NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

P. D. LAX, New York University

No one will quarrel with the statement that computing is bound to revolutionize applied mathematics by providing new and powerful tools for getting numerical answers, the life and blood of the subject. The versatility of the new computers has already stimulated the systematic study of numerical methods; this has attracted a number of pure mathematicians who, before, had been repelled by the *ad hoc* nature of the methods used by the hapless applied mathematician.

This brief article is about the numerical solution of partial differential equations. A very general (although by no means the only) method for solving these is to convert them into difference equations through the replacement of derivatives by difference quotients of some sort. The effectiveness of this method is attested to by the vast literature on it which has sprung up during the last 15 years—a literature so extensive that it is hopeless to try to summarize it. Instead of trying to present some of the highlights of the new developments I shall try to convey its flavor. My aim is to convince a skeptical reader who may regard using finite differences as the last resort of a scoundrel that the theory of difference equations is a rather sophisticated affair, more sophisticated than the corresponding theory of partial differential equations. My argument will be based on two contentions:

1) In order to prove that solutions of a sequence of difference equations converge one needs estimates for difference operators which are analogous to the estimates needed in the existence and uniqueness theory for solutions of differential equations.
2) Estimates for difference operators are much harder to derive than the corresponding estimates for differential operators.

These contentions will be illustrated on a very simple linear equation; the reader interested in the general theory should consult R. D. Richtmyer’s book (and the research papers listed at the end, most of which have been incorporated into the second edition of his book).

I regret that there is no room to describe difference approximations to nonlinear equations; this part of the theory has many surprises and many unsolved problems. An intriguing approach to a certain class of such problems is discussed by Frank Harlow in an article appearing in this issue.

1. In problems involving partial differential equations we want to determine a function or set of functions \( u \) which satisfies a partial differential equation

\[
Lu = f
\]

in some domain \( D \) and which satisfies certain conditions on a subset of \( D \). These additional conditions, usually called boundary (or initial) values of \( u \), together with the given function \( f \) constitute the so-called data of \( u \).

According to the modern theory of linear partial differential equations, the data determine a unique solution \( u \) if and only if \( u \) can be estimated in terms of its data, i.e. an inequality of the form

\[
\|u\| \leq \text{constant} \|Lu\|
\]

holds, and a similar inequality holds for the so-called adjoint problem. (For simplicity we have assumed that the boundary data of \( u \) are zero; the symbol \( \| \| \) denotes some norm which measures magnitude.) For linear operators \( L \) it follows from (2) that

\[
\|u_1 - u_2\| \leq \text{constant} \|Lu_1 - Lu_2\|.
\]

This clearly implies that \( u \) is uniquely determined by \( Lu = f \); in fact it says substantially more; it says that \( u \) is a continuous function of \( f \) in the sense of the norm employed. For problems of mathematical physics nothing short of this continuous dependence of solutions on the data will suffice since, as Hadamard pointed out, the data here come from observations which are never exact; thus observations (i.e. approximate data) must suffice to determine approximate solutions. For this reason an operator which satisfies inequality (2) is sometimes called stable.

We turn now to difference approximations. The differential operator \( L \) is replaced by a difference operator, denoted by \( L_\Delta \), where \( \Delta \) stands for the width of the mesh to be employed in the difference scheme. Denote by \( u_\Delta \) the solution of

\[
L_\Delta u_\Delta = f.
\]

Since \( L_\Delta \) approximates \( L \), \( L_\Delta u = f_\Delta \) approximates \( Lu = f \), that is \( \|f - f_\Delta\| \) is small
for small $\Delta$. Therefore, if we knew that $L_\Delta$ satisfies an inequality analogous to (2): 

\[(2_\Delta) \quad \|v\| \leq \text{constant} \|L_\Delta v\|\]

with a constant that does not depend on the mesh width $\Delta$, then by applying $(2_\Delta)$ to $v = u - u_\Delta$, we would obtain 

$$\|u - u_\Delta\| \leq \text{constant} \|f - f_\Delta\|$$

or, in other words, that $u_\Delta$ is a good approximation of $u$ for small $\Delta$.

The converse proposition also holds: if for each $f$ with finite norm the solution $u_\Delta$ of $(1_\Delta)$ tends to the solution $u$ of $(1)$ then according to the principle of uniform boundedness $(2_\Delta)$ holds with a constant independent of $\Delta$.

A difference scheme for which inequalities $(2_\Delta)$ are satisfied uniformly is called stable; we can summarize the result obtained above by saying that a difference scheme is convergent for all data $f$ if and only if it is stable. This bears out the first contention made in the introduction—that in order to prove the convergence of a difference scheme one has to derive the same kind of inequalities for the difference operators as those needed for the differential operator in order to prove the existence and uniqueness of solutions of the differential equation.

The second contention made in the introduction is that the inequalities for the difference operators are harder to prove than the corresponding inequality for the differential operator. We shall illustrate this for the operator

\[(3) \quad L = \partial_t + a \partial_x\]

acting on a scalar function $u$ of two variables $x$ and $t$; the coefficient $a$ is a given function of $x$ and $t$. The domain in question is some slab $0 \leq t \leq T$ and the prescribed initial data are the values of $u$ at $t = 0$,

\[(4) \quad u(x, 0) = g(x)\]

The differential equation

\[(5) \quad Lu = u_t + au_x = f\]

states that the directional derivative of $u$ in the direction

\[
\frac{dx}{dt} = a
\]

is equal to $f$. Given the initial values $(4)$ and the function $f$, our task is to determine $u$ uniquely. To find $u$, we have to solve the ordinary differential equation $(6)$ and then integrate $f$ along the trajectories of $(6)$. But even without performing any integrations, we can give a priori estimates of $u$ in terms of its data, $f$ and $g$. For simplicity we take the case when $f$ is zero; then the differential equation $(5)$ asserts that $u$ is constant along the trajectories of $(6)$. Since any point in the upper half plane can be connected to some point of the initial line
by a trajectory, it follows that the range of values of $u$ for $t \geq 0$ is the same as the range of values of $u$ at $t = 0$. So

$$\| u \|_{\text{max}} = \| g \|_{\text{max}}.$$  

We shall also derive an \textit{a priori} estimate for $u$ in the $L_2$ norm. We multiply (5) by $u$ and integrate, getting

$$\int u u_t dx + \int a uu_x dx = 0. \quad (8)$$

Now introduce the quantity $E$, sometimes called \textit{energy}, as

$$E(t) = \frac{1}{2} \int u^2 dx;$$

then the first term in (8) can be written as $E_t$. In the second term we write $uu_x$ as $\frac{1}{2}(u^2)_x$ and integrate by parts; assuming that $u$ tends to zero as $|x|$ becomes large we can rewrite (8) as

$$E_t - \frac{1}{2} \int a_x u^2 dx = 0. \quad (9)$$

Denote by $M$ the supremum of $a_x$:

$$a_x \leq M. \quad (10)$$

It follows from (9) that $E_t - ME \leq 0$; multiplying this by $e^{-Mt}$ we obtain $(e^{-Mt}E)_t \leq 0$. This means that $e^{-Mt}E$ is a nonincreasing function of $t$; so

$$E(t) \leq e^{Mt}E(0); \quad (11)$$

this is called an energy inequality.

We shall set up and study three different difference analogues of the differential equation (5). In the first scheme we replace $u_t$ by a forward difference quotient and $u_x$ by a centered difference quotient:

$$u_t \cong [u(x, t + \Delta t) - u(x, t)]/\Delta t,$$

$$u_x \cong [u(x + \Delta x, t) - u(x - \Delta x, t)]/2\Delta x. \quad (12)$$

We introduce the abbreviations $u(h\Delta x, t) = u_k$, $u(k\Delta x, t + \Delta t) = v_k$ and

$$a \frac{\Delta t}{\Delta x} = b. \quad (13)$$

Using the approximations (12) in (5) where we assume that $f = 0$, we obtain the difference equation

$$v_k = \frac{b}{2} u_{k-1} + u_k - \frac{b}{2} u_{k+1}. \quad (I)$$
The second method is very much like the first one except that in the forward difference formula for \( u_t \) we replace \( u(x, t) \) by the average quantity \( [u(x+\Delta x, t) + u(x-\Delta x, t)]/2 \). The resulting difference equation can be written as

\[
v_k = \frac{1 + b}{2} u_{k-1} + \frac{1 - b}{2} u_{k+1}.
\]

The third method is a little more complicated; we start with the truncated Taylor series

\[
u(x, t + \Delta t) \approx u + \Delta t u_t + \frac{(\Delta t)^2}{2} u_{tt}.
\]

Differentiating (5) with respect to \( x \) and \( t \) gives

\[
u_{tt} = -a u_{tt} - a_t u_x, \quad u_{xt} = -a u_{xx} - a_x u_x.
\]

By substituting the expression for \( u_{xt} \) into that for \( u_{tt} \), we deduce that for all solutions of (5) \( u_{tt} = a^2 u_{xx} - \dot{a} u_x \), where \( \dot{a} = a_t - aa_x \). Using this and (5) we can express the \( t \) derivatives in (14) by \( x \) derivatives; we replace these by difference quotients, \( u_x \) as in (12) and \( u_{xx} \) by

\[
u_{xx} \approx \frac{u_{k+1} - 2u_k + u_{k-1}}{(\Delta x)^2}.
\]

The resulting difference equation is

\[
v_k = \frac{b^2 + b}{2} u_{k-1} + \frac{1 - b^2}{2} u_{k+1}.
\]

All three difference equations can be written in the form

\[
v = u(t + \Delta t) = S_{\Delta} u(t),
\]

where \( S_{\Delta} \) is a spatial difference operator

\[
S_{\Delta} u = v_k = \sum c_j u_{k-j}.
\]

If we assume for simplicity that the coefficients of the operator \( S_{\Delta} \) depend on \( x \) but not on \( t \) then the solution of the initial value problem for (15) can be written as

\[
u_{\Delta}(t) = S_{\Delta}^n u(0), \quad n = t/\Delta t.
\]

According to estimates (7) and (11) the solutions \( u(t) \) of (5) at time \( t \) depend boundedly on the initial values in the sense of the maximum norm and the \( L_2 \) norm. We called a difference scheme stable in one of these norms if \( u_{\Delta}(t) \) is uniformly bounded in its dependence on the initial data. In view of the operator expression (17) for \( u_{\Delta}(t) \) we can express this as follows:
The difference scheme (15) is stable if and only if the norms of the operators

\[
S^n_\Delta
\]

are uniformly bounded in the range \( n\Delta t \leq T \).

We shall regard the solutions of the difference equation (15) as being defined on the discrete set of points \( x = k\Delta x \), with \( k \) an integer, and shall use the discrete maximum and \( L_2 \) norms

\[
|u|_{\text{max}} = \text{Sup} |u_k|, \quad \|u\|^2 = \sum |u_k|^2.
\]

**Theorem:** (A) The difference scheme (I) is stable if and only if

\[
\frac{\Delta t}{(\Delta x)^2} \leq \text{const.}
\]

is satisfied.

(B) The difference scheme (II) is stable with respect to both the maximum and \( L_2 \) norms if

\[
(a \frac{\Delta t}{\Delta x}) \leq 1
\]

and is unstable if \( (a(\Delta t/\Delta x)) > 1 + \delta, \delta > 0 \).

(C) The difference scheme (III) is stable or unstable in the \( L_2 \) norm under the same conditions as (II).

In order to evaluate the efficiency of a given numerical method we have to know the number of operations which have to be carried out. For a scheme of type (15), (16) the operations are the calculation of the coefficients \( c_j \) and the indicated multiplications in (16); these have to be carried out \( N \) times, \( N \) being the number of lattice points used in the calculation. \( N \) is proportional to \((\Delta x \Delta t)^{-1}\).

Condition (20) forces us to choose a much smaller value for \( \Delta t \) than condition (21), and therefore the evaluation of the approximation (17) by method (I) needs many more iterations of the operator \( S_\Delta \) than by methods (II) or (III). This makes (I) quite impractical and in fact (I) is never used; of the other two, (III) is much more accurate than (II) and only moderately more cumbersome. Therefore (III) would be used in preference to (II) except when only a very crude answer is required.

We turn now to part A of the theorem; we shall show using a method due to von Neumann that if condition (20) is violated then scheme (I) is unstable in the maximum norm by exhibiting a sequence of initial data \( g_\Delta \) such that \( |g_\Delta|_{\text{max}} \) remains bounded but \( |S^n_\Delta g_\Delta|_{\text{max}}, \Delta t n = 1 \), tends to infinity. Clearly it pays to
choose $g_\Delta$ as an element for which $S_\Delta^a g_\Delta$ is easy to evaluate; such an element is an eigenvector of $S_\Delta$. We consider now the special case when $a$ is constant; then the operator $S_\Delta$ commutes with translation and so its eigenvectors are exponentials. Taking $g_\Delta(k) = e^{i\xi k}$ we have by (I)

$$
(S_\Delta g_\Delta)(k) = \frac{b}{2} e^{i\xi (k-1)} + e^{i\xi k} - \frac{b}{2} e^{i\xi (k+1)} = \lambda g_\Delta(k),
$$

where $\lambda = 1 - ib \sin \xi$. Then

$$
S_\Delta^n g_\Delta = \lambda^n g_\Delta;
$$

clearly $|g_\Delta|_{\max} = 1$ and $|S_\Delta^n g_\Delta|_{\max} = |\lambda|^n$. The largest value of $|\lambda|^2 = 1 + b^2 \sin^2 \xi$ occurs when $\xi = \pi/2$. For this value of $\xi$

$$
|\lambda|^n = (1 + b^2)^{n/2} \leq e^{(b^2)^{n/2}}.
$$

Using the definition (13) of $b$ and $n = 1/\Delta t$ we can write

$$
b^2 n = \frac{a^2 \Delta t}{2(\Delta x)^2} ;
$$

clearly if (20) is violated, $b^2 n$ tends to infinity and so does $|S_\Delta^n g_\Delta|_{\max}$. This proves the negative portion of part A of the theorem; since this discredits scheme (I) we don't bother to prove the positive side but go on to part (B). We shall use the following stability criterion:

*If there exists a constant $M$ such that for all $\Delta$

$$
||S_\Delta|| \leq 1 + M \Delta t
$$

then the scheme is stable.*

The validity of this criterion follows from the chain of inequalities

$$
||S_\Delta^n|| \leq ||S_\Delta||^n \leq (1 + M \Delta t)^n \leq e^{Mn\Delta t} = e^{Mt} .
$$

We turn to scheme (II); here $v_k$ is a linear combination of $u_{k-1}$ and $u_{k+1}$ with weights $(1 \pm b)/2$; the sum of the weights is one and if condition (21) is satisfied the weights are nonnegative. So it follows that

$$
\max |v_k| \leq \max |u_k|,
$$

which is inequality (22) in the maximum norm with $M = 1$.

Next we prove the stability of (II) in the $L_2$ norm; multiplying (II) by $v_k$ we get

$$
v_k = \frac{1 + b}{2} u_{k-1} v_k + \frac{1 - b}{2} u_{k+1} v_k .
$$
Using the inequality

\[ uv \leq \frac{1}{2} (u^2 + v^2) \]

twice on the right we obtain

\[ v_k^2 \leq \frac{1 + b}{4} (u_{k-1}^2 + v_k^2) + \frac{1 - b}{4} (u_{k+1}^2 + v_k^2), \]

which is equivalent to

\[ v_k^2 \leq \frac{1 + b_k}{2} u_{k-1}^2 + \frac{1 - b_k}{2} u_{k+1}^2, \]

where \( b_k \) stands for \( a(k \Delta x)(\Delta t/\Delta x) \). Sum with respect to \( k \); re-indexing on the right we get

\[ \sum v_k^2 \leq \sum \left( 1 + \frac{b_{k+1} - b_{k-1}}{2} \right) u_k. \]

Recalling the definition of \( M \) as \( \sup a \) we see that \( (b_{k+1} - b_{k-1}) \leq 2 M \Delta t \); using this in (25) we get

\[ \|v\|^2 = \sum v_k^2 \leq (1 + M \Delta t) \sum u_k^2 = (1 + M \Delta t) \|u\|^2. \]

This is inequality (22) in the \( L_2 \) norm.

The estimates (23) and (26) derived here are clear-cut analogues of the estimates (7) and (11) for solutions of the differential equation. Similar estimates can be derived by the same arguments for all difference operators of the form (16) where the weights \( c_j \) are nonnegative, add up to one and vary Lipschitz continuously with \( x \). Friedrichs has observed that the \( L_2 \) estimates can be derived also in the much more interesting case when \( u \) is a vector-valued function and the coefficients \( c_j \) in (16) are matrices which are positive in the sense of quadratic forms. The only change necessary in the reasoning is to replace inequality (24) by the Schwarz inequality \( u \cdot cv \leq \frac{1}{2} (u \cdot cu + v \cdot cv) \).

The necessity of condition (21) for convergence has been demonstrated by Courant, Friedrichs and Lewy in their classical paper in the following elegant fashion: The difference scheme (II) expresses the value of \( u \) at \( (x, t + \Delta t) \) in terms of its values at \( (x \pm \Delta x, t) \). When we express \( u(t, x) \) in terms of the initial values, therefore, we need the initial values only at points inside the interval \( (x - (\Delta x/\Delta t)t, x + (\Delta x/\Delta t)t) \). On the other hand the value \( u(x, t) \) of the exact solution equals the value of the initial function at that point \( y \) which can be connected to \( (x, t) \) by a trajectory of (6). If \( \Delta x/\Delta t < a - \delta \) for arbitrarily small \( \Delta t \) then that point \( y \) will fall outside the above interval at least for some \( x \) and \( t \). In this case the value of \( u(x, t) \), and therefore also its limit, does not depend on the value of the initial function at \( y \) and therefore cannot in general be equal
to \(u(x, t)\). This shows the divergence of scheme (II) and also of (III) if (21) is violated.

Lastly we prove the stability of III in the \(L_2\) norm; we treat first the case when \(a\) is constant. We introduce the Fourier transform \(\hat{u}\) of \(\{u_k\}\) as

\[
\hat{u}(\xi) = \sum u_k e^{i\xi k}.
\]

For constant \(b\) the operator \(S_\delta\) defined by III is a convolution; a brief calculation shows that \(\hat{\varphi}(\xi) = \sum \hat{v}_k e^{i\xi k}\) is given by

\[
\hat{\varphi}(\xi) = \lambda(\xi) \hat{a}(\xi),
\]

where \(\lambda(\xi) = 1 - b^2 + b^2 \cos \xi - ib \sin \xi\). Another brief (≈15 minute) calculation gives

\[
|\lambda(\xi)|^2 = 1 - (\cos \xi - 1)^2(b^2 - b^2);
\]

this formula shows that if condition (21) is satisfied then \(|\lambda(\xi)| \leq 1\); therefore it follows from (27) that then \(|\hat{u}(\xi)| \leq |\hat{\varphi}(\xi)|\) and a fortiori

\[
\int |\hat{\varphi}(\xi)|^2 d\xi \leq \int |\hat{u}(\xi)|^2 d\xi.
\]

According to the Parseval identity

\[
|\|u\|^2 = \sum |u_k|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{u}(\xi)|^2 d\xi,
\]

\[
|\|\varphi\|^2 = \sum |v_k|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{\varphi}(\xi)|^2 d\xi.
\]

From these and (29) it follows that \(\|\varphi\|^2 \leq \|u\|^2\), which proves the stability of scheme (III) when \(b\) is constant.

For variable coefficients we prove stability as follows: squaring (III) gives

\[
v_k^2 = \left(\frac{b^2 + b}{2} u_{k-1} + (1 - b^2) u_k + \frac{b^2 - b}{2} u_{k+1}\right)^2.
\]

We add to the right side the quantity

\[
\left(\frac{b^2 - b^4}{4}\right)(u_{k-1} - 2u_k + u_{k+1})^2.
\]

If condition (21) is satisfied (30) is positive, so we get an inequality which, after some algebraic manipulation, can be written in the form

\[
v_k^2 \leq \frac{b^2 + b^3}{2} u_{k-1} + (1 - b^2) u_k + \frac{b^2 - b^3}{2} u_{k+1} + (b^2 - b)(u_{k+1}u_k - u_k u_{k-1}).
\]
The right side is a quadratic form in \( u_{k-1}, u_k, u_{k+1} \); the matrix of this quadratic form is
\[
\begin{bmatrix}
\frac{b^3 + b^3}{2} & \frac{b - b^3}{2} & 0 \\
\frac{b - b^3}{2} & 1 - b^2 & \frac{b^3 - b}{2} \\
\frac{b^3 - b}{2} & \frac{b^2 - b^3}{2} & 0
\end{bmatrix}.
\]

This matrix has the following properties which are crucial for deriving the \( L_2 \) inequality:

The sum of the elements along the diagonal equals one.
The sum of the elements along a subdiagonal is zero.

We sum (31) over all \( k \) and re-index the right-side, obtaining

\[
\sum v_k^2 \leq \sum \left[ \frac{1}{2} [b_{k+1}^3 + b_{k+1}^2 + 2(1 - b_k^2) + b_{k-1}^2 - b_{k-1}^3]u_k^2 + \sum [b_k^3 - b_k - b_{k+1}^3 + b_{k+1}]u_{k+1}u_k \right] (32)
\]

It follows from the properties of the matrix listed above that if \( b \) were independent of \( k \) then the right side of (32) would be just \( \sum u_k^2 \). If \( b \) is variable then, denoting \( \sup |a_x| \) by \( M_1 \) we have, just as before, the inequality \( |b_{k+1} - b_k| \leq M_1 \Delta t \). Using this we deduce that the right side of (31) is less than

\[
(1 + \text{const. } M_1 \Delta t) \sum u_k^2;
\]

this completes the proof of the stability of (III).

The reader may justly ask how the quantity (30) was picked; we give away the trade secret by pointing out that the crucial property of the resulting quadratic form follows by Fourier transformation of the trigonometric identity (28). Identity (28) itself may be regarded as a special instance of the Fejér-Riesz representation for nonnegative trigonometric polynomials; all this is explained fully in [6].

What about the stability of scheme (III) in the maximum norm? Recently Thomée has shown, by a subtle analysis, that (III) is unstable in the maximum norm. The instability is very mild and doesn't interfere with the effectiveness of (III) except possibly for very wild initial data.

We hope that the foregoing bears out our second contention.

References


NUMERICAL FLUID DYNAMICS

F. H. HARLOW, University of California, Los Alamos

1. Computer representation of fluid dynamics. It has been suggested that the numerical solution of a fluid dynamics problem is a good example of "experimental arithmetic." We shall see that this is true in several respects, but first it is necessary to consider how the computer can represent a fluid, and to see how both the instantaneous representation and the calculation of changes involve approximations.

For each instant in the running of a calculation, the computer must contain a complete account of everything required to describe the fluid configuration, together with all data necessary for calculating the changes which currently are occurring. Certainly the computer cannot remember the necessary information for every point within the fluid; infinite storage capability would be required. This, then, introduces the first approximation: The computer memory can have stored in it the values of the field variables (such as velocity, temperature, and density) at only a finite number of points. The greater the number, the more detailed will be the resolution and the more accurate the results.

It is convenient to imagine each of these points within the fluid to be the center of a tiny cell. Then the value of each field variable at the center position