This error may be further reduced by choosing a special value for k/h^2 because the equation for $T_{i,j}$ can be written as

$$T_{i,j} = \tfrac{1}{12}h^2\left(6\frac{k}{h^2}\frac{\partial^2 U}{\partial t^2} - \frac{\partial^4 U}{\partial x^4}\right)_{i,j} + O(k^2) + O(h^4).$$

By the differential equation,

$$\frac{\partial}{\partial t} = \frac{\partial^2}{\partial x^2},$$

so

$$\frac{\partial}{\partial t}\left(\frac{\partial U}{\partial t}\right) = \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 U}{\partial x^2}\right),$$

assuming that these derivatives exist. If we put $6k/h^2 = 1$, the expression in the brackets is then zero and $T_{i,j}$ is $O(k^2) + O(h^4)$. This is of little use in practice because $k = \frac{1}{6}h^2$ is very small for small h so the volume of arithmetic needed to advance the solution to a large time-level is substantial.

Consistency or compatibility

It is sometimes possible to approximate a parabolic or hyperbolic equation by a finite-difference scheme that is stable, (i.e. limits the amplification of all the components of the initial conditions), but which has a solution that converges to the solution of a different differential equation as the mesh lengths tend to zero. Such a difference scheme is said to be *inconsistent* or *incompatible* with the partial differential equation and an example is given in Worked Example 2.7.

The real importance of the concept of consistency lies in a theorem by Lax (reference 25), which states that if a linear finite-difference equation is consistent with a properly posed linear initial-value problem then stability guarantees convergence of u to U as the mesh lengths tend to zero. Consistency can be defined in either of two equivalent but slightly different ways.

The more general definition is as follows. Let $L(U) = 0$ represent the partial differential equation in the independent variables x and t, with exact solution U.

Let $F(u) = 0$ represent the approximating finite-difference equation with exact solution u.

Let v be a continuous function of x and t with a sufficient number of continuous derivatives to enable $L(v)$ to be evaluated at the point (ih, jk).

Then the truncation error $T_{i,j}(v)$ at the point (ih, jk) is defined by

$$T_{i,j}(v) = F_{i,j}(v) - L(v_{i,j}).$$

If $T_{i,j}(v) \to 0$ as $h \to 0$, $k \to 0$, the difference equation is said to be consistent or compatible with the partial differential equation. With this definition $T_{i,i}$ gives an indication of the error resulting from the replacement of $L(v_{i,j})$ by $F_{i,j}(v)$.

Most authors put $v = U$ because $L(U) = 0$. It then follows that

$$T_{i,j}(U) = F_{i,j}(U),$$

and the truncation error coincides with the local truncation error. The difference equation is then consistent if the limiting value of the local truncation error is zero as $h \to 0$, $k \to 0$. This is the definition that will be adopted in this book. By the Worked Example 2.6 it follows that the classical explicit approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$ is consistent with the differential equation.

Example 2.7

The equation

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$$

is approximated at the point (ih, jk) by the difference equation

$$\frac{u_{i,j+1} - u_{i,j-1}}{2k} - \frac{u_{i+1,j} - 2\{\theta u_{i,j+1} + (1-\theta)u_{i,j-1}\} + u_{i-1,j}}{h^2} = 0.$$

Show that the local truncation error at this point is

$$\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4} + (2\theta - 1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2} + O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Discuss the consistency of this scheme with the partial differential equation when:

$$\text{(i) } k = rh \quad \text{and} \quad \text{(ii) } k = rh^2,$$

where r is a positive constant and θ a variable parameter.

Expansion of the terms $U_{i,j+1}$, $U_{i,j-1}$, $U_{i+1,j}$, and $U_{i-1,j}$ about the point (ih, jk) by Taylor's series, as in Example 2.6, and substitution into

$$T_{i,j}=\frac{U_{i,j+1}-U_{i,j-1}}{2k}-\frac{U_{i+1,j}-2\{\theta U_{i,j+1}+(1-\theta)U_{i,j-1}\}+U_{i-1,j}}{h^2}$$

leads to

$$T_{i,j}=\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{i,j}+\left\{\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3}-\frac{h^2}{12}\frac{\partial^4 U}{\partial x^4}\right.$$
$$\left.+(2\theta-1)\frac{2k}{h^2}\frac{\partial U}{\partial t}+\frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2}\right\}+O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Hence the result since $\partial U/\partial t-\partial^2 U/\partial x^2=0$.

Case (i) $k=rh$

As $h\to 0$,

$$T_{i,j}=F_{i,j}(U)\to\left\{\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+(2\theta-1)\frac{2r}{h}\frac{\partial U}{\partial t}+r^2\frac{\partial^2 U}{\partial t^2}\right\}_{i,j}.$$

When $\theta\neq\frac{1}{2}$ the third term tends to infinity. When $\theta=\frac{1}{2}$ the limiting value of $T_{i,j}$ is

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+r^2\frac{\partial^2 U}{\partial t^2}.$$

In this case the finite-difference equation is consistent with the *hyperbolic* equation

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+r^2\frac{\partial^2 U}{\partial t^2}=0.$$

Hence the difference equation is always inconsistent with $\partial U/\partial t-\partial^2 U/\partial x^2=0$ when $k=rh$.

Case (ii) $k=rh^2$

As $h\to 0$,

$$T_{i,j}\to\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+2(2\theta-1)r\frac{\partial U}{\partial t}.$$

When $\theta\neq\frac{1}{2}$ the difference scheme is consistent with the parabolic

equation

$$\{1+2(2\theta-1)r\}\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0.$$

It is only when $\theta=\frac{1}{2}$ that the difference scheme is consistent with the given differential equation. This is then the well-known Du Fort and Frankel three-level explicit scheme which is also stable for all $r>0$. (See Worked Example 3.2). It was devised to overcome the unconditional instability of the early Richardson explicit scheme

$$\frac{u_{i,j+1}-u_{i,j-1}}{2k}-\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}=0,$$

but to retain the advantage of the central-difference approximation to the time-derivative which gives a local truncation error of $O(k^2)+O(h^2)$ as opposed to $O(k)+O(h^2)$ for the classical explicit approximation to $\partial U/\partial t-\partial^2 U/\partial x^2=0$.

Convergence and stability

The following sections are concerned with the conditions that must be satisfied if the solution of the finite-difference equations is to be a reasonably accurate approximation to the solution of the corresponding parabolic or hyperbolic partial differential equation.

These conditions are associated with two different but interrelated problems. The first concerns the convergence of the exact solution of the approximating difference equations to the solution of the differential equation; the second concerns the unbounded growth, or controlled decay or boundedness of the exact solution of the finite-difference equations, and therefore of all rounding errors introduced during the computation because the errors and exact solution are processed by the same arithmetic operations. (The stability problem.)

Descriptive treatment of convergence

Let U represent the *exact* solution of a partial differential equation with independent variables x and t, and u the *exact* solution of the difference equations used to approximate the partial

differential equation. Then the finite-difference equation is said to be convergent when u tends to U at a fixed point or along a fixed t-level as δx and δt both tend to zero.

Although the conditions under which u converges to U have been established for linear elliptic, parabolic and hyperbolic second-order partial differential equations with solutions satisfying fairly general boundary and initial conditions, they are not yet known for non-linear equations except in a few particular cases. (The equation,

$$a\frac{\partial^2 U}{\partial x^2}+b\frac{\partial^2 U}{\partial x\partial t}+c\frac{\partial^2 U}{\partial t^2}+d\frac{\partial U}{\partial x}+e\frac{\partial U}{\partial t}+fU+g=0,$$

is linear when the coefficients $a, b, \ldots, g$, are constants or functions of x and t only. Otherwise it is non-linear. If the coefficients of the second-order derivatives are functions of x, t, U, $\partial U/\partial x$ and $\partial U/\partial t$ but not of second-order derivatives the equation is described as quasi-linear even though it is non-linear. The important feature of linear equations is that the sum of separate solutions is also a solution.)

The difference $(U-u)$ is called the *discretization error*. Some texts call it the truncation error but in this book the latter term will be reserved for the difference between the differential equation and its approximating difference equation. The magnitude of the discretization error at any mesh point depends on the finite-sizes of the mesh lengths, δx and δt, i.e. on the distances between consecutive, discrete grid-points, and on the number of terms in the truncated series of differences used to approximate the derivatives. Readers familiar with the calculus of finite-differences will have recognized the approximation used earlier for $\partial U/\partial t$ as the first term in either the series

$$(\delta t)\left(\frac{\partial U}{\partial t}\right)_{i,j}=(\Delta_t-\tfrac{1}{2}\Delta_t^2+\tfrac{1}{3}\Delta_t^3-\ldots)U_{i,j}$$

or the series

$$(\delta t)\left(\frac{\partial U}{\partial t}\right)_{i,j+\frac{1}{2}}=(\delta_t-\tfrac{1}{24}\delta_t^3+\tfrac{3}{640}\delta_t^5+\ldots)U_{i,j+\frac{1}{2}},$$

and the approximation for $\partial^2 U/\partial x^2$ as the first term in the series

$$(\delta x)^2\frac{\partial^2 U_{i,j}}{\partial x^2}=(\delta_x^2-\tfrac{1}{12}\delta_x^4+\tfrac{1}{90}\delta_x^6-\ldots)U_{i,j},$$

where the subscripts t and x denote the directions in which the differences are calculated. The symbols Δ and δ are the forward and central difference operators defined by $\Delta_t u_{i,j} = u_{i,j+1} - u_{i,j}$, so that $\delta_t u_{i,j+\frac{1}{2}} = u_{i,j+1} - u_{i,j}$, and $\delta_x^2 u_{i,j} = \delta_x(\delta_x u_{i,j}) = \delta_x(u_{i+\frac{1}{2},i} - u_{i-\frac{1}{2},j}) = u_{i+1,j} - 2u_{i,j} + u_{i-1,j}$. Better approximations can be obtained by truncating the series after two or more terms but have the disadvantage of involving more pivotal values of u. It will be shown later that the discretization error can be analysed in terms of preceding local truncation errors. (See p. 73.)

The discretization error can usually be diminished by decreasing δx and δt, subject invariably to some relationship between them, but as this leads to an increase in the number of equations to be solved, this method of improvement is limited by such factors as cost of computation and computer storage requirements, etc.

In general, the problem of convergence is a difficult one to investigate usefully because the final expression for the discretization error is usually in terms of unknown derivatives for which no bounds can be estimated. Fortunately, however, the convergence of difference equations approximating *linear* parabolic and hyperbolic differential equations can be investigated in terms of stability and consistency, which are easier to deal with. (Reference Lax's equivalence theorem, p. 72).

Analytical treatment of convergence (*A direct method*)

The convergence of the solution of an approximating set of *linear*-difference equations to the solution of a *linear* partial differential equation is dealt with most easily via Lax's equivalence theorem. Explicit difference schemes however, can be investigated directly by deriving a difference equation for the discretization error e. Denote the exact solution of the partial differential equation by U and the exact solution of the finite-difference equation by u. Then $e = U - u$.

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1, t > 0, \tag{2.26}$$

where U is known for $0 \leq x \leq 1$ when $t = 0$, and at $x = 0$ and 1 when $t > 0$.

The simplest explicit finite-difference approximation to (2.26)

is
$$\frac{u_{i,j+1}-u_{i,j}}{k}=\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}. \tag{2.27}$$

At the mesh points,
$$u_{i,j}=U_{i,j}-e_{i,j},\quad u_{i,j+1}=U_{i,j+1}-e_{i,j+1},\text{ etc.}$$

Substitution into (2.27) leads to
$$e_{i,j+1}=re_{i-1,j}+(1-2r)e_{i,j}+re_{i+1,j}+U_{i,j+1}-U_{i,j}+r(2U_{i,j}-U_{i-1,j}-U_{i+1,j}). \tag{2.28}$$

By Taylor's theorem,
$$U_{i+1,j}=U(x_i+h,t_j)=U_{i,j}+h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\frac{h^2}{2!}\frac{\partial^2 U}{\partial x^2}(x_i+\theta_1 h,t_j),$$
$$U_{i-1,j}=U(x_i-h,t_j)=U_{i,j}-h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\frac{h^2}{2!}\frac{\partial^2 U}{\partial x^2}(x_i-\theta_2 h,t_j),$$
$$U_{i,j+1}=U(x_i,t_j+k)=U_{i,j}+k\frac{\partial U}{\partial t}(x_i,t_j+\theta_3 k),$$

where $0<\theta_1<1$, $0<\theta_2<1$ and $0<\theta_3<1$. Substitution into eqn (2.28) gives
$$e_{i,j+1}=re_{i-1,j}+(1-2r)e_{i,j}+re_{i+1,j}+k\left\{\frac{\partial U}{\partial t}(x_i,t_j+\theta_3 k)-\frac{\partial^2 U}{\partial x^2}(x_i+\theta_4 h,t_j)\right\}, \tag{2.29}$$

where $-1<\theta_4<1$.

This is a difference equation for $e_{i,j}$ which fortunately we need not solve.

Let E_j denote the maximum value of $|e_{i,j}|$ along the jth time-row and M the maximum modulus of the expression in the braces for all i and j. When $r\leqslant\frac{1}{2}$, all the coefficients of e in eqn (2.29) are positive or zero, so
$$\begin{aligned}|e_{i,j+1}|&\leqslant r|e_{i-1,j}|+(1-2r)|e_{i,j}|+r|e_{i+1,j}|+kM\\&\leqslant rE_j+(1-2r)E_j+rE_j+kM\\&=E_j+kM.\end{aligned}$$

As this is true for all values of i it is true for $\max|e_{i,j+1}|$. Hence
$$E_{j+1}\leqslant E_j+kM\leqslant(E_{j-1}+kM)+kM=E_{j-1}+2kM,$$

etc., from which it follows that

$$E_j \leqslant E_0 + jkM = tM,$$

because the initial values for u and U are the same, i.e. $E_0 = 0$. When h tends to zero, $k = rh^2$ also tends to zero and M tends to

$$\left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j}.$$

Since U is a solution of eqn (2.26) the limiting value of M and therefore of E_j is zero. As $|U_{i,j} - u_{i,j}| \leqslant E_j$, this proves that u converges to U as h tends to zero when $r \leqslant \frac{1}{2}$ and t is finite.

When $r > \frac{1}{2}$ it can be shown that the complementary function of the difference equation (2.29) tends to infinity as h tends to zero. There is no need, however, to do this when our main purpose is to find the conditions necessary for a useful numerical solution because we shall prove later that this finite-difference scheme is stable for $r \leqslant \frac{1}{2}$ but unstable for $r > \frac{1}{2}$.

The proof above implies that $\partial U/\partial t$ and $\partial^2 U/\partial x^2$ are uniformly continuous and bounded throughout the solution domain. This was so in the Worked Example 2.1 in which $\partial^2 U/\partial x^2$ was initially zero in spite of the discontinuity in $\partial U/\partial x$. If it is assumed that U possesses continuous bounded derivatives up to order three in t and order six in x, Exercise 13 shows that the discretization error is of order h^2, except when $r = \frac{1}{6}$ in which case it is of order h^4.

Descriptive treatment of stability

The equations that are actually solved are, of course, the finite-difference equations, and their application and solution to successive time-rows advances the finite-difference solution from the initial line, on which initial values are known, to time-levels k, $2k, \ldots Jk = T$, say, where T is finite. If no rounding errors were introduced into this numerical process then the exact solution $u_{i,j}$ of the finite-difference equations would be obtained at each mesh point (i, j), $i = 0(1)N$, $0 < j \leqslant J$.

The essential idea defining stability is that this numerical process, applied exactly, should limit the amplification of all components of the initial conditions.

For linear initial-value boundary-value problems, Lax and

Richtmyer have related stability to convergence via Lax's Equivalence Theorem (p. 72) by defining stability, in effect, in terms of the boundedness of the solution of the finite-difference equations at a fixed time-level T as $k \to 0$, i.e. as $J \to \infty$, it being assumed that $\delta x = h$ is related to k in such a way that $h \to 0$ as $k \to 0$.

Assume that the vector of solution values $\mathbf{u}_{j+1} = [u_{1,j+1}, u_{2,j+1}, \ldots, u_{N-1,j+1}]^T$ of the finite-difference equations at the $(j+1)$th time-level is related to the vector of solution values at the jth time-level by the equation

$$\mathbf{u}_{j+1} = \mathbf{A}\mathbf{u}_j + \mathbf{b}_j,$$

where $\mathbf{b}_j$ is a column vector of known boundary-values and zeros, and matrix $\mathbf{A}$ an $(N-1)\times(N-1)$ matrix of known elements. Then it will be shown that the practical consequence of this definition of stability is that a norm of matrix $\mathbf{A}$ compatible with a norm of $\mathbf{u}$ must satisfy

$$\|\mathbf{A}\| \leqslant 1$$

when the solution of the partial differential equation does not increase as t increases, or

$$\|\mathbf{A}\| \leqslant 1 + O(k)$$

when the solution of the partial differential increases as t increases.

These conditions also ensure the boundedness of all rounding errors because they are subject to the same arithmetic operations as the finite-difference solution.

In an actual computation, however, k and h are normally kept constant as the solution is propagated forward time-level by time-level from $t = 0$ to $t_j = jk$, and in many textbooks and papers stability is defined in terms of the boundedness of this numerical solution as $j \to \infty$, k fixed. In this process the order $(N-1)$ of matrix $\mathbf{A}$ remains constant, unlike the matrix $\mathbf{A}$ associated with Lax and Richtmyer's definition. The matrix method of analysis then shows that the equations are stable if the largest of the moduli of the eigenvalues of matrix $\mathbf{A}$, i.e. the spectral radius $\rho(\mathbf{A})$ of $\mathbf{A}$, satisfies

$$\rho(\mathbf{A}) \leqslant 1,$$

when the solution of the differential equation does not increase with increasing t.

Although this condition ensures the boundedness of the computed solution it does not guarantee convergence unless the eigenvalues of **A** are restricted to satisfy $\rho(\mathbf{A}) \leqslant \|\mathbf{A}\| \leqslant 1$, as $N \to \infty$. In practice, assuming that the difference equations are consistent, it is usually only in the immediate neighbourhood of $\rho(\mathbf{A}) = 1$ that non-convergence might occur. An illuminating discussion of these points is given in reference 19. (If the solution of the partial differential equation increases as $t \to \infty$, the condition for stability with fixed h and k is then $\rho(\mathbf{A}) \leqslant 1 + O(k)$. See p. 66.)

Vector and matrix norms

This section is needed for the Lax–Richtmyer definition of stability.

Vector norms

The norm of vector **x** is a real positive number giving a measure of the 'size' of the vector and is denoted by $\|\mathbf{x}\|$. It must satisfy the following axioms.

(i) $\|\mathbf{x}\| > 0$ if $\mathbf{x} \neq \mathbf{0}$ and $\|\mathbf{x}\| = 0$ if $\mathbf{x} = \mathbf{0}$.

(ii) $\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\|$ for a real or complex scalar c.

(iii) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

If the $n \times 1$ vector **x** has components $x_1, x_2, \ldots, x_n$, then the three most commonly used norms are defined as follows.

The 1-norm of **x** is the sum of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n| = \sum_{i=1}^{n} |x_i|.$$

The infinity norm of **x** is the maximum of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_\infty = \max_i |x_i|.$$

The 2-norm of **x** is the square root of the sum of the squares of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_2 = (|x_1|^2 + |x_2|^2 + \cdots |x_n|^2)^{\frac{1}{2}} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{\frac{1}{2}}.$$

The 2-norm gives the 'length' of the vector.