbf{x}} = \mathbf{Q}^T \cdot \mathbf{x}$, so that Eq. (9.189) becomes

$$\hat{\mathbf{x}}^{k+1} = \left(\mathbf{I} + \underbrace{\Delta t \mathbf{\Lambda}_s}_{\text{stretching}} + \underbrace{\Delta t \mathbf{Q}^T \cdot \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H \cdot \mathbf{Q}}_{\text{rotation}} \right) \cdot \hat{\mathbf{x}}^k. \tag{9.190}$$

This rotated coordinate system is aligned with the principal axes of deformation associated with $\mathbf{A}_s$. We see that the new value, $\hat{\mathbf{x}}^{k+1}$, is composed of the sum of three terms: 1) the old value, due to the action of $\mathbf{I}$, 2) a stretching along the coordinate axes by the term $\Delta t \mathbf{\Lambda}_s$, and 3) a rotation, normal to the coordinate axes by the term $\Delta t \mathbf{Q}^T \cdot \mathbf{U} \cdot \mathbf{\Lambda}_a \cdot \mathbf{U}^H \cdot \mathbf{Q}$. We note that since both $\mathbf{Q}$ and $\mathbf{U}$ have a norm of unity, that it is the magnitude of the eigenvalues, along with Δt that determines the amount of stretching and rotation that occurs. Note that although $\mathbf{\Lambda}_a$ and $\mathbf{U}$ have imaginary components, when combined together, they yield a real result.

9.5.1.5 Fundamental matrix

If $\mathbf{x}_n$, $n = 1, \cdots, N$, are linearly independent solutions of $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$, then

$$\mathbf{\Omega} = \begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N \\ \vdots & \vdots & \dots & \vdots \end{pmatrix}, \tag{9.191}$$

is called a *fundamental matrix*. The general solution is

$$\mathbf{x} = \mathbf{\Omega} \cdot \mathbf{c}, \tag{9.192}$$

where

$$\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix}. \tag{9.193}$$

The term $e^{\mathbf{A}t} = \mathbf{\Omega}(t) \cdot \mathbf{\Omega}^{-1}(0)$ is a fundamental matrix.

Example 9.9

Find the fundamental matrix of the previous example problem.

The fundamental matrix is

$$\mathbf{\Omega} = e^{4t} \begin{pmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & -3+t \\ 0 & 0 & 1 \end{pmatrix}, \tag{9.194}$$

so that

$$\mathbf{x} = \mathbf{\Omega} \cdot \mathbf{c} = e^{4t} \begin{pmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & -3+t \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}. \tag{9.195}$$

9.5.2 Inhomogeneous equations

If $\mathbf{A}$ is a constant matrix that is diagonalizable, the system of differential equations represented by

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x} + \mathbf{f}(t), \tag{9.196}$$

can be decoupled into a set of scalar equations, each of which is in terms of a single dependent variable. From Eq. (8.296), let $\mathbf{S}$ be such that $\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. Taking $\mathbf{x} = \mathbf{S} \cdot \mathbf{z}$, we get

$$\frac{d(\mathbf{S} \cdot \mathbf{z})}{dt} = \mathbf{A} \cdot \mathbf{S} \cdot \mathbf{z} + \mathbf{f}(t), \tag{9.197}$$

$$\mathbf{S} \cdot \frac{d\mathbf{z}}{dt} = \mathbf{A} \cdot \mathbf{S} \cdot \mathbf{z} + \mathbf{f}(t). \tag{9.198}$$

Applying $\mathbf{S}^{-1}$ to both sides,

$$\frac{d\mathbf{z}}{dt} = \underbrace{\mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}}_{=\mathbf{\Lambda}} \cdot \mathbf{z} + \mathbf{S}^{-1} \cdot \mathbf{f}(t), \tag{9.199}$$

$$\frac{d\mathbf{z}}{dt} = \mathbf{\Lambda} \cdot \mathbf{z} + \mathbf{g}(t), \tag{9.200}$$

where $\mathbf{\Lambda} = \mathbf{S}^{-1} \cdot \mathbf{A} \cdot \mathbf{S}$ and $\mathbf{g}(t) = \mathbf{S}^{-1} \cdot \mathbf{f}(t)$. This is the decoupled form of the original equation.

Example 9.10

For $x \in \mathbb{R}^2, t \in \mathbb{R}^1$, solve

$$\frac{dx_1}{dt} = 2x_1 + x_2 + 1, \tag{9.201}$$

$$\frac{dx_2}{dt} = x_1 + 2x_2 + t. \tag{9.202}$$

This can be written as

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ t \end{pmatrix}. \tag{9.203}$$

We have

$$\mathbf{S} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \qquad \mathbf{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \tag{9.204}$$

so that

$$\frac{d}{dt}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1-t \\ 1+t \end{pmatrix}. \tag{9.205}$$

The solution is

$$z_1 = ae^t + \frac{t}{2}, \tag{9.206}$$

$$z_2 = be^{3t} - \frac{2}{9} - \frac{t}{6}, \tag{9.207}$$

$$\tag{9.208}$$

which, using $x_1 = z_1 + z_2$ and $x_2 = -z_1 + z_2$ transforms to

$$x_1 = ae^t + be^{3t} - \frac{2}{9} + \frac{t}{3}, \tag{9.209}$$

$$x_2 = -ae^t + be^{3t} - \frac{2}{9} - \frac{2t}{3}. \tag{9.210}$$

Example 9.11

Solve the system

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot (\mathbf{x} - \mathbf{x}_o) + \mathbf{b}, \qquad \mathbf{x}(t_o) = \mathbf{x}_o. \tag{9.211}$$

Such a system arises naturally when one linearizes a non-linear system of the form $d\mathbf{x}/dt = \mathbf{f}(\mathbf{x})$ about a point $\mathbf{x} = \mathbf{x}_o$. Here then, $\mathbf{A}$ is the Jacobian matrix $\mathbf{A} = \partial\mathbf{f}/\partial\mathbf{x}|_{\mathbf{x}=\mathbf{x}_o}$. Note that the system is in equilibrium when

$$\mathbf{A} \cdot (\mathbf{x} - \mathbf{x}_o) = -\mathbf{b}, \tag{9.212}$$

or

$$\mathbf{x} = \mathbf{x}_o - \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.213}$$

Further note that if $\mathbf{b} = \mathbf{0}$, the initial condition $\mathbf{x} = \mathbf{x}_o$ is also an equilibrium condition, and is the unique solution to the differential equation.

First define a new dependent variable $\mathbf{z}$:

$$\mathbf{z} \equiv \mathbf{x} - \mathbf{x}_o + \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.214}$$

So we have

$$\mathbf{x} = \mathbf{z} + \mathbf{x}_o - \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.215}$$

At $t = t_o$, we then get

$$\mathbf{z}(t_o) = \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.216}$$

Then substitute into the original differential equation system to get

$$\frac{d}{dt}\left(\mathbf{z} + \mathbf{x}_o - \mathbf{A}^{-1} \cdot \mathbf{b}\right) = \mathbf{A} \cdot \left(\mathbf{z} - \mathbf{A}^{-1} \cdot \mathbf{b}\right) + \mathbf{b}, \qquad \mathbf{z}(t_o) = \mathbf{A}^{-1} \cdot \mathbf{b}, \tag{9.217}$$

$$\frac{d\mathbf{z}}{dt} = \mathbf{A} \cdot \mathbf{z}, \qquad \mathbf{z}(t_o) = \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.218}$$

Now assume that the Jacobian is fully diagonalizable so that we can take $\mathbf{A} = \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{S}^{-1}$. Thus, we have

$$\frac{d\mathbf{z}}{dt} = \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{S}^{-1} \cdot \mathbf{z}, \qquad \mathbf{z}(t_o) = \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.219}$$

Take now

$$\mathbf{w} \equiv \mathbf{S}^{-1} \cdot \mathbf{z}, \qquad \mathbf{z} = \mathbf{S} \cdot \mathbf{w}, \tag{9.220}$$

so that the differential equation becomes

$$\frac{d}{dt}\left(\mathbf{S} \cdot \mathbf{w}\right) = \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{w}, \qquad \mathbf{S} \cdot \mathbf{w}(t_o) = \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.221}$$

Since $\mathbf{S}$ and $\mathbf{S}^{-1}$ are constant, we can apply the operator $\mathbf{S}^{-1}$ to both sides of the differential equation system to get

$$\mathbf{S}^{-1} \cdot \frac{d}{dt}\left(\mathbf{S} \cdot \mathbf{w}\right) = \mathbf{S}^{-1} \cdot \mathbf{S} \cdot \mathbf{\Lambda} \cdot \mathbf{w}, \qquad \mathbf{S}^{-1} \cdot \mathbf{S} \cdot \mathbf{w}(t_o) = \mathbf{S}^{-1} \cdot \mathbf{A}^{-1} \cdot \mathbf{b}, \tag{9.222}$$

$$\frac{d}{dt}\left(\mathbf{S}^{-1} \cdot \mathbf{S} \cdot \mathbf{w}\right) = \mathbf{I} \cdot \mathbf{\Lambda} \cdot \mathbf{w}, \qquad \mathbf{I} \cdot \mathbf{w}(t_o) = \mathbf{S}^{-1} \cdot \mathbf{A}^{-1} \cdot \mathbf{b}, \tag{9.223}$$

$$\frac{d\mathbf{w}}{dt} = \mathbf{\Lambda} \cdot \mathbf{w}, \qquad \mathbf{w}(t_o) = \mathbf{S}^{-1} \cdot \mathbf{A}^{-1} \cdot \mathbf{b}, \tag{9.224}$$

$$\tag{9.225}$$

This is in diagonal form and has solution

$$\mathbf{w}(t) = e^{\mathbf{\Lambda}(t-t_o)} \cdot \mathbf{S}^{-1} \cdot \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.226}$$

In terms of $\mathbf{z}$, then the solution has the form

$$\mathbf{z}(t) = \mathbf{S} \cdot e^{\mathbf{\Lambda}(t-t_o)} \cdot \mathbf{S}^{-1} \cdot \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.227}$$

Then using the definition of $\mathbf{z}$, one can write the solution in terms of the original $\mathbf{x}$ as

$$\mathbf{x}(t) = \mathbf{x}_o + \left(\mathbf{S} \cdot e^{\mathbf{\Lambda}(t-t_o)} \cdot \mathbf{S}^{-1} - \mathbf{I}\right) \cdot \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.228}$$

Note that the time scales of evolution are entirely determined by $\mathbf{\Lambda}$; in particular the time scales of each mode, τ_i, are $\tau_i = 1/\lambda_i$, where λ_i is an entry in $\mathbf{\Lambda}$. The constant vector $\mathbf{b}$ plays a secondary role in determining the time scales.

Lastly, one infers from the discussion of the matrix exponential, Eq. (8.461), that $e^{\mathbf{A}(t-t_o)} = \mathbf{S} \cdot e^{\mathbf{\Lambda}(t-t_o)} \cdot \mathbf{S}^{-1}$, so we get the final form of

$$\mathbf{x}(t) = \mathbf{x}_o + \left(e^{\mathbf{A}(t-t_o)} - \mathbf{I} \right) \cdot \mathbf{A}^{-1} \cdot \mathbf{b}. \tag{9.229}$$

9.5.2.1 Undetermined coefficients

This method is similar to that presented for scalar equations.

Example 9.12

For $\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}^1, \mathbf{A} \in \mathbb{R}^3 \times \mathbb{R}^3, \mathbf{f} : \mathbb{R}^1 \to \mathbb{R}^3$, solve $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x} + \mathbf{f}(t)$ with

$$\mathbf{A} = \begin{pmatrix} 4 & 1 & 3 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} 3e^t \\ 0 \\ 0 \end{pmatrix}. \tag{9.230}$$

The homogeneous part of this problem has been solved before. Let the particular solution be

$$\mathbf{x}_P = \mathbf{c}e^t. \tag{9.231}$$

Substituting into the equation, we get

$$\mathbf{c}e^t = \mathbf{A} \cdot \mathbf{c}e^t + \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} e^t. \tag{9.232}$$

We can cancel the exponential to get

$$(\mathbf{I} - \mathbf{A}) \cdot \mathbf{c} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}, \tag{9.233}$$

which can be solved to get

$$\mathbf{c} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}. \tag{9.234}$$

Therefore,

$$\mathbf{x} = \mathbf{x}_H + \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} e^t. \tag{9.235}$$

The method must be modified if $\mathbf{f} = \mathbf{c}e^{\lambda t}$, where λ is an eigenvalue of $\mathbf{A}$. Then the particular solution must be of the form $\mathbf{x}_P = (\mathbf{c}_0 + t\mathbf{c}_1 + t^2\mathbf{c}_2 + \cdots)e^{\lambda t}$, where the series is finite, and we take as many terms as necessary.

9.5.2.2 Variation of parameters

This follows the general procedure explained in Section 3.3.2, page 90.

9.6 Non-linear systems

Non-linear systems can be difficult to solve. Even for algebraic systems, general solutions do not exist for polynomial equations of arbitrary degree. Non-linear differential equations, both ordinary and partial, admit analytical solutions only in special cases. Since these equations are quite common in engineering applications, many techniques for approximate numerical and analytical solutions have been developed. Our purpose here is more restricted; it is to analyze the long-time stability of the solutions as a function of a system parameter. We will first develop some of the basic ideas of stability, and then illustrate them through examples.

9.6.1 Definitions

With $x \in \mathbb{R}^N, t \in \mathbb{R}^1, f : \mathbb{R}^N \to \mathbb{R}^N$, consider a system of N non-linear first-order ordinary differential equations

$$\frac{dx_n}{dt} = f_n(x_1, x_2, \cdots, x_N), \qquad n = 1, \cdots, N. \tag{9.236}$$

where t is time, and f_n is a vector field. The system is *autonomous* since f_n is not a function of t. The coordinates $x_1, x_2, \cdots, x_N$ form a *phase* or *state* space. The divergence of the vector field, $\text{div} f_n = \sum_{n=1}^{N} \partial f_n / \partial x_n$, indicates the change of a given volume of initial conditions in phase space. If the divergence is zero, the volume remains constant, and the system is said to be *conservative*. If the divergence is negative, the volume shrinks with time, and the system is *dissipative*. The volume in a dissipative system eventually goes to zero. This final state to which some initial set of points in phase space goes is called an *attractor*. Attractors may be points, closed curves, tori, or fractals (strange). A given dynamical system may have several attractors that co-exist. Each attractor has its own *basin of attraction* in $\mathbb{R}^N$; initial conditions that lie on this basin tend to that particular attractor.

The steady state solutions $x_n = \overline{x}_n$ of Eq. (9.236) are called *critical* (or *fixed*, *singular* or *stationary*) points. Thus, by definition

$$f_n(\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_N) = 0, \qquad n = 1, \cdots, N, \tag{9.237}$$

which is an algebraic, potentially transcendental, set of equations. The dynamics of the system are analyzed by studying the stability of the critical point. For this we perturb the system so that

$$x_n = \overline{x}_n + \tilde{x}_n, \tag{9.238}$$

where the $\sim$ denotes a perturbation. If $||\tilde{x}_n||$ is bounded for $t \to \infty$, the critical point is said to be *stable*, otherwise it is *unstable*. As a special case, if $||\tilde{x}_n|| \to 0$ as $t \to \infty$, the critical point is *asymptotically* stable.

Example 9.13

Evaluate some of the properties of non-linear systems for the degenerate case of the linear system

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{9.239}$$

This is of the form $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$. This particular alibi mapping $\mathbf{f} = \mathbf{A}\cdot\mathbf{x}$ was studied in an earlier example in Sec. 8.4. Here $f_1 = -x_2$ and $f_2 = x_1 - x_2$ defines a vector field in phase space. Its divergence is

$$\text{div}\ \mathbf{f} = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} = 0 - 1 = -1, \tag{9.240}$$

so the system is dissipative; that is, a volume composed of a set of points shrinks with time. In this case the equilibrium state, $f_i = 0$, exists at a unique point, the origin, $x_1 = \overline{x}_1 = 0$, $x_2 = \overline{x}_2 = 0$. The eigenvalues of $\mathbf{A} = \partial f_i/\partial x_j$ are $-1/2 \pm \sqrt{3}i/2$. Thus, $\rho(\mathbf{A}) = |-1/2 \pm \sqrt{3}i/2| = 1$, the equilibrium is stable, and the basin of attraction is the entire x_1, x_2 plane.

Note that $\det \mathbf{A} = 1$, and thus the mapping $\mathbf{A}\cdot\mathbf{x}$ is volume- and orientation-preserving. We also find from Eq. (7.301) that $||\mathbf{A}||_2 = \sqrt{(3+\sqrt{5})/2} = 1.61803$, so $\mathbf{A}$ operating on $\mathbf{x}$ tends to lengthen $\mathbf{x}$. This seems to contradict the dissipative nature of the dynamical system, which is volume-shrinking! A way to reconcile this is to consider that the mapping of a vector $\mathbf{x}$ by the dynamical system is more complicated. Returning to the definition of the derivative, the dynamical system can also be expressed, using the so-called "implicit" formulation, as

$$\lim_{\Delta t\to 0}\begin{pmatrix} \frac{x_1^{k+1}-x_1^k}{\Delta t} \\ \frac{x_2^{k+1}-x_2^k}{\Delta t} \end{pmatrix} = \lim_{\Delta t\to 0}\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix}. \tag{9.241}$$

Had the right-hand side been evaluated at k instead of $k+1$, the formulation would be known as "explicit." We have selected the implicit formulation so as to maintain the proper dissipative property of the continuous system, which for this problem would not be obtained with an explicit scheme. We demand here that $\lim_{\Delta t\to 0} x_i^k = x_i, i = 1, 2$. We focus small finite Δt, though our analysis allows for large Δt as well, and rearrange Eq. (9.241) to get

$$\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & -\Delta t \\ \Delta t & -\Delta t \end{pmatrix}\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} + \begin{pmatrix} x_1^n \\ x_2^n \end{pmatrix}, \tag{9.242}$$

$$\begin{pmatrix} 1 & \Delta t \\ -\Delta t & 1+\Delta t \end{pmatrix}\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix}, \tag{9.243}$$

$$\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1+\Delta t}{1+\Delta t+\Delta t^2} & \frac{-\Delta t}{1+\Delta t+\Delta t^2} \\ \frac{\Delta t}{1+\Delta t+\Delta t^2} & \frac{1}{1+\Delta t+\Delta t^2} \end{pmatrix}}_{=\mathbf{B}}\begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix}. \tag{9.244}$$

So our dynamical system, for finite Δt, is appropriately considered as an iterated map of the form

$$\mathbf{x}^{k+1} = \mathbf{B}\cdot\mathbf{x}^k, \tag{9.245}$$

where

$$\mathbf{B} = \begin{pmatrix} \frac{1+\Delta t}{1+\Delta t+\Delta t^2} & \frac{-\Delta t}{1+\Delta t+\Delta t^2} \\ \frac{\Delta t}{1+\Delta t+\Delta t^2} & \frac{1}{1+\Delta t+\Delta t^2} \end{pmatrix}. \tag{9.246}$$

The matrix $\mathbf{B}$ has

$$\det \mathbf{B} = \frac{1}{1+\Delta t+\Delta t^2}. \tag{9.247}$$

For $\Delta t > 0$, $\det \mathbf{B} < 1$ indicating a shrinking of the volume element, consistent with div $\mathbf{f} < 0$. The eigenvalues of $\mathbf{B}$ are

$$\frac{1 + \frac{\Delta t}{2} \pm \frac{\sqrt{3}}{2} i}{1 + \Delta t + \Delta t^2}, \tag{9.248}$$

which for small Δt expand as

$$1 - \left(1 \pm \sqrt{3} i\right) \frac{\Delta t}{2} + \ldots \tag{9.249}$$

More importantly, the spectral norm of $\mathbf{B}$ is the square root of the largest eigenvalue of $\mathbf{B} \cdot \mathbf{B}^T$. Detailed calculation reveals this, and its series expansion in two limits, to be

$$||\mathbf{B}||_2 = \frac{1 + \Delta t + \frac{3}{2}\Delta t^2 + \Delta t \sqrt{1 + \Delta t + \frac{5}{4}\Delta t^2}}{1 + 2\Delta t + 3\Delta t^2 + 2\Delta t^3 + \Delta t^4}, \tag{9.250}$$

$$\lim_{\Delta t \to 0} ||\mathbf{B}||_2 = 1 - \frac{\Delta t^2}{2} + \ldots, \tag{9.251}$$

$$\lim_{\Delta t \to \infty} ||\mathbf{B}||_2 = \sqrt{\frac{3 + \sqrt{5}}{2}} \frac{1}{\Delta t} = \frac{||\mathbf{A}||_2}{\Delta t}. \tag{9.252}$$

In both limits of Δt, we see that $||\mathbf{B}||_2 < 1$; this can be shown to hold for all Δt. It takes on a value of unity only for $\Delta t = 0$. Then, since $||\mathbf{B}||_2 \leq 1$, $\forall \Delta t$, the action of $\mathbf{B}$ on any $\mathbf{x}$ is to diminish its norm; thus, the system is dissipative. Now $\mathbf{B}$ has a non-zero anti-symmetric part, which is typically associated with rotation. One could show via a variety of decompositions that the action of $\mathbf{B}$ on a vector is to compress and rotate it.

9.6.2 Linear stability

The *linear* stability of the critical point is determined by restricting the analysis to a small neighborhood of the critical point, i.e. for small values of $||\tilde{x}_i||$. We substitute Eq. (9.238) into Eq. (9.236), and linearize by keeping only the terms that are linear in $\tilde{x}_i$ and neglecting all products of $\tilde{x}_i$. Thus, Eq. (9.236) takes a linearized *local form*

$$\frac{d\tilde{x}_n}{dt} = \sum_{j=1}^{N} A_{nj} \tilde{x}_j. \tag{9.253}$$

Another way of obtaining the same result is to expand the vector field in a Taylor series around $x_j = \overline{x}_j$ so that

$$f_n(x_j) = \sum_{j=1}^{N} \left.\frac{\partial f_n}{\partial x_j}\right|_{x_j = \overline{x}_j} \tilde{x}_j + \ldots, \tag{9.254}$$

which has neglecting the higher order terms. Thus, in Eq. (9.253)

$$A_{nj} = \left.\frac{\partial f_n}{\partial x_j}\right|_{x_j = \overline{x}_j}, \tag{9.255}$$

is the Jacobian of f_n evaluated at the critical point. In matrix form the linearized equation for the perturbation $\tilde{\mathbf{x}}$ is

$$\frac{d\tilde{\mathbf{x}}}{dt} = \mathbf{A} \cdot \tilde{\mathbf{x}}. \tag{9.256}$$

The real parts of the eigenvalues of $\mathbf{A}$ determine the linear stability of the critical point $\tilde{\mathbf{x}} = \mathbf{0}$, and the behavior of the solution near it:

- If all eigenvalues have real parts < 0, the critical point is asymptotically stable.
- If *at least one* eigenvalue has a real part > 0, the critical point is unstable.
- If all eigenvalues have real parts ≤ 0, and some have zero real parts, then the critical point is stable if $\mathbf{A}$ has k linearly independent eigenvectors for each eigenvalue of multiplicity k. Otherwise it is unstable.

The following are some terms used in classifying critical points according to the real and imaginary parts of the eigenvalues of $\mathbf{A}$.

Classification	*Eigenvalues*
Hyperbolic	Non-zero real part
Saddle	Some real parts negative, others positive
Stable node or sink	All real parts negative
ordinary sink	All real parts negative, imaginary parts zero
spiral sink	All real parts negative, imaginary parts non-zero
Unstable node or source	All real parts positive
ordinary source	All real parts positive, imaginary parts zero
spiral source	All real parts positive, imaginary parts non-zero
Center	All purely imaginary and non-zero

Figures 9.6 and 9.7 show examples of phase planes for simple systems which describe an ordinary source node, a spiral sink node, an ordinary center node, and a saddle node. Figure 9.8 gives a phase plane, vector field, and trajectories for a complex system with many nodes present. Here the nodes are spiral and saddle nodes.

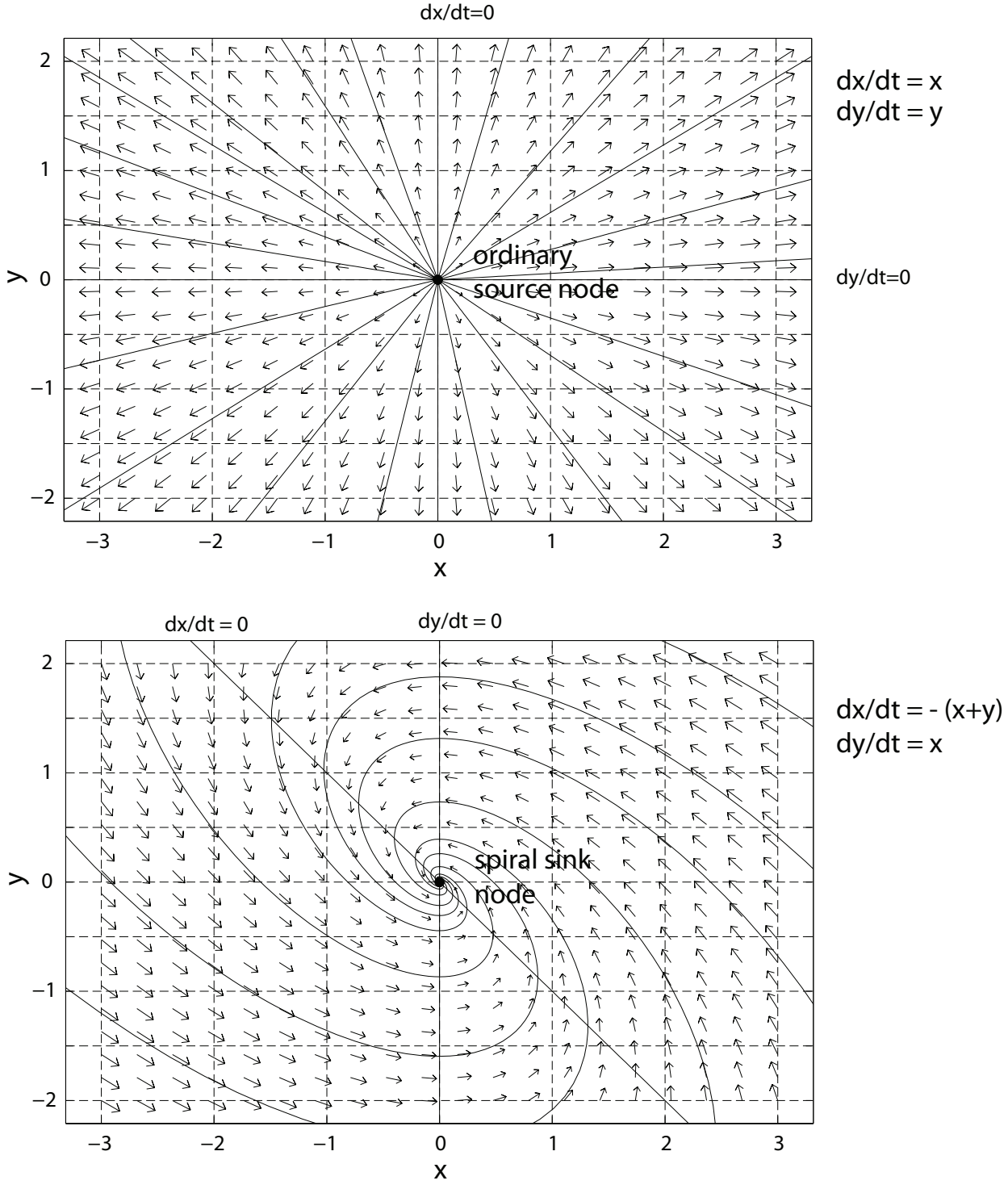

Figure 9.6: Phase plane for system with ordinary source node and spiral sink node.

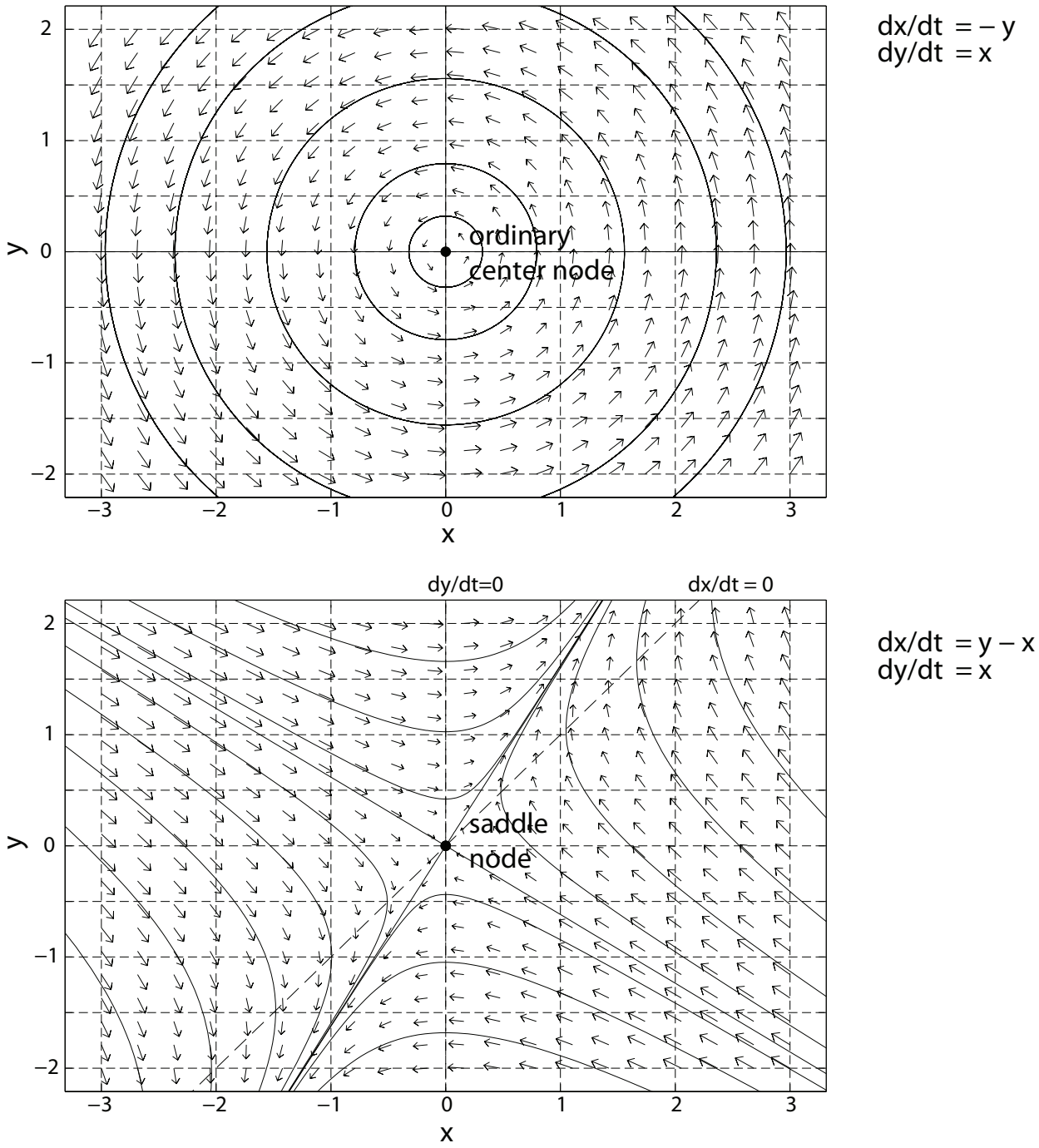

Figure 9.7: Phase plane for systems with center node and saddle node.

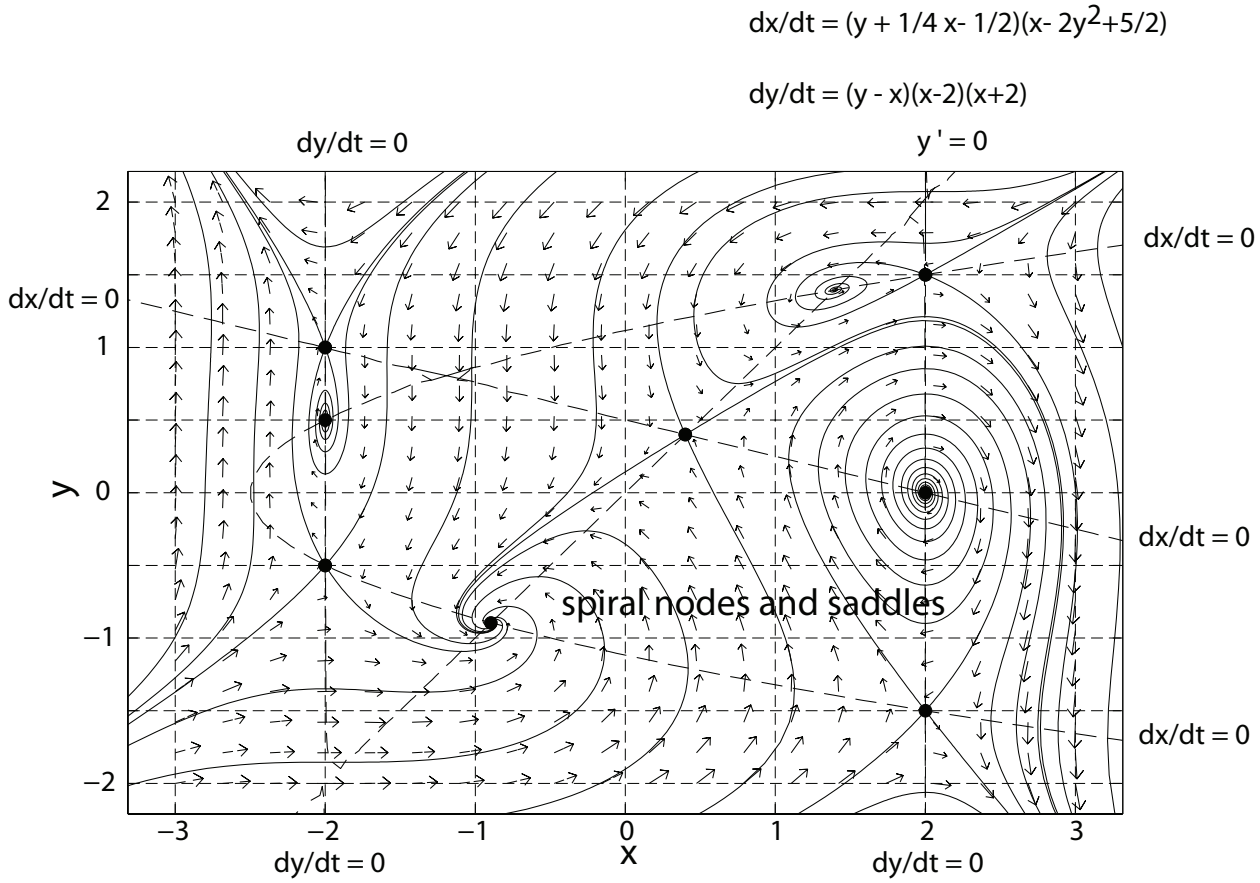

Figure 9.8: Phase plane for system with many nodes.

9.6.3 Lyapunov functions

For $x \in \mathbb{R}^N, t \in \mathbb{R}^1, f : \mathbb{R}^N \to \mathbb{R}^N$ Consider the system of differential equations

$$\frac{dx_n}{dt} = f_n(x_1, x_2, \cdots, x_N), \qquad n = 1, 2, \cdots, N, \tag{9.257}$$

with $x_n = 0$ as a critical point. If there exists a $V(x_1, x_2, \cdots, x_N) : \mathbb{R}^N \to \mathbb{R}^1$ such that

- $V > 0$ for $x_n \neq 0$,
- $V = 0$ for $x_n = 0$,
- $dV/dt < 0$ for $x_n \neq 0$, and
- $dV/dt = 0$ for $x_n = 0$,

then the equilibrium point of the differential equations, $x_i = 0$, is globally stable to all perturbations, large or small. The function $V(x_1, x_2, \cdots, x_N)$ is called a Lyapunov[3] function.

Although one cannot always find a Lyapunov function for a given system of differential equations, we can pose a method to seek a Lyapunov function given a set of autonomous ordinary differential equations. While the method lacks robustness, it is always straightforward to guess a functional form for a Lyapunov function and test whether or not the proposed function satisfies the criteria:

1. Choose a test function $V(x_1, \cdots, x_N)$. The function should be chosen to be strictly positive for $x_n \neq 0$ and zero for $x_n = 0$.

2. Calculate

$$\frac{dV}{dt} = \frac{\partial V}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial V}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial V}{\partial x_N}\frac{dx_N}{dt}, \tag{9.258}$$

$$\frac{dV}{dt} = \frac{\partial V}{\partial x_1}f_1(x_1, \cdots, x_N) + \frac{\partial V}{\partial x_2}f_2(x_1, \cdots, x_N) + \cdots + \frac{\partial V}{\partial x_N}f_N(x_1, \cdots, x_N). \tag{9.259}$$

   It is this step where the differential equations actually enter into the calculation.

3. Determine if for the proposed $V(x_1, \cdots, x_N)$ whether or not $dV/dt < 0, x_n \neq 0; dV/dt = 0, x_n = 0$. If so, then it is a Lyapunov function. If not, there may or may not be a Lyapunov function for the system; one can guess a new functional form and test again.

Example 9.14

Show that $x = 0$ is globally stable, if

$$m\frac{d^2x}{dt^2} + \beta\frac{dx}{dt} + k_1 x + k_2 x^3 = 0, \text{ where } m, \beta, k_1, k_2 > 0. \tag{9.260}$$

[3] Alexandr Mikhailovich Lyapunov, 1857-1918, Russian mathematician.

This system models the motion of a mass-spring-damper system when the spring is non-linear. Breaking the original second order differential equation into two first order equations, we get

$$\frac{dx}{dt} = y, \tag{9.261}$$

$$\frac{dy}{dt} = -\frac{\beta}{m}y - \frac{k_1}{m}x - \frac{k_2}{m}x^3. \tag{9.262}$$

Here x represents the position, and y represents the velocity. Let us guess that the Lyapunov function has the form

$$V(x,y) = ax^2 + by^2 + cx^4, \text{ where } a,b,c > 0. \tag{9.263}$$

Note that $V(x,y) \geq 0$ and that $V(0,0) = 0$. Then

$$\frac{dV}{dt} = \frac{\partial V}{\partial x}\frac{dx}{dt} + \frac{\partial V}{\partial y}\frac{dy}{dt}, \tag{9.264}$$

$$= 2ax\frac{dx}{dt} + 4cx^3\frac{dx}{dt} + 2by\frac{dy}{dt}, \tag{9.265}$$

$$= (2ax + 4cx^3)y + 2by\left(-\frac{\beta}{m}y - \frac{k_1}{m}x - \frac{k_2}{m}x^3\right), \tag{9.266}$$

$$= 2\left(a - \frac{bk_1}{m}\right)xy + 2\left(2c - \frac{bk_2}{m}\right)x^3y - \frac{2b}{m}\beta y^2. \tag{9.267}$$

If we choose $b = m/2$, $a = 1/2k_1$, $c = k_2/4$, then the coefficients on xy and x^3y in the expression for dV/dt are identically zero, and we get

$$\frac{dV}{dt} = -\beta y^2, \tag{9.268}$$

which for $\beta > 0$ is negative for all $y \neq 0$ and zero for $y = 0$. Further, with these choices of a, b, c, the Lyapunov function itself is

$$V = \frac{1}{2}k_1x^2 + \frac{1}{4}k_2x^4 + \frac{1}{2}my^2 \geq 0. \tag{9.269}$$

Checking, we see

$$\frac{dV}{dt} = k_1x\frac{dx}{dt} + k_2x^3\frac{dx}{dt} + my\frac{dy}{dt}, \tag{9.270}$$

$$= k_1xy + k_2x^3y + my\left(-\frac{\beta}{m}y - \frac{k_1}{m}x - \frac{k_2}{m}x^3\right), \tag{9.271}$$

$$= k_1xy + k_2x^3y - \beta y^2 - k_1xy - k_2x^3y, \tag{9.272}$$

$$= -\beta y^2 \leq 0. \tag{9.273}$$

Thus, V is a Lyapunov function, and $x = y = 0$ is globally stable. Actually, in this case, $V =$ (kinetic energy + potential energy), where kinetic energy $= (1/2)my^2$, and potential energy $= (1/2)k_1x^2 + (1/4)k_2x^4$. Note that $V(x,y)$ is just an algebraic function of the system's state variables. When we take the time derivative of V, we are forced to invoke our original system, which defines the differential equations. We note for this system that precisely since V is strictly positive or zero for all x, y, and moreover that it is decaying for all time, that this necessarily implies that $V \to 0$, hence $x, y \to 0$.

9.6.4 Hamiltonian systems

Closely related to the Lyapunov function of a system is the Hamiltonian, which exists for systems which are non-dissipative, that is those systems for which $dV/dt = 0$. In such a case we define the Hamiltonian H to be the Lyapunov function $H = V$ with $dH/dt \equiv 0$. For such systems, we integrate once to find that $H(x_i, y_i)$ must be a constant for all x_i, y_i. Such systems are said to be conservative.

With $x \in \mathbb{R}^N, y \in \mathbb{R}^N, t \in \mathbb{R}^1, f : \mathbb{R}^{2N} \to \mathbb{R}^N, g : \mathbb{R}^{2N} \to \mathbb{R}^N$ We say a system of equations of the form

$$\frac{dx_n}{dt} = f_n(x_1, \cdots, x_N, y_1, \cdots, y_N), \qquad \frac{dy_n}{dt} = g_n(x_1, \cdots, x_N, y_1, \cdots, y_N), \qquad n = 1, \cdots, N, \tag{9.274}$$

is *Hamiltonian* if we can find a function $H(x_n, y_n) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^1$ such that

$$\frac{dH}{dt} = \frac{\partial H}{\partial x_n}\frac{dx_n}{dt} + \frac{\partial H}{\partial y_n}\frac{dy_n}{dt} = 0, \tag{9.275}$$

$$\frac{dH}{dt} = \frac{\partial H}{\partial x_n} f_n(x_1, \cdots, x_N, y_1, \cdots, y_N) + \frac{\partial H}{\partial y_n} g_n(x_1, \cdots, x_N, y_1, \cdots, y_N) = 0. \tag{9.276}$$

This differential equation can at times be solved directly by the method of separation of variables in which we assume a specific functional form for $H(x_i, y_i)$.

Alternatively, we can also determine H by demanding that

$$\frac{\partial H}{\partial y_n} = \frac{dx_n}{dt}, \qquad \frac{\partial H}{\partial x_n} = -\frac{dy_n}{dt}. \tag{9.277}$$

Substituting from the original differential equations, we are led to equations for $H(x_i, y_i)$

$$\frac{\partial H}{\partial y_i} = f_i(x_1, \cdots, x_N, y_1, \cdots, y_N), \qquad \frac{\partial H}{\partial x_i} = -g_i(x_1, \cdots, x_N, y_1, \cdots, y_N). \tag{9.278}$$

Example 9.15

Find the Hamiltonian for a linear mass spring system:

$$m\frac{d^2x}{dt^2} + kx = 0, \qquad x(0) = x_0, \left.\frac{dx}{dt}\right|_0 = \dot{x}_0. \tag{9.279}$$

Taking $dx/dt = y$ to reduce this to a system of two first order equations, we have

$$\frac{dx}{dt} = f(x, y) = y, \qquad x(0) = x_0, \tag{9.280}$$

$$\frac{dy}{dt} = g(x, y) = -\frac{k}{m}x, \qquad y(0) = y_0. \tag{9.281}$$

For this system $N = 1$.

We seek $H(x,y)$ such that $dH/dt = 0$. That is

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial y}\frac{dy}{dt} = 0. \tag{9.282}$$

Substituting from the given system of differential equations we have

$$\frac{\partial H}{\partial x}y + \frac{\partial H}{\partial y}\left(-\frac{k}{m}x\right) = 0. \tag{9.283}$$

As with all partial differential equations, one has to transform to a system of ordinary equations in order to solve. Here we will take the approach of the method of separation of variables and assume a solution of the form

$$H(x,y) = A(x) + B(y), \tag{9.284}$$

where A and B are functions to be determined. With this assumption, we get

$$y\frac{dA}{dx} - \frac{k}{m}x\frac{dB}{dy} = 0. \tag{9.285}$$

Rearranging, we get

$$\frac{1}{x}\frac{dA}{dx} = \frac{k}{my}\frac{dB}{dy}. \tag{9.286}$$

Now the term on the left is a function of x only, and the term on the right is a function of y only. The only way this can be generally valid is if both terms are equal to the same constant, which we take to be C. Hence,

$$\frac{1}{x}\frac{dA}{dx} = \frac{k}{my}\frac{dB}{dy} = C, \tag{9.287}$$

from which we get two ordinary differential equations:

$$\frac{dA}{dx} = Cx, \qquad \frac{dB}{dy} = \frac{Cm}{k}y. \tag{9.288}$$

The solution is

$$A(x) = \frac{1}{2}Cx^2 + K_1, \qquad B(y) = \frac{1}{2}\frac{Cm}{k}y^2 + K_2. \tag{9.289}$$

A general solution is

$$H(x,y) = \frac{1}{2}C\left(x^2 + \frac{m}{k}y^2\right) + K_1 + K_2. \tag{9.290}$$

While this general solution is perfectly valid, we can obtain a common physical interpretation by taking $C = k, K_1 + K_2 = 0$. With these choices, the Hamiltonian becomes

$$H(x,y) = \frac{1}{2}kx^2 + \frac{1}{2}my^2. \tag{9.291}$$

The first term represents the potential energy of the spring, the second term represents the kinetic energy. Since by definition $dH/dt = 0$, this system conserves its mechanical energy. Verifying the properties of a Hamiltonian, we see

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial y}\frac{dy}{dt}, \tag{9.292}$$

$$= kxy + my\left(-\frac{k}{m}x\right), \tag{9.293}$$

$$= 0. \tag{9.294}$$

Since this system has $dH/dt = 0$, then $H(x, y)$ must be constant for all time, including $t = 0$, when the initial conditions apply. So

$$H(x(t), y(t)) = H(x(0), y(0)) = \frac{1}{2}\left(kx_0^2 + my_0^2\right). \tag{9.295}$$

Thus, the system has the integral

$$\frac{1}{2}\left(kx^2 + my^2\right) = \frac{1}{2}\left(kx_0^2 + my_0^2\right). \tag{9.296}$$

We can take an alternate solution approach by consideration of Eq. (9.278) as applied to this problem:

$$\frac{\partial H}{\partial y} = f = y, \qquad \frac{\partial H}{\partial x} = -g = \frac{k}{m}x. \tag{9.297}$$

Integrating the first of these, we get

$$H(x, y) = \frac{1}{2}y^2 + F(x). \tag{9.298}$$

Differentiating with respect to x, we get

$$\frac{\partial H}{\partial x} = \frac{dF}{dx}, \tag{9.299}$$

and this must be

$$\frac{dF}{dx} = \frac{k}{m}x. \tag{9.300}$$

So

$$F(x) = \frac{k}{2m}x^2 + K. \tag{9.301}$$

Thus,

$$H(x, y) = \frac{1}{2m}\left(kx^2 + my^2\right) + K. \tag{9.302}$$

We can choose $K = 0$, and since $dH/dt = 0$, we have H as a constant which is set by the initial conditions, thus giving

$$\frac{1}{2m}\left(kx^2 + my^2\right) = \frac{1}{2m}\left(kx_0^2 + my_0^2\right), \tag{9.303}$$

which gives identical information as does Eq. (9.296).

9.7 Differential-algebraic systems

Many dynamic systems are better considered as differential-algebraic systems of equations of the general form given in Eq. (9.48). There is a rich theory on such systems, which we will not be able to fully exploit here. Instead, we shall consider briefly certain types of linear and non-linear differential-algebraic systems.

9.7.1 Linear homogeneous

Consider the system of homogeneous differential-algebraic equations of the form

$$\mathbf{B} \cdot \frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}. \tag{9.304}$$

Here $\mathbf{A}$ and $\mathbf{B}$ are constant matrices, and we take $\mathbf{B}$ to be singular; thus, it cannot be inverted. We will assume $\mathbf{A}$ is invertible. There is an apparent equilibrium when $\mathbf{x} = \mathbf{0}$, but the singularity of $\mathbf{B}$ gives us concern that this may not always hold. In any case, we can assume solutions of the type $\mathbf{x} = \mathbf{e}e^{\lambda t}$ and substitute into Eq. (9.304) to get

$$\begin{aligned} \mathbf{B} \cdot \mathbf{e}\lambda e^{\lambda t} &= \mathbf{A} \cdot \mathbf{e}e^{\lambda t}, & (9.305)\\ \mathbf{B} \cdot \mathbf{e}\lambda &= \mathbf{A} \cdot \mathbf{e}, & (9.306)\\ (\mathbf{A} - \lambda\mathbf{B}) \cdot \mathbf{e} &= \mathbf{0}. & (9.307) \end{aligned}$$

Eq. (9.307) is a generalized eigenvalue problem in the second sense, as considered in Sec. 8.3.2.

Example 9.16

Solve the linear homogeneous differential-algebraic system

$$\begin{aligned} \frac{dx_1}{dt} + 2\frac{dx_2}{dt} &= x_1 + x_2, & (9.308)\\ 0 &= 2x_1 - x_2. & (9.309) \end{aligned}$$

While this problem is simple enough to directly eliminate x_2 in favor of x_1, other problems are not that simple, so let us illustrate the general method. In matrix form, we can say

$$\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{9.310}$$

Taking $x_1 = e_1 e^{\lambda t}$ and $x_2 = e_2 e^{\lambda t}$ gives

$$\begin{aligned} \lambda \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} e^{\lambda t} &= \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} e^{\lambda t}, & (9.311)\\ \begin{pmatrix} \lambda & 2\lambda \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, & (9.312)\\ \begin{pmatrix} 1-\lambda & 1-2\lambda \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}. & (9.313) \end{aligned}$$

The determinant of the coefficient matrix must be zero, giving

$$\begin{aligned} -(1-\lambda) - 2(1-2\lambda) &= 0, & (9.314)\\ -1 + \lambda - 2 + 4\lambda &= 0, & (9.315)\\ \lambda &= \frac{3}{5}. & (9.316) \end{aligned}$$

With this generalized eigenvalue, our generalized eigenvectors in the second sense are found via

$$\begin{aligned} \begin{pmatrix} 1-\frac{3}{5} & 1-2\left(\frac{3}{5}\right) \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & (9.317)\\ \begin{pmatrix} \frac{2}{5} & -\frac{1}{5} \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}. & (9.318) \end{aligned}$$

By inspection, the non-unique solution must be of the form

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{9.319}$$

So the general solution is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} C_1 e^{3t/5} \\ 2C_1 e^{3t/5} \end{pmatrix}. \tag{9.320}$$

There is only one arbitrary constant for this system.

A less desirable approach to differential algebraic systems is to differentiate the constraint. This requires care in that an initial condition must be imposed which is consistent with the original constraint. Applying this method to our example problem gives rise to the system

$$\frac{dx_1}{dt} + 2\frac{dx_2}{dt} = x_1 + x_2, \tag{9.321}$$

$$2\frac{dx_1}{dt} - \frac{dx_2}{dt} = 0. \tag{9.322}$$

In matrix form, this gives

$$\begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{9.323}$$

$$\begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{pmatrix} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{2}{5} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{9.324}$$

The eigenvectors of the coefficient matrix are $\lambda = 0$ and $\lambda = 3/5$. Whenever one finds an eigenvalue of zero in a dynamic system, there is actually a hidden algebraic constraint within the system. Diagonalization allows us to write the system as

$$\begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{pmatrix}^{-1} \begin{pmatrix} \frac{3}{5} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{9.325}$$

$$\begin{pmatrix} \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{9.326}$$

Regrouping, we can say

$$\frac{d}{dt}(x_1 + x_2) = \frac{3}{5}(x_1 + x_2), \tag{9.327}$$

$$\frac{d}{dt}(-2x_1 + x_2) = 0. \tag{9.328}$$

Solving gives

$$x_1 + x_2 = C_1 e^{3t/5}, \tag{9.329}$$

$$-2x_1 + x_2 = C_2. \tag{9.330}$$

So the problem with the differentiated constraint yields two arbitrary constants. For consistency with the original formulation, we must take $C_2 = 0$, thus $x_2 = 2x_1$. Thus,

$$x_1 = \frac{1}{3} C_1 e^{3t/5}, \tag{9.331}$$

$$x_2 = \frac{2}{3} C_1 e^{3t/5}. \tag{9.332}$$

Because C_1 is arbitrary, this is fully consistent with our previous solution.

9.7.2 Non-linear

Let us consider two simple non-linear examples for differential-algebraic equation systems.

Example 9.17

Solve

$$\frac{dx}{dt} = -y, \tag{9.333}$$

$$x^2 + y^2 = 1, \tag{9.334}$$

$$x(0) = 0.99 \tag{9.335}$$

The system is non-linear because of the non-linear constraint. However, we can also view this system as a Hamiltonian system for a linear oscillator. The non-linear constraint is the Hamiltonian. We recognize that if we differentiate the non-linear constraint, the system of non-linear differential algebraic equations reduces to a linear system of differential equations, $dx/dt = -y$, $dy/dt = x$, which is that of a linear oscillator.

Formulated as a differential-algebraic system, we can say

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} -y \\ x^2 + y^2 - 1 \end{pmatrix}, \qquad x(0) = 0.99. \tag{9.336}$$

We might imagine an equilibrium to be located at $(x, y) = (\pm 1, 0)$. Certainly at such a point $dx/dt = 0$, and the constraint is satisfied. However, at such a point, $dy/dt \neq 0$, so it is not a true equilibrium. Linearization near $(\pm 1, 0)$ would induce another generalized eigenvalue problem in the second sense. For the full problem, the form presented is suitable for numerical integration by many appropriate differential-algebraic software packages. We do so and find the result plotted in Fig. 9.9. For this

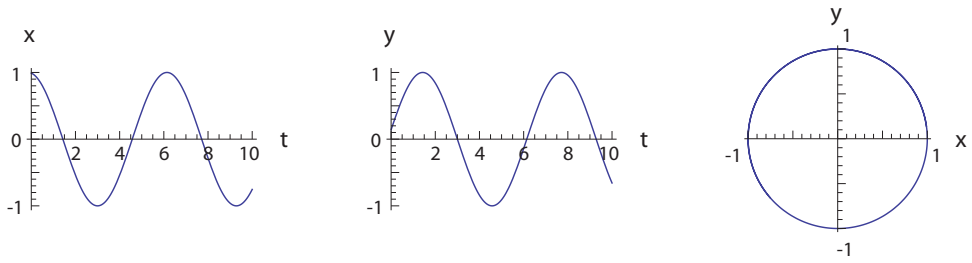

Figure 9.9: Solution to the differential-algebraic system of Eq. (9.336).

system, what is seen to be a pseudo-equilibrium at $(x, y) = (\pm 1, 0)$ is realized periodically. The point is not a formal equilibrium, since it does not remain there as $t \to \infty$. We also clearly see that the trajectory in the (x, y) plane is confined to the unit circle, as required by the constraint.

Example 9.18

Solve

$$\frac{dx}{dt} = y^2 + xy, \tag{9.337}$$

$$2x^2 + y^2 = 1, \tag{9.338}$$
$$x(0) = 0. \tag{9.339}$$

Formulated as a differential-algebraic system, we can say

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} y^2 + xy \\ 2x^2 + y^2 - 1 \end{pmatrix}, \qquad x(0) = 0. \tag{9.340}$$

We could linearize near the potential equilibria, located at $(x, y) = (\pm 1/\sqrt{3}, \mp 1/\sqrt{3})$, $(\pm\sqrt{1/2}, 0)$. This would induce another generalized eigenvalue problem in the second sense. For the full problem, the form presented is suitable for numerical integration by many appropriate differential-algebraic software packages. We do so and find the result plotted in Fig. 9.10. For this system, a true equilibrium at

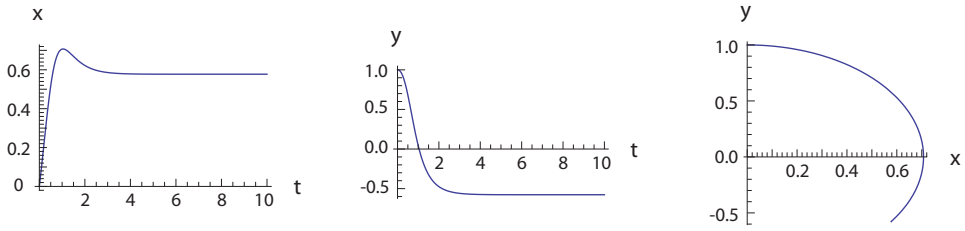

Figure 9.10: Solution to the differential-algebraic system of Eq. (9.340).

$(x, y) = (1/\sqrt{3}, -1/\sqrt{3})$ is realized. We also clearly see that the trajectory in the (x, y) plane is confined to the ellipse, as required by the constraint.

9.8 Fixed points at infinity

Often in dynamic systems there are additional fixed points, not readily seen in finite phase space. These fixed points are actually at infinity, and such points can play a role in determining the dynamics of a system as well as aiding in finding basins of attraction. Fixed points at infinity can be studied in a variety of ways. One method involves the so-called Poincaré sphere. Another method uses what is called projective space.

9.8.1 Poincaré sphere

For two-dimensional dynamic systems, a good way is to transform the doubly-infinite phase plane onto the surface of a sphere with radius unity. The projection will be such that points at infinity are mapped onto the equator of the sphere. One can then view the sphere from the north pole and see more clearly how the dynamics develop on the surface of the sphere.

Example 9.19

Using the Poincaré sphere, find the global dynamics, including at infinity, for the simple system

$$\frac{dx}{dt} = x, \tag{9.341}$$

$$\frac{dy}{dt} = -y. \tag{9.342}$$

Obviously the equilibrium point is at $(x, y) = (0, 0)$, and that point is a saddle node. Let us project the two state variables x and y into a three-dimensional space by the mapping $\mathbb{R}^2 \to \mathbb{R}^3$:

$$X = \frac{x}{\sqrt{1 + x^2 + y^2}}, \tag{9.343}$$

$$Y = \frac{y}{\sqrt{1 + x^2 + y^2}}, \tag{9.344}$$

$$Z = \frac{1}{\sqrt{1 + x^2 + y^2}}. \tag{9.345}$$

We actually could alternatively analyze this system with a closely related mapping from $\mathbb{R}^2 \to \mathbb{R}^2$, but this makes some of the analysis less geometrically transparent.

Note that

$$\lim_{x \to \infty} X = 1 \quad \forall y < \infty, \tag{9.346}$$

$$\lim_{y \to \infty} Y = 1 \quad \forall x < \infty. \tag{9.347}$$

Note further if both x and y go to infinity, say on the line $y = mx$, then

$$\lim_{x \to \infty, y = mx} X = \frac{1}{\sqrt{m^2 + 1}}, \tag{9.348}$$

$$\lim_{x \to \infty, y = mx} Y = \frac{m}{\sqrt{m^2 + 1}}, \tag{9.349}$$

$$\lim_{x \to \infty, y = mx} X^2 + Y^2 = 1. \tag{9.350}$$

So points at infinity are mapping onto a unit circle in (X, Y) space. Also, going into the saddle node at $(x, y) = (0, 0)$ along the same line gives

$$\lim_{x \to 0, y = mx} X = x + \ldots, \tag{9.351}$$

$$\lim_{x \to 0, y = mx} Y = y + \ldots. \tag{9.352}$$

So the original and transformed space have the same essential behavior near the finite equilibrium point. Last, note that

$$X^2 + Y^2 + Z^2 = \frac{x^2 + y^2 + 1}{1 + x^2 + y^2} = 1. \tag{9.353}$$

Thus, in fact, the mapping takes one onto a unit sphere in (X, Y, Z) space. The surface $X^2 + Y^2 + Z^2 = 1$ is called the Poincaré sphere. One can actually view this in the same way one does an actual map of the surface of the Earth. Just as a Mercator[4] projection map is a representation of the spherical surface

[4]Geradus Mercator, 1512-1594, Flemish cartographer.

of the earth projected onto a flat surface (and vice versa), the original (x, y) phase space is a planar representation of the surface of the Poincaré sphere.

Let us find the inverse transformation. By inspection, it is seen that

$$x = \frac{X}{Z}, \tag{9.354}$$

$$y = \frac{Y}{Z}. \tag{9.355}$$

Now apply the transformation, Eqs. (9.354,9.355) to our dynamical system, Eqs. (9.341,9.342):

$$\underbrace{\frac{d}{dt}\left(\frac{X}{Z}\right)}_{dx/dt} = \underbrace{\frac{X}{Z}}_{x}, \tag{9.356}$$

$$\underbrace{\frac{d}{dt}\left(\frac{Y}{Z}\right)}_{dy/dt} = \underbrace{-\frac{Y}{Z}}_{-y}. \tag{9.357}$$

Expand using the quotient rule to get

$$\frac{1}{Z}\frac{dX}{dt} - \frac{X}{Z^2}\frac{dZ}{dt} = \frac{X}{Z}, \tag{9.358}$$

$$\frac{1}{Z}\frac{dY}{dt} - \frac{Y}{Z^2}\frac{dZ}{dt} = -\frac{Y}{Z}. \tag{9.359}$$

Now on the unit sphere $X^2 + Y^2 + Z^2 = 1$, we must have

$$2X dX + 2Y dY + 2Z dZ = 0, \tag{9.360}$$

so dividing by dt and solving for dZ/dt, we must have

$$\frac{dZ}{dt} = -\frac{X}{Z}\frac{dX}{dt} - \frac{Y}{Z}\frac{dY}{dt}. \tag{9.361}$$

Using Eq. (9.361) to eliminate dZ/dt in Eqs. (9.358,9.359), our dynamical system can be written as

$$\frac{1}{Z}\frac{dX}{dt} - \frac{X}{Z^2}\underbrace{\left(-\frac{X}{Z}\frac{dX}{dt} - \frac{Y}{Z}\frac{dY}{dt}\right)}_{dZ/dt} = \frac{X}{Z}, \tag{9.362}$$

$$\frac{1}{Z}\frac{dY}{dt} - \frac{Y}{Z^2}\underbrace{\left(-\frac{X}{Z}\frac{dX}{dt} - \frac{Y}{Z}\frac{dY}{dt}\right)}_{dZ/dt} = -\frac{Y}{Z}. \tag{9.363}$$

Multiply Eqs. (9.362,9.363) by Z^3 to get

$$Z^2\frac{dX}{dt} + X\left(X\frac{dX}{dt} + Y\frac{dY}{dt}\right) = Z^2 X, \tag{9.364}$$

$$Z^2\frac{dY}{dt} + Y\left(X\frac{dX}{dt} + Y\frac{dY}{dt}\right) = -Z^2 Y. \tag{9.365}$$

Regroup to find

$$(X^2 + Z^2)\frac{dX}{dt} + XY\frac{dY}{dt} = Z^2 X, \tag{9.366}$$

$$XY\frac{dX}{dt} + (Y^2 + Z^2)\frac{dY}{dt} = -Z^2 Y. \tag{9.367}$$

Now, eliminate Z by demanding $X^2 + Y^2 + Z^2 = 1$ to get

$$(1 - Y^2)\frac{dX}{dt} + XY\frac{dY}{dt} = (1 - X^2 - Y^2)X, \tag{9.368}$$

$$XY\frac{dX}{dt} + (1 - X^2)\frac{dY}{dt} = -(1 - X^2 - Y^2)Y. \tag{9.369}$$

Solve this quasi-linear system for dX/dt and dY/dt to get

$$\frac{dX}{dt} = X - X^3 + XY^2, \tag{9.370}$$

$$\frac{dY}{dt} = -Y + Y^3 - X^2Y. \tag{9.371}$$

The five equilibrium points, and their stability, for this system are easily verified to be

$$(X, Y) = (0, 0), \quad \text{saddle}, \tag{9.372}$$

$$(X, Y) = (1, 0), \quad \text{sink}, \tag{9.373}$$

$$(X, Y) = (-1, 0), \quad \text{sink}, \tag{9.374}$$

$$(X, Y) = (0, 1), \quad \text{source}, \tag{9.375}$$

$$(X, Y) = (0, -1), \quad \text{source}. \tag{9.376}$$

Note that in this space, four new equilibria have appeared. As we are also confined to the Poincaré sphere on which $X^2 + Y^2 + Z^2 = 1$, we can also see that each of the new equilibria has $Z = 0$; that is, the new equilibrium points lie on the equator of the Poincaré sphere. Transforming back to the original space, we find the equilibria are at

$$(x, y) = (0, 0), \quad \text{saddle}, \tag{9.377}$$

$$(x, y) = (\infty, 0), \quad \text{sink}, \tag{9.378}$$

$$(x, y) = (-\infty, 0), \quad \text{sink}, \tag{9.379}$$

$$(x, y) = (0, \infty), \quad \text{source}, \tag{9.380}$$

$$(x, y) = (0, -\infty), \quad \text{source}. \tag{9.381}$$

Phase portraits showing several trajectories projected into (X, Y) and (X, Y, Z) space are shown in Fig. 9.11. Fig. 9.11a represents the Poincaré sphere from above the north pole; Fig. 9.11b depicts the entire Poincaré sphere. On the sphere itself there are some additional complexities due to so-called anti-podal equilibrium points. In this example, both the north pole and the south pole are saddle equilibria, when the entire sphere is considered. For more general problems, one must realize that this projection induces pairs of equilibria, and that usually only one member of the pairs needs to be considered in detail.

Additionally, one notes in the global phase portraits two interesting features for two-dimensional phase spaces:

- except at critical points, individual trajectories never cross each other,
- all trajectories connect one critical point to another, and
- it formally takes an infinite amount of time to reach a critical point.

Any trajectory can also be shown to be a so-called *invariant manifold*. An invariant manifold is a set of points with the special property that if any one of them is used as an initial condition for the dynamic system, the time-evolution due to the dynamic system restricts the system to the invariant manifold. Certain of these manifolds are so-called *slow invariant manifolds* in that nearby trajectories

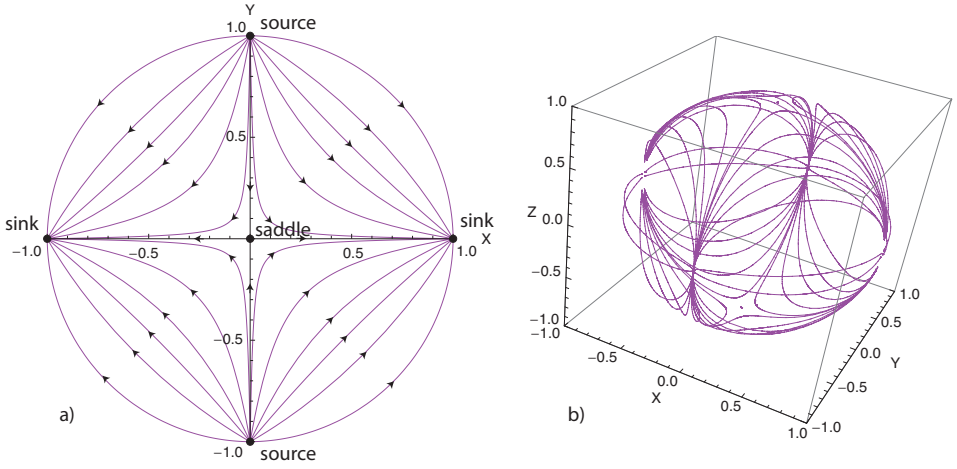

Figure 9.11: Global phase portraits of the system $dx/dt = x$, $dy/dt = -y$: a) projection from the Poincaré sphere onto the (X, Y) plane, b) full projection onto the Poincaré sphere in (X, Y, Z) space.

are attracted to them. The line $Y = 0$, and so $y = 0$, represents a slow invariant manifold for this system. Note that a finite initial condition can only approach two fixed points at infinity. But the curve representing points at infinity, $Z = 0$, is an invariant manifold. Except for trajectories that originate at the two source points, a point at infinity must remain at infinity.

9.8.2 Projective space

When extended to higher dimension, the Poincaré sphere approach becomes lengthy. A more efficient approach is provided by projective space. This approach does not have the graphical appeal of the Poincaré sphere.

Example 9.20

Using projective space, find the global dynamics, including at infinity, for the same simple system

$$\frac{dx}{dt} = x, \tag{9.382}$$

$$\frac{dy}{dt} = -y. \tag{9.383}$$

Again, it is obvious that the equilibrium point is at $(x, y) = (0, 0)$, and that point is a saddle node. Let us project the two state variables x and y into a new two-dimensional space by the mapping $\mathbb{R}^2 \to \mathbb{R}^2$:

$$X = \frac{1}{x}, \tag{9.384}$$

$$Y = \frac{y}{x}. \tag{9.385}$$

Note along the line $y = mx$, as $x \to \infty$, we get $X \to 0$, $Y \to m$. So for $x \neq 0$, a point at infinity in (x, y) space maps to a finite point in (X, Y) space. By inspection, the inverse mapping is

$$x = \frac{1}{X}, \tag{9.386}$$

$$y = \frac{Y}{X}. \tag{9.387}$$

Under this transformation, Eqs. (9.382-9.383) become

$$\frac{d}{dt}\left(\frac{1}{X}\right) = \frac{1}{X}, \tag{9.388}$$

$$\frac{d}{dt}\left(\frac{Y}{X}\right) = -\frac{Y}{X}. \tag{9.389}$$

Expanding, we find

$$-\frac{1}{X^2}\frac{dX}{dt} = \frac{1}{X}, \tag{9.390}$$

$$\frac{1}{X}\frac{dY}{dt} - \frac{Y}{X^2}\frac{dX}{dt} = -\frac{Y}{X}. \tag{9.391}$$

Simplifying gives

$$\frac{dX}{dt} = -X, \tag{9.392}$$

$$X\frac{dY}{dt} - Y\frac{dX}{dt} = -XY. \tag{9.393}$$

Solving for the derivatives, the system reduces to

$$\frac{dX}{dt} = -X, \tag{9.394}$$

$$\frac{dY}{dt} = -2Y. \tag{9.395}$$

By inspection, there is a sink at $(X, Y) = (0, 0)$. At such a point, the inverse mapping tells us $x \to \pm\infty$ depending on whether X is positive or negative, and y is indeterminate. If we approach $(X, Y) = (0, 0)$ along the line $Y = mX$, then y approaches the finite number m. This is consistent with trajectories being swept away from the origin towards $x \to \pm\infty$ in the original phase space, indicating an attraction at $x \to \pm\infty$. But it does not account for the trajectories emanating from $y \to \pm\infty$. This is because the transformation selected obscured this root.

To recover it, we can consider the alternate transformation $\hat{X} = x/y$, $\hat{Y} = 1/y$. Doing so leads to the system $d\hat{X}/dt = 2\hat{X}$, $d\hat{Y}/dt = \hat{Y}$, which has a source at $(\hat{X}, \hat{Y}) = (0, 0)$, which is consistent with the source-like behavior in the original x, y space as $y \to \pm\infty$. This transformation, however, obscures the sink like behavior at $x \to \pm\infty$.

To capture both points at infinity, we can consider a non-degenerate transformation, of which there are infinitely many. One is $\tilde{X} = 1/(x + y)$, $\tilde{Y} = (x - y)/(x + y)$. Doing so leads to the system $d\tilde{X}/dt = -\tilde{X}\tilde{Y}$, $d\tilde{Y}/dt = 1 - \tilde{Y}^2$. This system has two roots, a source at $(\tilde{X}, \tilde{Y}) = (0, -1)$ and a sink at $(\tilde{X}, \tilde{Y}) = (0, 1)$. The source corresponds to $y \to \pm\infty$. The sink corresponds to $x \to \pm\infty$.

9.9 Fractals

In the discussion on attractors in Section 9.6.1, we included geometrical shapes called fractals. These are objects that are not smooth, but occur frequently in the dynamical systems literature either as attractors or as boundaries of basins of attractions.

A fractal can be defined as a geometrical shape in which the parts are in some way similar to the whole. This self-similarity may be exact, i.e. a piece of the fractal, if magnified, may look exactly like the whole fractal. Before discussing examples we need to put forward a working definition of dimension. Though there are many definitions in current use, we present here the Hausdorff-Besicovitch[5] dimension D. If N_ϵ is the number of 'boxes' of side length ϵ needed to cover an object, then

$$D = \lim_{\epsilon \to 0} \frac{\ln N_\epsilon}{\ln(1/\epsilon)}. \tag{9.396}$$

We can check that this definition corresponds to the common geometrical shapes.

1. Point: $N_\epsilon = 1, D = 0$ since $D = \lim_{\epsilon \to 0} \frac{\ln 1}{-\ln \epsilon} = 0$,
2. Line of length l: $N_\epsilon = l/\epsilon, D = 1$ since $D = \lim_{\epsilon \to 0} \frac{\ln(l/\epsilon)}{-\ln \epsilon} = \frac{\ln l - \ln \epsilon}{-\ln \epsilon} = 1$,
3. Surface of size l^2: $N_\epsilon = (l/\epsilon)^2, D = 2$ since $D = \lim_{\epsilon \to 0} \frac{\ln(l^2/\epsilon^2)}{-\ln \epsilon} = \frac{2\ln l - 2\ln \epsilon}{-\ln \epsilon} = 2$,
4. Volume of size l^3: $N_\epsilon = (l/\epsilon)^3, D = 3$ since $D = \lim_{\epsilon \to 0} \frac{\ln(l^3/\epsilon^3)}{-\ln \epsilon} = \frac{3\ln l - 3\ln \epsilon}{-\ln \epsilon} = 3$.

A fractal has a dimension that is not an integer. Many physical objects are fractal-like, in that they are fractal within a range of length scales. Coastlines are among the geographical features that are of this shape. If there are N_ϵ units of a measuring stick of length ϵ, the measured length of the coastline will be of the power-law form $\epsilon N_\epsilon = \epsilon^{1-D}$, where D is the dimension.

9.9.1 Cantor set

Consider the line corresponding to $k = 0$ in Fig. 9.12. Take away the middle third to leave

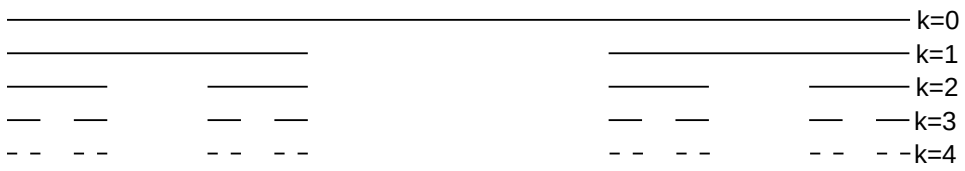

Figure 9.12: Cantor set.

[5] after Felix Hausdorff, 1868-1942, German mathematician, and Abram Samoilovitch Besicovitch, 1891-1970, Russian mathematician.

the two portions; this is shown as $k = 1$. Repeat the process to get $k = 2, 3, \ldots$. If $k \to \infty$, what is left is called the Cantor[6] set. Let us take the length of the line segment to be unity when $k = 0$. Since $N_\epsilon = 2^k$ and $\epsilon = 1/3^k$, the dimension of the Cantor set is

$$D = \lim_{\epsilon \to 0} \frac{\ln N_\epsilon}{\ln(1/\epsilon)} = \lim_{k \to \infty} \frac{\ln 2^k}{\ln 3^k} = \frac{k \ln 2}{k \ln 3} = \frac{\ln 2}{\ln 3} = 0.6309 \ldots . \tag{9.397}$$

It can be seen that the endpoints of the removed intervals are never removed; it can be shown the Cantor set contains an infinite number of points, and it is an uncountable set. It is totally disconnected and has a Lebesgue measure zero.

9.9.2 Koch curve

Here we start with an equilateral triangle shown in Fig. 9.13 as $k = 0$. Each side of the

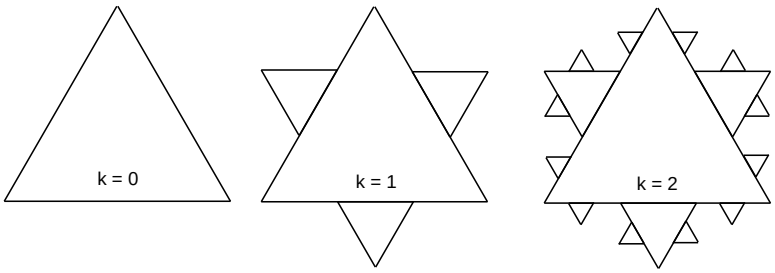

Figure 9.13: Koch curve.

original triangle has unit length. The middle third of each side of the triangle is removed, and two sides of a triangle drawn on that. This is shown as $k = 1$. The process is continued, and in the limit gives a continuous, closed curve that is nowhere smooth. Since $N_\epsilon = 3 \times 4^k$ and $\epsilon = 1/3^k$, the dimension of the Koch[7] curve is

$$D = \lim_{\epsilon \to 0} \frac{\ln N_\epsilon}{\ln(1/\epsilon)} = \lim_{k \to \infty} \frac{\ln(3)4^k}{\ln 3^k} = \lim_{k \to \infty} \frac{\ln 3 + k \ln 4}{k \ln 3} = \frac{\ln 4}{\ln 3} = 1.261 \ldots . \tag{9.398}$$

The limit curve itself has infinite length, it is nowhere differentiable, and it surrounds a finite area.

9.9.3 Menger sponge

An example of a fractal which is an iterate of an object which starts in three-dimensional space is a "Menger sponge."[8] A Menger sponge is depicted in Fig. 9.14.

[6]Georg Ferdinand Ludwig Philipp Cantor, 1845-1918, Russian-born, German-based mathematician.

[7]Niels Fabian Helge von Koch, 1870-1924, Swedish mathematician.

[8]Karl Menger, 1902-1985, Austrian-born mathematician and member of the influential "Vienna Circle." He served on the faculties of the Universities of Amsterdam, Vienna, Notre Dame, and the Illinois Institute of Technology.

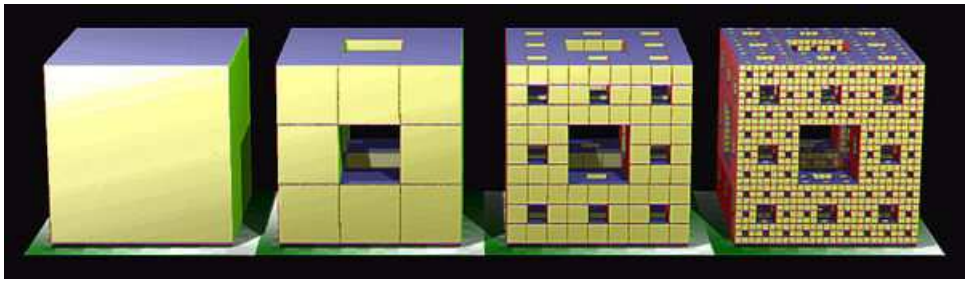

Figure 9.14: Menger sponge.

9.9.4 Weierstrass function

For $a,\ b,\ t \in \mathbb{R}^1, W : \mathbb{R}^1 \to \mathbb{R}^1$, the Weierstrass[9] function

$$W(t) = \sum_{k=1}^{\infty} a^k \cos b^k t, \tag{9.399}$$

where a is real, b is odd, and $ab > 1 + 3\pi/2$. It is everywhere continuous, but nowhere differentiable! Both require some effort to prove. A Weierstrass function is plotted in Fig. 9.15. Its

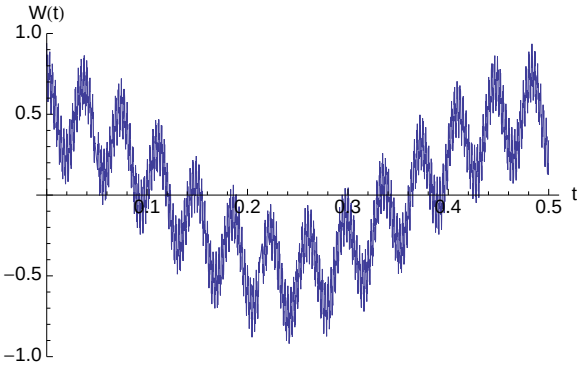

Figure 9.15: Four term ($k = 1, \ldots, 4$) approximation to the Weierstrass function, $W(t)$ for $b = 13$, $a = 1/2$.

fractal character can be seen when one recognizes that cosine waves of ever higher frequency are superposed onto low frequency cosine waves.

9.9.5 Mandelbrot and Julia sets

For $z \in \mathbb{C}^1, c \in \mathbb{C}^1$, the Mandelbrot[10] set is the set of all c for which

$$z_{k+1} = z_k^2 + c, \tag{9.400}$$

[9] Karl Theodor Wilhelm Weierstrass, 1815-1897, Westphalia-born German mathematician.
[10] Benoît Mandelbrot, 1924-2010, Polish-born mathematician based mainly in France.

stays bounded as $k \to \infty$, when $z_0 = 0$. The boundaries of this set are fractal. A Mandelbrot set is sketched in Fig. 9.16.

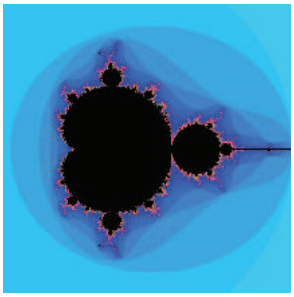

Figure 9.16: Mandelbrot set. Black regions stay bounded; colored regions become unbounded with shade indicating how rapidly the system becomes unbounded. Image generated from `http://cs.clarku.edu/~djoyce/julia/explorer.html`.

Associated with each c for the Mandelbrot set is a Julia[11] set. In this case, the Julia set is the set of complex initial seeds z_0 which allow $z_{k+1} = z_k^2 + c$ to converge for fixed complex c. A Julia set for $c = 0.49 + 0.57i$ is plotted in Fig. 9.17.

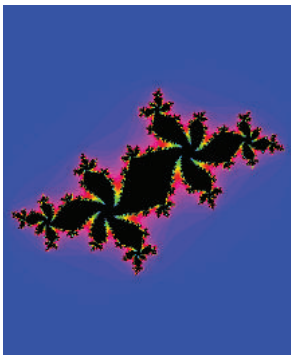

Figure 9.17: Julia set for $c = 0.49+0.57i$. Black regions stay bounded; colored regions become unbounded with shade of color indicating how rapidly the system becomes unbounded. Image generated from `http://cs.clarku.edu/~djoyce/julia/explorer.html`.

9.10 Bifurcations

Dynamical systems representing some physical problem frequently have parameters associated with them. Thus, for $x \in \mathbb{R}^N, t \in \mathbb{R}^1, r \in \mathbb{R}^1, f : \mathbb{R}^N \to \mathbb{R}^N$, we can write

$$\frac{dx_n}{dt} = f_n(x_1, x_2, \cdots, x_N; r) \ (n = 1, \cdots, N), \tag{9.401}$$

[11] Gaston Maurice Julia, 1893-1978, Algerian-born French mathematician.

where r is a parameter. The theory can easily be extended if there is more than one parameter.

We would like to consider the changes in the behavior of $t \to \infty$ solutions as the real number r, called the *bifurcation parameter*, is varied. The nature of the critical point may change as the parameter r is varied; other critical points may appear or disappear, or its stability may change. This is a bifurcation, and the r at which it happens is the bifurcation point. The study of the solutions and bifurcations of the steady state falls under *singularity theory*.

Let us look at some of the bifurcations obtained for different vector fields. Some of the examples will be one-dimensional, i.e. $x \in \mathbb{R}^1, r \in \mathbb{R}^1, f : \mathbb{R}^1 \to \mathbb{R}^1$.

$$\frac{dx}{dt} = f(x; r). \tag{9.402}$$

Even though this can be solved exactly in most cases, we will assume that such a solution is not available so that the techniques of analysis can be developed for more complicated systems. For a coefficient matrix that is a scalar, the eigenvalue is the coefficient itself. The eigenvalue will be real and will cross the imaginary axis of the complex plane through the origin as r is changed. This is called a simple bifurcation.

9.10.1 Pitchfork bifurcation

For $x \in \mathbb{R}^1, t \in \mathbb{R}^1, r \in \mathbb{R}^1, r_0 \in \mathbb{R}^1$, consider

$$\frac{dx}{dt} = -x(x^2 - (r - r_0)). \tag{9.403}$$

The critical points are $\overline{x} = 0$, and $\pm\sqrt{r - r_0}$. $r = r_0$ is a *bifurcation point*; for $r < r_0$ there is only one critical point, while for $r > r_0$ there are three.

Linearizing around the critical point $\overline{x} = 0$, we get

$$\frac{d\tilde{x}}{dt} = (r - r_0)\tilde{x}. \tag{9.404}$$

This has solution

$$\tilde{x}(t) = \tilde{x}(0) \exp\left((r - r_0)t\right). \tag{9.405}$$

For $r < r_0$, the critical point is asymptotically stable; for $r > r_0$ it is unstable.

Notice that the function $V(x) = x^2$ satisfies the following conditions: $V > 0$ for $x \neq 0$, $V = 0$ for $x = 0$, and $dV/dt = (dV/dx)(dx/dt) = -2x^2(x^2 - (r - r_0)) \leq 0$ for $r < r_0$. Thus, $V(x)$ is a Lyapunov function and $\overline{x} = 0$ is globally stable for all perturbations, large or small, as long as $r < r_0$.

Now let us examine the critical point $\overline{x} = \sqrt{r - r_0}$ which exists only for $r > r_0$. Putting $x = \overline{x} + \tilde{x}$, the right side of Eq. (9.403) becomes

$$f(x) = -\left(\sqrt{r - r_0} + \tilde{x}\right)\left(\left(\sqrt{r - r_0} + \tilde{x}\right)^2 - (r - r_0)\right). \tag{9.406}$$

Linearizing for small $\tilde{x}$, we get

$$\frac{d\tilde{x}}{dt} = -2(r - r_0)\tilde{x}. \tag{9.407}$$

This has solution

$$\tilde{x}(t) = \tilde{x}(0) \exp\left(-2(r - r_0)t\right). \tag{9.408}$$

For $r > r_0$, this critical point is stable. The other critical point $\overline{x} = -\sqrt{r - r_0}$ is also found to be stable for $r > r_0$. The results are summarized in the bifurcation diagram sketched in Figure 9.18. At the bifurcation point, $r = r_0$, we have

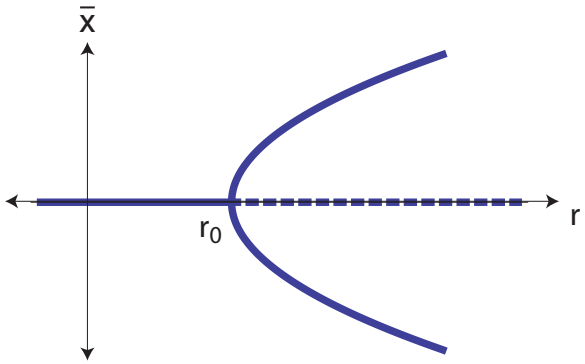

Figure 9.18: Sketch of a pitchfork bifurcation. Heavy lines are stable equilibria; dashed lines are unstable equilibria.

$$\frac{dx}{dt} = -x^3. \tag{9.409}$$

This equation has a critical point at $x = 0$ but has no linearization. We must do a non-linear analysis to determine the stability of the critical point. In this case it is straightforward. Solving directly and applying an initial condition, we obtain

$$x(t) = \pm \frac{x(0)}{\sqrt{1 + 2x(0)^2 t}}, \tag{9.410}$$

$$\lim_{t \to \infty} x(t) = 0. \tag{9.411}$$

Since the system approaches the critical point as $t \to \infty$ for all values of $x(0)$, the critical point $x = 0$ is unconditionally stable.

9.10.2 Transcritical bifurcation

For $x \in \mathbb{R}^1, t \in \mathbb{R}^1, r \in \mathbb{R}^1, r_0 \in \mathbb{R}^1$, consider

$$\frac{dx}{dt} = -x(x - (r - r_0)). \tag{9.412}$$

The critical points are $\overline{x} = 0$ and $r - r_0$. The bifurcation occurs at $r = r_0$. Once again the linear stability of the solutions can be determined. Near $\overline{x} = 0$, the linearization is

$$\frac{d\tilde{x}}{dt} = (r - r_0)\tilde{x}, \tag{9.413}$$

which has solution

$$\tilde{x}(t) = \tilde{x}(0) \exp\left((r - r_0)t\right). \tag{9.414}$$

So this solution is stable for $r < r_0$. Near $\overline{x} = r - r_0$, we take $\tilde{x} = x - (r - r_0)$. The resulting linearization is

$$\frac{d\tilde{x}}{dt} = -(r - r_0)\tilde{x}, \tag{9.415}$$

which has solution

$$\tilde{x}(t) = \tilde{x}(0) \exp\left(-(r - r_0)t\right). \tag{9.416}$$

So this solution is stable for $r > r_0$.

At the bifurcation point, $r = r_0$, there is no linearization, and the system becomes

$$\frac{dx}{dt} = -x^2, \tag{9.417}$$

which has solution

$$x(t) = \frac{x(0)}{1 + x(0)t}. \tag{9.418}$$

Here the asymptotic stability depends on the initial condition! For $x(0) \geq 0$, the critical point at $x = 0$ is stable. For $x(0) < 0$, there is a blowup phenomena at $t = -1/x(0)$. The results are summarized in the bifurcation diagram sketched in Figure 9.19.

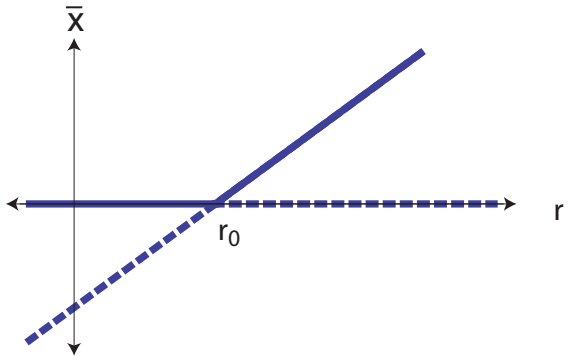

Figure 9.19: Sketch of a transcritical bifurcation. Heavy lines are stable equilibria; dashed lines are unstable equilibria.

9.10.3 Saddle-node bifurcation

For $x \in \mathbb{R}^1, t \in \mathbb{R}^1, r \in \mathbb{R}^1, r_0 \in \mathbb{R}^1$, consider

$$\frac{dx}{dt} = -x^2 + (r - r_0). \tag{9.419}$$

The critical points are $\overline{x} = \pm\sqrt{r - r_0}$. Taking $\tilde{x} = x \mp \sqrt{r - r_0}$ and linearizing, we obtain

$$\frac{d\tilde{x}}{dt} = \mp 2\sqrt{r - r_0}\tilde{x}, \tag{9.420}$$

which has solution

$$\tilde{x}(t) = \tilde{x}(0) \exp\left(\mp\sqrt{r - r_0}t\right). \tag{9.421}$$

For $r > r_0$, the root $x = +\sqrt{r - r_0}$ is asymptotically stable. The root $x = -\sqrt{r - r_0}$ is asymptotically unstable.

At the point, $r = r_0$, there is no linearization, and the system becomes

$$\frac{dx}{dt} = -x^2, \tag{9.422}$$

which has solution

$$x(t) = \frac{x(0)}{1 + x(0)t}. \tag{9.423}$$

Here the asymptotic stability again depends on the initial condition For $x(0) \geq 0$, the critical point at $x = 0$ is stable. For $x(0) < 0$, there is a blowup phenomena at $t = -1/x(0)$. The results are summarized in the bifurcation diagram sketched in Figure 9.20.

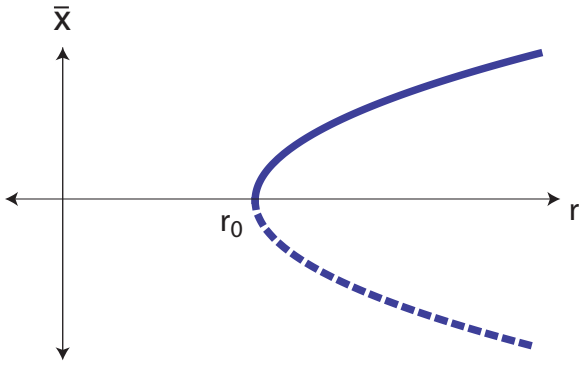

Figure 9.20: Sketch of saddle-node bifurcation. Heavy lines are stable equilibria; dashed lines are unstable equilibria.

9.10.4 Hopf bifurcation

To give an example of complex eigenvalues, one must go to a two-dimensional vector field.

Example 9.21

With $x, y, t, r, r_0 \in \mathbb{R}^1$, take

$$\frac{dx}{dt} = (r - r_0)x - y - x(x^2 + y^2), \tag{9.424}$$

$$\frac{dy}{dt} = x + (r - r_0)y - y(x^2 + y^2). \tag{9.425}$$

The origin (0,0) is a critical point. The linearized perturbation equations are

$$\frac{d}{dt}\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} r - r_0 & -1 \\ 1 & r - r_0 \end{pmatrix}\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}. \tag{9.426}$$

The eigenvalues λ of the coefficient matrix are $\lambda = (r - r_0) \pm i$. For $r < r_0$, the real part is negative, and the origin is stable. At $r = r_0$ there is a Hopf[12] bifurcation as the eigenvalues cross the imaginary axis of the complex plane as r is changed. For $r > r_0$, a periodic orbit in the (x, y) phase plane appears. The linear analysis will not give the amplitude of the motion. Writing the given equation in polar coordinates (ρ, θ) yields

$$\frac{d\rho}{dt} = \rho(r - r_0) - \rho^3, \tag{9.427}$$

$$\frac{d\theta}{dt} = 1. \tag{9.428}$$

This is a pitchfork bifurcation in the amplitude of the oscillation ρ.

9.11 Lorenz equations

For independent variable $t \in \mathbb{R}^1$, dependent variables $(x, y, z)^T \in \mathbb{R}^3$, and parameters $\sigma, r, b \in \mathbb{R}^1$, $\sigma > 0$, $r > 0$, $b > 0$, the Lorenz[13] equations are

$$\frac{dx}{dt} = \sigma(y - x), \tag{9.429}$$

$$\frac{dy}{dt} = rx - y - xz, \tag{9.430}$$

$$\frac{dz}{dt} = -bz + xy. \tag{9.431}$$

The bifurcation parameter will be taken to be r.

[12] Eberhard Frederich Ferdinand Hopf, 1902-1983, Austrian-born, German mathematician.

[13] Edward Norton Lorenz, 1917-2008, American meteorologist.

9.11.1 Linear stability

The critical points are obtained from

$$\begin{aligned} \overline{y} - \overline{x} &= 0, & (9.432)\\ r\overline{x} - \overline{y} - \overline{x}\,\overline{z} &= 0, & (9.433)\\ -b\overline{z} + \overline{x}\,\overline{y} &= 0, & (9.434) \end{aligned}$$

which gives

$$\begin{pmatrix} \overline{x} \\ \overline{y} \\ \overline{z} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sqrt{b(r-1)} \\ \sqrt{b(r-1)} \\ r-1 \end{pmatrix}, \begin{pmatrix} -\sqrt{b(r-1)} \\ -\sqrt{b(r-1)} \\ r-1 \end{pmatrix}. \tag{9.435}$$

Note when $r = 1$, there is only one critical point at the origin. For more general r, a linear stability analysis of each of the three critical points follows.

- $\overline{x} = \overline{y} = \overline{z} = 0$. Small perturbations around this point give

$$\frac{d}{dt}\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix}. \tag{9.436}$$

  The characteristic equation is

$$(\lambda + b)(\lambda^2 + \lambda(\sigma + 1) - \sigma(r - 1)) = 0, \tag{9.437}$$

  from which we get the eigenvalues

$$\lambda = -b, \qquad \lambda = \frac{1}{2}\left(-(1+\sigma) \pm \sqrt{(1+\sigma)^2 - 4\sigma(1-r)}\right). \tag{9.438}$$

  For $0 < r < 1$, the eigenvalues are real and negative, since $(1 + \sigma)^2 > 4\sigma(1 - r)$. At $r = 1$, there is a pitchfork bifurcation with one zero eigenvalue. For $r > 1$, the origin becomes unstable.

- $\overline{x} = \overline{y} = \sqrt{b(r-1)}$, $\overline{z} = r - 1$. We first note we need $r \geq 1$ for a real solution. Small perturbations give

$$\frac{d}{dt}\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\sqrt{b(r-1)} \\ \sqrt{b(r-1)} & \sqrt{b(r-1)} & -b \end{pmatrix}\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix}. \tag{9.439}$$

  The characteristic equation is

$$\lambda^3 + (\sigma + b + 1)\lambda^2 + (\sigma + r)b\lambda + 2\sigma b(r - 1) = 0. \tag{9.440}$$

This system is difficult to fully analyze. Detailed analysis reveals of a critical value of r:

$$r = r_c = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}. \tag{9.441}$$

At $r = r_c$ the characteristic equation, Eq. (9.440), can be factored to give the eigenvalues

$$\lambda = -(\sigma + b + 1), \qquad \lambda = \pm i\sqrt{\frac{2b\sigma(\sigma + 1)}{\sigma - b - 1}}, \tag{9.442}$$

If $\sigma > b + 1$, two of the eigenvalues are purely imaginary, and this corresponds to a Hopf bifurcation. The periodic solution which is created at this value of r can be shown to be unstable so that the bifurcation is subcritical.

If $r = r_c$ and $\sigma < b+1$, one can find all real eigenvalues, including at least one positive eigenvalue, which tells us this is unstable.

We also find instability if $r > r_c$. If $r > r_c$ and $\sigma > b + 1$, we can find one negative real eigenvalue and two complex eigenvalues with positive real parts; hence, this is unstable. If $r > r_c$, and $\sigma < b+1$, we can find three real eigenvalues, with at least one positive; this is unstable.

For $1 < r < r_c$ and $\sigma < b + 1$, we find three real eigenvalues, one of which is positive; this is unstable.

For stability, we can take

$$1 < r < r_c, \quad \text{and} \quad \sigma > b + 1. \tag{9.443}$$

In this case, we can find one negative real eigenvalue and two eigenvalues (which could be real or complex) with negative real parts; hence, this is stable.

- $\overline{x} = \overline{y} = -\sqrt{b(r-1)}$, $\overline{z} = r - 1$. Analysis of this critical point is essentially identical to that of the previous point.

For a particular case, these results are summarized in the bifurcation diagram of Fig. 9.21. Shown here are results when $\sigma = 10$, $b = 8/3$. For these values, Eq. (9.441) tells us $r_c = 24.74$. Note also that $\sigma > b + 1$. For real equilibria, we need $r > 0$. The equilibrium at the origin is stable for $r \in [0, 1]$ and unstable for $r > 1$; the instability is denoted by the dashed line. At $r = 1$, there is a pitchfork bifurcation, and two new real equilibria are available. These are both linearly stable for $r \in [1, r_c]$. For $r \in [1, 1.34562]$, the eigenvalues are both real and negative. For $r \in [1.134562, r_c]$, two of the eigenvalues become complex, but all three have negative real parts, so local linear stability is maintained. For $r > r_c$, all three equilibria are unstable and indicated by dashed lines. As an aside, we note that because of non-linear effects, some initial conditions in fact yield trajectories which do not relax to a stable equilibrium for $r < r_c$. It can be shown, for example, that if $x(0) = y(0) = z(0) = 1$, that $r = 24 < r_c$ gives rise to a trajectory which never reaches either of the linearly stable critical points.

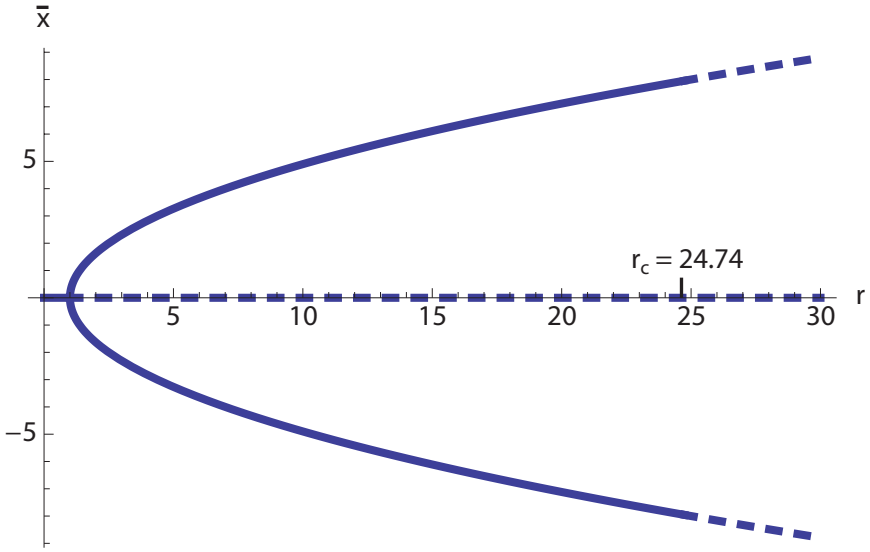

Figure 9.21: Bifurcation diagram for Lorenz equations, with $\sigma = 10$, $b = 8/3$.

9.11.2 Non-linear stability: center manifold projection

This is a procedure for obtaining the non-linear behavior near an eigenvalue with zero real part. As an example we will look at the Lorenz system at the bifurcation point $r = 1$. Recall when $r = 1$, the Lorenz equations have a single equilibrium at the origin. Linearization of the Lorenz equations near the equilibrium point at $(0, 0, 0)$ gives rise to a system of the form $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$, where

$$\mathbf{A} = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}. \tag{9.444}$$

The matrix $\mathbf{A}$ has eigenvalues and eigenvectors

$$\lambda_1 = 0, \qquad \mathbf{e}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \tag{9.445}$$

$$\lambda_2 = -(\sigma + 1), \qquad \mathbf{e}_2 = \begin{pmatrix} -\sigma \\ 1 \\ 0 \end{pmatrix}, \tag{9.446}$$

$$\lambda_3 = -b, \qquad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{9.447}$$

The fact that $\lambda_1 = 0$ suggests that there is a local algebraic dependency between at least two of the state variables, and that locally, the system behaves as a differential-algebraic system, such as studied in Sec. 9.7.

We use the eigenvectors as a basis to define new coordinates (u, v, w) where

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & -\sigma & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix}. \tag{9.448}$$

This linear transformation has a Jacobian whose determinant is $J = 1+\sigma$; thus, for $\sigma > -1$, it is orientation-preserving. It is volume-preserving only if $\sigma = 0$ or -2. Inversion shows that

$$u = \frac{x + \sigma y}{1+\sigma}, \tag{9.449}$$

$$v = \frac{y - x}{1+\sigma}, \tag{9.450}$$

$$w = z. \tag{9.451}$$

In terms of the new variables, the derivatives are expressed as

$$\frac{dx}{dt} = \frac{du}{dt} - \sigma\frac{dv}{dt}, \tag{9.452}$$

$$\frac{dy}{dt} = \frac{du}{dt} + \frac{dv}{dt}, \tag{9.453}$$

$$\frac{dz}{dt} = \frac{dw}{dt}, \tag{9.454}$$

so that original non-linear Lorenz equations (9.429-9.431) become

$$\frac{du}{dt} - \sigma\frac{dv}{dt} = \sigma(1+\sigma)v, \tag{9.455}$$

$$\frac{du}{dt} + \frac{dv}{dt} = -(1+\sigma)v - (u - \sigma v)w, \tag{9.456}$$

$$\frac{dw}{dt} = -bw + (u - \sigma v)(u + v). \tag{9.457}$$

Solving directly for the derivatives so as to place the equations in autonomous form, we get

$$\frac{du}{dt} = 0u - \frac{\sigma}{1+\sigma}(u - \sigma v)w = \lambda_1 u + \text{non-linear terms}, \tag{9.458}$$

$$\frac{dv}{dt} = -(1+\sigma)v - \frac{1}{1+\sigma}(u - \sigma v)w = \lambda_2 v + \text{non-linear terms}, \tag{9.459}$$

$$\frac{dw}{dt} = -bw + (u - \sigma v)(u + v) = \lambda_3 w + \text{non-linear terms}. \tag{9.460}$$

The objective of using the eigenvectors as basis vectors is to change the original system to diagonal form in the linear terms. Notice that the linear portion of the system is in diagonal form with the coefficients on each linear term as a distinct eigenvalue. Furthermore, the

eigenvalues $\lambda_2 = -(1+\sigma)$ and $\lambda_3 = -b$ are negative ensuring that the linear behavior $v = e^{-(1+\sigma)t}$ and $w = e^{-bt}$ takes the solution very quickly to zero in these variables.

It would appear then that we are only left with an equation in $u(t)$ for large t. However, if we put $v = w = 0$ in the right side, dv/dt and dw/dt would be zero if it were not for the u^2 term in dw/dt, implying that the dynamics is confined to $v = w = 0$ only if we ignore this term. According to the center manifold theorem it is possible to find a manifold (called the center manifold) which is tangent to $u = 0$, but is not necessarily the tangent itself, to which the dynamics is indeed confined.

We can get as good an approximation to the center manifold as we want by choosing new variables. Expanding Eq. (9.460), which has the potential problem, we get

$$\frac{dw}{dt} = -bw + u^2 - (\sigma - 1)uv - \sigma v^2. \tag{9.461}$$

Letting

$$\tilde{w} = w - \frac{u^2}{b}, \tag{9.462}$$

so that $-bw + u^2 = -b\tilde{w}$, we can eliminate the potential problem with the derivative of w.

In the new variables $(u, v, \tilde{w})$, the full Lorenz equations are written as

$$\frac{du}{dt} = -\frac{\sigma}{1+\sigma}(u - \sigma v)\left(\tilde{w} + \frac{u^2}{b}\right), \tag{9.463}$$

$$\frac{dv}{dt} = -(1+\sigma)v - \frac{1}{1+\sigma}(u - \sigma v)\left(\tilde{w} + \frac{u^2}{b}\right), \tag{9.464}$$

$$\frac{d\tilde{w}}{dt} = -b\tilde{w} - (\sigma - 1)uv - \sigma v^2 + \frac{2\sigma}{b(1+\sigma)}u(u - \sigma v)\left(\tilde{w} + \frac{u^2}{b}\right). \tag{9.465}$$

Once again, the variables v and $\tilde{w}$ go to zero quickly. Formally setting them to zero, we recast Eqs. (9.463-9.465) as

$$\frac{du}{dt} = -\frac{\sigma}{b(1+\sigma)}u^3, \tag{9.466}$$

$$\frac{dv}{dt} = -\frac{1}{b(1+\sigma)}u^3, \tag{9.467}$$

$$\frac{d\tilde{w}}{dt} = \frac{2\sigma}{b^2(1+\sigma)}u^4. \tag{9.468}$$

Here, dv/dt and $d\tilde{w}/dt$ approach zero if u approaches zero. Now the equation for the evolution of u, Eq. (9.466), suggests that this is the case. Simply integrating Eq. (9.466) and applying an initial condition, we get

$$u(t) = \pm(u(0))\sqrt{\frac{b(1+\sigma)}{b(1+\sigma) + 2\sigma(u(0))^2 t}}, \tag{9.469}$$

which is asymptotically stable as $t \to \infty$. So to this approximation the dynamics is confined to the $v = \tilde{w} = 0$ line. The bifurcation at $r = 1$ is said to be *supercritical*. Higher order terms can be included to obtain improved accuracy, if necessary.

We next focus attention on a particular case where the parameters were chosen to be $r = 1$, $\sigma = 1$, and $b = 8/3$. Figure 9.22 gives the projection onto the (u, w) phase space of several solution trajectories calculated in (u, v, w) phase space for a wide variety of initial conditions along with the center manifold, $\tilde{w} = w - u^2/b = 0$. It is seen that a given solution

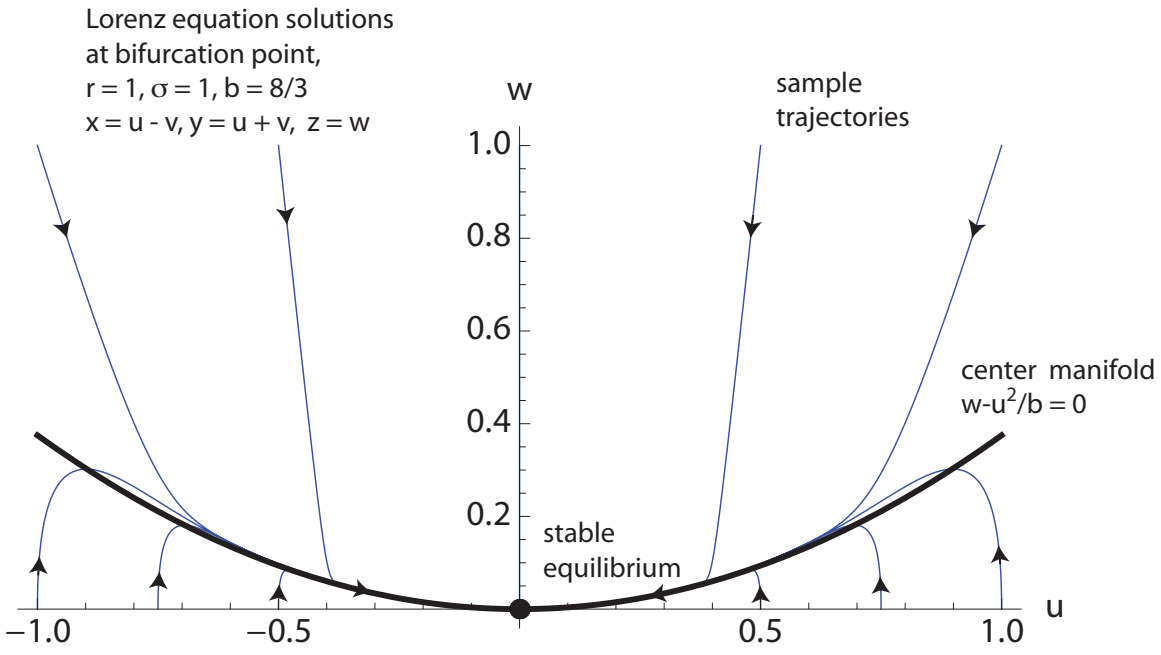

Figure 9.22: Projection onto the (u, w) plane of solution trajectories (blue curves) and center manifold (black curve) for Lorenz equations at the bifurcation point; $r = 1$, $\sigma = 1$, $b = 8/3$.

trajectory indeed approaches the center manifold on its way to the equilibrium point at the origin. The center manifold approximates the solution trajectories well in the neighborhood of the origin. Far from the origin, not shown here, it is not an attracting manifold.

We can gain more insight into the center manifold by transforming back into (x, y, z) space. Figure 9.23 shows in that space several solution trajectories, a representation of the surface which constitutes the center manifold, as well as a curve embedded within the center manifold to which trajectories are further attracted. We can in fact decompose the motion of a trajectory into the following regimes, for the parameters $r = 1$, $\sigma = 1$, $b = 8/3$.

- *Very fast attraction to the two-dimensional center manifold,* $\tilde{w} = 0$: Because $b > \sigma + 1$, for this case, $\tilde{w}$ approaches zero faster than v approaches zero, via exponential decay dictated by Eqs. (9.464, 9.465). So on a time scale of $1/b$, the trajectory first approaches $\tilde{w} = 0$, which means it approaches $w - u^2/b = 0$. Transforming back to (x, y, z) via

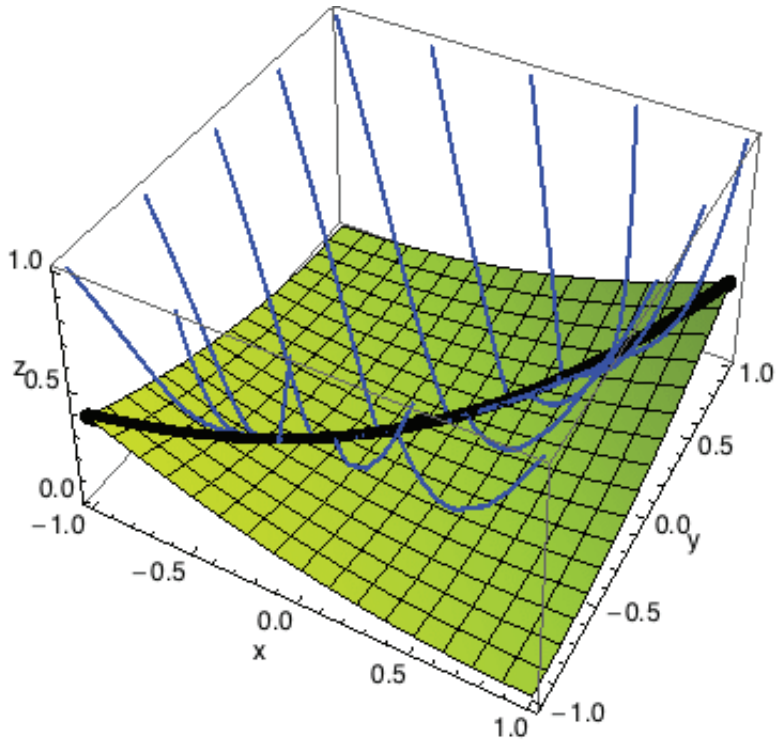

Figure 9.23: Solution trajectories (blue curves) and center manifold (green surface and black curve) for Lorenz equations at the bifurcation point; $r = 1$, $\sigma = 1$, $b = 8/3$.

Eqs. (9.449-9.451), a trajectory thus approaches the surface

$$z = \frac{1}{b}\left(\underbrace{\frac{x}{1+\sigma} + \frac{\sigma y}{1+\sigma}}_{u}\right)^2\Bigg|_{\sigma=1,b=8/3} = \frac{3}{8}\left(\frac{x+y}{2}\right)^2. \tag{9.470}$$

- *Fast attraction to the one-dimensional curve, $v = 0$:* Once on the two dimensional manifold, the slower time scale relaxation with time constant $1/(\sigma+1)$ to the curve given by $v = 0$ occurs. When $v = 0$, we also have $x = y$, so this curve takes the parametric form

$$x(s) = s, \tag{9.471}$$
$$y(s) = s, \tag{9.472}$$
$$z(s) = \frac{1}{b}\left(\frac{s}{1+\sigma} + \frac{\sigma s}{1+\sigma}\right)^2\Bigg|_{\sigma=1,b=8/3} = \frac{3}{8}s^2. \tag{9.473}$$

- *Slow attraction to the zero-dimensional equilibrium point at* $(0,0,0)$: This final relaxation brings the system to rest.

For different parameters, this sequence of events is modified, as depicted in Fig. 9.24. In

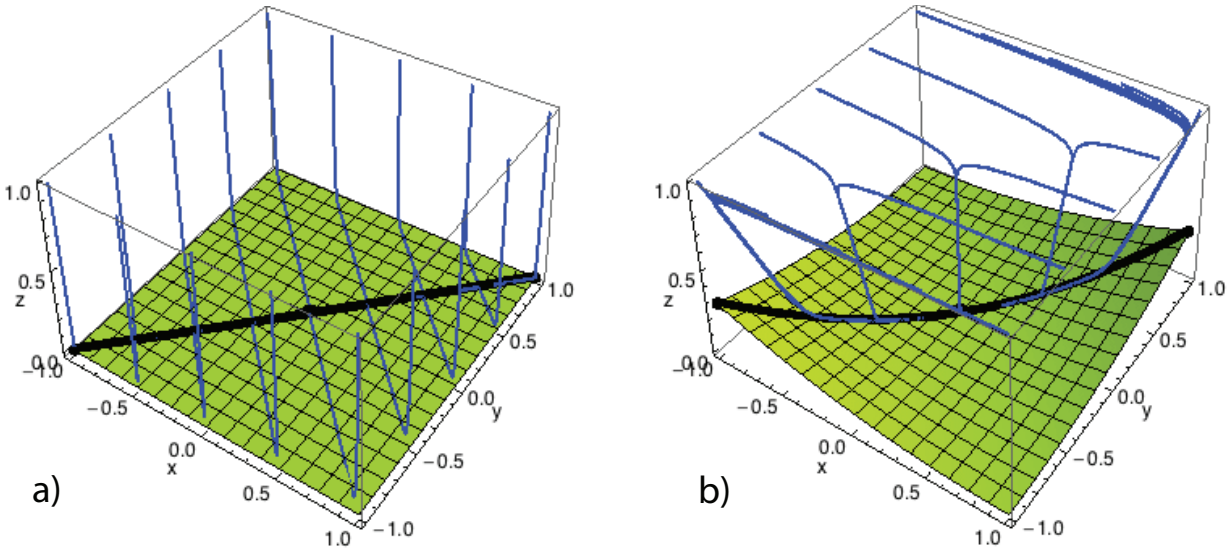

Figure 9.24: Solution trajectories (blue curves) and center manifold (green surface and black curve) for Lorenz equations at the bifurcation point; a) $r = 1$, $\sigma = 1$, $b = 100$, and b) $r = 1$, $\sigma = 100$, $b = 8/3$.

Fig. 9.24a, we take $r = 1$, $\sigma = 1$, $b = 100$. By Eqs. (9.464,9.465), these parameters induce an even faster relaxation to $\tilde{w} = 0$; as before, this is followed by a fast relaxation to $v = 0$, where $x = y$, and a final slow relaxation to equilibrium. One finds that the center manifold surface $\tilde{w} = 0$ has less curvature and that the trajectories, following an initial nearly vertical descent, have sharp curvature as they relax onto the flatter center manifold, where the again approach equilibrium at the origin.

In Fig. 9.24b, we take $r = 1$, $\sigma = 100$, $b = 8/3$. By Eqs. (9.464,9.465), these parameters induce an initial very fast relaxation to $v = 0$, where $x = y$. This is followed by a fast relaxation to the center manifold where $\tilde{w} = 0$, and then a slow relaxation to equilibrium at the origin.

9.11.3 Transition to chaos

By varying the bifurcation parameter r, we can predict what is called a transition to chaos. We illustrate this transition for two sets of parameters for the Lorenz equations. The first will have trajectories which relax to a stable fixed point; the second will have so-called *chaotic* trajectories which relax to what is known as a *strange attractor*.

Example 9.22

Consider the solution to the Lorenz equations for conditions: $\sigma = 10$, $r = 10$, $b = 8/3$ with initial conditions $x(0) = y(0) = z(0) = 1$.

We first note that $r > 1$, so we expect the origin to be unstable. Next note from Eq. (9.441) that

$$r_c = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1} = \frac{10(10 + \frac{8}{3} + 3)}{10 - \frac{8}{3} - 1} = \frac{470}{19} = 24.74. \tag{9.474}$$

So we have $1 < r < r_c$. We also have $\sigma > b + 1$. Thus, by Eq. (9.443), we expect the other equilibria to be stable. From Eq. (9.435), the first fixed point we examine is the origin $(\overline{x}, \overline{y}, \overline{z}) = (0, 0, 0)$. We find its stability by solving the characteristic equation, Eq. (9.437):

$$(\lambda + b)(\lambda^2 + \lambda(\sigma + 1) - \sigma(r - 1) = 0, \tag{9.475}$$

$$\left(\lambda + \frac{8}{3}\right)(\lambda^2 + 11\lambda - 90) = 0. \tag{9.476}$$

Solution gives

$$\lambda = -\frac{8}{3}, \qquad \lambda = \frac{1}{2}\left(-11 \pm \sqrt{481}\right), \tag{9.477}$$

Numerically, this is $\lambda = -2.67, -16.47, 5.47$. Since one of the eigenvalues is positive, the origin is unstable. From Eq. (9.435), a second fixed point is given by

$$\overline{x} = \sqrt{b(r-1)} = \sqrt{\frac{8}{3}(10-1)} = 2\sqrt{6} = 4.90, \tag{9.478}$$

$$\overline{y} = \sqrt{b(r-1)} = \sqrt{\frac{8}{3}(10-1)} = 2\sqrt{6} = 4.90, \tag{9.479}$$

$$\overline{z} = r - 1 = 10 - 1 = 9. \tag{9.480}$$

Consideration of the roots of Eq. (9.440) shows the second fixed point is stable:

$$\lambda^3 + (\sigma + b + 1)\lambda^2 + (\sigma + r)b\lambda + 2\sigma b(r-1) = 0, \tag{9.481}$$

$$\lambda^3 + \frac{41}{3}\lambda^2 + \frac{160}{3}\lambda + 480 = 0. \tag{9.482}$$

Solution gives

$$\lambda = -12.48, \qquad \lambda = -0.595 \pm 6.17\ i. \tag{9.483}$$

From Eq. (9.435), a third fixed point is given by

$$\overline{x} = -\sqrt{b(r-1)} = -\sqrt{\frac{8}{3}(10-1)} = -2\sqrt{6} = -4.90, \tag{9.484}$$

$$\overline{y} = -\sqrt{b(r-1)} = -\sqrt{\frac{8}{3}(10-1)} = -2\sqrt{6} = -4.90, \tag{9.485}$$

$$\overline{z} = r - 1 = 10 - 1 = 9. \tag{9.486}$$

The stability analysis for this point is essentially identical as that for the second point. The eigenvalues are identical $\lambda = -12.48, -0.595 \pm 6.17i$; thus, the root is linearly stable. Because we have two stable roots, we might expect some initial conditions to induce trajectories to one of the stable roots, and other initial conditions to induce trajectories to the other.

Figure 9.25 shows the phase space trajectories in (x, y, z) space and the behavior in the time domain, $x(t), y(t), z(t)$. Examination of the solution reveals that for this set of initial conditions, the second equilibrium is attained.

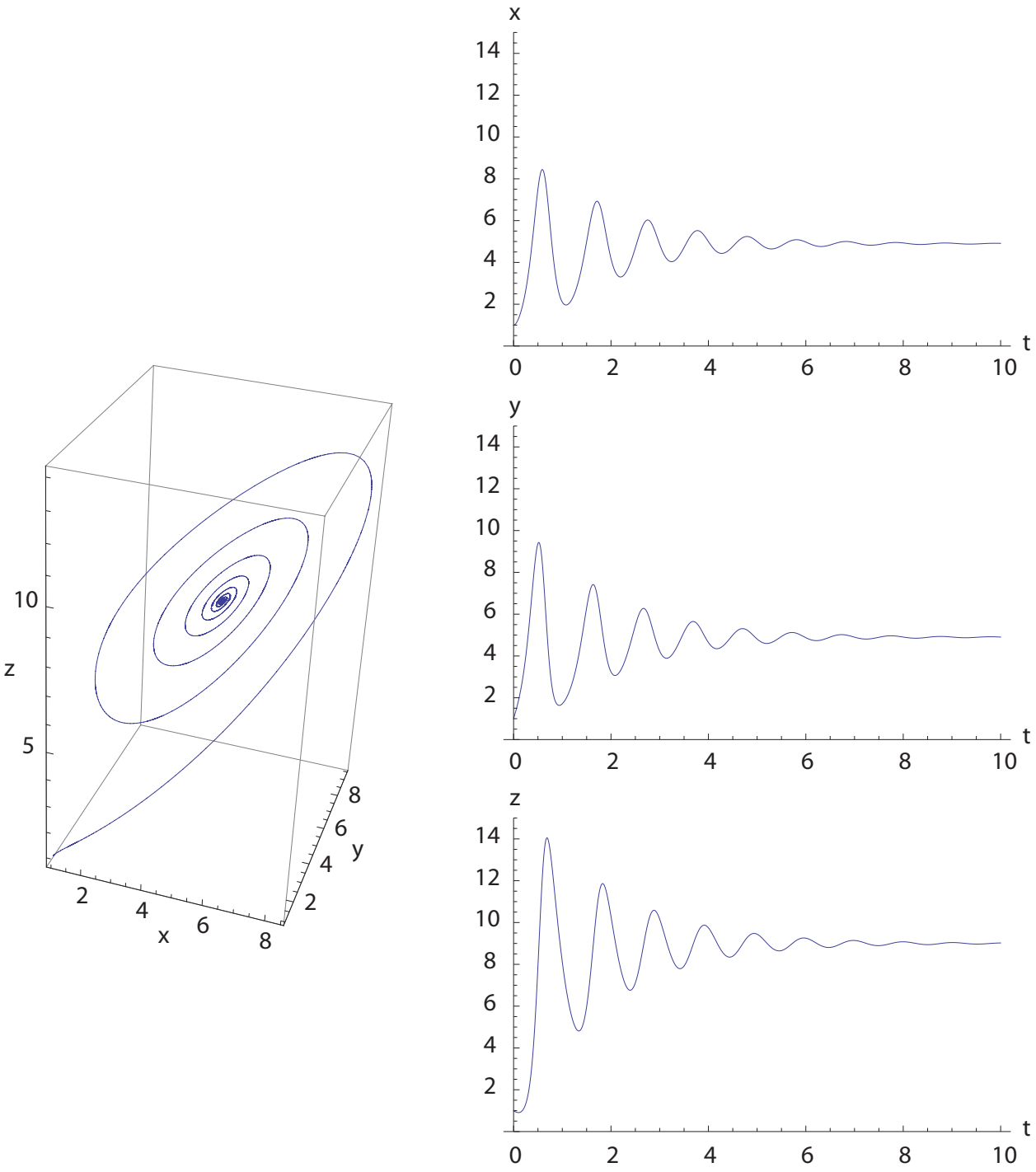

Figure 9.25: Solution to Lorenz equations, $\sigma = 10$, $r = 10$, $b = 8/3$. Initial conditions are $x(0) = y(0) = z(0) = 1$.

Example 9.23

Now consider the conditions: $\sigma = 10$, $r = 28$, $b = 8/3$. Initial conditions remain $x(0) = y(0) = z(0) = 1$.

The analysis is very similar to the previous example, except that we have changed the bifurcation parameter r. We first note that $r > 1$, so we expect the origin to be an unstable equilibrium. We next note from Eq. (9.441) that

$$r_c = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1} = \frac{10(10 + \frac{8}{3} + 3)}{10 - \frac{8}{3} - 1} = \frac{470}{19} = 24.74, \tag{9.487}$$

remains unchanged from the previous example. So we have $r > r_c$. Thus, we expect the other equilibria to be unstable as well.

From Eq. (9.435), the origin is again a fixed point, and again it can be shown to be unstable. From Eq. (9.435), the second fixed point is now given by

$$\overline{x} = \sqrt{b(r-1)} = \sqrt{\frac{8}{3}(28-1)} = 8.485, \tag{9.488}$$

$$\overline{y} = \sqrt{b(r-1)} = \sqrt{\frac{8}{3}(28-1)} = 8.485, \tag{9.489}$$

$$\overline{z} = r - 1 = 28 - 1 = 27. \tag{9.490}$$

Now, consideration of the roots of the characteristic equation, Eq. (9.440), shows the second fixed point here is unstable:

$$\lambda^3 + (\sigma + b + 1)\lambda^2 + (\sigma + r)b\lambda + 2\sigma b(r-1) = 0, \tag{9.491}$$

$$\lambda^3 + \frac{41}{3}\lambda^2 + \frac{304}{3}\lambda + 1440 = 0. \tag{9.492}$$

Solution gives

$$\lambda = -13.8546, \qquad \lambda = 0.094 \pm 10.2\ i. \tag{9.493}$$

Moreover, the third fixed point is unstable in exactly the same fashion as the second. The consequence of this is that there is no possibility of achieving an equilibrium as $t \to \infty$. More importantly, numerical solution reveals the solution to approach what is known as a strange attractor. Moreover, numerical experimentation would reveal an extreme, exponential sensitivity of the solution trajectories to the initial conditions. That is, a small change in initial conditions would induce a large deviation of a trajectory in a finite time. Such systems are known as chaotic.

Figure 9.26 shows the phase space trajectory, the strange attractor, and the behavior in the time domain of this system which has underwent a transition to a chaotic state.

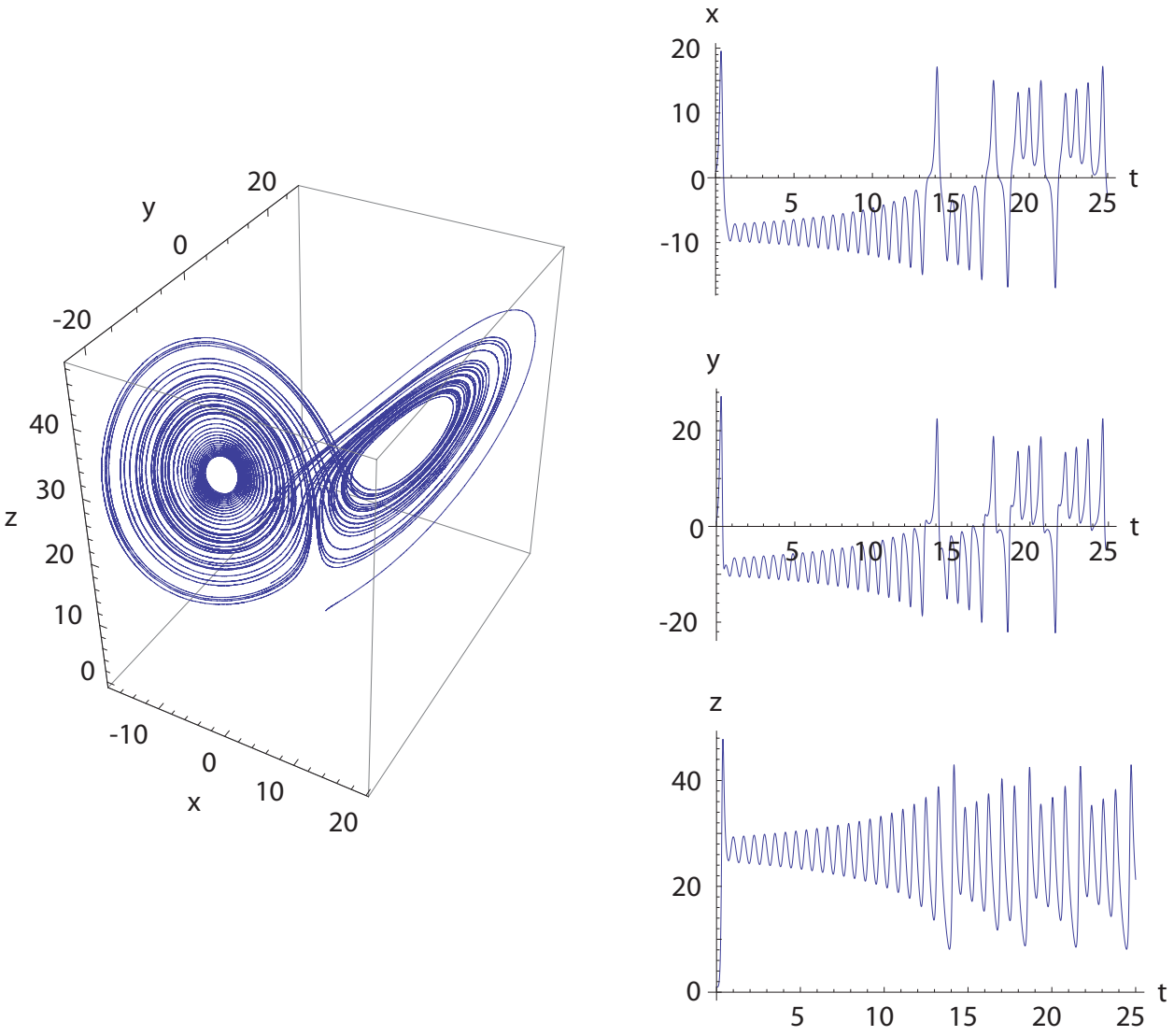

Figure 9.26: Phase space trajectory and time domain plots for solution to Lorenz equations, $\sigma = 10$, $r = 28$, $b = 8/3$. Initial conditions are $x(0) = y(0) = z(0) = 1$.

Problems

1. For the logistics equation: $x_{k+1} = rx_k(1 - x_k); 0 < x_k < 1, 0 < r < 4$, write a short program which determines the value of x as $k \to \infty$. Plot the bifurcation diagram, that is the limiting value of x as a function of r for $0 < r < 4$. If r_i is the i^{th} bifurcation point, that is the value at which the number of fixed points changes, make an estimate of Feigenbaum's constant,

$$\delta = \lim_{n\to\infty} \frac{r_{n-1} - r_n}{r_n - r_{n+1}}.$$

2. If

$$\begin{aligned} x\frac{dx}{dt} + xy\frac{dy}{dt} &= x - 1, \\ (x+y)\frac{dx}{dt} + x\frac{dy}{dt} &= y + 1, \end{aligned}$$

write the system in autonomous form,

$$\begin{aligned} \frac{dx}{dt} &= f(x,y), \\ \frac{dy}{dt} &= g(x,y). \end{aligned}$$

Plot curves on which $f = 0, g = 0$ in the x, y phase plane. Also plot in this plane the vector field defined by the differential equations. With a combination of analysis and numerics, find a path in phase space from one critical point to another critical point. For this path, also known as *heteroclinic orbit*[14], plot $x(t)$, $y(t)$ and include the path in the (x, y) phase plane.

3. Show that for all initial conditions the solutions of

$$\begin{aligned} \frac{dx}{dt} &= -x + x^2y - y^2, \\ \frac{dy}{dt} &= -x^3 + xy - 6z, \\ \frac{dz}{dt} &= 2y, \end{aligned}$$

tend to $x = y = z = 0$ as $t \to \infty$.

4. Draw the bifurcation diagram of

$$\frac{dx}{dt} = x^3 + x\left((r-3)^2 - 1\right),$$

where r is the bifurcation parameter, indicating stable and unstable branches.

5. A two-dimensional dynamical system expressed in polar form is

$$\begin{aligned} \frac{d\rho}{dt} &= \rho(\rho - 2)(\rho - 3), \\ \frac{d\theta}{dt} &= 2. \end{aligned}$$

Find the (a) critical point(s), (b) periodic solution(s), and (c) analyze their stability.

[14] In contrast, a *homoclinic orbit* starts near a critical point, travels away from the critical point, and then returns to the same critical point.

6. Find a critical point of the following system, and show its local and global stability.

$$\begin{aligned}\frac{dx}{dt} &= (x-2)\left((y-1)^2-1\right),\\ \frac{dy}{dt} &= (2-y)\left((x-2)^2+1\right),\\ \frac{dz}{dt} &= (4-z).\end{aligned}$$

7. Find the general solution of $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & -3 & 1 \\ 2 & -1 & -2 \\ 2 & -3 & 0 \end{pmatrix}.$$

8. Find the solution of $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$ where

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

and

$$\mathbf{x}(0) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

9. Find the solution of $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$ where

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

and

$$\mathbf{x}(0) = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

10. Find the solution of $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix},$$

and

$$\mathbf{x}(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

11. Express

$$\begin{aligned}\frac{dx_1}{dt} + x_1 + \frac{dx_2}{dt} + 3x_2 &= 0,\\ \frac{dx_1}{dt} + 3\frac{dx_2}{dt} + x_2 &= 0,\end{aligned}$$

in the form $d\mathbf{x}/dt = \mathbf{A}\cdot\mathbf{x}$ and solve. Plot the some solution trajectories in x_1, x_2 phase plane and as well as the vector field defined by the system of equations.

12. Classify the critical points of

$$\begin{aligned} \frac{dx}{dt} &= x - y - 3, \\ \frac{dy}{dt} &= y - x^2 + 1, \end{aligned}$$

and analyze their stability. Plot the global (x, y) phase plane including critical points and vector fields.

13. The following equations arise in a natural circulation loop problem

$$\begin{aligned} \frac{dx}{dt} &= y - x, \\ \frac{dy}{dt} &= a - zx, \\ \frac{dz}{dt} &= xy - b, \end{aligned}$$

where a and b are nonnegative parameters. Find the critical points and analyze their linear stability. Find numerically the attractors for (i) $a = 2$, $b = 1$, (ii) $a = 0.95$, $b = 1$, and (iii) $a = 0$, $b = 1$.

14. Sketch the steady state bifurcation diagrams of the following equations. Determine and indicate the linear stability of each branch.

$$\begin{aligned} \frac{dx}{dt} &= -\left(\frac{1}{x} - r\right)(2x - r), \\ \frac{dx}{dt} &= -x\left((x-2)^2 - (r-1)\right). \end{aligned}$$

15. The motion of a freely spinning object in space is given by

$$\begin{aligned} \frac{dx}{dt} &= yz, \\ \frac{dy}{dt} &= -2xz, \\ \frac{dz}{dt} &= xy, \end{aligned}$$

where x, y, z represent the angular velocities about the three principal axes. Show that $x^2+y^2+z^2$ is a constant. Find the critical points and analyze their linear stability. Check by throwing a non-spherical object (a book?) in the air.

16. A bead moves along a smooth circular wire of radius a which is rotating about a vertical axis with constant angular speed ω. Taking gravity and centrifugal forces into account, the motion of the bead is given by

$$a\frac{d^2\theta}{dt^2} = -g\sin\theta + a\omega^2\cos\theta\sin\theta,$$

where θ is the angular position of the bead with respect to the downward vertical position. Find the equilibrium positions and their stability as the parameter $\mu = a\omega^2/g$ is varied.

17. Find a Lyapunov function of the form $V = ax^2 + by^2$ to investigate the global stability of the critical point $x = y = 0$ of the system of equations

$$\begin{aligned} \frac{dx}{dt} &= -2x^3 + 3xy^2, \\ \frac{dy}{dt} &= -x^2y - y^3. \end{aligned}$$

18. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Solve the equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}.$$

Determine the critical points and their stability.

19. Draw the bifurcation diagram of

$$\frac{dx}{dt} = (x^2 - 2)^2 - 2(x^2 + 1)(r - 1) + (r - 1)^2,$$

where r is the bifurcation parameter, indicating the stability of each branch.

20. Show that for all initial conditions the solutions of

$$\begin{aligned} \frac{dx}{dt} &= -x + x^2 y - y^2, \\ \frac{dy}{dt} &= -x^3 + xy - 6z, \\ \frac{dz}{dt} &= 2y, \end{aligned}$$

tend to $x = y = z = 0$ as $t \to \infty$.

21. Draw the bifurcation diagram of

$$\frac{dx}{dt} = x^3 + x\left((r - 2)^3 - 1\right),$$

where r is the bifurcation parameter, indicating stable and unstable branches.

22. Solve the system of equations $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} -3 & 0 & 2 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

23. Find a Lyapunov function for the system

$$\begin{aligned} \frac{dx}{dt} &= -x - 2y^2, \\ \frac{dy}{dt} &= xy - y^3. \end{aligned}$$

24. Analyze the local stability of the origin in the following system

$$\begin{aligned} \frac{dx}{dt} &= -2x + y + 3z + 8y^3, \\ \frac{dy}{dt} &= -6y - 5z + 2z^3, \\ \frac{dz}{dt} &= z + x^2 + y^3. \end{aligned}$$

25. Show that the origin is linearly stable

$$\begin{aligned} \frac{dx}{dt} &= (x-by)(x^2+y^2-1), \\ \frac{dy}{dt} &= (ax+y)(x^2+y^2-1), \end{aligned}$$

where $a, b > 0$. Show also that the origin is stable to large perturbations, as long as they satisfy $x^2 + y^2 < 1$.

26. Draw the bifurcation diagram and analyze the stability of

$$\frac{dx}{dt} = -x(x^3 - r - 1) - \frac{1}{10},$$

where r is the bifurcation parameter.

27. Find the dynamical system corresponding to the Hamiltonian $H(x, y) = x^2 + 2xy + y^2$ and then solve it.

28. Show that solutions of the system of differential equations

$$\begin{aligned} \frac{dx}{dt} &= -x + y^3 - z^3, \\ \frac{dy}{dt} &= = -y + z^3 - x^3, \\ \frac{dz}{dt} &= -z + x^3 - y^3, \end{aligned}$$

eventually approach the origin for all initial conditions.

29. Find and plot all critical points $(\overline{x}, \overline{y})$ of

$$\begin{aligned} \frac{dx}{dt} &= (r-1)x - 3xy^2 - x^3, \\ \frac{dy}{dt} &= (r-1)y - 3x^2y - y^3. \end{aligned}$$

as functions of r. Determine the stability of $(\overline{x}, \overline{y}) = (0, 0)$, and of *one* post-bifurcation branch.

30. Write in matrix form and solve

$$\begin{aligned} \frac{dx}{dt} &= y + z, \\ \frac{dy}{dt} &= z + x, \\ \frac{dz}{dt} &= x + y. \end{aligned}$$

31. Find the critical point (or points) of the Van der Pol equation

$$\frac{d^2x}{dt^2} - a(1-x^2)\frac{dx}{dt} + x = 0, \ a > 0,$$

and determine its (or their) stability to small perturbations. For $a = 1$, plot the $dx/dt, x$ phase plane including critical points and vector fields.

32. Consider a straight line between $x = 0$ and $x = l$. Remove the middle half (i.e. the portion between $x = l/4$ and $x = 3l/4$). Repeat the process on the two pieces that are left. Find the dimension of what is left after an infinite number of iterations.

33. Classify the critical points of

$$\begin{aligned} \frac{dx}{dt} &= x + y - 2, \\ \frac{dy}{dt} &= 2y - x^2 + 1, \end{aligned}$$

and analyze their stability.

34. Determine if the origin is stable if $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$, where

$$\mathbf{A} = \begin{pmatrix} 3 & -3 & 0 \\ 0 & -5 & -2 \\ -6 & 0 & -3 \end{pmatrix}.$$

35. Find a Lyapunov function of the form $V = ax^2 + by^2$ to investigate the global stability of the critical point $x = y = 0$ of the system of equations

$$\begin{aligned} \frac{dx}{dt} &= -2x^3 + 3xy^2, \\ \frac{dy}{dt} &= -x^2y - y^3. \end{aligned}$$

36. Draw a bifurcation diagram for the differential equation

$$\frac{dx}{dt} = (x - 3)(x^2 - r),$$

where r is the bifurcation parameter. Analyze linear stability and indicate stable and unstable branches.

37. Solve the following system of differential equations using generalized eigenvectors

$$\begin{aligned} \frac{dx}{dt} &= -5x + 2y + z, \\ \frac{dy}{dt} &= -5y + 3z, \\ \frac{dz}{dt} &= -5z. \end{aligned}$$

38. Analyze the linear stability of the critical point of

$$\begin{aligned} \frac{dx}{dt} &= 2y + y^2, \\ \frac{dy}{dt} &= -r + 2x^2. \end{aligned}$$

39. Show that the solutions of

$$\begin{aligned} \frac{dx}{dt} &= y - x^3 \\ \frac{dy}{dt} &= -x - y^3 \end{aligned}$$

tend to (0,0) as $t \to \infty$.

40. Sketch the bifurcation diagram showing the stable and unstable steady states of

$$\frac{dx}{dt} = rx(1-x) - x,$$

where r is the bifurcation parameter.

41. Show in parameter space the different possible behaviors of

$$\begin{aligned} \frac{dx}{dt} &= a + x^2 y - 2bx - x, \\ \frac{dy}{dt} &= bx - x^2 y, \end{aligned}$$

where $a, b > 0$.

42. Show that the Hénon-Heiles system

$$\begin{aligned} \frac{d^2x}{dt^2} &= -x - 2xy, \\ \frac{d^2y}{dt^2} &= -y + y^2 - x^2, \end{aligned}$$

is Hamiltonian. Find the Hamiltonian of the system, and determine the stability of the critical point at the origin.

43. Solve $d\mathbf{x}/dt = \mathbf{A} \cdot \mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix},$$

using the exponential matrix.

44. Sketch the steady state bifurcation diagrams of

$$\frac{dx}{dt} = (x-r)(x+r)((x-3)^2 + (r-1)^2 - 1),$$

where r is the bifurcation parameter. Determine the linear stability of each branch; indicate the stable and unstable ones differently on the diagram.

45. Classify the critical point of

$$\frac{d^2x}{dt^2} + (r - r_0)x = 0.$$

46. Show that $x = 0$ is a stable critical point of the differential equation

$$\frac{dx}{dt} = -\sum_{n=0} N a_n x^{2n+1}$$

where $a_n \geq 0,\ n = 0, 1, \cdots, N$.

47. Find the stability of the critical points of the Duffing equation

$$\frac{d^2x}{dt^2} = a\frac{dx}{dt} - bx + x^3 = 0,$$

for positive and negative values of a and b. Sketch the flow lines.

48. Find a Lyapunov function to investigate the critical point $x = y = 0$ of the system of equations

$$\begin{aligned} \frac{dx}{dt} &= -2x^3 + 3xy^2, \\ \frac{dy}{dt} &= -x^2 y - y^3. \end{aligned}$$

49. The populations x and y of two competing animal species are governed by

$$\begin{aligned} \frac{dx}{dt} &= x - 2xy, \\ \frac{dy}{dt} &= -y + xy. \end{aligned}$$

What are the steady-state populations? Is the situation stable?

50. For the Lorenz equations with $b = 8/3$, $r = 28$, and initial conditions $x(0) = 2$, $y(0) = 1$, $z(0) = 3$, numerically integrate the Lorenz equations for two cases, $\sigma = 1$, $\sigma = 10$. For each case plot the trajectory in (x, y, z) phase space and plot $x(t), y(t), z(t)$ for $t \in [0, 50]$. Change the initial condition on x to $x(0) = 2.002$ and plot the difference in the predictions of x versus time for both values of σ.

51. Use the Poincaré sphere to find all critical points, finite and infinite of the system

$$\begin{aligned} \frac{dx}{dt} &= 2x - 2xy, \\ \frac{dy}{dt} &= 2y - x^2 + y^2. \end{aligned}$$

Plot families of trajectories in the x, y phase space and the X, Y projection of the Poincaré sphere.

52. For the Lorenz equations with $\sigma = 10$, $b = 8/3$, and initial conditions $x(0) = 0$, $y(0) = 1$, $z(0) = 0$, numerically integrate the Lorenz equations for three cases, $r = 10$, $r = 24$, and $r = 28$. For each case, plot the trajectory in (x, y, z) phase space and plot $x(t)$, $y(t)$, $z(t)$ for $t \in [0, 50]$. Change the initial condition on x to $x(0) = 0.002$ and plot the difference in the predictions of x versus time for all three values of r.

Chapter 10

Appendix

The material in this section is not covered in detail; some is review from undergraduate classes.

10.1 Taylor series

The Taylor series of $y(x)$ about the point $x = x_o$ is

$$\begin{aligned} y(x) &= y(x_o) + \left.\frac{dy}{dx}\right|_{x=x_o} (x - x_o) + \frac{1}{2} \left.\frac{d^2y}{dx^2}\right|_{x=x_o} (x - x_o)^2 + \frac{1}{6} \left.\frac{d^3y}{dx^3}\right|_{x=x_o} (x - x_o)^3 + \dots \\ &\quad + \frac{1}{n!} \left.\frac{d^ny}{dx^n}\right|_{x=x_o} (x - x_o)^n + \dots \end{aligned} \tag{10.1}$$

Example 10.1

For a Taylor series of $y(x)$ about $x = 0$ if

$$y(x) = \frac{1}{(1+x)^n}. \tag{10.2}$$

Direct substitution reveals that the answer is

$$y(x) = 1 - nx + \frac{(-n)(-n-1)}{2!}x^2 + \frac{(-n)(-n-1)(-n-2)}{3!}x^3 + \dots \tag{10.3}$$

10.2 Trigonometric relations

$$\sin x \sin y = \frac{1}{2}\cos(x-y) - \frac{1}{2}\cos(x+y), \tag{10.4}$$

$$\sin x \cos y = \frac{1}{2}\sin(x+y) + \frac{1}{2}\sin(x-y), \tag{10.5}$$

$$\cos x \cos y = \frac{1}{2}\cos(x-y) + \frac{1}{2}\cos(x+y), \tag{10.6}$$

$$\sin^2 x = \frac{1}{2} - \frac{1}{2}\cos 2x, \tag{10.7}$$

$$\sin x \cos x = \frac{1}{2}\sin 2x, \tag{10.8}$$

$$\cos^2 x = \frac{1}{2} + \frac{1}{2}\cos 2x, \tag{10.9}$$

$$\sin^3 x = \frac{3}{4}\sin x - \frac{1}{4}\sin 3x, \tag{10.10}$$

$$\sin^2 x \cos x = \frac{1}{4}\cos x - \frac{1}{4}\cos 3x, \tag{10.11}$$

$$\sin x \cos^2 x = \frac{1}{4}\sin x + \frac{1}{4}\sin 3x, \tag{10.12}$$

$$\cos^3 x = \frac{3}{4}\cos x + \frac{1}{4}\cos 3x, \tag{10.13}$$

$$\sin^4 x = \frac{3}{8} - \frac{1}{2}\cos 2x + \frac{1}{8}\cos 4x, \tag{10.14}$$

$$\sin^3 x \cos x = \frac{1}{4}\sin 2x - \frac{1}{8}\sin 4x, \tag{10.15}$$

$$\sin^2 x \cos^2 x = \frac{1}{8} - \frac{1}{8}\cos 4x, \tag{10.16}$$

$$\sin x \cos^3 x = \frac{1}{4}\sin 2x + \frac{1}{8}\sin 4x, \tag{10.17}$$

$$\cos^4 x = \frac{3}{8} + \frac{1}{2}\cos 2x + \frac{1}{8}\cos 4x, \tag{10.18}$$

$$\sin^5 x = \frac{5}{8}\sin x - \frac{5}{16}\sin 3x + \frac{1}{16}\sin 5x, \tag{10.19}$$

$$\sin^4 x \cos x = \frac{1}{8}\cos x - \frac{3}{16}\cos 3x + \frac{1}{16}\cos 5x, \tag{10.20}$$

$$\sin^3 x \cos^2 x = \frac{1}{8}\sin x + \frac{1}{16}\sin 3x - \frac{1}{16}\sin 5x, \tag{10.21}$$

$$\sin^2 x \cos^3 x = -\frac{1}{8}\cos x - \frac{1}{16}\cos 3x - \frac{1}{16}\cos 5x, \tag{10.22}$$

$$\sin x \cos^4 x = \frac{1}{8}\sin x + \frac{3}{16}\sin 3x + \frac{1}{16}\sin 5x, \tag{10.23}$$

$$\cos^5 x = \frac{5}{8}\cos x + \frac{5}{16}\cos 3x + \frac{1}{16}\cos 5x. \tag{10.24}$$

10.3 Hyperbolic functions

The hyperbolic functions are defined as follows:

$$\sinh\theta = \frac{e^\theta - e^{-\theta}}{2}, \tag{10.25}$$

$$\cosh\theta = \frac{e^\theta + e^{-\theta}}{2}. \tag{10.26}$$

10.4 Routh-Hurwitz criterion

Here we consider the Routh-Hurwitz[1] criterion. The polynomial equation

$$a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n = 0, \tag{10.27}$$

has roots with negative real parts if and only if the following conditions are satisfied:

- $a_1/a_0, a_2/a_0, \ldots, a_n/a_0 > 0$,
- $D_i > 0$, $i = 1, \ldots, n$.

The Hurwitz determinants D_i are defined by

$$D_1 = a_1, \tag{10.28}$$

$$D_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix}, \tag{10.29}$$

$$D_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix}, \tag{10.30}$$

$$D_n = \begin{vmatrix} a_1 & a_3 & a_5 & \ldots & a_{2n-1}, \\ a_0 & a_2 & a_4 & \ldots & a_{2n-2}, \\ 0 & a_1 & a_3 & \ldots & a_{2n-3}, \\ 0 & a_0 & a_2 & \ldots & a_{2n-4}, \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_n \end{vmatrix}, \tag{10.31}$$

with $a_i = 0$, if $i > n$.

[1] Edward John Routh, 1831-1907, Canadian-born English mathematician, and Adolf Hurwitz, 1859-1919, German mathematician.

10.5 Infinite series

Definition: A *power* series is of the form

$$\sum_{n=0}^{\infty} a_n (x-a)^n. \tag{10.32}$$

The series converges if $|x-a| < R$, where R is the *radius of convergence.*

Definition: A function $f(x)$ is said to be *analytic* at $x = a$ if f and all its derivatives exist at this point.

An analytic function can be expanded in a *Taylor* series:

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2 + \cdots \tag{10.33}$$

where the function and its derivatives on the right side are evaluated at $x = a$. This is a power series for $f(x)$. We have used primes to indicate derivatives.

Example 10.2

Expand $(1+x)^n$ about $x = 0$.

$$\begin{aligned} f(x) &= (1+x)^n, && (10.34)\\ f(0) &= 1, && (10.35)\\ f'(0) &= n, && (10.36)\\ f''(0) &= n(n-1), && (10.37)\\ &\vdots && (10.38)\\ (1+x)^n &= 1 + nx + \frac{1}{2} n(n-1) x^2 + \cdots && (10.39) \end{aligned}$$

A function of two variables $f(x,y)$ can be similarly expanded

$$\begin{aligned} f(x,y) &= f\Big|_{a,b} + \frac{\partial f}{\partial x}\Big|_{a,b} (x-a) + \frac{\partial f}{\partial y}\Big|_{a,b} (y-b) \\ &\quad + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}\Big|_{a,b} (x-a)^2 + \frac{\partial^2 f}{\partial x \partial y}\Big|_{a,b} (x-a)(y-b) + \\ &\quad \frac{1}{2} \frac{\partial^2 f}{\partial y^2}\Big|_{a,b} (y-b)^2 + \cdots \end{aligned} \tag{10.40}$$

if f and all its partial derivatives exist and are evaluated at $x = a$, $y = b$.

10.6 Asymptotic expansions

Definition: Consider two function $f(x)$ and $g(x)$. We write that

$f(x) \sim g(x)$, if $\lim_{x\to a} \frac{f(x)}{g(x)} = 1$.

$f(x) = o(g(x))$, if $\lim_{x\to a} \frac{f(x)}{g(x)} = 0$;

$f(x) = O(g(x))$, if $\lim_{x\to a} \left|\frac{f(x)}{g(x)}\right|$ =constant;

10.7 Special functions

10.7.1 Gamma function

The Gamma function may be thought of as an extension to the factorial function. Recall the factorial function requires an integer argument. The Gamma function admits real arguments; when the argument of the Gamma function is an integer, one finds it is directly related to the factorial function. The Gamma function is defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt. \tag{10.41}$$

Generally, we are interested in $x > 0$, but results are available for all x. Some properties are:

1. $\Gamma(1) = 1$.
2. $\Gamma(x) = (x-1)\Gamma(x-1)$, $x > 1$.
3. $\Gamma(x) = (x-1)(x-2)\cdots(x-r)\Gamma(x-r)$, $x > r$.
4. $\Gamma(n) = (n-1)!$, where n is a positive integer.
5. $\Gamma(x) \sim \sqrt{\frac{2\pi}{x}} x^x e^{-x} \left(1 + \frac{1}{12x} + \frac{1}{288x^2} + \ldots\right)$, (Stirling's formula).

Bender and Orszag show that Stirling's[2] formula is a divergent series. It is an asymptotic series, but as more terms are added, the solution can actually get worse. The Gamma function and its amplitude are plotted in Fig. 10.1.

10.7.2 Beta function

The beta function is defined by

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\, dx. \tag{10.42}$$

Property:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{10.43}$$

[2]James Stirling, 1692-1770, Scottish mathematician and member of a prominent Jacobite family.

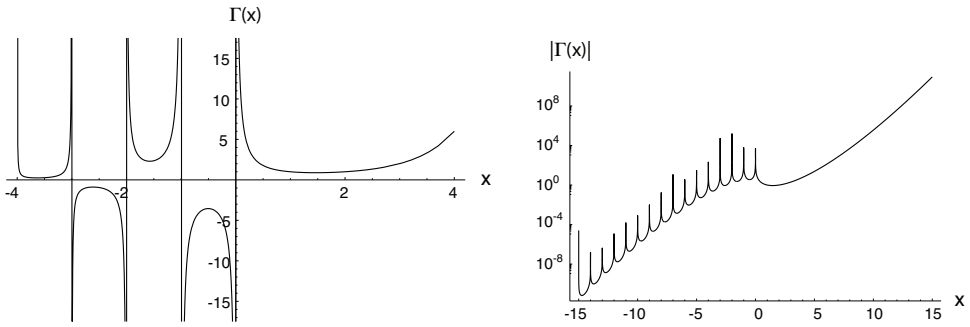

Figure 10.1: Gamma function and amplitude of Gamma function.

10.7.3 Riemann zeta function

This is defined as

$$\zeta(x) = \sum_{n=1}^{\infty} n^{-x}. \tag{10.44}$$

The function can be evaluated in closed form for even integer values of x. It can be shown that $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$, $\zeta(6) = \pi^6/945$, ..., $\zeta(2n) = (-1)^{n+1} B_{2n}(2\pi)^{2n}/2/(2n)!$, where B_{2n} is a so-called Bernoulli number, which can be found via a complicated recursion formula. All negative even integer values of x give $\zeta(x) = 0$. Further $\lim_{x\to\infty} \zeta(x) = 1$. For large negative values of x, the Riemann zeta function oscillates with increasing amplitude. Plots of the Riemann zeta function for $x \in [-1, 3]$ and the amplitude of the Riemann zeta function over a broader domain on a logarithmic scale as shown in Fig. 10.2.

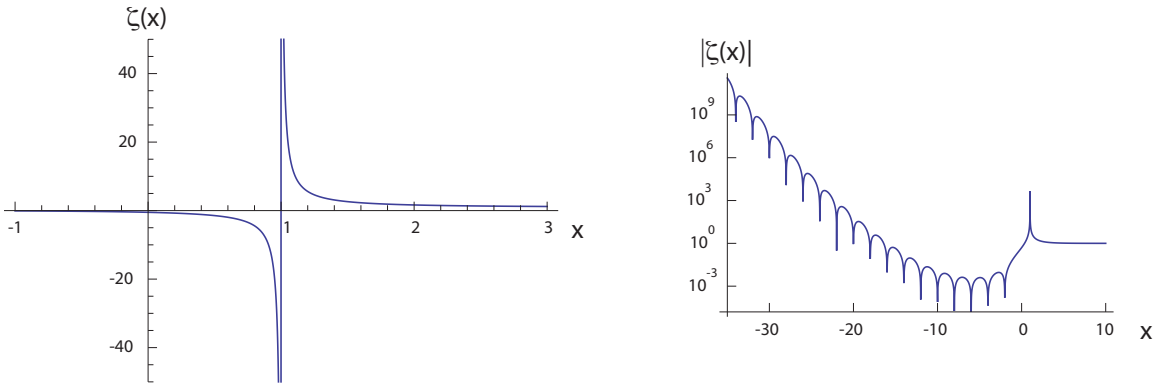

Figure 10.2: Riemann zeta function and amplitude of Riemann zeta function.

10.7.4 Error functions

The error function is defined by

$$\text{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2}\, d\xi, \tag{10.45}$$

and the complementary error function by

$$\text{erfc}\,(x) = 1 - \text{erf}\, x. \tag{10.46}$$

The error function and the error function complement are plotted in Fig. 10.3.

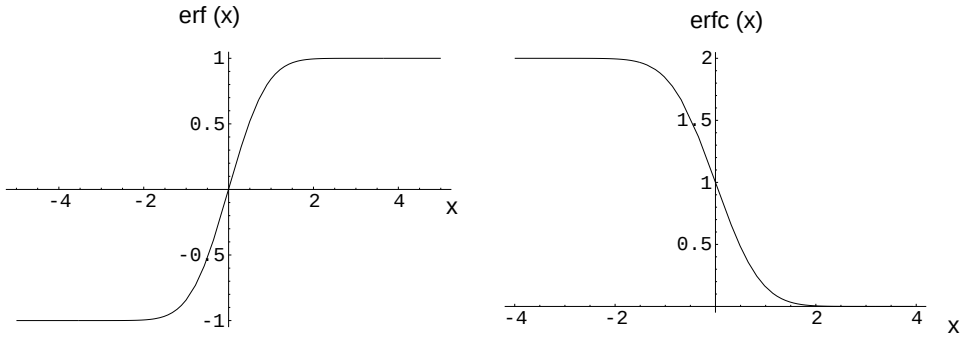

Figure 10.3: Error function and error function complement.

The imaginary error function is defined by

$$\text{erfi}\,(z) = -i\,\text{erf}\,(iz), \tag{10.47}$$

where $z \in \mathbb{C}^1$. For real arguments, $x \in \mathbb{R}^1$, it can be shown that $\text{erfi}\,(x) = -i\,\text{erf}\,(ix) \in \mathbb{R}^1$. The imaginary error function is plotted in Fig. 10.4 for a real argument, $x \in \mathbb{R}^1$.

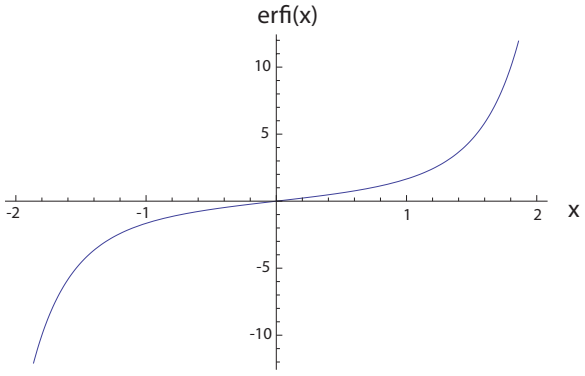

Figure 10.4: Imaginary error function, $\text{erfi}\,(x)$, for real argument, $x \in \mathbb{R}^1$.

10.7.5 Fresnel integrals

The Fresnel[3] integrals are defined by

$$\mathrm{C}(x) = \int_0^x \cos\frac{\pi t^2}{2}\, dt, \tag{10.48}$$

$$\mathrm{S}(x) = \int_0^x \sin\frac{\pi t^2}{2}\, dt. \tag{10.49}$$

The Fresnel cosine and sine integrals are plotted in Fig. 10.5.

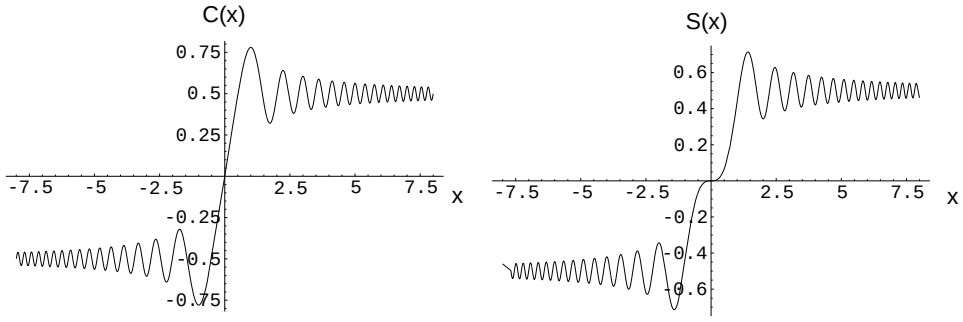

Figure 10.5: Fresnel cosine, $\mathrm{C}(x)$, and sine, $\mathrm{S}(x)$, integrals.

10.7.6 Sine-, cosine-, and exponential-integral functions

The sine-integral function is defined by

$$\mathrm{Si}(x) = \int_0^x \frac{\sin\xi}{\xi}\, d\xi, \tag{10.50}$$

and the cosine-integral function by

$$\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos\xi}{\xi}\, d\xi. \tag{10.51}$$

The sine integral function is real valued for $x \in (-\infty, \infty)$. The cosine integral function is real valued for $x \in [0, \infty)$. We also have $\lim_{x\to 0_+} \mathrm{Ci}(x) \to -\infty$. The cosine integral takes on a value of zero at discrete positive real values, and has an amplitude which slowly decays as $x \to \infty$. The sine integral and cosine integral functions are plotted in Fig. 10.6.

The exponential-integral function is defined by

$$\mathrm{Ei}(x) = -\!\!\!\!\!\!\int_{-x}^{\infty} \frac{e^{-\xi}}{\xi}\, d\xi = \!\!\!\!\!\!\int_{-\infty}^{x} \frac{e^{\xi}}{\xi}\, d\xi. \tag{10.52}$$

The exponential integral function is plotted in Fig. 10.7. Note we must use the Cauchy principal value of the integral if $x > 0$.

[3]Augustin-Jean Fresnel, 1788-1827, French physicist noted for work in optics.

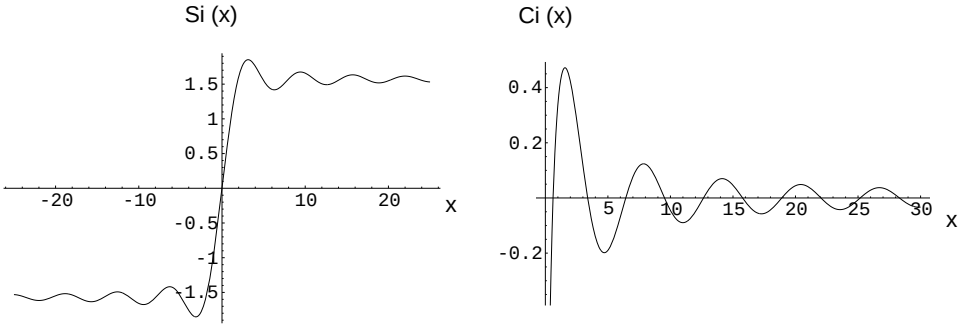

Figure 10.6: Sine integral function, $\mathrm{Si}(x)$, and cosine integral function $\mathrm{Ci}(x)$.

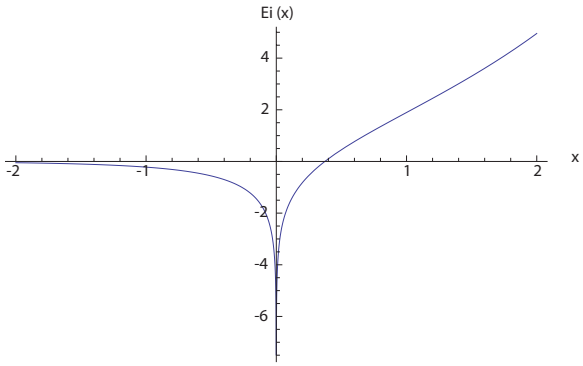

Figure 10.7: Exponential integral function, $\mathrm{Ei}(x)$.

10.7.7 Elliptic integrals

The *Legendre elliptic integral of the first kind* is

$$F(y,k) = \int_0^y \frac{d\eta}{\sqrt{(1-\eta^2)(1-k^2\eta^2)}}. \tag{10.53}$$

Another common way of writing the elliptic integral is to take $\eta = \sin\phi$, so that

$$F(\phi,k) = \int_0^\phi \frac{d\phi}{\sqrt{(1-k^2\sin^2\phi)}}. \tag{10.54}$$

The *Legendre elliptic integral of the second kind* is

$$E(y,k) = \int_0^y \frac{(1-k^2\eta^2)}{\sqrt{(1-\eta^2)}}\, d\eta, \tag{10.55}$$

which, on again using $\eta = \sin\phi$, becomes

$$E(\phi,k) = \int_0^\phi \sqrt{1-k^2\sin^2\phi}\, d\phi. \tag{10.56}$$

The *Legendre elliptic integral of the third kind* is

$$\Pi(y,n,k) = \int_0^\phi \frac{d\phi}{(1+n\sin^2\phi)\sqrt{(1-k^2\sin^2\phi)}}, \tag{10.57}$$

which is equivalent to

$$\Pi(\phi,n,k) = \int_0^\phi \sqrt{1-k^2\sin^2\phi}\; d\phi. \tag{10.58}$$

For $\phi = \pi/2$, we have the *complete elliptic integrals*:

$$F\left(\frac{\pi}{2},k\right) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1-k^2\sin^2\phi)}}, \tag{10.59}$$

$$E\left(\frac{\pi}{2},k\right) = \int_0^{\pi/2} \sqrt{1-k^2\sin^2\phi}\; d\phi, \tag{10.60}$$

$$\Pi\left(\frac{\pi}{2},n,k\right) = \int_0^{\pi/2} \sqrt{1-k^2\sin^2\phi}\; d\phi. \tag{10.61}$$

10.7.8 Hypergeometric functions

A generalized *hypergeometric function* is defined by

$${}_pF_q\left(\{a_1,\ldots,a_p\},\{b_1,\ldots,b_q\};x\right) = \sum_{k=1}^{\infty} \frac{(a_1)_k(a_2)_k\cdots(a_p)_k}{(b_1)_k(b_2)_k\cdots(b_q)_k}\frac{x^k}{k!}, \tag{10.62}$$

where the rising factorial notation, $(s)_k$, is defined by

$$(s)_k \equiv \frac{\Gamma(s+k)}{\Gamma(s)}. \tag{10.63}$$

There are many special hypergeometric functions. If $p = 2$ and $q = 1$, we have Gauss's hypergeometric function ${}_2F_1\left(\{a_1,a_2\},\{b_1\};x\right)$. Since there are only three parameters, Gauss's hypergeometric function is sometimes denoted as ${}_2F_1\left(a,b,c,x\right)$. An integral representation of Gauss's *hypergeometric* function is

$${}_2F_1\left(a,b,c,x\right) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tx)^{-a}\; dt. \tag{10.64}$$

For special values of parameters, hypergeometric functions can reduce to other functions such as $\tanh^{-1}$.

10.7.9 Airy functions

The Airy functions Ai (x), and Bi (x) are most compactly defined as the two linearly independent solutions to the second order differential equation $y'' - xy = 0$, yielding $y = C_1 \text{Ai}\,(x) + C_2 \text{Bi}\,(x)$. They can be expressed in a variety of other forms. In terms of the so-called confluent hypergeometric limit function ${}_0F_1$, we have

$$\text{Ai}\,(x) = \frac{1}{3^{2/3}\Gamma\left(\frac{2}{3}\right)} {}_0F_1\left(\{\}\,;\left\{\frac{2}{3}\right\};\frac{1}{9}x^3\right) - \frac{x}{3^{1/3}\Gamma\left(\frac{1}{3}\right)} {}_0F_1\left(\{\}\,;\left\{\frac{4}{3}\right\};\frac{1}{9}x^3\right), \quad (10.65)$$

$$\text{Bi}\,(x) = \frac{1}{3^{1/6}\Gamma\left(\frac{2}{3}\right)} {}_0F_1\left(\{\}\,;\left\{\frac{2}{3}\right\};\frac{1}{9}x^3\right) - \frac{3^{1/6}x}{\Gamma\left(\frac{1}{3}\right)} {}_0F_1\left(\{\}\,;\left\{\frac{4}{3}\right\};\frac{1}{9}x^3\right), \quad (10.66)$$

The Airy functions are plotted in Fig. 10.8. In integral form, the Airy functions are, for

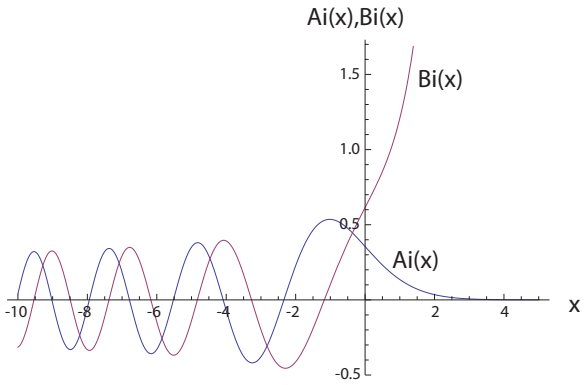

Figure 10.8: Airy functions Ai (x) and Bi (x).

$x \in \mathbb{R}^1$,

$$\text{Ai}\,(x) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{1}{3}t^3 + xt\right) dt, \quad (10.67)$$

$$\text{Bi}\,(x) = \frac{1}{\pi}\int_0^\infty \left(\exp\left(-\frac{1}{3}t^3 + xt\right) + \sin\left(\frac{1}{3}t^3 + xt\right)\right) dt. \quad (10.68)$$

10.7.10 Dirac δ distribution and Heaviside function

Definition: The Dirac[4] *δ-distribution* (or *generalized* function, or simply function), is defined by

$$\int_\alpha^\beta f(x)\delta(x-a)dx = \begin{cases} 0 & \text{if } a \notin [\alpha,\beta], \\ f(a) & \text{if } a \in [\alpha,\beta]. \end{cases} \quad (10.69)$$

[4]Paul Adrien Maurice Dirac, 1902-1984, English physicist.

From this it follows that

$$\delta(x-a) = 0 \text{ if } x \neq a, \tag{10.70}$$

$$\int_{-\infty}^{\infty} \delta(x-a)dx = 1. \tag{10.71}$$

The δ-distribution may be imagined in a limiting fashion as

$$\delta(x-a) = \lim_{\epsilon \to 0^+} \Delta_\epsilon(x-a), \tag{10.72}$$

where $\Delta_\epsilon(x-a)$ has one of the following forms:

1.
$$\Delta_\epsilon(x-a) = \begin{cases} 0 & \text{if } x < a - \frac{\epsilon}{2}, \\ \frac{1}{\epsilon} & \text{if } a - \frac{\epsilon}{2} \leq x \leq a + \frac{\epsilon}{2}, \\ 0 & \text{if } x > a + \frac{\epsilon}{2}, \end{cases} \tag{10.73}$$

2.
$$\Delta_\epsilon(x-a) = \frac{\epsilon}{\pi((x-a)^2 + \epsilon^2)}, \tag{10.74}$$

3.
$$\Delta_\epsilon(x-a) = \frac{1}{\sqrt{\pi\epsilon}} e^{-(x-a)^2/\epsilon}. \tag{10.75}$$

The derivative of the function

$$h(x-a) = \begin{cases} 0 & \text{if } x < a - \frac{\epsilon}{2}, \\ \frac{1}{\epsilon}(x-a) + \frac{1}{2} & \text{if } a - \frac{\epsilon}{2} \leq x \leq a + \frac{\epsilon}{2}, \\ 1 & \text{if } x > a + \frac{\epsilon}{2}, \end{cases} \tag{10.76}$$

is $\Delta_\epsilon(x-a)$ in Eq. (10.73). If we define the *Heaviside*[5] function, $H(x-a)$, as

$$H(x-a) = \lim_{\epsilon \to 0^+} h(x-a), \tag{10.77}$$

then

$$\frac{d}{dx} H(x-a) = \delta(x-a). \tag{10.78}$$

The generator of the Dirac function $\Delta_\epsilon(x-a)$ and the generator of the Heaviside function $h(x-a)$ are plotted for $a = 0$ and $\epsilon = 1/5$ in Fig. 10.9. As $\epsilon \to 0$, Δ_ϵ has its width decrease and its height increase in such a fashion that its area remains constant; simultaneously h has its slope steepen in the region where it jumps from zero to unity as $\epsilon \to 0$.

[5]Oliver Heaviside, 1850-1925, English mathematician.

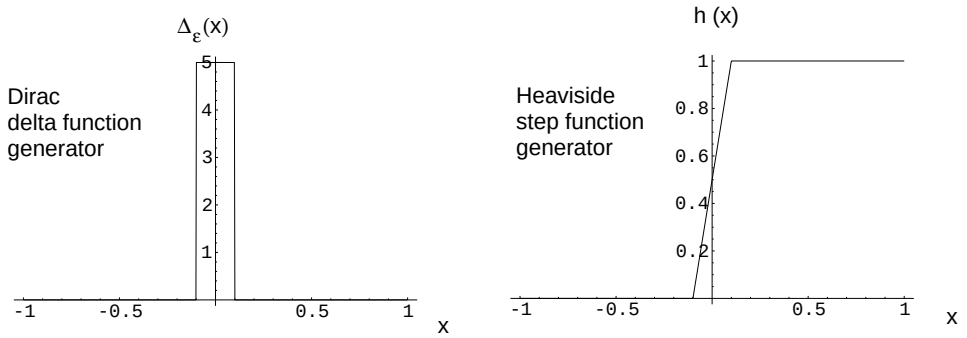

Figure 10.9: Generators of Dirac delta function and Heaviside function, $\Delta_\epsilon(x-a)$ and $h(x-a)$ plotted for $a=0$ and $\epsilon = 1/5$.

10.8 Total derivative

A function of several variables $f(x_1, x_2, \cdots, x_n)$ may be differentiated via the total derivative

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt}. \tag{10.79}$$

Multiplying through by dt, we get the useful formula

$$df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_n}dx_n. \tag{10.80}$$

10.9 Leibniz's rule

Differentiation of an integral is done using Leibniz's rule, Eq. (1.293):

$$y(x) = \int_{a(x)}^{b(x)} f(x,t)dt, \tag{10.81}$$

$$\frac{dy(x)}{dx} = \frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)dt = f(x,b(x))\frac{db(x)}{dx} - f(x,a(x))\frac{da(x)}{dx} + \int_{b(x)}^{a(x)} \frac{\partial f(x,t)}{\partial x}dt. \tag{10.82}$$

10.10 Complex numbers

Here we briefly introduce some basic elements of complex variable theory. Recall that the imaginary number i is defined such that

$$i^2 = -1, \qquad i = \sqrt{-1}. \tag{10.83}$$

10.10.1 Euler's formula

We can get a very useful formula *Euler's formula*, by considering the following Taylor expansions of common functions about $t = 0$:

$$e^t = 1 + t + \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \frac{1}{4!}t^4 + \frac{1}{5!}t^5 \ldots, \tag{10.84}$$

$$\sin t = 0 + t + 0\frac{1}{2!}t^2 - \frac{1}{3!}t^3 + 0\frac{1}{4!}t^4 + \frac{1}{5!}t^5 \ldots, \tag{10.85}$$

$$\cos t = 1 + 0t - \frac{1}{2!}t^2 + 0\frac{1}{3!}t^3 + \frac{1}{4!}t^4 + 0\frac{1}{5!}t^5 \ldots. \tag{10.86}$$

With these expansions now consider the following combinations: $(\cos t + i \sin t)|_{t=\theta}$ and $e^t|_{t=i\theta}$:

$$\cos\theta + i\sin\theta = 1 + i\theta - \frac{1}{2!}\theta^2 - i\frac{1}{3!}\theta^3 + \frac{1}{4!}\theta^4 + i\frac{1}{5!}\theta^5 + \ldots, \tag{10.87}$$

$$e^{i\theta} = 1 + i\theta + \frac{1}{2!}(i\theta)^2 + \frac{1}{3!}(i\theta)^3 + \frac{1}{4!}(i\theta)^4 + \frac{1}{5!}(i\theta)^5 + \ldots, \tag{10.88}$$

$$= 1 + i\theta - \frac{1}{2!}\theta^2 - i\frac{1}{3!}\theta^3 + \frac{1}{4!}\theta^4 + i\frac{1}{5!}\theta^5 + \ldots \tag{10.89}$$

As the two series are identical, we have Euler's formula

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{10.90}$$

Powers of complex numbers can be easily obtained using *de Moivre's*[6] formula:

$$e^{in\theta} = \cos n\theta + i\sin n\theta. \tag{10.91}$$

10.10.2 Polar and Cartesian representations

Now if we take x and y to be real numbers and define the complex number z to be

$$z = x + iy, \tag{10.92}$$

we can multiply and divide by $\sqrt{x^2 + y^2}$ to obtain

$$z = \sqrt{x^2 + y^2}\left(\frac{x}{\sqrt{x^2 + y^2}} + i\frac{y}{\sqrt{x^2 + y^2}}\right). \tag{10.93}$$

Noting the similarities between this and the transformation between Cartesian and polar coordinates suggests we adopt

$$r = \sqrt{x^2 + y^2}, \qquad \cos\theta = \frac{x}{\sqrt{x^2 + y^2}}, \qquad \sin\theta = \frac{y}{\sqrt{x^2 + y^2}}. \tag{10.94}$$

[6] Abraham de Moivre, 1667-1754, French mathematician.

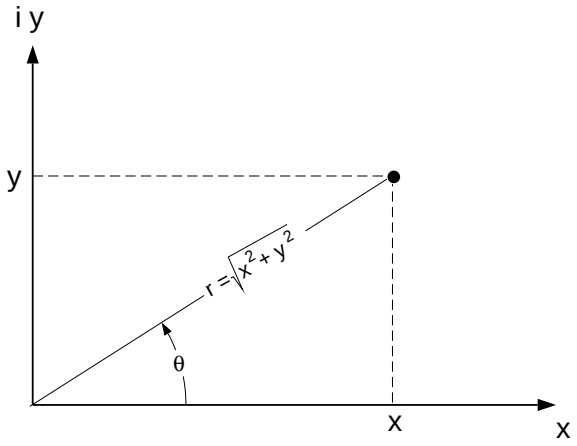

Figure 10.10: Polar and Cartesian representation of a complex number z.

Thus, we have

$$z = r\left(\cos\theta + i\sin\theta\right), \tag{10.95}$$
$$z = re^{i\theta}. \tag{10.96}$$

The polar and Cartesian representation of a complex number z is shown in Fig. 10.10. Now we can define the *complex conjugate* $\overline{z}$ as

$$\overline{z} = x - iy, \tag{10.97}$$
$$\overline{z} = \sqrt{x^2+y^2}\left(\frac{x}{\sqrt{x^2+y^2}} - i\frac{y}{\sqrt{x^2+y^2}}\right), \tag{10.98}$$
$$\overline{z} = r\left(\cos\theta - i\sin\theta\right), \tag{10.99}$$
$$\overline{z} = r\left(\cos(-\theta) + i\sin(-\theta)\right), \tag{10.100}$$
$$\overline{z} = re^{-i\theta}. \tag{10.101}$$

Note now that

$$z\overline{z} = (x+iy)(x-iy) = x^2+y^2 = |z|^2, \tag{10.102}$$
$$= re^{i\theta}re^{-i\theta} = r^2 = |z|^2. \tag{10.103}$$

We also have

$$\sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}, \tag{10.104}$$
$$\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}. \tag{10.105}$$

10.10.3 Cauchy-Riemann equations

Now it is possible to define complex functions of complex variables $W(z)$. For example take a complex function to be defined as

$$\begin{aligned} W(z) &= z^2 + z, && (10.106)\\ &= (x+iy)^2 + (x+iy), && (10.107)\\ &= x^2 + 2xyi - y^2 + x + iy, && (10.108)\\ &= \left(x^2 + x - y^2\right) + i\left(2xy + y\right). && (10.109) \end{aligned}$$

In general, we can say

$$W(z) = \phi(x,y) + i\psi(x,y). \tag{10.110}$$

Here ϕ and ψ are *real* functions of *real* variables.

Now $W(z)$ is defined as *analytic* at z_o if dW/dz exists at z_o and is independent of the direction in which it was calculated. That is, using the definition of the derivative

$$\left.\frac{dW}{dz}\right|_{z_o} = \frac{W(z_o+\Delta z) - W(z_o)}{\Delta z}. \tag{10.111}$$

Now there are many paths that we can choose to evaluate the derivative. Let us consider two distinct paths, $y = C_1$ and $x = C_2$. We will get a result which can be shown to be valid for arbitrary paths. For $y = C_1$, we have $\Delta z = \Delta x$, so

$$\left.\frac{dW}{dz}\right|_{z_o} = \frac{W(x_o + iy_o + \Delta x) - W(x_o + iy_o)}{\Delta x} = \left.\frac{\partial W}{\partial x}\right|_y. \tag{10.112}$$

For $x = C_2$, we have $\Delta z = i\Delta y$, so

$$\left.\frac{dW}{dz}\right|_{z_o} = \frac{W(x_o + iy_o + i\Delta y) - W(x_o + iy_o)}{i\Delta y} = \frac{1}{i}\left.\frac{\partial W}{\partial y}\right|_x = -i\left.\frac{\partial W}{\partial y}\right|_x. \tag{10.113}$$

Now for an analytic function, we need

$$\left.\frac{\partial W}{\partial x}\right|_y = -i\left.\frac{\partial W}{\partial y}\right|_x. \tag{10.114}$$

or, expanding, we need

$$\begin{aligned} \frac{\partial \phi}{\partial x} + i\frac{\partial \psi}{\partial x} &= -i\left(\frac{\partial \phi}{\partial y} + i\frac{\partial \psi}{\partial y}\right), && (10.115)\\ &= \frac{\partial \psi}{\partial y} - i\frac{\partial \phi}{\partial y}. && (10.116) \end{aligned}$$

For equality, and thus path independence of the derivative, we require

$$\frac{\partial \phi}{\partial x} = \frac{\partial \psi}{\partial y}, \qquad \frac{\partial \phi}{\partial y} = -\frac{\partial \psi}{\partial x}. \tag{10.117}$$

These are the well known *Cauchy-Riemann* equations for analytic functions of complex variables.

Now most common functions are easily shown to be analytic. For example for the function $W(z) = z^2 + z$, which can be expressed as $W(z) = (x^2 + x - y^2) + i(2xy + y)$, we have

$$\phi(x,y) = x^2 + x - y^2, \qquad \psi(x,y) = 2xy + y, \tag{10.118}$$

$$\frac{\partial \phi}{\partial x} = 2x + 1, \qquad \frac{\partial \psi}{\partial x} = 2y, \tag{10.119}$$

$$\frac{\partial \phi}{\partial y} = -2y, \qquad \frac{\partial \psi}{\partial y} = 2x + 1. \tag{10.120}$$

Note that the Cauchy-Riemann equations are satisfied since $\partial\phi/\partial x = \partial\psi/\partial y$ and $\partial\phi/\partial y = -\partial\psi/\partial x$. So the derivative is independent of direction, and we can say

$$\frac{dW}{dz} = \left.\frac{\partial W}{\partial x}\right|_y = (2x+1) + i(2y) = 2(x+iy) + 1 = 2z + 1. \tag{10.121}$$

We could get this result by ordinary rules of derivatives for real functions.

For an example of a non-analytic function consider $W(z) = \overline{z}$. Thus

$$W(z) = x - iy. \tag{10.122}$$

So $\phi = x$ and $\psi = -y$, $\partial\phi/\partial x = 1$, $\partial\phi/\partial y = 0$, and $\partial\psi/\partial x = 0$, $\partial\psi/\partial y = -1$. Since $\partial\phi/\partial x \neq \partial\psi/\partial y$, the Cauchy-Riemann equations are not satisfied, and the derivative depends on direction.

Problems

1. Find the limit as $x \to 0$ of
$$\frac{4\cosh x + \sinh(\arctan \ln \cos 2x) - 4}{e^{-x} + \arcsin x - \sqrt{1+x^2}}.$$

2. Find $d\phi/dx$ in two different ways, where
$$\phi = \int_{x^2}^{x^4} x\sqrt{y}\,dy.$$

3. Determine
   (a) $\sqrt[4]{i}$,
   (b) $i^{\sqrt[i]{i}}$.

4. Write three terms of a Taylor series expansion for the function $f(x) = \exp(\tan x)$ about the point $x = \pi/4$. For what range of x is the series convergent?

5. Find all complex numbers $z = x + iy$ such that $|z + 2i| = |1 + i|$.

6. Determine $\lim_{n\to\infty} z_n$ for $z_n = \frac{3}{n} + ((n+1)/(n+2))i$.

7. A particle is constrained to a path which is defined by the function $s(x,y) = x^2 + y - 5 = 0$. The velocity component in the x-direction, $dx/dt = 2y$. What are the position and velocity components in the y-direction when $x = 4$.

8. The error function is defined as $\text{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\, du$. Determine its derivative with respect to x.

9. Verify that
$$\lim_{n\to\infty} \int_\pi^{2\pi} \frac{\sin nx}{nx}\, dx = 0.$$

10. Write a Taylor series expansion for the function $f(x,y) = x^2 \cos y$ about the point $x = 2$, $y = \pi$. Include the x^2, y^2 and xy terms.

11. Show that
$$\phi = \int_0^\infty e^{-t^2} \cos 2tx\, dt,$$
satisfies the differential equation
$$\frac{d\phi}{dx} + 2\phi x = 0.$$

12. Evaluate the Dirichlet discontinuous integral
$$I = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin ax}{x} dx,$$
for $a \in (-\infty, \infty)$. You can use the results of example 3.11, Greenberg.

13. Defining
$$u(x,y) = \frac{x^3 - y^3}{x^2 + y^2},$$
except at $x = y = 0$, where $u = 0$, show that $u_x(x,y)$ exists at $x = y = 0$ but is not continuous there.

14. Using complex numbers show that
    (a) $\cos^3 x = \frac{1}{4}(\cos 3x + 3\cos x)$,
    (b) $\sin^3 x = \frac{1}{4}(3\sin x - \sin 3x)$.

Bibliography

M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1964.

V. I. Arnold, *Ordinary Differential Equations*, MIT Press, Cambridge, MA, 1973.

V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer, New York, NY, 1983.

A. A. Andronov, *Qualitative Theory of Second Order Dynamical Systems*, Wiley, New York, NY, 1973.

R. Aris, *Vectors, Tensors, and the Basic Equations of Fluid Mechanics*, Dover, New York, NY, 1962.

N. H. Asmar, *Applied Complex Analysis with Partial Differential Equations*, Prentice-Hall, Upper Saddle River, NJ, 2002.

G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge University Press, Cambridge, UK, 1996.

R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, NY, 1963.

C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Springer-Verlag, New York, NY, 1999.

M. L. Boas, *Mathematical Methods in the Physical Sciences*, Third Edition, Wiley, New York, NY, 2005.

A. I. Borisenko and I. E. Tarapov, *Vector and Tensor Analysis with Applications*, Dover, New York, NY, 1968.

M. Braun, *Differential Equations and Their Applications*, Springer-Verlag, New York, NY, 1983.

I. N. Bronshtein and K. A. Semendyayev, *Handbook of Mathematics*, Springer, Berlin, 1998.

C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, NY, 1988.

G. F. Carrier and C. E. Pearson, *Ordinary Differential Equations*, SIAM, Philadelphia, PA, 1991.

P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge University Press, Cambridge, UK, 1989.

J. A. Cochran, H. C. Wiser and B. J. Rice, *Advanced Engineering Mathematics*, Second Edition, Brooks/Cole, Monterey, CA, 1987.

R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vols. 1 and 2, Wiley, New York, NY, 1989.

I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

L. Debnath and P. Mikusinski, *Introduction to Hilbert Spaces with Applications*, Third Edition, Elsevier, Amsterdam, Netherlands, 2005.

P. G. Drazin, *Nonlinear Systems*, Cambridge University Press, Cambridge, UK, 1992.

R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, New York, NY, 1977.

J. Feder, *Fractals*, Plenum Press, New York, NY, 1988.

B. A. Finlayson, *The Method of Weighted Residuals and Variational Principles*, Academic Press, New York, NY, 1972.

B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge, New York, NY, 1998.

B. Friedman, *Principles and Techniques of Applied Mathematics*, Dover Publications, New York, NY, 1956.

I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Dover, New York, NY, 2000.

J. Gleick, *Chaos*, Viking, New York, NY, 1987.

G. H. Golub and C. F. Van Loan, *Matrix Computations*, Third Edition, Johns Hopkins, Baltimore, MD, 1996.

S. W. Goode, *An Introduction to Differential Equations and Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, PA, 1977.

M. D. Greenberg, *Foundations of Applied Mathematics*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

J. Guckenheimer and P. H. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, NY, 1983.

J. Hale and H. Koçak, *Dynamics and Bifurcations*, Springer-Verlag, New York, NY, 1991.

F. B. Hildebrand, *Advanced Calculus for Applications*, 2nd Ed., Prentice-Hall, Englewood Cliffs, NJ, 1976.

M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, Boston, MA, 1974.

M. H. Holmes, *Introduction to Perturbation Methods*, Springer-Verlag, New York, NY, 1995.

M. H. Holmes, *Introduction to the Foundations of Applied Mathematics*, Springer-Verlag, New York, NY, 2009.

M. Humi and W. Miller, *Second Course in Ordinary Differential Equations for Scientists and Engineers*, Springer-Verlag, New York, NY, 1988.

E. J. Hinch, *Perturbation Methods*, Cambridge, Cambridge, UK, 1991.

D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, Clarendon Press, Oxford, UK, 1987.

P. B. Kahn, *Mathematical Methods for Engineers and Scientists*, Dover, New York, NY, 2004.

W. Kaplan, *Advanced Calculus*, Fifth Edition, Addison-Wesley, Boston, MA, 2003.

D. C. Kay, *Tensor Calculus*, Schaum's Outline Series, McGraw-Hill, New York, NY, 1988.

J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, NY, 1981.

J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods*, Springer-Verlag, New York, NY, 1996.

A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Dover, New York, NY, 1999.

L. D. Kovach, *Advanced Engineering Mathematics*, Addison-Wesley, Reading, MA,1982.

E. Kreyszig, *Advanced Engineering Mathematics*, Ninth Edition, Wiley, New York, NY, 2005.

E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, New York, NY, 1978.

P. D. Lax, *Functional Analysis*, Wiley, New York, NY, 2002.

P. D. Lax, *Linear Algebra and its Applications*, Second Edition, Wiley, Hoboken, NJ, 2007.

A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics*, Second Edition, Springer, Berlin, 1992.

C. C. Lin and L. A. Segel, *Mathematics Applied to Deterministic Problems in the Natural Sciences*, SIAM, Philadelphia, PA, 1988.

J. R. Lee, *Advanced Calculus with Linear Analysis*, Academic Press, New York, NY, 1972.

J. D. Logan, *Applied Mathematics*, Third Edition, Wiley, Hoboken, NJ, 2006.

R. J. Lopez, *Advanced Engineering Mathematics*, Addison Wesley Longman, Boston, MA, 2001.

J. Mathews and R. L. Walker, *Mathematical Methods of Physics*, Addison-Wesley, Redwood City, CA, 1970.

A. J. McConnell, *Applications of Tensor Analysis*, Dover, New York, NY, 1957.

A. N. Michel and C. J. Herget, *Applied Algebra and Functional Analysis*, Dover, New York, NY, 1981.

P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vols. 1 and 2, McGraw-Hill, New York, NY, 1953.

J. A. Murdock, *Perturbations, Theory and Methods*, John Wiley and Sons, New York, NY, 1991.

J. T. Oden and L. F. Demkowicz, *Applied Functional Analysis*, CRC, Boca Raton, FL, 1996.

P. V. O'Neil, *Advanced Engineering Mathematics*, Sixth Edition, CL-Engineering, 2006.

L. Perko, *Differential Equations and Dynamical Systems*, Third Edition, Springer, Berlin, 2006.

J. N. Reddy, *Applied Functional Analysis and Variational Methods in Engineering*, McGraw-Hill, New York, NY, 1986.

J. N. Reddy and M. L. Rasmussen, *Advanced Engineering Analysis*, Wiley, New York, NY, 1982.

F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, NY, 1990.

K. F. Riley, M. P. Hobson, and S. J. Bence, *Mathematical Methods for Physics and Engineering*, Third Edition, Cambridge, Cambridge, UK, 2006.

M. Rosenlicht, *Introduction to Analysis*, Dover, New York, NY, 1968.

H. M. Schey, *Div, Grad, Curl, and All That*, Fourth Edition, W.W. Norton, London, 2005.

M. J. Schramm, *Introduction to Real Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1996.

I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, 2nd Ed., McGraw-Hill, New York, NY, 1966.

G. Stephenson and P. M. Radmore, *Advanced Mathematical Methods for Engineering and Science Students*, Cambridge University Press, Cambridge, UK, 1990.

G. Strang, *Linear Algebra and its Applications*, 2nd Ed., Academic Press, New York, NY, 1980.

G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge, Wellesley, MA, 1986.

S. H. Strogatz, *Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry, and Engineering*, Westview, Boulder, CO, 2001.

M. Van Dyke, *Perturbation Methods in Fluid Mechanics*, Parabolic Press, Stanford, CA, 1975.

S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer Verlag, New York, NY, 1990.

C. R. Wylie and L. C. Barrett, *Advanced Engineering Mathematics*, 6th Ed., McGraw-Hill, New York, NY, 1995.

D. Xiu, *Numerical Methods for Stochastic Computations*, Princeton, Princeton, NJ, 2010.

E. Zeidler, *Applied Functional Analysis*, Springer Verlag, New York, NY, 1995.

D. G. Zill and M. R. Cullen, *Advanced Engineering Mathematics*, Third Edition, Jones and Bartlett, Boston, MA, 2006.