1

ODE Integration Methods

1.1 INTRODUCTION

Before we start to solve *differential equations* (DEs) by *numerical integration*, we will briefly review some background topics prior to discussing a selection of the available methods. We will do this in a nonrigorous way, as this aspect is covered in detail in any textbook dealing with the *calculus*; for example, see [Joh-08] for an excellent basic introduction to differentiation and integration and [Kre-11] for a more in-depth and broader coverage of the calculus and its application. The latter book is considered by many science, technology, engineering, and mathematics (STEM) students to be their math bible.

When we first start to study the calculus, we learn a little about the historical context in which it was conceived and the priority dispute between Newton[1] and Leibniz[2] over who invented it. Today, this dispute is considered moot, and historians credit both with independent discovery. We are then introduced to the concept of a *limit*, whereby we evaluate a *continuous* function, $f(x)$, over the interval $[a, b]$—written as $f(x) \in C[a, b]$—as x moves closer to a particular value, say, $x = x_0$. This is, of course, only necessary when $f(x_0)$ is indeterminate, as otherwise we could just plug $x = x_0$ into $f(x)$. Later in the course, we learn advanced methods for dealing with indeterminate situations by obtaining limits that require the use of *derivatives*.

The study of limits leads naturally to the idea of a *derivative*, being equal in one-dimensional space to the *slope* of a *tangent* to the curve of our function $f(x)$. We write the derivative of $f(x)$ with respect to x as $\frac{df(x)}{dx}$ or, alternatively, as $f'(x)$, where the leading d in the numerator and denominator is an *operator* representing an *infinitesimal* change in $f(x)$ and x, respectively. Usually, our first attempt at obtaining an approximation to the tangent is to consider a straight line constructed between two points: one located at the point on the curve where we wish to evaluate the derivative, x_0, and the other a short

[1] Sir Isaac Newton, English mathematician and scientist (1642–1727), also famous for discovering the inverse law of gravity, which he published in 1687 in his famous book *Philosophi Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy), commonly known as the *Principia*.

[2] Gottfried Wilhelm von Leibniz, German mathematician and philosopher (1646–1716). He was also linguist and wrote extensively on a wide range of subjects, including philosophy, politics, law, ethics, theology, history, and philology.

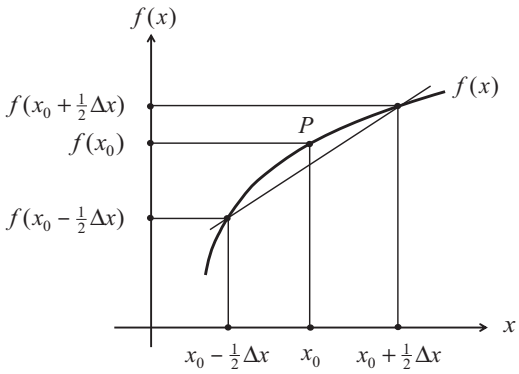

Figure 1.1. Diagram of a two-sided derivative construction. In the limit as $\Delta x \to 0$, the slope at point P will be equal to the derivative of $f(x)$ evaluated at $x = x_0$, that is, $\frac{df(x)}{dx}\big|_{x=x_0}$.

increment away. This is called a *single-sided* approximation, and the tangent slope, s, can be calculated in two different ways:

$$\begin{aligned} s_R &= \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}, \quad \text{right-sided} \\ s_L &= \frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x}, \quad \text{left-sided,} \end{aligned} \tag{1.1}$$

where the symbol Δx means a small increment in x.

A more accurate approximation is the *two-sided* approximation, where the points are located equidistantly either side of the point $f(x_0)$. The slope of this approximate tangent is given by

$$s = \frac{f(x_0 + \Delta x/2) - f(x_0 - \Delta x/2)}{\Delta x}; \tag{1.2}$$

see Fig. (1.1).

Clearly the approximation of eqn. (1.2) will become increasingly more accurate as Δx becomes smaller and the points move closer to x_0. In the limit when $\Delta x = 0$, we obtain a value for the actual tangent slope at x_0 and, consequently, the derivative. This is why we were introduced to the concept of a limit prior to discussing derivatives. A two-sided derivative of $f(x)$ at $x = x_0$ can be defined as

$$\frac{df(x = x_0)}{dx} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x/2) - f(x_0 - \Delta x/2)}{\Delta x}, \quad f(x) \in C^1, \tag{1.3}$$

where the symbol C^1 means that this definition applies to a function $f(x)$ that is both continuous and has a continuous first derivative at $x = x_0$.

Figure 1.2 includes a plot showing how a two-sided derivative approximation converges rapidly as $\Delta x \to 0$. The function under consideration is $f(x) = x^{\frac{1}{3}}$, and the tangent slope is evaluated at $x = 0.5$.

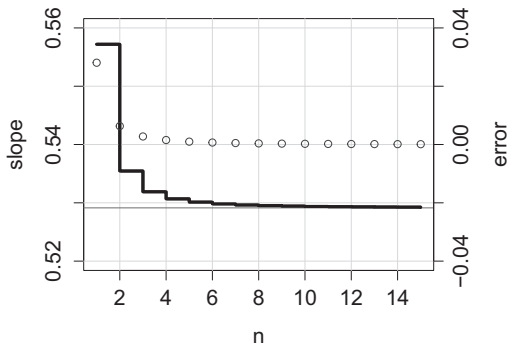

Figure 1.2. Plot of two-sided derivative approximate values of the function $f(x) = x^{\frac{1}{3}}$ evaluated as $x = 0.5$ (stepped) with corresponding errors (circles). Values for the denominator in eqn. (1.3) are $\Delta x = 0.5/i,\ i = 1, \ldots, 15$. The slope approximation converges rapidly toward the true value of 0.529 (thin horizontal line) as i increases.

The R code that produced Fig. (1.2) is shown in Listing 1.1.

```
# File: derivativeApprox.R
# Two-sided derivative approx. of f(x)=x^(1/3) evaluated at x=0.5
N <- 15; x <- seq(0,1,len=N); slope <- rep(0,N); err <- rep(0,N)
Nrange <- 1:N
x0 <- 0.5; x1 <- 0.25; x2 <- 0.75
Dx <- x0-x1 # equal spacing about x=0.5
s_anal <- 0.5291337 # derivative of f(x) at x=0.5
for(n in Nrange){
  dx <- Dx/n                      # Next spacing of x
  xa <- x0-dx; xb <- x0+dx        # x points for new slope
  fa <- xa^(1/3); fb <- xb^(1/3)  # new function values
  slope[n] <- (fb-fa)/(xb-xa)     # derivative approx.
}
err <- slope-s_anal               # error
txt <- sprintf("tangent slope=%5.3e at x0=%3.1f for dx=%5.3e\n",slope[N],x0,dx)
cat(txt) # print summary of final result
#
par(mar=c(4,5,2,5))
plot(Nrange,slope[Nrange],type="s",lwd=3, xlim=c(1,N),
     ylim=c(0.52,0.56), xlab="n",ylab="slope") # type="s" means step plot
grid(lty=1); abline(h=s_anal,col="red") # Analytical value
lines(Nrange,slope[Nrange],lwd=3,type="s")
#
par(new = T) # Plot error on second y=axis
plot(Nrange,err[Nrange], col = "blue", axes = F,
     xlab = NA, ylab = NA,ylim=c(-0.04,0.04))
axis(side = 4); mtext ( text ="error", side = 4, line = 3 )
```

Listing 1.1. File: `derivativeApprox.R`—Code to generate two-sided derivative approximations

Table 1.1. Basic derivative pairs

Function, $f(x)$	Derivative, $f'(x) = \dfrac{df(x)}{dx}$
x^{an}	nax^{an-1}
$\sin ax$	$a\cos ax$
$\tan ax$	$a\sec^2 ax$
$\exp ax$	$a\exp ax$
$\vdots$	$\vdots$
etc.	

Once we have grasped the concept of the derivative of a function, we then learn how to differentiate analytically a whole series of functions, including trigonometrical functions, quotients, powers, and so on. We also learn to differentiate a function $f(x)$ with respect to x multiple times, where the nth derivative is written as $\frac{d^n f(x)}{dx^n}$. From this point on, we are able to use a lookup table to obtain derivatives of standard functions, such as those listed in Table 1.1.

At this point, we are equipped to apply differentiation to solving real-world problems such as min/max of functions, obtaining velocity from distance traveled, and acceleration from velocity. We also learn more advanced methods of obtaining limits that require the use of derivatives, for example, *l'Hôpital's rule*.[3]

The next stage in learning the calculus is an introduction to the concept of *integration* of a function. We learn that *indefinite integration* is the opposite of differentiation (i.e., the *antiderivative*) and needs to include a constant of integration and that *definite integration* (also known as *quadrature*) is equal to the area under the curve of a function $f(x)$ from $x = x_a$ to $x = x_b$. We are shown how to obtain an approximation to the definite integral of our function $f(x)$ with respect to x, written as $\int_a^b f(x)\,dx$, by dividing the area into a set of rectangles and summing the areas of the individual rectangles. This is known as a *Riemann sum*.[4] A Riemann sum can be formed in three ways: left, center, and right (see Fig. 1.3).

The area of one rectangle at, say, $x = x_i$ is given by $A_i = (x_{i+1} - x_i) f(x_i)$. Therefore, using a left Riemann sum of n rectangles, the total area is given by $A = \sum_{i=0}^{n-1} A_i = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$. Similar to the derivative discussed earlier, the approximation will become increasingly more accurate as n increases and x_i and x_{i+1} approach each other and the rectangle widths reduce. Using our knowledge of limits, we can now define a definite integral of $f(x)$ with respect to x from $x = x_a$ to $x = x_b$. This can be done using the left and right Riemann sums, as follows:

$$
\begin{aligned}
I_L &= \int_{x=x_a}^{x=x_b} f(x)\,dx = \lim_{n\to\infty} \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i), \quad x_a = x_1,\ x_b = x_n \\
I_R &= \int_{x=x_a}^{x=x_b} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} (x_{i+1} - x_i) f(x_i), \quad x_a = x_1,\ x_b = x_n,
\end{aligned}
\tag{1.4}
$$

[3] First published by French mathematician Guillaume François Antoine, Marquis de l'Hôpital (1661–1704). If $\lim_{x\to x_0} \frac{f(x)}{g(x)}$ is indeterminate, but $\lim_{x\to x_0} \frac{f'(x)}{g'(x)}$ exists, then the method states that the first limit also exists and is equal to the second limit.

[4] After German mathematician Georg Friedrich Bernhard Riemann (1826–1863).

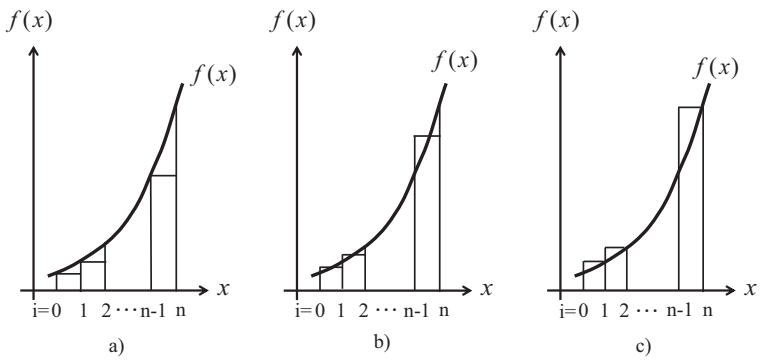

Figure 1.3. Diagram showing three possible Riemann sums for the area under the curve of function $f(x)$ between $x = x_0$ and $x = x_n$: (a) A *left* Riemann sum where the area $A = \sum_{i=0}^{n-1} f_i(x_{i+1} - xi)$; (b) a *center* Riemann sum where the area $A = \sum_{i=0}^{n-1}(f_i + f_{i+1})(x_{i+1} - xi)/2$; and (c) a *right* Riemann sum where the area $A = \sum_{i=1}^{n} f_i(x_{i+1} - xi)$. Here we have used the symbol f_i to represents $f(x_i)$.

where I_L represents the *left* Riemann sum and I_R represents the *right* Riemann sum. Note that it is not a requirement that the x_i be spaced equally along the x-axis.

A more accurate approximation is obtained from the *center* Riemann sum, where the height of the rectangle at $f(x_i)$ is equal to the average height of the two values $f(x_i)$ and $f(x_{i+1})$. The integral approximation by this method is given by

$$I_C = \int_{x=x_a}^{x=x_b} f(x)\,dx = \lim_{n\to\infty} \sum_{i=0}^{n-1} \frac{1}{2}(x_{i+1} - x_i)(f(x_i) + f(x_{i+1}), \quad x_a = x_1, x_b = x_n \tag{1.5}$$

(see Fig. 1.1).

Figure 1.4 includes a plot of the center Riemann sum approximation to the definite integral of $f(x) = x^3$ evaluated between $x = 0$ and $x = 1$. The approximation converges toward the true integral value of 0.25 as the number of rectangles n increases. The center sum approximation requires 12 rectangles to reduce the error to below 1%, whereas left

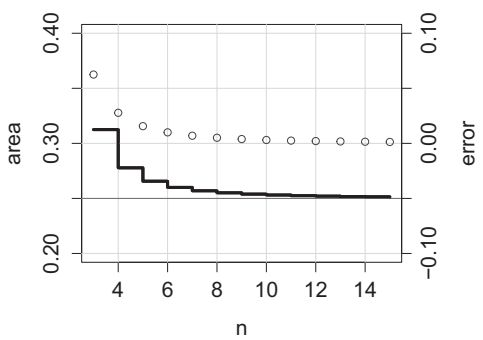

Figure 1.4. Plot of *center* Riemann sum approximate values of the integral of $f(x) = x^3$ evaluated between $x = 0$ and $x = 1$ (stepped) with corresponding errors (circles). The integral approximation converges to the true value of 0.25 (thin horizontal line) as n increases.

and right approximations would require more that 200 rectangles to achieve the same accuracy.

The R code that produced Fig. 1.4 is shown in Listing 1.2.

```
# File: RiemannSum.R
# Center Riemann sum of f(x)=x^3 evaluated from x=0 to x=1
N <- 15; Nrange <- seq(3,N,by=1)
area = rep (0,N)
I_anal <- 0.25            # Integration of f(x) between x=0 and x=1
for (n in Nrange ){       # calculate Riemann sum for 3 to N intervals
  x  <- seq (0,1, len=n)  # Next sequence from x=0 to x=1
  dx <- x[2] -x [1]       # equal spacing
  f  <- x^3               # evaluate f(x), x=0,...,1
  area[n] <- dx*(sum(f[2:(n-1)])+(f[1]+f[n])/2)
}
err <- area-I_anal        # error
txt <- sprintf("Center Riemann sum =%5.3e for n=%d rectangles \n", area[n],n)
cat(txt) # print summary of final result
# Plot results
par(mar=c(4,5,2,5))
plot ( Nrange , area [Nrange], type ="s", lwd =3, ylim =c (0.2 ,0.4),
       xlab ="n", ylab ="area") # type ="s" means step plot
grid ( lty =1); abline (h =0.25 , col =" red ") # Analytical value
#
par(new = T) # Plot error on second y=axis
plot(Nrange,err[Nrange], col = "blue", axes = F,
     xlab = NA, ylab = NA,ylim=c(-0.1,0.1))
axis(side = 4); mtext ( text ="error", side = 4, line = 3 )
```

Listing 1.2. File: `RiemannSum.R`—Code to generate Riemann sums

Once we have grasped the concept of integration, we learn how to integrate analytically a wide range of functions, including trigonometrical and logarithmic functions, quotients, and powers. As for derivatives, we are now able to use a lookup table to obtain integrals of standard functions, such as those listed in Table 1.2.

With our new knowledge of integration, we can now solve additional real-world problems to those we tackled using differentiation, such as distance traveled from velocity, volumes of revolution, areas between curves, and lengths of curves.

At this stage in our study, we have learned the basics of the calculus and how to use it to solve otherwise intractable problems. We are then equipped and ready to be exposed to the concept of *differential equations* (DEs), and this opens up a whole new field of study. Historically, differential equations arose from the analysis of physical system dynamics. Subsequently, the ideas have moved into other disciplines, such as finance and the social sciences.

Table 1.2. Basic integral pairs

Function, $f(x)$	Integral, $F(x) = \int f(x)\,dx$
x^{an}	$\frac{x^{an+1}}{an+1}$
$\sin ax$	$-\frac{1}{a}\cos ax$
$\tan ax$	$\frac{1}{a}\ln(\lvert\sec ax\rvert)$
$\exp ax$	$\frac{1}{a}\exp ax$
$\vdots$	$\vdots$
etc.	

Note: (a) An *indefinite integral* is equal to $F(x)+k$, where k represents a constant of integration. (b) A *definite integral* that is evaluated between the limits $x = x_a$ and $x = x_b$ is equal to $F(x_b) - F(x_a)$. This is known as the *fundamental theorem of calculus* and establishes a formal relationship between differentiation and integration.

A differential equation is an equation that can include constants, functions, and, of course, derivatives; hence use of the term *differential*. It is appropriate here to clarify the following terms that are used to classify differential equations:

- A DE must include one or more derivatives of any order, for example, first, second, or third and the order of a DE is determined by the highest-order derivative.
- A DE is called *linear* if all terms containing the dependent variable or its derivatives are of the first degree and do not contain higher powers or products.
- A DE is defined as *homogeneous* if all the terms contain the dependent variable and/or its derivative of any order. Thus a constant term destroys homogeneity and the DE becomes *nonhomogeneous*.

For example, the following represents a linear, nonhomogeneous, second-order differential equation:

$$d\,\frac{d^2 f(x)}{dx^2} + c\,\frac{df(x)}{dx} + b\,f(x) + a = 0, \tag{1.6}$$

where a, b, c, and d are constants and $f(x)$ is an unknown function that we wish to determine, that is, the solution. However, for a sufficiently smooth function, there exists an infinite family of functions that will satisfy eqn. (1.6). Depending on the type and order of DE, to obtain a particular solution, $f(x)$, we also need to specify appropriate initial conditions (ICs) and/or boundary conditions (BCs). The number of these auxiliary conditions is equal to the *highest-order derivative with respect to the independent variable*. For example, a first-order initial value ODE with dependent variable $u(t)$ requires one IC at the initial value of t, that is, $u_0 = u(t_0)$. Similarly, the second-order ODE given in eqn. (1.6)

requires one IC for the dependent variable at the initial value of x, that is, $f_0 = f(x_0)$, plus an initial value for the first derivative of the dependent variable, that is, $f_0' = f'(x_0)$, and so on. These requirements are illustrated by example in subsequent sections and are also discussed in more detail for PDEs in Chapter 3.

Equation (1.6) is defined as an *ordinary differential equation* (ODE) as there is only one independent variable, x. If a DE includes more than one independent variable, say, x and t, then this DE is called a *partial differential equation* (PDE). For example, the following represents a second-order PDE, the so-called *diffusion equation*:

$$\frac{\partial u(x,t)}{\partial t} = D\frac{\partial^2 u(x,t)}{\partial x^2}, \tag{1.7}$$

where D is the diffusion constant and the unknown function is represented by $u(x,t)$. Note that the derivatives are now called *partial derivatives* and the operator symbol has changed from d to ∂. A partial derivative is defined as

$$\frac{\partial u(x,t)}{\partial x} = \left.\frac{du(x,t)}{dx}\right|_t, \tag{1.8}$$

which means that $\frac{\partial u(x,t)}{\partial x}$ is equal to $\frac{du(x,t)}{dx}$ with all independent variables, other than x, held constant—in this case, just t. Hence use of the term *partial*.

Partial derivatives still represent the slope of a tangent, but instead of the tangent being to a line, it is a tangent line on a surface. In this instance, the tangent line is aligned in the direction of the variable x. PDEs are discussed in more detail in Chapter 3.

The process of solving a differential equation involves integration, either directly or indirectly. Hence we refer to *integration of differential equations*. We usually start learning about differential equations by studying certain types of ODEs, some of which can be solved by substituting an *ansatz*[5] into the equation and solving for the unknowns. For example, letting $a = -1$, $b = c = 1$, $d = 0$, and $f(x) = y(x)$ in eqn. (1.6) yields the first-order linear nonhomogeneous ODE

$$\frac{dy}{dx} + y = 1, \quad y = y(x). \tag{1.9}$$

In our DE course, we learned that for a linear nonhomogeneous ODE, the solution consists of two parts, a *homogeneous* solution, y_h, to the homogeneous part of the ODE and a *particular* solution, y_p, to the nonhomogeneous part, with the *total* solution being equal to $y = y_h + y_p$. First we substitute the ansatz $y = Ae^{mx}$ into the homogeneous part of eqn. (1.9) (by setting the nonhomogeneous right-hand side to zero), giving

$$Ame^{mx} + Ae^{mx} = 0. \tag{1.10}$$

If we now divide through by Ae^{mx}, we obtain the *characteristic* equation

$$m + 1 = 0, \quad \therefore m = -1. \tag{1.11}$$

Thus $y_h = Ae^{-x}$ is a *general* solution to the homogeneous part of eqn. (1.9), with A being an arbitrary constant.

[5] The term *ansatz* is used to describe an assumed form of solution, a trial solution, or an educated guess.

We now attempt to solve for y_p trying the ansatz $y_p = C \times 1$, a constant multiplied by the nonhomogeneous right-hand side of eqn. (1.9),[6] which yields

$$0 + C = 1. \tag{1.12}$$

Therefore $C = 1$ and $y = y_h + y_p = Ae^{-x} + 1$. However, if we know a value of y for a particular value of x, we can solve for A. This is known as the *particular* solution to eqn. (1.9)—because particular values are assigned to the constants A and m. Letting the known value be the initial condition $y = 0$ when $x = 0$ yields $A = -1$, and the overall particular solution satisfying the initial condition becomes

$$y = 1 - e^{-x}. \tag{1.13}$$

Of course, for this case, eqn. (1.9) is a *separable equation* and can be solved by a simple integration, that is,

$$\begin{gathered} \int \frac{1}{1-y} dy = \int dx, \\ \therefore -\ln(1-y) + k_1 = x + k_2, \quad \Rightarrow \quad y = 1 + Ae^{-x}, \end{gathered} \tag{1.14}$$

where $k = k_1 - k_2$ is the combined constant of integration and $A = -e^k$. Thus we have obtained the same general solution; and using the same initial condition, we obtain the same overall particular solution, $y = 1 - e^{-x}$.

Higher-order ODEs can, in general, be transformed into a set of coupled first-order ODEs. By "coupled," we mean that the dependent variables occur in more than one equation. Thus transformation often facilitates analysis and can simplify the process of obtaining a general or particular solution. Consider the following second-order ODE:

$$\frac{d^2y}{dx^2} = a\frac{dy}{dx} + by + c, \quad y = y(x). \tag{1.15}$$

This equation can be transformed into two coupled first-order equations by letting $y_1 = y$ and $y_2 = \frac{dy_1}{dx}$, when we obtain, in matrix form,

$$\begin{bmatrix} \dfrac{dy_1}{dx} \\ \dfrac{dy_2}{dx}, \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ c \end{bmatrix}. \tag{1.16}$$

In the same way, a third-order system can be transformed into a set of three coupled first-order equations, and so on for higher-order systems. This form of mathematical presentation is very useful for analytical analysis when eigenvalues can be obtained that provide insight into the stability and dynamic behavior of the system being investigated. This approach can also be applied to the solution of PDEs and is discussed in later chapters.

The analysis of *nonlinear* differential equations is beyond the scope of this brief introductory overview. A wide discussion on DEs in general may be found in [Zwi-98], and an

[6] If the homogeneous solution contains the nonhomogeneous part, then we try letting y_p equal the nonhomogeneous part multiplied by x^n, where n is the minimum integer that makes $y_h + y_p$ linearly independent. There are many tricks for solving nonhomogeneous ODEs, for example, see [Kre-11].

interesting book that discusses nonlinear DEs and provides an introduction to nonlinear dynamics is [Tel-06].

An amazing characteristic of DEs is that very similar equations crop up in many different disciplines, for example, rate of reaction and continuity in chemistry, celestial mechanics and diffusion in physics, natural modes of vibration and inertial systems in mechanical engineering, and resistance–capacitance–inductance circuits and electromagnetic waves in electrical engineering. Even in financial analysis, there is the *Black–Scholes* PDE model[7] for describing stochastic[8] processes involved in the financial instruments known as *derivative options*. These PDEs include convection- and diffusion-like terms. Consequently, from their wide application, it is clear that the solution of DEs is an immensely important area of applied mathematics.

For many problems, ODEs cannot be solved analytically by the preceding or similar methods. This can be due to the presence of nonlinear terms, parameters that vary over the solution domain, difficult boundary conditions, or just the sheer complexity of the problem. In these situations, we have to resort to *numerical methods*. The process of solving ODEs/PDEs numerically is called *numerical integration* and is performed by a *numerical integrator*, alternatively referred to as a *numerical solver*. The matrix form of eqn. (1.16) is particularly useful in numerical analysis as standard numerical solvers usually expect ODEs in this form. A numerical solution to a differential equation is one where the differentials are replaced by a set of algebraic equations and whose solution provides an approximation to the true solution. The job of the numerical analyst is to find an approximate solution that is sufficiently close to the true solution.

The idea of numerical integration is to use one or more past values to approximate the derivative of the dependent variable, and then to use this to project the solution forward, either in time or space. The new solution in turn becomes a past value that is used to project forward to the next solution value, and so on. In this step-by-step approach, a solution trajectory gradually evolves until the problem is solved to the satisfaction of the analyst. Of course, this is a rather simplistic overview, and many considerations need to be taken into account to obtain a good numerical solution. For example, in later sections, we discuss integration errors and methods for keeping the solution within desired accuracy bounds, and in the next chapter, we learn about methods to investigate solution stability.

Numerical integration methods is a vast subject that we can only discuss briefly here, and we shall only cover a small selection of the more common integrators to give some insight into the numerical solution of initial value problems. The integrator methods we consider are suitable for solving initial value problems described by ordinary differential equations (ODEs) of the form

$$\frac{dy}{dt} = f(y,t), \quad [y(t_0) = y_0], \quad y \in \mathbb{R}^m, t > 0. \tag{1.17}$$

[7] The so-called Black–Scholes model, shown here without discussion, was published by Fischer Black and Myron Scholes in their 1973 paper "The Pricing of Options and Corporate Liabilities," published in the *Journal of Political Economy*:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV - rS\frac{\partial V}{\partial S}.$$

[8] A stochastic process is a *random* process that is governed by its statistics. It can be considered to be the opposite of a *deterministic* process.