Numerical Methods for Differential Equations and Applications

J. C. Butcher
The University of Auckland

January 18, 1997

Abstract

This paper surveys a number of aspects of numerical methods for ordinary differential equations. The discussion includes the method of Euler and introduces Runge-Kutta methods and linear multistep methods as generalizations of Euler. Stability considerations arising from stiffness lead to a discussion of implicit methods and implementation issues. To the extent possible within this short survey, numerical methods are looked at in the context of problems arising in practical applications.

1 Introduction

Differential equations play a role in the modelling of almost every scientific discipline. However, it is relatively rare for a differential equation to have a solution that can be written in terms of elementary functions. Usually, the only information about the solution is that it is known to exist and to be unique, on theoretical grounds, and that it can be approximated more or less accurately using computational techniques. In this review paper, we will consider some aspects of numerical methods for the solution of initial value problems in systems of ordinary differential equations. There are two standard forms for expressing such problems. The first of these is

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0. \tag{1}$$

Here the solution y is assumed to be a differentiable function on an interval $[x_0, \overline{x}]$ to a finite dimensional Euclidean space $\mathbb{R}^N$. The formulation (1) is very general and includes, for example, second and higher order differential equations; these are easily recast in this way. By introducing an additional variable, if necessary, which always remains exactly equal to x, it is possible to reformulate the general problem as an 'autonomous' system of equations. This is the second standard form.

$$y'(x) = f(y(x)), \quad y(x_0) = y_0. \tag{2}$$

Computer software for solving ordinary differential equations exists for both formulations and there is no advantage to either, except that many problems are most naturally expressed in the non-autonomous form (1) rather than autonomous form (2). However, for many theoretical investigations, the autonomous form is to be preferred. We will see an example of this in Section 3 where the exposition is much simpler and more transparent than would have been possible for the non-autonomous formulation.

The methods under study in this paper are introduced in Section 2 in the context of the central role of the Euler method and the relationship that many other methods have as generalizations of Euler. This leads on to Section 3 which discusses the one-step Runge-Kutta methods and this leads on to Section 4 where linear multistep methods are reviewed.

Section 5 discusses the important property of stiffness and the effect it has on the performance of traditional numerical methods. The stability considerations that become essential in the light of stiffness are the subject of Section 6.

In Section 7 issues concerned with the implementation of methods for stiff and non-stiff problems are discussed. This is followed by Section 8 which considers some applications for which special care is required. In Section 9 we make some concluding remarks.

2 The Euler method and related methods

The most natural physical interpretation of a differential equation and its solution is in terms of distance and velocity. If x is interpreted as 'time' and $y(x)$ as the position of a moving particle at a particular time then the value of $f(x, y(x))$ is the velocity at this time. Hence, we can interpret the solution at $x_0 + h$, where h is a small time interval as being the value of y_0 to which has been added the product of the width of this interval and the average velocity over this interval.

Of course, if we are given only the differential equation (1) we have no obvious means of calculating the average velocity; as a very first approximation, the average can be replaced by the velocity at the beginning of the interval; that is $f(x_0, y_0)$. Hence, if y_1 denotes an approximate solution at $x_1 = x_0 + h$, we have a possible numerical method based on the formula for its first step

$$y_1 + hf(x_0, y_0). \tag{3}$$

The numerical method originally proposed by Euler extends this idea by generating approximations at a sequence of points, $x_i = x_0 + hi$, $i = 1, 2, \dots$ where each point is related to the one next before it by the formula

$$y_i = y_{i-1} + hf(x_{i-1}, y_{i-1}), \quad i = 1, 2, \dots . \tag{4}$$

If it is required to find the solution at a known output point $\overline{x}$ then it is convenient to choose $h = (\overline{x} - x_0)/n$, where n is an integer.

Because (3) can be looked at as the first two terms of a Taylor expansion

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{h^2}{2}y''(x_0) + \cdots,$$

it should be expected that, at least for small values of h, the error in completing each step is approximately proportional to h^2. Since the total number of steps is proportional to h^{-1}, this would mean that the total error committed by the time n steps have been completed to produce an approximation to $y(\overline{x})$, would be approximately proportional to h.

The fact that errors behave like the first power of h is regarded as a serious limitation of the Euler method, because reducing h, and therefore increasing the total computational effort, leads to only a modest improvement in accuracy. What are preferred are 'higher order' methods for which the error in a step behaves approximately like h^{p+1} and the total or global error behaves like h^p, where the order p is 2 or more.

Using the velocity-distance interpretation of a differential equation and its solution, we can first consider how to overcome the limitation of having to use, instead of the average velocity in a time interval, the value at the beginning of the interval.

Several approaches have been used for approximating the average velocity more accurately than in the Euler method. One idea is to somehow use an approximation at the mid-point of the interval, instead of at the left-hand end. A second idea is to use the mean of the values at the two ends of the interval. The Runge-Kutta method, which we will explore in more detail in the next section, was originally based by Runge on each of these ideas. To obtain an approximation of the derivative at the centre or the end of the interval, it is possible to take a temporary step to the desired point and calculate the derivative using this approximate value as the second argument of the function f. Denote the temporary approximation as $y^*_{i-1/2}$ or y^*_i for the 'mid-point rule' and the 'trapezoidal rule' versions of Runge's methods, and we have the following two sequences of calculations to find the approximation after a single step.

Mid-point rule method:

$$\begin{aligned} y^*_{i-1/2} &= y_{i-1} + \frac{1}{2} h f(x_{i-1}, y_{i-1}) && (5) \\ y_i &= y_{i-1} + h f\left(x_{i-1} + \frac{1}{2} h, y^*_{i-1/2}\right) && (6) \end{aligned}$$

Trapezoidal rule method:

$$\begin{aligned} y^*_i &= y_{i-1} + h f(x_{i-1}, y_{i-1}) && (7) \\ y_i &= y_{i-1} + \frac{h}{2} \left(f(x_{i-1}, y_{i-1}) + f(x_{i-1} + h, y^*_i)\right) && (8) \end{aligned}$$

Each of the 'mid-point rule Runge-Kutta method', (5) and (6), and the 'trapezoidal rule Runge-Kutta method', (7) and (8), is of order $p = 2$ and is thus more accurate than the simple Euler method, as long as h is sufficiently small.

Another approach to approximating the average velocity within the interval $[x_{i-1}, x_i]$ is to note that between the beginning and end of the most recently completed step, from x_{i-2} to x_{i-1}, the velocity had increased from approximately $f(x_{i-1}, y_{i-1})$ to $f(x_{i-2}, y_{i-2})$; that is, $f(x_{i-2}, y_{i-2}) - f(x_{i-1}, y_{i-1})$ per step. This suggests that within the next half step it will increase further by approximately $\frac{1}{2}\left(f(x_{i-2}, y_{i-2}) - f(x_{i-1}, y_{i-1})\right)$. Hence, it might be a reasonable approximation to the value of the average derivative within a step to add this quantity onto the derivative at the beginning: $f(x_{i-1}, y_{i-1})$. Hence the method constructed in this way can be written in the form

$$y_i = y_{i-1} + h\left(\frac{3}{2} f(x_{i-1}, y_{i-1}) - \frac{1}{2} f(x_{i-2}, y_{i-2})\right), \quad i = 2, 3, \ldots . \qquad (9)$$

Of course, in the very first step, from x_0 to $x_1 = x_0 + h$, this method cannot be used. However, once the first step has been completed, it can be used in every later step.

The method given by (9), and known as an Adams-Bashforth method is also of order 2 but it differs from the two Runge-Kutta methods that have been given in two different ways. The first is that the value of the function f is evaluated only once in each step; thus it is less computationally expensive. The second difference is that it is a multistep method; this means that the value computed in a step depends on two previous values; the method is therefore more complicated to use as we have already seen.

The order condition for (9) may be verified by expanding the difference of the two sides in Taylor series. That is,

$$\begin{aligned}
&y(x_i) - y(x_{i-1}) - \frac{3}{2}hy'(x_{i-1}) + \frac{1}{2}hy'(x_{i-2}) \\
&\quad= y(x_{i-1}+h) - y(x_{i-1}) - \frac{3}{2}hy'(x_{i-1}) + \frac{1}{2}hy'(x_{i-1}-h) \\
&\quad= \Big(y(x_{i-1} + hy'(x_{i-1}) + \frac{1}{2}y''(x_{i-1} + O(h^3)\Big) - y(x_{i-1}) \\
&\qquad\quad -\frac{3}{2}hy'(x_{i-1}) + \frac{1}{2}\left(hy'(x_{i-1} + y''(x_{i-1} + O(h^3)\right) \\
&\quad= O(h^3)
\end{aligned}$$

Because the Adams-Bashforth methods give the value of y_i as a linear combination of y_{i-1}, $hf(x_{i-1}, y_{i-1})$ and the values of the same quantities but with other subscripts, they are known as 'linear multistep methods'. If the average derivative within step number i is computed as the mean of $hf(x_{i-1}, y_{i-1})$ and $hf(x_i, y_i)$, we get the second order example of what is known as an 'Adams-Moulton method'. Like all methods in this family, this method is *implicit* (because y_i is given as the solution to an algebraic equation, rather than by an explicit formula).

$$y_i = y_{i-1} + \frac{h}{2}\left(f(x_i, y_i) + f(x_{i-1}, y_{i-1})\right), \quad i = 1, 2, \ldots \tag{10}$$

Another special example of a linear multistep method, also of implicit type is written in the form $y_i = ay_{i-1} + by_{i-2} + chf(x_i, y_i)$. By expanding by Taylor series and matching coefficients, it is found that the unique choice of a, b and c which gives order 2 is $a = \frac{4}{3}$, $b = -\frac{1}{3}$, $c = \frac{2}{3}$. The method that results from this choice:

$$y_i = \frac{4}{3}y_{i-1} - \frac{1}{3}y_{i-2} + \frac{2}{3}hf(x_i, y_i) \tag{11}$$

is known as a BDF or backward difference method.

In the next sections we will explore some aspects of the Runge-Kutta and linear multistep methods in further detail.

3 Runge-Kutta methods

Runge-Kutta methods, as we have seen, are 'one step' in the sense that the result found at the end of a step is functionally dependent only on the result given at the end of the previous step. That is, if y_n denotes a computed approximation to $y(x_n)$, then y_n is given by a formula of the form

$$y_n = y_{n-1} + h\sum_{i=1}^{s} b_i F_i,$$

where the quantities F_1, F_2, …, F_s are derivatives computed from approximations Y_1, Y_2, …, Y_s to the solution at $x_{n-1} + hc_1$, $x_{n-1} + hc_2$, …, $x_{n-1} + hc_s$. That is, $F_i = f(x_{n-1} + hc_i, Y_i)$, $i = 1, 2, \ldots, s$ for the differential equation system (1) or $F_i = f(Y_i)$, $i = 1, 2, \ldots, s$ for the autonomous system (2). The values of Y_i, $i = 1, 2, \ldots, s$ are found from the equation

$$y_i = y_{n-1} + h\sum_{j=1}^{s} a_{ij} F_j, \quad i = 1, 2, \ldots, s.$$

It turns out that the components of the c vector are related to the elements of the A matrix by

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, 2, \ldots, s.$$

The number of stages s is the number of Y vectors needed to compute the solution in a method of this form and is a measure of the complexity of a particular method.

The characteristic coefficients of a specific Runge-Kutta method are conveniently displayed in a tableau as follows

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

Even though we have allowed for the possibility of implicit methods in this formulation, the traditional Runge-Kutta methods, for example those due to Runge, Kutta, Nyström and other early contributors, have been explicit. This means that a_{ij} is precisely zero unless $i > j$ (and consequently $c_1 = 0$) because each quantity used in the computation should be functionally dependent only on other quantities already computed.

The midpoint and trapezoidal rule methods introduced in Section 2 are given by the tableaus

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \qquad\qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Note that here, as is usual for explicit methods, the zero elements on and above the diagonal of A have been omitted.

The next two example tableaus are of orders 3 and 4 respectively

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad\qquad \begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

The fourth order method has become very popular since it was proposed by Kutta and is sometimes referred to as 'the Runge-Kutta method', as though it were the only one available.

To verify the order of these and other Runge-Kutta methods it is necessary to expand the exact and computed solutions in powers of h and to check that the terms up to and including those with an exponent p agree with each other. For example, it can be shown that the exact solution for (2) near an initial point (x_0, y_0) is given by the expansion

$$\begin{aligned} y(x_0 + h) &= y_0 + h\mathbf{f} + \tfrac{h^2}{2}\mathbf{f}'(\mathbf{f}) + \tfrac{h^3}{6}\left(\mathbf{f}''(\mathbf{f},\mathbf{f}) + \mathbf{f}'(\mathbf{f}'(\mathbf{f}))\right) \\ &\quad + \tfrac{h^4}{24}\left(\mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f}) + 3\,\mathbf{f}''(\mathbf{f},\mathbf{f}'(\mathbf{f})) + \mathbf{f}'(\mathbf{f}''(\mathbf{f},\mathbf{f})) + \mathbf{f}'(\mathbf{f}'(\mathbf{f}'(\mathbf{f})))\right) + O(h^5), \end{aligned} \tag{12}$$

where $\mathbf{f} = f(y_0)$, $\mathbf{f}' = f'(y_0)$, $\mathbf{f}'' = f''(y_0)$, $\mathbf{f}''' = f'''(y_0)$. On the other hand, the Taylor expansion for the numerical solution computed by an explicit Runge-Kutta with $s = 4$ is equal to

$$
\begin{aligned}
y_1 \;=\; & y_0 + h(b_1+b_2+b_3+b_4)\,\mathbf{f} + \tfrac{h^2}{2}(b_2c_2+b_3c_3+b_4c_4)\,\mathbf{f}'(\mathbf{f}) \\
& + \tfrac{h^3}{6}\,(2(b_2c_2^2+b_3c_3^2+b_4c_4^2)\,\mathbf{f}''(\mathbf{f},\mathbf{f}) + (b_3a_{32}c_2+b_4a_{42}c_2+b_4a_{43}c_3)\,\mathbf{f}'(\mathbf{f}'(\mathbf{f}))) \\
& + \tfrac{h^4}{24}\Big(6(b_2c_2^3+b_3c_3^3+b_4c_4^3)\,\mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f}) \\
& \qquad + (b_3c_3a_{32}c_2+b_4c_4a_{42}c_2+b_4c_4a_{43}c_3)\,\mathbf{f}''(\mathbf{f},\mathbf{f}'(\mathbf{f})) \\
& \qquad + (b_3a_{32}c_2^2+b_4a_{42}c_2^2+b_4a_{43}c_3^2)\,\mathbf{f}'(\mathbf{f}''(\mathbf{f},\mathbf{f})) \\
& \qquad + b_4a_{43}a_{32}c_2\,\mathbf{f}'(\mathbf{f}'(\mathbf{f}'(\mathbf{f})))\Big) + O(h^5),
\end{aligned}
\tag{13}
$$

A comparison of the terms in (12) and (13) shows that agreement up to h^4 terms occurs if and only if

$$
\begin{array}{rcl@{\qquad}rcl}
b_1+b_2+b_3+b_4 &=& 1, & b_2c_2+b_3c_3+b_4c_4 &=& \frac{1}{2}, \\
b_2c_2^2+b_3c_3^2+b_4c_4^2 &=& \frac{1}{3}, & b_3a_{32}c_2+b_4a_{42}c_2+b_4a_{43}c_3 &=& \frac{1}{6}, \\
b_2c_2^3+b_3c_3^3+b_4c_4^3 &=& \frac{1}{4}, & b_3c_3a_{32}c_2+b_4c_4a_{42}c_2+b_4c_4a_{43}c_3 &=& \frac{1}{8}, \\
b_3a_{32}c_2^2+b_4a_{42}c_2^2+b_4a_{43}c_3^2 &=& \frac{1}{12}, & b_4a_{43}a_{32}c_2 &=& \frac{1}{24}.
\end{array}
$$

These are easily seen to be satisfied by the values $a_{21} = a_{32} = c_2 = c_3 = \frac{1}{2}$, $a_{31} = a_{41} = a_{42} = 0$, $a_{43} = c_4 = 1$, $b_1 = b_4 = \frac{1}{6}$, $b_2 = b_3 = \frac{1}{3}$ to give the classical fourth order method. Explicit methods are known at least as high as order 10 but these require increasingly more stages. For example, for $p = 5$, $s = 6$ is necessary and for $p = 8$, $s = 11$ is necessary.

In contrast to the complicated relationship between the value of p and the minimal value of s to achieve this order for explicit methods, for implicit methods there *is* a simple relationship between s and p. This is that for any positive integer s there exists an implicit Runge-Kutta method with order $p = 2s$ (but no higher). In fact methods with these high orders are a generalization of Gaussian quadrature formulas and reduce to them in the special case of the trivial differential equation $y'(x) = f(x)$.

We give a single example of one of the Gauss family of methods, the method with $s = 2$ and $p = 4$.

$$
\begin{array}{c|cc}
\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\
\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\tag{14}
$$

4 Linear multistep methods

As we saw in Section 2, linear multistep methods make use of already computed solution values and derivatives over several previous steps but only evaluate the function f once in each step. The general form of these methods is thus

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} + \beta_0 f_n + \beta_1 f_{n-1} + \beta_2 f_{n-2} + \cdots \beta_k f_{n-k},$$

where f_i is defined as $f(x_i, y_i)$ for problem (1) and as $f(y_i)$ for problem (2). Note that if $\beta_0 \neq 0$, the method is implicit and the approximation y_n has to be determined as the solution of an algebraic equation. Assuming either $\alpha_k \neq 0$ or $\beta_k \neq 0$, the integer k is a measure of the complexity of these methods and the method is sometimes known as a k-step method.

The order of linear multistep methods is easy to determine by Taylor series analyses. In conditions for various orders, the α and β coefficients all occur linearly and therefore this sort of question is much simpler than for Runge-Kutta methods.

In particular, for order at least 1, the following conditions are necessary and sufficient.

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1, \tag{15}$$

$$\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k = \beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k. \tag{16}$$

These conditions are of central importance to the study of linear multistep methods and are usually known as the 'consistency conditions'. It turns out that for a method to be capable of producing a sequence of approximations which converge to the exact solution as $h \to 0$, it is necessary and sufficient that the method be both consistent and 'stable' where a stable method is one for which the polynomial

$$z^k - \alpha_1 z^{k-1} - \cdots - \alpha_k$$

has zeros only in the closed unit disc and repeated zeros only in the open unit disc.

Amongst explicit methods, the most important are the Adams-Bashforth methods. For these $\alpha_1 = 1$ and each of α_2, ..., α_k is zero. Furthermore, the values of β_1, β_2, $\cdots$, β_k are chosen in such a way as to give an order of $p = k$. Corresponding implicit Adams-Moulton methods have the same values of the αs and, because β_0 is an additional parameter, it is possible to obtain an order $p = k + 1$.

There are good reasons, which we will discuss later for combining an Adams-Bashforth with an Adams-Moulton method of the same order (or possibly with an order greater by 1) into a single algorithm. In these so-called 'predictor-corrector' pairs, the Adams-Bashforth predictor is used to obtain an approximate 'predicted' value of y_n from which an approximate value of f_n is computed. The approximation to t_n is then 'corrected' using the Adams-Moulton formula but with f_n replaced by the value computed in the predictor stage of the algorithm. Many variants of this scheme are used in practical programs.

It might be thought that the loss of generality in assuming the special form for the α coefficients is a disadvantage of the Adams methods. Even though by allowing a more general form for the method so that formally orders as high as $2k$ are actually possible, the stability restrictions generally makes these methods useless. The greatest order for a linear multistep method that is also stable and therefore convergent is $k + 2$ (or only $k + 1$ if k is an odd integer).

Instead of the choice of coefficients made in the Adams methods, it is also possible to use the values $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ with β_0, α_1, α_2, ..., α_k selected to give order $p = k$. These backward difference methods, of which an example is (11), have a special role in the solution of stiff problems.

5 Stiff problems

For many important problems, a phenomenon arises which makes numerical computation complicated and difficult. To illustrate this effect, we will confine ourselves to a simple linear problem but the difficulty is by no means confined to such problems. Suppose the $N \times N$ matrix M has distinct eigenvalues λ_1, λ_2, ..., λ_N and consider the differential equation

$$y'(x) = My(x).$$

It is well-known that the solution to this equation takes the form

$$y(x) = \sum_{i=1}^{N} C_i \exp(\lambda_i x),$$

where the constant vectors C_1, C_2, ..., C_N depend on the information given in the initial value.

The special difficulty, known as 'stiffness' arises when some of the eigenvalues are close to zero, or when some of them have a real part close to zero, whereas other eigenvalues have very negative real parts. Because of the characteristic behaviour of the exponential function, it is the very negative eigenvalues that cause trouble. Even though they ultimately have a negligible effect on the exact solution, they can have an overwhelming effect on the result of a numerical computation using traditional methods. For example, if a single step is taken using the Euler method, the result computed is found to be

$$y_1 = y_0 + hMy_0 = (I + hM)y_0.$$

Suppose the nonsingular matrix V is such that $V^{-1}MV = \text{diag}(q_1, q_2, \ldots, q_N)$, then

$$y_1 = V^{-1}\text{diag}(1 + hq_1, 1 + hq_2, \ldots, 1 + hq_N)Vy_0.$$

Repeated use of this one-step operation will be unstable unless each of the (possibly complex) numbers $1 + hq_1$, $1 + hq_2$, ..., $1 + hq_N$ has a magnitude not exceeding 1. This can only be achieved by making h so small that the 'non-stiff' components take many steps to make progress in modelling the physically interesting behaviour of the problem.

Problems that have this stiffness property arise in many common situations. They are particularly common in applications of the method of lines in the solution of many partial differential equations. Linear and non-linear stiff problems arise also, for example, in the analysis of electrical circuits and in chemical kinetics.

6 Stability questions

In the example of a stiff problem discussed in the previous section, the behaviour of a single component of the solution was identified as being of significance. Hence, we consider a problem in only one dimension of the simple linear form

$$y' = qy. \tag{17}$$

Since in the analysis we will undertake, the factor q always arises with h as a multiplicative factor, it is convenient to define $z = hq$ and, because we will be modelling just one component from a possibly larger system, we will allow z to be complex. For the Euler method we saw in the last section that a single step applied to (17) results in the numerical approximation being multiplied by $1 + z$. The situations that interest us are when h is determined by criteria that have nothing to do with q; we can then think of h as being large in magnitude. On the other hand, because q corresponds to a 'stiff' component, we may think of q as being large in magnitude and having a negative real part. What we want is for the stiff components to simply die away, or at very least to remain bounded; in this way they will not affect the significance of the dominant part of the solution of a large stiff system. Hence, we would like $1 + z$ to have magnitude less than 1 but this is impossible under the conditions we have specified.

A contrasting method defines y_n by the implicit equation

$$y_n = y_{n-1} + hf(x_n, y_n).$$

If we carry out a similar analyis for this 'implicit Euler method', it is found that the computed solution is multiplied in each step by $(1-z)^{-1}$. If q, and therefore z is a complex number with negative real part then the magnitude of this factor is less than 1. This means that the stiff components in a large stiff system solved by this implicit method have a desirable property not possessed by the traditional Euler method. This property is known as A-stability and refers to the boundedness of any sequence of approximations computed by a method when the problem being solved is (17) and z is in the left-half complex plane.

Unfortunately, linear multistep methods cannot possess this property unless, like the implicit Euler method, the implicit mid-point rule, the implicit trapezoidal rule and the second order backward difference method (11), their orders do not exceed 2.

For Runge-Kutta methods, A-stability is available for any required order. For example the Gauss methods, such as the fourth order method given by (14), are all A-stable.

7 Implementation considerations

Even though we have introduced several numerical methods and classes of numerical methods in a constant stepsize setting, it is usually advisable to vary the stepsize in an actual computation. The reason for this is that, for many problems, the behaviour of the solution varies in such a way that the contribution to the overall inaccuracy of the computed solution is much greater from some parts of the trajectory than from others. From the parts where the contribution will be small, large stepsizes are appropriate because the greater local truncation errors will ultimately do little harm. On the other hand, for those parts of the trajectory on which the final errors are most dependent, small stepsizes must be used.

Using a calculus of variations argument it can be concluded that to achieve a suitable balance in the errors arising from different parts of the trajectory, it is best to make the local truncation error approximately the same for all steps. Hence, one of the central implementation questions will be how best to estimate the local truncation error.

In the case of predictor-corrector methods a very simple device makes this task very easy to accomplish. Consider for example the Adams-Bashforth (9), Adams-Moulton (10) pair written together in the form

$$\begin{aligned} y_i^* &= y_{i-1} + h\left(\frac{3}{2}f(x_{i-1}, y_{i-1}) - \frac{1}{2}f(x_{i-2}, y_{i-2})\right), & (18) \\ y_i &= y_{i-1} + h\left(\frac{1}{2}f(x_i, y_i^*) + \frac{1}{2}f(x_{i-1}, y_{i-1})\right). & (19) \end{aligned}$$

Note that the predicted value y_i^* found in (18) is used only to permit the evaluation of an approximate value of the derivative at the end of the current step for use in the corrector (19).

If it is assumed that all previously computed values are exactly correct, so as to restrict consideration to the new errors introduced in step number i, then by Taylor series we easily find that

$$\begin{aligned} y(x_i) - y_i^* &= \frac{5}{12}h^3 y'''(x_i) + O(h^4), & (20) \\ y(x_i) - y_i &= -\frac{1}{12}h^3 y'''(x_i) + O(h^4). & (21) \end{aligned}$$

To obtain an approximation to the local truncation error the difference $y_i^* - y_i$ can be calculated and divided by 6. This is easily verified to be an asymptotically correct estimate from (20) and (21).

A similar approximation to the local truncation error is available for any Adams-Bashforth, Adams-Moulton pair of the same orders.

For explicit Runge-Kutta methods the computation of local truncation errors is much more difficult and it is usual to add additional stages to make it possible to obtain two embedded methods with comparable orders; the difference of the results computed from these gives useful guidance on the errors, but not the asymptotically correct errors as provided by predictor corrector pairs.

Related to error estimation is the question of actually adjusting stepsizes upwards or downwards in accordance with the requirement of keeping local errors approximately constant from step to step. For Runge-Kutta methods this is trivial, because only one value is passed from one step to the next. However, for linear multistep methods there is no single accepted answer. We will discuss only one of these, the 'Nordsieck technique'; this approach has become as popular as any other. In the Nordsieck method the information is carried from step to step, not in its natural form as single y value and a set of k f values, but as linear combinations of these quantities which approximate the sequence of scaled derivatives

$$y(x_i), \quad hy'(x_i), \quad \frac{h^2}{2!}y''(x_i), \quad \dots \quad , \frac{h^k}{k!}y^{(k)}(x_i).$$

For example, in the case of the Adams-Bashforth, Adams-Moulton pair given by (18), (19), the approximations to $y(x_i)$, $hy'(x_i - 1)$ and $\frac{1}{2}h^2y''(x_{i-1})$ are, respectively, y_i, $f(x_{i-1}, y_{i-1})$ and $\frac{1}{2}(f(x_{i-1}, y_{i-1}) - f(x_{i-2}, y_{i-2}))$. If, before step number i is taken, the step size is to be changed from h to rh, then the three incoming approximations can be scaled respectively by 1, r and r^2. If r is always equal to 1, then the reformulation in terms of scaled derivatives makes no difference to the computed result but the variable stepsize generalization is also possible and is simple and convenient only in the scaled derivative formulation.

Implementation difficulties for implicit methods applied to stiff problems centre round the solution of the algebraic system defining the stage values, or the final step result in the case of a linear multistep method. The algebraic system is usually solved by a variant of the Newton-Raphson method. The cost of this solution increases sharply with the dimension of the differential equation system to be solved and, equally significant, it rises sharply with the number of stages. Avoiding the s factors by using backward difference methods (for which $s = 1$) limits the order to 2 or implies that the advantages of A-stabilty are sacrificed.

In the case of Runge-Kutta methods, the cost can be lowered considerably by selecting methods with special structures. The most important of these special structures is for the coefficient matrix A to have a one-point spectrum (the resulting methods are said to be 'singly implicit').

8 Some special applications

As we have mentioned, one of the most commonly-occurring problems leading to stiff problems, is the the set of ordinary differential equations formed by applying the method of lines to partial differential equations. Some other typical stiff problems are associated with the kinetics of chemical reactions in situations in which there is wide variation amongst the reaction rates.

Many physical problems are more appropriately modelled using not differential equations alone, but differential equations combined with algebraic constraints. In some cases these can be viewed as limiting cases of singular perturbation problems but they are often so-called ‘differential-algebraic equations’ in their own right. Many of the methods associated with stiff problems can be adapted to these more difficult problems.

Another type of physical problem for which differential equations alone are not an adequate means of description, is where the rate of change of some or all of the dependent variables needs to be written in terms of the variable values at one or more times in the past, as well as at the present time. These ‘delay-differential equations’ are often non-stiff and the complications arise from the need to interpolate already computed solutions to evaluate approximations to the delay values. The solution of these problems by explicit Runge-Kutta methods imposes additional design demands on the methods because accurate interpolation formulae must be provided; these are not readily available unless additional stages are added to the method. Another characteristic difficulty with delay problems is the occurrence of discontinuous behaviour in the solution. Special techniques are needed to locate and to pass through the discontinuities without destroying too greatly the efficient numerical performance. Delay differential equations arise, for example, in economic and biological modelling.

9 Concluding remarks

In this short survey of numerical methods for ordinary differential equations it has been possible to mention only selected aspects of this very active research area. We have not mentioned methods that make use of higher derivatives of the solution, for example Taylor series methods. We have also not mentioned methods that are both multistage (as in Runge-Kutta methods) and multivalue (as in linear multistep methods); these have many of the desirable properties of each of the main traditional classes which they generalize. It has also not been possible to discuss recent exciting work on the solution of problems arising from a Hamiltonian formulation of mechanical problems. It is important to identify numerical methods with a ‘symplectic’ character because these are capable of accurately preserving theoretical invariants that, in general, cannot be held constant with other numerical methods.

It has not been convenient to give detailed references to all the work that has been discussed in the paper. However, a number of references are listed. Some of these are textbooks and reference books on the subject and some are historically important papers from the literature of this subject.

The work by Cauchy [6], Coriolis [7] and Euler [12] is fundamental both to the theory of differential equations and to their numerical solution. The papers by Runge [24], Heun [18], Kutta [19] and Nyström [23] contain the early work on Runge-Kutta methods while the papers by Gill [14], Merson [21] and Butcher [3] represent the modern phase of the theory of these methods. Some papers on some of the aspects of implicit Runge-Kutta methods that we have discussed here are those by Butcher [4], Ehle [11] and Burrage, Butcher and Chipman [2].

Some of the fundamental publications on linear multistep methods are Adams and Bashforth [1], Moulton [22], Dahlquist [9] and Curtiss and Hirschfelder [8]. In addition to [8], fundamental reading on stiffness and A-stability includes a further paper by Dahlquist [10] and the paper by Wanner, Hairer and Nørsett, [26], where the idea of ‘order stars’ was first introduced.

Some of the early and successful techniques used in the implementation of linear multistep methods, particularly involving the Nordsieck approach, are presented in the book by Gear [13]. Many other monographs and textbooks on the subject of this survey have been produced within the last 35 years. Amongst them are those by Henrici [17], Lambert [20], Stetter [25], Butcher [5] and the two-volume set by Hairer, Nørsett and Wanner [15] and by Hairer and Wanner [16].

References

[1] J.C. Adams, *Appendix* in F. Bashforth, An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid. With an explanation of the method of integration employed in constructing the tables which give the theoretical form of such drops, by J.C.Adams. *Cambridge Univ. Press.* (1883).

[2] K. Burrage, J. C. Butcher and F. H. Chipman, An implementation of singly-implicit methods, *BIT*, **20** (1980), 326-340.

[3] J. C. Butcher, Coefficients for the study of Runge-Kutta integration processes, *J. Austral. math. Soc.*, **3** (1963), 185-201.

[4] J. C. Butcher, Implicit Runge-Kutta Processes, *Math. Comput.,* **18** (1964), 50-64.

[5] J. C. Butcher, The Numerical Analysis of Ordinary Differential Equations, *John Wiley & Sons* (1986).

[6] A.L. Cauchy, Résumé des Leçons données à l'Ecole Royale Polytechnique. Suite du Calcul Infinitésimal; published: *Equations différentielles ordinaires,* ed. Chr. Gilain, Johnson 1981.

[7] G. Coriolis, Mémoire sur le degré d'approximation qu'on obtient pour les valeurs numériques d'une variable qui satisfait à une équation différentielle, en employant pour calculer ces valeurs diverses équations aux différences plus ou moins approchées, *J. de Mathématiques pures et appliquées (Liouville),* **2** (1837), 229-244.

[8] C.F. Curtiss & J.O. Hirschfelder, Integration of stiff equations. *Proc. Nat. Acad. Sci.,* **38** (1952), 235-243.

[9] G. Dahlquist, Convergence and stability in the numerical solution of ordinary differential equations, *Math. Scand.* **4** (1956), 33-53.

[10] G. Dahlquist, A special stability problem for linear multistep methods, *BIT* **3** (1963), 27-43.

[11] B.L. Ehle, On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems, *Research Report CSRR 2010*, Dept. AACS, Univ. of Waterloo, Ontario, Canada (1969).

[12] L. Euler, Institutionum Calculi Integralis. Volumen Primum, (1768), *Opera Omnia*, Vol.XI.

[13] C. W. Gear, Numerical initial value problems in ordinary differential equations, *Prentice-Hall* (1971).

[14] S. Gill, A process for the step-by-step integration of differential equations in an automatic digital computing machine. *Proc. Cambridge Philos. Soc.,* **47** (1951), 95-108.

[15] E. Hairer, S.P. Nørsett & G. Wanner, Solving ordinary differential equations I. Nonstiff problems, *Springer* (1987, Second edition 1993).

[16] E. Hairer & G. Wanner, Solving ordinary differential equations II. Stiff and differential-algebraic problems, *Springer* (1991, Second edition 1996).

[17] P. Henrici, Discrete variable methods in ordinary differential equations. *John Wiley & Sons* (1962).

[18] K. Heun, Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Zeitschr. für Math. u. Phys.,* **45** (1900), 23-38.

[19] W. Kutta, Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. *Zeitschr. für Math. u. Phys.,* **46** (1901), 435-453.

[20] J. D. Lambert, Numerical methods for ordinary differential equations, *John Wiley & Sons* (1991).

[21] R.H. Merson, An operational method for the study of integration processes. *Proc. Symp. Data Processing,* Weapons Research Establishment, Salisbury, Australia, (1957) 110-1 to 110-25.

[22] F.R. Moulton, New methods in exterior ballistics. *Univ. Chicago Press,* (1926).

[23] E.J. Nyström, Ueber die numerische Integration von Differentialgleichungen. *Acta Soc. Sci. Fenn.,* **50** (1925), p.1-54.

[24] C. Runge, Ueber die numerische Auflösung von Differentialgleichungen. *Math. Ann.,* **46** (1895), 167-178.

[25] H. J. Stetter, Analysis of discretization methods for ordinary differential equations, *Springer* (1973).

[26] G. Wanner, E. Hairer & S.P. Nørsett, Order stars and stability theorems, *BIT* **18** (1978), 475-489.