COMMUNICATIONS ON PURE AND APPLIED MATHEMATICS, VOL. IX, 267–293 (1956)

Survey of the Stability of Linear Finite Difference Equations*

P. D. LAX and R. D. RICHTMYER

PART I

AN EQUIVALENCE THEOREM

1. Introduction

Beginning with the discovery by Courant, Friedrichs and Lewy [1] of the conditional stability of certain finite difference approximations to partial differential equations, the subject of stability has been variously discussed in the literature (see bibliography at end). The present paper is concerned with the numerical solution of initial value problems by finite difference methods, generally for a finite time interval, by a sequence of calculations with increasingly finer mesh. Thus if t is the time variable and Δt its increment, we are concerned with limits as $\Delta t \to 0$ for fixed t, not with limits as $t \to \infty$ for fixed Δt (although often the stability considerations are similar). The basic question is whether the solution converges to the true solution of the initial value problem as the mesh is refined. The term *stability*, as usually understood, refers to a property of the finite difference equations, or rather of the above mentioned sequence of finite difference equations with increasingly finer mesh. We shall give a definition of stability in terms of the uniform boundedness of a certain set of operators and then show that under suitable circumstances, for linear initial value problems, stability is necessary and sufficient for convergence in a certain uniform sense for arbitrary initial data. The circumstances are first that a certain consistency condition must be satisfied which essentially insures that the difference equations approximate the differential equations under study, rather than for example some other differential equations, and secondly that the initial value problem be properly posed, in a sense to be defined later.

We shall not be concerned with rounding errors, and in fact assume that all arithmetic steps are carried out with infinite precision. But it will

* The work for this paper was done under Contract AT-(30-1)-1480 of the Atomic Energy Commission.

be evident to the reader that there is an intimate connection between stability and practicality of the equations from the point of view of the growth and amplification of rounding errors. Indeed, O'Brien, Hyman and Kaplan [8] defined stability in terms of the growth of rounding errors. However, we have a slight preference for the definition given below, because it emphasizes that stability still has to be considered, even if rounding errors are negligible, unless, of course, the initial data are chosen with diabolical care so as to be exactly free of those components that would be unduly amplified if they were present.

The basic notions will be spelled out in considerable detail below in an attempt to motivate the definitions given and to justify the approach via the theory of linear operators in Banach space. We shall then give the usual definition of a properly posed initial value problem, define the consistency of a finite difference approximation, define the stability of a sequence of finite difference equations, and prove the equivalence theorem.

2. The Function Space of an Initial-Value Problem

In the solution of an initial-value problem the time variable t plays a special role. An instantaneous state of the physical system is described by one or more functions of certain other variables which we shall call space variables. At any stage of a machine- or hand-calculation one has at hand a numerical representation (e.g. in tabular form) of these functions, that is, of the state of the system at some time t. As time goes on, the state of the system changes according to certain differential or integro-differential equations. It is convenient to think of these functions, for a fixed t, as an element or point in a function space $\mathscr{B}$ and to denote them by a single symbol u.

The initial-value problems under consideration are linear and we suppose $\mathscr{B}$ to be linear also. This may force us to accept as elements of $\mathscr{B}$ some functions not having direct significance as states of a physical system, e.g., functions having negative values for inherently positive quantities like temperature and particle density. But it is convenient to admit such functions as representing *generalized* states of the system, and also to admit complex valued functions. If sums and differences of elements of $\mathscr{B}$ are defined in the obvious manner by sums and differences of the corresponding functions, and if multiplication of an element of $\mathscr{B}$ by a number is defined in the equally obvious manner as multiplication of the corresponding functions by that number, it is clear that $\mathscr{B}$ is a linear vector space.

For a discussion of approximation and errors, one needs a measure of the difference of two states u and v, and it is clear that this measure should have the properties of a norm of the element $w = u - v$; we therefore denote this quantity by $\| w \|$ and suppose that $\mathscr{B}$ is a Banach space. The specific choice of norm may vary from one application to another; in many cases it can be identified with energy. Our assumption that $\mathscr{B}$ is *complete* with respect to the norm plays an important role in the equivalence theorem of Section 8.

3. The Initial Value Problem

Let A denote a linear operator that transforms the element u into the element Au by spatial differentiations, matrix-vector multiplications and the like. The initial value problem is to find a one-parameter set of elements $u(t)$ such that

$$\frac{d}{dt}u(t) = Au(t), \qquad 0 \leqq t \leqq T, \tag{1}$$

$$u(0) = u_0 \tag{2}$$

where u_0 represents a preassigned initial state of the system.

Systems involving higher order derivatives with respect to t can be put into the above form in the usual way by introducing the lower order derivatives as further unknown functions.

All the general considerations in the present discussion apply as well when the operator A depends explicitly on t, and in fact were originally presented in that generality[1], but in the interest of simplicity of the formulas we discuss here only the case of an operator A not depending on the parameter t.

If there are boundary conditions in the problem, it is assumed that they are linear homogeneous and are taken care of by restricting the domain of A to functions satisfying the conditions.

By a ***genuine solution*** of (1) we mean a one-parameter set $u(t)$ such that first, $u(t)$ is in the domain of A for $0 \leqq t \leqq T$ and secondly

$$\text{as } \tau \to 0, \quad \left\| \frac{u(t+\tau)-u(t)}{\tau} - Au(t) \right\| \to 0 \text{ uniformly in } t, \quad 0 \leqq t \leqq T. \tag{3}$$

If we pick an element u_0 not in the domain of A (e.g., if A is a differential operator and the functions represented by u_0 are nondifferentiable at one or more points), we obviously cannot find a genuine solution satisfying (2), but we assume that u_0 can always be approximated, as closely as one desires, by an element u_0, for which a unique genuine solution exists. That

[1] P. D. Lax, Seminar, New York University, January 1954.

is, if we define an operator $E_0(t)$ — really a one-parameter family of operators — so that

$$u(t) = E_0(t)u(0), \qquad 0 \leqq t \leqq T,$$

for any genuine solution of (1) depending uniquely on $u(0)$, we assume that *the domain of $E_0(t)$ is dense in $\mathscr{B}$.*

It is also desirable that the solution depend continuously on the initial data. If we alter the initial date u_0 by addition of v_0, we want to guarantee that the alteration of the solution is small if v_0 is small, i.e., that there should be a constant K such that

$$\| E_0(t)v_0 \| \leqq K \| v_0 \|, \qquad 0 \leqq t \leqq T.$$

We therefore assume that *the operators $E_0(t)$ are uniformly bounded, for $0 \leqq t \leqq T$.*

The foregoing assumptions characterize a *properly posed* problem. For such a problem, $E_0(t)$ has a bounded linear extension $E(t)$ whose domain is the entire space $\mathscr{B}$ and whose bound is the same as that of $E_0(t)$, because a bounded linear operator with a dense domain can always be so extended. Then, for arbitrary u_0, the one-parameter set of elements of $\mathscr{B}$, $u(t)$, given by

$$u(t) = E(t)u_0$$

is interpreted as a generalized solution of the initial value problem (1), (2).

4. Finite Difference Approximations

When an approximate solution is obtained by finite difference methods, the time variable t, in the first place, assumes discrete values $t = t^0, t^1, \cdots, t^n, \cdots$, where $t^n = n\Delta t$, and correspondingly, one deals with a discrete sequence $u^0, u^1, \cdots, u^n, \cdots$, of states of the physical system.

In the second place, the space variables are also discrete so that the functions describing a state of the system are specified only at the points of a lattice or net of values of the space variables. However, we may still regard such a specification (although imperfect) as represented by a point in the same function space $\mathscr{B}$, by adopting some rule for specifying function values between the points of the space lattice, for example linear interpolation. Such a rule, if chosen with reasonable care, will not interfere with the linearity or boundedness of the operators dealt with. (Some authors, such as L. V. Kantorovitch [5] prefer to represent the sates u^n in a different Banach space $\mathscr{B}'$, and to establish suitable homomorphisms between $\mathscr{B}$ and $\mathscr{B}'$.)

The finite difference equations are:

$$u^{n+1} = B(\Delta t, \Delta x, \Delta y, \cdots)u^n, \tag{4}$$

where u^n is (it is hoped) an approximation to $u(t^n)$, and B denotes a linear finite difference operator which depends, as indicated, on the size of the time increment Δt and on the sizes of the space increments Δx, $\Delta y, \cdots$.

Contrary to possible appearance, this formulation is not restricted to explicit difference systems. If the system is implicit, the operator B will contain the inverse of a (possibly infinite) matrix, but for present purposes it is not necessary to suppose that B can be easily written in explicit form. Whatever the calculation procedure may be which leads to u^{n+1} when u^n is known, it results in a transformation in $\mathscr{B}$ and this transformation is denoted by B.

We do assume, however, that the calculation procedure is a definite one which can be applied to any function u^n and that the result u^{n+1} depends linearly and continuously on u^n, as is clearly the case for any reasonable scheme. In other words, for any fixed Δt, Δx and Δy, B is a bounded linear transformation whose domain is the whole Banach space.

The concepts of stability and convergence with which we deal here suppose an infinite sequence of calculations with increasingly finer mesh. We assume relations

$$\begin{aligned} \Delta x &= g_1(\Delta t), \\ \Delta y &= g_2(\Delta t), \\ &\cdots\cdots\cdots \end{aligned}$$

which tell how the space increments approach zero as the time increment goes to zero along the sequence, and we set

$$B(\Delta t,\ g_1(\Delta t),\ g_2(\Delta t), \cdots) = C(\Delta t),$$

so that

$$u^{n+1} = C(\Delta t)u^n. \tag{5}$$

5. The Consistency Condition

Since

$$\frac{u^{n+1} - u^n}{\Delta t}$$

is to be an approximation to the time derivative,

$$\frac{C(\Delta t)u - u}{\Delta t}$$

must be an approximation, in some sense, to Au. We cannot expect this to be true for all u in $\mathscr{B}$, because in general Au is not even defined for all u in $\mathscr{B}$ But we want it to be true for nearly all u that can appear in a genuine solution of the initial value problem; and for any particular genuine solution we want the approximation to be uniformly good for all t in $0 \leqq t \leqq T$. Specifically, we shall call the family of operators $C(\Delta t)$ a *consistent approximation* for the initial value problem, if for some class U of genuine solutions it is true that, for any $u(t)$ in this class,

$$\lim_{\Delta t \to 0} \left\| \left\{ \frac{C(\Delta t) - I}{\Delta t} - A \right\} u(t) \right\| = 0 \text{ uniformly in } t, \qquad 0 \leqq t \leqq T, \tag{6}$$

provided that the class U is sufficiently wide and that its initial elements $u(0)$ are dense in $\mathscr{B}$. (6) is called the *consistency condition.*

In applications, A is usually a differential and C a difference operator in the space variables. To verify the consistency condition (6), Au has to be compared to $(C(\Delta t)-I)u/\Delta t$; to carry out this comparison expand each term in $C(\Delta t)u$ into a finite Taylor series (take two or three terms, depending on the order of the differential operator A), obtaining a differential operator. The error in replacing Cu by such a differential expression can be estimated, by Taylor's theorem, for sufficiently smooth functions. Therefore the comparison can be carried out for all sufficiently smooth solutions, and it is well known that the smooth solutions are dense among all solutions.

6. Convergence

Operating n times on u_0 with $C(\Delta t)$ gives $u^n = C(\Delta t)^n u_0$ which, it is hoped, approximates $u(n\Delta t)$. Since $u(t) = E(t)u_0$, we therefore make the following definition: the family of operators $C(\Delta t)$ provides a *convergent approximation* for the initial value problem if for any u_0 in $\mathscr{B}$ and for any sequences $\Delta_j t$, n_j such that $\Delta_j t$ tends to zero and $n_j \Delta_j t \to t$ where $0 \leqq t \leqq T$ then

$$\left\| \left\{ C(\Delta_j t) \right\}^{n_j} u_0 - E(t)u_0 \right\| \to 0, \qquad 0 \leqq t \leqq T. \tag{7}$$

Note that we require (7) to hold for every u_0 in $\mathscr{B}$ if $C(\Delta t)$ is to be called a convergent approximation.

7. Stability

In a sequence of calculations with $\Delta_j t \to 0$, if each calculation is carried from $t = 0$ to $t \approx T$, the operators which are used are those belonging to the set

$$\{C(\Delta_j t)\}^n, \qquad \begin{array}{l} j = 1, 2, 3, \cdots \\ 0 \leq n\Delta_j t < T \text{ for each } j, \end{array} \tag{8}$$

all applied to u_0. The idea of stability is that there should be a limit to the extent to which any component of an initial function can be amplified in the numerical procedure. Therefore the approximation $C(\Delta_j t)$ is said to be *stable* if the operators of the above set are uniformly bounded. Note that we make no reference here to the differential equation whose solution is desired so that stability, as defined, is a property solely of a sequence of difference equation systems.

In practice the bound of $\{C(\Delta t)\}^n$ is generally a continuous function of Δt in some interval, $0 < \Delta t \leqq \tau$, so that we may equivalently define the approximation $C(\Delta t)$ to be stable if for some $\tau > 0$, the set of operators

$$\{C(\Delta t)\}^n, \qquad \begin{array}{l} 0 < \Delta t \leqq \tau \\ 0 \leqq n\Delta t \leqq T \end{array} \tag{9}$$

is uniformly bounded.

8. The Equivalence Theorem

Given a properly posed initial value problem (1), (2) *and a finite difference approximation* $C(\Delta t)$ *to it that satisfies the consistency condition, stability is a necessary and sufficient condition that* $C(\Delta t)$ *be a convergent approximation.*

According to the definition of Section 6, this involves convergence for an arbitrary initial element u_0. In principle, an unstable scheme can sometimes give convergence for special initial elements. (Such schemes are not generally very useful in practise, because the initial data seldom have the required properties, and even if they do, round-off errors are likely to perturb the calculation enough to throw it into a neighboring divergent situation.)

We now prove the first part of the theorem: a convergent scheme is necessarily stable.

We start by showing that for a convergent scheme, the set of elements

$$C^n(\Delta t)u_0, \qquad n\Delta t \leqq T \tag{10}$$

are bounded for each fixed u_0 in $\mathscr{B}$. For, assume to the contrary that for a sequence n_j, $\Delta_j t$, $n_j\Delta_j t \leqq T$, the norms of the elements $C^{n_j}(\Delta_j t)u_0$ tend to infinity. Select a subsequence such that $n_j\Delta_j t$ tends to some limit t; since the scheme was assumed convergent, $C^{n_j}(\Delta_j t)u_0$ would have to tend to $E(t)u_0$, which it couldn't if it were unbounded.

We now appeal to the principle of uniform boundedness, which says that if each operator L of a set is bounded and if there exists a function $K(u)$ such that $\| Lu \| \leqq K(u)$ for all L in the set and all u in $\mathscr{B}$, then the set is uniformly bounded. Applying this to the present case, we see that the set (8) is uniformly bounded, and the approximation is stable.

To prove that, conversely, stability implies convergence, let $u(t)=E(t)u_0$ be a genuine solution belonging to the set U referred to in the definition of consistency. Then, for any positive ε,

$$\left\| \left\{ \frac{C(\Delta t)-I}{\Delta t} - A \right\} u(t) \right\| < \frac{\varepsilon}{2}, \qquad 0 \leqq t \leqq T,$$

for sufficiently small Δt. Also, from the definition of a genuine solution,

$$\left\| \left\{ \frac{E(\Delta t)-I}{\Delta t} - A \right\} u(t) \right\| < \frac{\varepsilon}{2}, \qquad 0 \leqq t \leqq T,$$

for sufficiently small Δt, so that by the triangle inequality,

$$\| \{C(\Delta t) - E(\Delta t)\} u(t) \| < \varepsilon \Delta t, \qquad 0 \leqq t \leqq T, \tag{11}$$

for sufficiently small Δt. This last inequality might have been taken as the basis of the definition of consistency, but the definition given in Section 5 is preferred for practical applications because it involves the operator A rather than the generally unknown solution operator $E(t)$. Set

$$\begin{aligned} \psi_j &= [\{C(\Delta_j t)\}^{n_j} - E(n_j \Delta_j t)] u_0 \\ &= \sum_{0}^{n_j-1}{}_{(k)} \{C(\Delta_j t)\}^k [C(\Delta_j t) - E(\Delta_j t)] E((n_j-1-k)\Delta_j t)\, u_0 . \end{aligned}$$

The equality of the second and third members of this equation results from cancellation of all except the first and last terms of the third member, when written out in full. The norm of ψ_j can be estimated by use of inequality (11) with the help of the triangle inequality:

$$\| \psi_j \| < K \sum_{0}^{n_j-1}{}_{(k)} \varepsilon \Delta_j t = K \varepsilon n_j \Delta_j t < K \varepsilon T,$$

for sufficiently small $\Delta_j t$, where K denotes the uniform bound of the set (8). Therefore, since ε was arbitrary,

$$\| \psi_j \| \to 0 \text{ as } \Delta_j t \to 0. \tag{12}$$

Now suppose that $n_j \Delta_j t \to t$ as $j \to \infty$, where t is a number in the interval $(0, T)$. The difference $\{E(n_j \Delta_j t) - E(t)\} u_0$ may be written in either of two ways, depending on which of the two arguments $n_j \Delta_j t$ and t is the larger, that is, as

$$(E(s) - I) E(t') u_0 \text{ if } s = n_j \Delta_j t - t \geqq 0, \qquad t' = t,$$

or as

$$-(E(s) - I) E(t') u_0 \text{ if } s = t - n_j \Delta_j t > 0, \qquad t' = n_j \Delta_j t.$$

(The reason for making the distinction is that the solution operator $E(t)$

is generally defined only for non-negative arguments.) In either case,

$$||\{E(n_j\Delta_jt) - E(t)\}u_0|| < K_E ||(E(s) - I)u_0||$$

which goes to zero as $s \to 0$ and therefore as $j \to \infty$. Thus, combining this result with (12),

$$||[\{C(\Delta_jt)\}^{n_j} - E(t)]u_0|| \to 0 \text{ as } j \to \infty \tag{13}$$

for any u_0 which can be the initial element of a genuine solution of the class U. But these initial elements are dense in $\mathscr{B}$, so that if u is any element of $\mathscr{B}$ there is a sequence $u_1, u_2, \cdots$ converging to u, each u_i the initial element of a genuine solution for which (13) holds. Then

$$[\{C(\Delta_jt)\}^{n_j} - E(t)]u = [\{C(\Delta_jt)\}^{n_j} - E(t)]u_m + \{C(\Delta_jt)\}^{n_j}(u - u_m) + E(t)(u - u_m).$$

The last two terms on the right of this equation can be made as small as one pleases by choosing m sufficiently large, on account of uniform boundedness of the operators C^n and of $E(t)$. Then the first term on the right can be made as small as one pleases by choosing Δ_jt sufficiently small. Therefore the left member of the above equation goes to zero as $j \to \infty$. Since u was arbitrary, it is now established that $C(\Delta t)$ is a convergent approximation as defined in Section 6, and the equivalence theorem is established.

The above sufficiency proof is an operator-theoretic analogue of Fritz John's result relating the uniform boundedness of the values of the approximate solution to convergence in the maximum norm.

PART II

PARTIAL DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS

9. Introduction

Here the stability requirement as defined in Part I, and whose significance is indicated by the equivalence theorem, is applied to a special class of linear initial value problems — those of partial differential equations with constant coefficients and with auxiliary conditions permitting the use of Fourier series or integrals. If the space variables are restricted to a finite domain and the boundary conditions are of such a nature that they can be represented as a periodicity condition, Fourier series are used. If the domain is infinite, but the functions are quadratically integrable, Fourier integrals are used, via Plancherel's theorem. Combinations are also possible, in which some of the space variables have finite domain and others are unlimited. All these cases lead to exactly the same results, and our discussion will be based on Fourier series.

We shall be dealing with the following Banach space $\mathscr{B}$: if p is the number of functions used to describe a state of the physical system, and d is the number of space variables, a point in $\mathscr{B}$ represents a p-vector function defined in a d-dimensional unit cube (or rectangular parallelopiped). We suppose that these functions are in L^2 over this cube and that the square of the Banach norm is given by equation (14).

The advantage of this norm over, for example, the maximum norm is that the Parseval equation then shows that the Fourier transform establishes a norm-preserving isomorphism between $\mathscr{B}$ and the space $\mathscr{B}'$ of the Fourier coefficients. The stability requirement takes on a particularly simple form in $\mathscr{B}'$ leading immediately to the Von Neumann condition as a necessary condition for stability.

Of course, the choice of a norm is restricted by the nature of the problem; i.e., the solution operators $E(t)$ have to be bounded with respect to the norm. In most problems of mathematical physics, the L^2 norm can be used.

We give several sufficient conditions for stability; these are mostly of the nature of an auxiliary condition under which the Von Neumann condition is also sufficient for stability.

One may perhaps surmise that in all practical cases (including problems with variable coefficients,[2] and even nonlinear problems) the Von Neumann condition is both necessary and sufficient for stability. Such a surmise has often been made (so far apparently without misfortune) by people who have to make actual calculations, and one can construct a good bit of heuristic evidence for it. But the purpose of the present discussion is to discuss only certain cases that can be treated rigorously. A few simple applications will be given.

10. Notation; Fourier Series

Let $\mathbf{x} = (x_1, x_2, \cdots, x_d)$ be a vector (vectors will be denoted by bold face type) whose components $x_1, x_2, \cdots, x_d$ are the space variables of the problem. Suppose that the functions with which we deal are periodic with periods $L_1, L_2, \cdots L_d$ in the space variables. Consider a series

$$\sum_{(\mathbf{k})} c(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}$$

[2] Fritz John has succeeded in proving in his important paper [13] that for parabolic equations a mildly strengthened form of Von Neumann's condition is sufficient for stability even for operators with variable coefficients. A similar result for a certain class of hyperbolic equations with variable coefficients has been obtained by Peter Lax [14].

where $\mathbf{k}$ is a d-component vector whose components are $2\pi l_1/L_1, \cdots, 2\pi l_d/L_d$ and where the summation is understood to be over all such vectors obtained by letting $l_1, l_2, \cdots, l_d$ run independently over all positive and negative integers and where $c(\mathbf{k})$ is a complex-valued function defined on the lattice of these vectors. This is a general trigonometrical series with the periodicity described above. In our applications there are p functions of the x's; we treat them as the components of a p-component vector $\mathbf{f}(\mathbf{x})$. For any vector $\mathbf{y}$ we denote by $|\,\mathbf{y}\,|$ the square root of the sum of the squares of the absolute values of the components. Therefore, if $\mathbf{f}(\mathbf{x})$ can be expanded as

$$\mathbf{f}(\mathbf{x}) = \textstyle\sum_{(k)} \mathbf{c}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}},$$

the Parseval equation is

$$\frac{1}{V}\int_0^{L_1} dx_1 \cdots \int_0^{L_d} dx_d\, |\,\mathbf{f}(\mathbf{x})\,|^2 = \textstyle\sum_{(\mathbf{k})} |\,\mathbf{c}(\mathbf{k})\,|^2 \tag{14}$$

where $V = L_1 L_2 \cdots L_d$.

Any periodic $\mathbf{f}(\mathbf{x})$ for which the left member of (14) exists will be called an element of $\mathscr{B}$ and the square root of that member will be called its norm. Similarly, any set of coefficients for which the right member of (14) exists will be called an element of $\mathscr{B}'$ and the square root of that member will be called its norm. Then the Fischer-Riesz theorem says that $\mathscr{B}$ is a complete space and the Riesz-Fischer theorem says that there is a one-to-one correspondence between elements of $\mathscr{B}$ and of $\mathscr{B}'$, if we adopt the usual agreement that functions $\mathbf{f}(\mathbf{x})$ which differ only on a set of measure zero are regarded as identical — this agreement is reasonable, because the corresponding states of the physical system would be physically indistinguishable. The Parseval equation (14) shows that the correspondence between $\mathscr{B}$ and $\mathscr{B}'$ is norm-preserving. Statements of convergence, boundedness and the like can be taken over directly from $\mathscr{B}$ to $\mathscr{B}'$ or from $\mathscr{B}'$ to $\mathscr{B}$.

11. Properly Posed Problems

The general linear differential operator with constant coefficients can be otained formally by taking a function $D(\mathbf{k})$ or $D(k_1, k_2, \cdots, k_d)$ which is a $p \times p$ matrix whose elements are polynomials in $k_1, k_2, \cdots, k_d$, and substituting $\partial/\partial x_1$ for k_1, $\partial/\partial x_2$ for k_2, etc. If A is such an operator and we apply it to the element $\mathbf{v}e^{i\mathbf{k}\cdot\mathbf{x}}$ where $\mathbf{v}$ is a constant vector, the result is simply the product of this element and $D(i\mathbf{k})$. Therefore the solution of the initial value problem

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) = A\mathbf{u}(\mathbf{x}, t), \tag{15}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \tag{16}$$

is

$$\mathbf{u}(\mathbf{x}, t) = \textstyle\sum_{(\mathbf{k})} e^{i\mathbf{k}\cdot\mathbf{x}} e^{tD(i\mathbf{k})} \mathbf{v}_0(\mathbf{k}) \tag{17}$$

where

$$\mathbf{v}_0(\mathbf{k}) = \frac{1}{V} \int_0^{L_1} dx_1 \cdots \int_0^{L_d} dx_d \mathbf{u}_0(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{18}$$

The first requirement for a properly posed problem (see Section 3), namely that the domain of the solution operator be dense in $\mathscr{B}$, is automatically satisfied for the problems considered here, because the above solution (equations (17) and (18)) is certainly valid whenever the initial element $\mathbf{u}_0(x)$ is a trigonometric polynomial and the trigonometric polynomials are dense in $\mathscr{B}$.

The second requirement for a properly posed problem takes the form that $\| e^{tD(i\mathbf{k})} \|$ should be a bounded function of $\mathbf{k}$ and that the bound should be uniform in t. (It should be obvious to the reader that if M is a $p \times p$ matrix and we write $\| M \|$ we mean the bound of the transformation corresponding to M in a p-dimensional vector space with complex Euclidean norm.) Whether this condition is satisfied must usually be investigated separately in each case.

12. Finite Difference Equations

Just as the differential operator A is represented, in the space $\mathscr{B}'$, by the matrix $D(i\mathbf{k})$, the finite-difference operator $B(\Delta t, \Delta\mathbf{x})$ will be represented in $\mathscr{B}'$ by a matrix $G(\Delta t, \Delta\mathbf{x}, \mathbf{k})$ whose elements are functions of the components of $\mathbf{k}$ as well as of the parameters Δt, $\Delta\mathbf{x}$. One reason for making the Fourier transformation is that the elements of $G(\Delta t, \Delta\mathbf{x}, \mathbf{k})$ can generally be found easily, even though $B(\Delta t, \Delta\mathbf{x})$ represents an implicit system of difference equations.

Each difference equation equates to zero a certain linear combination of the components of $\mathbf{u}^n$ and of $\mathbf{u}^{n+1}$ at a group of neighboring points of the net used for the numerical work. Specifially, let this group of points be referred to a particular point of the group with coordinates $x_1, x_2, \cdots x_d$ so that a typical neighbor of this point in the group has coordinates $x_1 + \beta_1 \Delta x_1, \cdots, x_d + \beta_d \Delta x_d$ where $\beta_1, \cdots, \beta_d$ are integers. The difference equations can then be written in the form

$$\begin{aligned}\textstyle\sum_{(\beta_1, \cdots, \beta_d)} [&A(\beta_1, \cdots, \beta_d)\mathbf{u}^{n+1}(x_1 + \beta_1 \Delta x_1, \cdots, x_d + \beta_d \Delta x_d) \\ &+ B(\beta_1, \cdots, \beta_d)\mathbf{u}^n(x_1 + \beta_1 \Delta x_1, \cdots, x_d + \beta_d \Delta x_d)] = 0,\end{aligned} \tag{19}$$

where A and B are $p \times p$ matrices whose elements depend on the β_i and on Δt and the Δx_i but not on t (i.e. n) or the x_i themselves. The summation is over a finite number of neighbors — that is over a finite number of sets of values of $\beta_1, \cdots, \beta_d$.

This system is in general implicit, because in the numerical work the unknowns are the values of the components of $\mathbf{u}^{n+1}$ at the various net points, and each equation contains generally several of the unknowns. We assume, however, that the system is such that if $\mathbf{u}^n(\mathbf{x})$ is given as any element of $\mathscr{B}$, then $\mathbf{u}^{n+1}(\mathbf{x})$ is uniquely determined by the difference equations (19) and the periodicity requirement.

If the Fourier series

$$\mathbf{u}^n(\mathbf{x}) = \textstyle\sum_{(\mathbf{k})} \mathbf{v}^n(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} \tag{20}$$

and a similar one for $\mathbf{u}^{n+1}(\mathbf{x})$ are substituted into (19), the typical term contains a factor

$$\exp\{i[k_1(x_1 + \beta_1 \Delta x_1) + \cdots + k_d(x_d + \beta_d \Delta x_d)]\}$$

from which we cancel out the common part $e^{i\mathbf{k}\cdot\mathbf{x}}$ from all the terms of the equation. What is left can be written as

$$H_1 \mathbf{v}^{n+1}(\mathbf{k}) + H_2 \mathbf{v}^n(\mathbf{k}) = 0$$

where H_1 is an abbreviation for the matrix

$$\textstyle\sum_{(\beta_1, \cdots, \beta_d)} A(\beta_1, \cdots, \beta_d) \exp\{i[k_1\beta_1\Delta x_1 + \cdots + k_d\beta_d\Delta x_d]\}$$

and H_2 is similar. The solvability assumption made in the preceding paragraph is tantamount to the assumption that H_1 has an inverse. Therefore we can write

$$\mathbf{v}^{n+1}(\mathbf{k}) = G\mathbf{v}^n(\mathbf{k}) \tag{21}$$

where the matrix G is given by

$$G = G(\Delta t, \Delta\mathbf{x}, \mathbf{k}) = -H_1^{-1} H_2 . \tag{22}$$

G will be called the *amplification matrix*: it is the representation in $\mathscr{B}'$ of the operator $B(\Delta t, \Delta\mathbf{x})$. Therefore the stability requirement is that if the manner of refinement of the mesh is given by $\Delta\mathbf{x} = \mathbf{g}(\Delta t)$, the set of matrices

$$\{G(\Delta_j t, \mathbf{g}(\Delta_j t), \mathbf{k})\}^n , \qquad \begin{array}{l} j = 1, 2, \cdots, \\ 0 \leqq n\Delta_j t \leqq T \end{array} \tag{23}$$

should be uniformly bounded for all $\mathbf{k}$ with real components, the bound being uniform in $\mathbf{k}$. As in Part I of this paper $\Delta_j t$ is a sequence tending to zero as $j \to \infty$ and corresponding to each j there is a net or grid of space points such that each $g_1(\Delta_j t) = \Delta x_1$, $g_2(\Delta_j t) = \Delta x_2$, etc. also tends to zero.

The problem of stability is thus reduced to that of finding estimates for the bounds of powers of the amplification matrix G.

13. The Von Neumann (Necessary) Condition for Stability

A lower limit for the bounds of the powers of G is easily given. Let the eigenvalues of G be $\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(p)}$ (not assumed real, not assumed distinct). The spectral radius of G is

$$r_1 = r_1(\varDelta t, \varDelta \mathbf{x}, \mathbf{k}) = \operatorname{Max}_{(i)} |\lambda^{(i)}|.$$

Suppose the λ's so ordered that $|\lambda^{(1)}| = r_1$ and let $v^{(1)}$ be an eigenvector corresponding to eigenvalue $\lambda^{(1)}$. Then

$$||G|| = \operatorname{Max}_{(\mathbf{v})} \frac{||G\mathbf{v}||}{||\mathbf{v}||} \geqq \frac{||G\mathbf{v}^{(1)}||}{||\mathbf{v}^{(1)}||} = r_1,$$

or generally the spectral radius is a lower bound for the bound of a matrix. If G is raised to any power, each of its eigenvalues gets raised to the same power, and therefore the spectral radius G^n is r_1^n. Therefore

$$||G^n|| \geqq r_1^n.$$

We call

$$R_1 = R_1(\varDelta t) = \operatorname{Max}_{(\mathbf{k})} r_1(\varDelta t, \mathbf{g}(\varDelta t), \mathbf{k}),$$

where the maximum is with respect to all $\mathbf{k}$ with real components. The stability requirement of uniform boundedness of the set (23) implies that for some K,

$$\{R_1(\varDelta t)\}^n \leqq K, \qquad \begin{array}{l}\varDelta t > 0,\\ 0 \leqq n\varDelta t \leqq T,\end{array}$$

but this is equivalent to the condition

$$R_1(\varDelta t) \leqq 1 + O(\varDelta t) \tag{24}$$

where $O(\varDelta t)$ denotes a quantity bounded by a constant times $\varDelta t$. This is the Von Neumann necessary condition for stability.

14. A Sufficient Condition for Stability

Let the eigenvalues of G^*G be denoted by $\mu^{(1)}, \cdots, \mu^{(p)}$. The bound of G is

$$r_2 = r_2(\varDelta t, \varDelta \mathbf{x}, \mathbf{k}) = ||G|| = \operatorname{Max}_{(i)} |\mu^{(i)}|^{1/2}.$$

Since $||G^n|| \leqq ||G||^n$, if we call

$$R_2 = R_2(\varDelta t) = \operatorname{Max}_{(\mathbf{k})} r_2(\varDelta t, \mathbf{g}(\varDelta t), \mathbf{k}),$$

the stability requirement is satisfied provided there is a K such that

$$\{R_2(\varDelta t)\}^n \leqq K, \qquad \begin{array}{l}\varDelta t > 0,\\ 0 \leqq n\varDelta t \leqq T.\end{array}$$

but this is equivalent to the condition

$$R_2(\Delta t) \leqq 1 + O(\Delta t). \tag{25}$$

Consequently, we have

THEOREM 1. *Condition* (25) *is sufficient for stability.*

If G is a normal matrix (i.e., one that commutes with its Hermitian conjugate), the eigenvalues of G^*G are just the squares of the absolute values of the eigenvalues of G (because G^* and G can be reduced to diagonal form by the same unitary transformation), so that $R_1(\Delta t) = R_2(\Delta t)$. Therefore, we can state the

COROLLARY. *If G is a normal matrix, the Von Neumann condition* (24) *is sufficient as well as necessary for stability.*

15. A Second Sufficient Condition for Stability

As noted in the corollary, the case in which G is a normal matrix is an important special case, and in that case there is a complete orthogonal set of eigenvectors of G. Even if G is not normal, there *may* be a complete set of linearly independent eigenvectors (not generally orthogonal). A stability condition will now be given for such cases.

Let $\phi^{(1)}, \cdots, \phi^{(p)}$ denote a set of normalized, linearly independent eigenvectors of G. Let T be the matrix having these eigenvectors as columns, so that $T_{ij} = \phi_j^{(i)}$; and let Δ denote the determinant of T. T provides a similarity transformation (not in general unitary) that diagonalizes G. That is,

$$G = T^{-1}\begin{pmatrix} \lambda^{(1)} & \cdot & 0 \\ & \cdot & \\ 0 & \cdot & \lambda^{(p)} \end{pmatrix} T,$$

and therefore

$$G^n = T^{-1}\begin{pmatrix} \lambda^{(1)} & \cdot & 0 \\ & \cdot & \\ 0 & \cdot & \lambda^{(p)} \end{pmatrix}^n T. \tag{26}$$

The inverse of T has elements given by

$$(T^{-1})_{ij} = \frac{\text{algebraic cofactor of } T_{ji}}{\Delta}.$$

If the columns (or rows) of any determinant are regarded as a set of vectors, the absolute value of the determinant does not exceed the product of the lengths of the vectors (corresponding to the interpretation of the determinant as the volume of a multidimensional parallelopiped of which the vectors form a set of coterminous edges). Each column of the cofactor mentioned above consists of $p - 1$ of the components of a normalized eigenvector of G and hence has length less than or equal to 1. Consequently,

$$|(T^{-1})_{ij}| \leqq \frac{1}{|\Delta|}.$$

Clearly, the absolute value of an element of T cannot exceed 1, so from (26),

$$|(G^n)_{ij}| \leqq \frac{p^2}{|\Delta|} r_1^n,$$

where the factor p^2 comes from the fact that there are p^2 terms in the expansion of the matrix product (26). Since the bound of a $p \times p$ matrix does not exceed p times its absolutely largest element,

$$\| G^n \| \leqq \frac{p^3}{|\Delta|} r_1^n.$$

The determinant Δ of course is a function of Δt and $\mathbf{k}$, but if it is bounded away from zero, we can replace $|\Delta|$ by its greatest lower bound in the above inequality and use the same reasoning that led to (25) in Section 14, to prove

THEOREM 2. *If there is a constant a such that* $|\Delta| > a > 0$ *for all real* $\mathbf{k}$ *and all sufficiently small* Δt, *where* Δ *is the determinant of the normalized eigenvectors of the amplification matrix* $G(\Delta t, \mathfrak{g}(\Delta t), \mathbf{k})$, *the Von Neumann condition* (24) *is sufficient as well as necessary for stability.*

16. A Third Sufficient Condition for Stability

In some cases of practical importance the determinant Δ vanishes for certain values of $\mathbf{k}$ so that a different criterion must be found. To find one, we start from Schur's theorem that any square matrix A can be reduced to triangular form by a unitary transformation

$$B = U^*AU$$

where B is the triangular matrix:

$$B = \begin{bmatrix} \lambda^{(1)} & B_{12} & B_{13} \cdots B_{1p} \\ 0 & \lambda^{(2)} & B_{23} \cdots B_{2p} \\ 0 & 0 & \lambda^{(3)} \cdots B_{3p} \\ \cdot & \cdot & \cdot \qquad \cdot \\ \cdot & \cdot & \cdot \quad \cdot \quad \cdot \\ \cdot & \cdot & \cdot \qquad \cdot \\ 0 & 0 & 0 \cdots \lambda^{(p)} \end{bmatrix}$$

whose diagonal elements are the eigenvalues of A and such that $B_{ij} = 0$

for $i > j$. Since no element of U can exceed 1 in absolute value

$$\text{Max}\,|B_{ij}| \leqq p^2\,\text{Max}\,|A_{ij}|. \tag{27}$$

The general element of the n-th power of B has the form

$$(B^n)_{ij} = \sum B_{ik_1} B_{k_1 k_2} \cdots B_{k_{n-1} j} \tag{28}$$

summed over the indices $k_1, k_2, \ldots, k_{n-1}$, arranged in every possible way, provided

$$i \leqq k_1 \leqq k_2 \leqq \cdots \leqq k_{n-1} \leqq j.$$

No matter how large n is, at most $j - i$ of the factors in the above product can be off-diagonal elements of B. This will enable us to obtain a satisfactory bound for B^n by imposing restrictions on the diagonal elements only. We assume that

$$\underset{i>1}{\text{Max}}\,|\lambda^{(i)}| = \gamma < 1, \tag{29}$$

and call

$$\text{Max}\,(|\lambda^{(1)}|,\, 1) = \lambda^* .$$

We focus our attention temporarily on those factors of the typical product in (28) which are off-diagonal elements of B, disregarding the diagonal elements occurring in the product. The number of such factors can be r where $0 \leqq r \leqq j - i$; let N_r^{j-i} be the number of distinct ways of choosing these r factors from among the off-diagonal elements of B, taking into account the chain rule for subscripts in matrix multiplication. $\Big($Except for the trivial cases in which $j - i = 0$ or $r = 0$, N_r^{j-i} is just the binomial coefficient $\binom{j-i-1}{r-1}\Big)$.

Having chosen the r off-diagonal elements, we consider the various ways in which they can be combined with diagonal elements to make the general typical product in (28) with the factors in the order shown there. One arrangement is with the off-diagonal elements crowded together at the right side of the product and preceded by a suitable power of $\lambda^{(i)}$ on their left. Other arrangements can be obtained from this one by decreasing the power of $\lambda^{(i)}$ and inserting suitable diagonal elements in positions to the right of the leftmost off-diagonal element. The first such factor can be inserted in any one of r positions, the next in any one of $r + 1$ positions, the next in any one of $r + 2$ positions, and so forth. But the order of insertion is irrelevant, so the number of distinct ways in which q such factors can be inserted is

$$\frac{r(r+1)(r+2)\cdots(r+q-1)}{q!}.$$

The inserted factors are all eigenvalues with index greater than i (because of their positions in the product) hence with index greater than 1. Therefore, by (29), the inserted factors, when multiplied together, are bounded by γ^q. Now the series

$$\sum_{0}^{\infty}{}_{(q)} \frac{r(r+1)\cdots(r+q-1)}{q!}\gamma^q$$

being a hypergeometric series, is convergent to a finite limit because $\gamma < 1$. Let that limit be denoted by $F_r(\gamma)$. The power of $\lambda^{(i)}$ occurring in the product is in any case bounded by $(\lambda^*)^n$, so, finally

$$|(B^n)_{ij}| \leqq (\lambda^*)^n \sum_{1}^{j-1}{}_{(r)} N_r^{j-1} (\operatorname*{Max}_{s,t} |B_{st}|)^r F_r(\gamma).$$

This expansion is clearly maximized by taking $j-1 = p-1$. Then, since the bound of a matrix does not exceed p times the absolute value of its largest element, and using (27) and the fact that the bound is invariant under a unitary transformation, we find

$$\|A^n\| \leqq (\lambda^*)^n p \sum_{1}^{p-1}{}_{(r)} N_r^{p-1} (p^2 \operatorname*{Max}_{s,t} |A_{st}|)^r F_r(\gamma).$$

To apply this result to the stability problem, we interpret A as the amplification matrix $G(\Delta t, \Delta\mathbf{x}, \mathbf{k})$. The factor $(\lambda^*)^n$ is then bounded for the set (23) if the Von Neumann condition is satisfied, and the other factors in the above expression are bounded if $\gamma < 1$ and the A_{st} are bounded. The following theorem results:

THEOREM 3. *If the elements of the amplification matrix* $G(\Delta t, \mathfrak{g}(\Delta t), \mathbf{k})$ *are bounded functions of* $\mathbf{k}$ *and* Δt *for all real* $\mathbf{k}$ *and all sufficiently small positive* Δt, *and if there is a constant* γ *such that*

$$|\gamma^{(i)}(\Delta t, \mathfrak{g}(\Delta t), \mathbf{k})| \leqq \gamma < 1, \qquad i = 2, 3, \cdots, p,$$

then the Von Neumann condition (24) *is sufficient as well as necessary for stability.*

Roughly speaking, one eigenvalue is permitted to get up to 1, or even $1 + \theta(\Delta t)$ provided the bound of the others is less than 1.

17. The Wave Equation

As a first example to illustrate the foregoing ideas we consider the wave equation

$$\frac{\partial^2\psi}{\partial t^2} - c^2\frac{\partial^2\psi}{\partial x^2} = 0.$$

A satisfactory formulation is obtained by making the further definition

$w = c\,\partial\psi/\partial x$, whereupon the equations become

$$\frac{\partial\phi}{\partial t} = c\frac{\partial w}{\partial x}, \qquad \frac{\partial w}{\partial t} = c\frac{\partial\phi}{\partial x}, \tag{31}$$

and we now have a properly posed problem. In this case the square of the norm is the energy of the wave motion and by conservation of energy the solution operator is bounded with bound unity.

The first choice of the finite difference equations that we wish to consider is

$$\begin{aligned} \phi^{n+1}(x) &= \phi^n(x) + \frac{c\Delta t}{\Delta x}\left[w^n\left(x+\frac{\Delta x}{2}\right) - w^n\left(x-\frac{\Delta x}{2}\right)\right], \\ w^{n+1}(x) &= w^n(x) + \frac{c\Delta t}{\Delta x}\left[\phi^n\left(x+\frac{\Delta x}{2}\right) - \phi^n\left(x-\frac{\Delta x}{2}\right)\right]. \end{aligned} \tag{32}$$

The amplification matrix is

$$G(\Delta t, \Delta x, k) = \begin{bmatrix} 1 & 2i\dfrac{c\Delta t}{\Delta x}\sin\dfrac{k\Delta x}{2} \\ 2i\dfrac{c\Delta t}{\Delta x}\sin\dfrac{k\Delta x}{2} & 1 \end{bmatrix}$$

as found by substituting a Fourier term

$$\begin{pmatrix} a^n \\ b^n \end{pmatrix} e^{ikx}$$

for $\begin{pmatrix} \phi^n \\ w^n \end{pmatrix}$ into (32) and solving for $\begin{pmatrix} a^{n+1} \\ b^{n+1} \end{pmatrix}$ in terms of $\begin{pmatrix} a^n \\ b^n \end{pmatrix}$ as

$$\begin{pmatrix} a^{n+1} \\ b^{n+1} \end{pmatrix} = G\begin{pmatrix} a^n \\ b^n \end{pmatrix}.$$

The quantity $R_1 = R_1(\Delta t)$ defined in Section 13 and appearing in the Von Neumann condition, namely the maximum with respect to k of the spectral radius of G, is

$$R_1 = \sqrt{1 + 4\left(\frac{c\Delta t}{\Delta x}\right)^2}$$

and we reach the well known conclusion that the difference equations (32) are unstable, at least if $c\Delta t/\Delta x$ is kept at any constant value as Δt and $\Delta x \to 0$.

In this example the Von Neumann condition could be satisfied by making Δt and $\Delta x \to 0$ in such a way that $\Delta t/(\Delta x)^2$ is constant, but there is not much point in pursuing this lead because we all know perfectly well that there is a much better difference scheme than (32).

According to the scheme usual in fluid dynamical calculations, the differential equations (21) are approximated by

$$\begin{aligned}\phi^{n+1}(x) &= \phi^n(x) + \frac{c\Delta t}{\Delta x}\left[w^n\left(x+\frac{\Delta x}{2}\right) - w^n\left(x-\frac{\Delta x}{2}\right)\right],\\ w^{n+1}(x) &= w^n(x) + \frac{c\Delta t}{\Delta x}\left[\phi^{n+1}\left(x+\frac{\Delta x}{2}\right) - \phi^{n+1}\left(x-\frac{\Delta x}{2}\right)\right].\end{aligned} \tag{33}$$

This scheme differs from (32) only in the superscripts on ϕ in the second equation. (The equations would have a more centered look if we had used the notation $\phi^{n+\frac{1}{2}}$ and $\phi^{n-\frac{1}{2}}$ in place of ϕ^{n+1} and ϕ^n). The amplification matrix is

$$G = \begin{pmatrix} 1 & ia \\ ia & 1-a^2 \end{pmatrix} \tag{34}$$

where a is an abbreviation for $\dfrac{2c\Delta t}{\Delta x}\sin\dfrac{k\Delta x}{2}$,

and
$$G^*G = \begin{pmatrix} 1+a^2 & ia^3 \\ -ia^3 & 1-a^2+a^4 \end{pmatrix}.$$

The characteristic equations of these matrices are

$$\lambda^2 - (2-a^2)\lambda + 1 = 0 \tag{35}$$

and

$$\mu^2 - (2+a^4)\mu + 1 = 0. \tag{36}$$

For each of these characteristic equations the product of the roots is 1. In (35) the sum of the roots is $2 - a^2$; consequently the roots lie on the unit circle if $a^2 \leqq 4$. In (36) the roots are real. We find, for the quantities $R_1(\Delta t)$ and $R_2(\Delta t)$ introduced in Sections 13 and 14

$$R_1(\Delta t)\begin{cases} = 1 \text{ if } \dfrac{c\Delta t}{\Delta x} \leqq 1, \\[2ex] > 1 \text{ if } \dfrac{c\Delta t}{\Delta x} > 1, \end{cases}$$

$$R_2(\Delta t) = \sqrt{1 + 8\left(\frac{c\Delta t}{\Delta x}\right)^4 + 4\left(\frac{c\Delta t}{\Delta x}\right)^2\sqrt{1+4\left(\frac{c\Delta t}{\Delta x}\right)^4}}.$$

The Von Neumann condition is satisfied for $c\Delta t/\Delta x \leqq 1$ but not for any other fixed value of $c\Delta t/\Delta x$.

The sufficient condition for stability given in Section 14, namely

$R_2(\Delta t) \leqq 1 + O(\Delta t)$, requires $\Delta t = O((\Delta x)^2)$ as Δt, $\Delta x \to 0$, which is much more stringent than the Von Neumann condition. But the sufficient condition given in Section 15 gives what we want. The normalized eigenvectors of the matrix (34) are easily found, and their determinant has the absolute value

$$|\Delta| = \sqrt{1 - \left(\frac{c\Delta t}{\Delta x}\sin\frac{k\Delta x}{2}\right)^2}.$$

This is bounded away from zero if $c\Delta t < \Delta x$. We arrive thus at the conclusion, first stated in the Courant-Friedrichs-Lewy paper, that equations (33) are stable if $c\Delta t/\Delta x =$ constant < 1 but not if $c\Delta t/\Delta x =$ constant > 1. The case $c\Delta t/\Delta x = 1$ (stable according to Courant, Friedrichs and Lewy) is not handled by our method.

Lastly, we consider the implicit system

$$\begin{aligned}\phi^{n+1}(x) &= \phi^n(x) + \frac{c\Delta t}{2\Delta x}\left[w^n\left(x + \frac{\Delta x}{2}\right)\right.\\ &\quad \left. + w^{n+1}\left(x + \frac{\Delta x}{2}\right) - w^n\left(x - \frac{\Delta x}{2}\right) - w^{n+1}\left(x - \frac{\Delta x}{2}\right)\right],\\ w^{n+1}(x) &= w^n(x) + \frac{c\Delta t}{2\Delta x}\left[\phi^n\left(x + \frac{\Delta x}{2}\right)\right.\\ &\quad \left. + \phi^{n+1}\left(x + \frac{\Delta x}{2}\right) - \phi^n\left(x - \frac{\Delta x}{2}\right) - \phi^{n+1}\left(x - \frac{\Delta x}{2}\right)\right]\end{aligned} \tag{37}$$

as approximation to the differential equations (31). (Equations of this type have been used, for example, by Arthur Carson of Los Alamos in studies of the dynamics of stellar interiors.) The amplification matrix is

$$G = \begin{bmatrix} \dfrac{1 - a^2/4}{1 + a^2/4} & \dfrac{ia}{1 + a^2/4} \\ \dfrac{ia}{1 + a^2/4} & \dfrac{1 - a^2/4}{1 + a^2/4} \end{bmatrix}$$

and G^*G is the unit matrix. The criterion of Section 14 is always satisfied and the equations (37) are stable as $\Delta t \to 0$, $\Delta x \to 0$, no matter what are the relative rates at which Δt and Δx approach zero.

18. Diffusion Equation; Two Level Formulas

Consider the equation

$$\frac{\partial u}{\partial t} = A\frac{\partial^2 u}{\partial x^2} + 2B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} \tag{38}$$

where the quadratic form (A, B, C) is required to be positive definite; this makes the differential equation parabolic. In consequence of this requirement, the differential equation provides a properly posed initial value problem; for example in the space $\mathscr{B}$ of functions $u(x, y)$ in L^2 over a rectangle in the x, y-plane.

We consider in this section a certain class of difference equations in which $u^n(x, y)$ and $u^{n+1}(x, y)$ are connected directly. (In the two following sections we will consider some schemes in which u^n, u^{n+1} and u^{n-1} all appear in the same equation; these will be referred to as three-level formulas.)

Introduce the following abbreviations:

$$\begin{aligned} &u^n_{ij} && \text{for } u^n(x, y), \\ &u^n_{i+1\,j} && \text{for } u^n(x + \Delta x, y), \\ &u^n_{i\;j+1} && \text{for } u^n(x, y + \Delta y), \text{ etc.,} \end{aligned}$$

and

$$\Phi^n_{ij} \text{ for } \Big[A \frac{u^n_{i+1\,j} - 2u^n_{ij} + u^n_{i-1\,j}}{(\Delta x)^2} + 2B \frac{u^n_{i+1\,j+1} - u^n_{i-1\,j+1} - u^n_{i+1\,j-1} + u^n_{i-1\,j-1}}{4\Delta x \Delta y} + C \frac{u^n_{i\;j+1} - 2u^n_{ij} + u^n_{i\;j-1}}{(\Delta y)^2} \Big].$$

The class of finite difference equations we wish to consider is

$$u^{n+1}_{ij} - u^n_{ij} = \Delta t[\theta \Phi^{n+1}_{ij} + (1 - \theta)\Phi^n_{ij}], \tag{39}$$

where θ is a non-negative constant. The choice $\theta = 0$ gives the usual explicit system and the choices $\theta = \frac{1}{2}$, $\theta = 1$ give the two favorite implicit systems.

(As is well known, if $B = C = 0$, so that the problem reduces to that of one space variable, the implicit equations can be readily solved by a simple algorithm. For two or more space variables it is wise to solve the implicit equations approximately by a relaxation technique; this is of course much more labor than required, per cycle, by the explicit equations, but it is nevertheless worthwhile, in some cases, to use the implicit equations provided the relaxation is done by some method like the extrapolated Liebmann method.)

Since there is only one dependent variable, the amplification matrix has just one element:

$$G(\Delta t, \Delta x, \Delta y, k_x, k_y) = \frac{1 + (1 - \theta)W}{1 - \theta W},$$

where

$$W = \Delta t \left[\frac{2A}{(\Delta x)^2} (\cos k_x \Delta x - 1) - \frac{2B}{\Delta x \Delta y} \sin k_x \Delta x \sin k_y \Delta y + \frac{2C}{(\Delta y)^2} (\cos k_y \Delta y - 1) \right].$$

From the positive definite character of the quadratic form (A, B, C), it follows, after a little calculation, that

$$-4\Delta t \left[\frac{A}{(\Delta x)^2} + \frac{C}{(\Delta y)^2} \right] \leqq W \leqq 0.$$

The expression $\frac{1 + (1-\theta)W}{1 - \theta W}$ is an increasing function of W in the above interval and has the value 1 at $W = 0$. Therefore we will have $|G| \leqq 1$ if this expression is $\geqq -1$ when W has its most negative value. From this the Von Neumann condition is found to be:

1) if $\frac{1}{2} \leqq \theta$, no restriction on the way Δt, Δx, Δy go to zero,

2) if $0 \leqq \theta \leqq \frac{1}{2}$ and if we suppose $\Delta t/(\Delta x)^2$ and $\Delta t/(\Delta y)^2$ kept constant as Δt, Δx, $\Delta y \to 0$, then

$$2\Delta t \left[\frac{A}{(\Delta x)^2} + \frac{C}{(\Delta y)^2} \right] \leqq \frac{1}{1 - 2\theta}. \tag{40}$$

Since the matrix G has only one element, G commutes with G^* (in this particular example $G = G^*$), so that the Von Neumann condition is sufficient as well as necessary for stability.

This example can be generalized in various ways. For example one may include lower order terms, $D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu$, where D, E, F are constants, in the differential equation (38), and investigate their influence on stability. This is easily done because G is still a one-element matrix, but we omit details. It is found that for any reasonable manner of treating these terms in the finite difference equation, the stability condition is the same as before, except that sometimes the sign $\leqq$ in (40) has to be replaced by $<$. We may note, however, that if $F > 0$ it is important to have the Von Neumann condition in the form $R_1(\Delta t) \leqq 1 + O(\Delta t)$ rather than merely $R_1(\Delta t) \leqq 1$, because there are then generally true solutions of the differential equation which increase exponentially as t increases, and clearly we cannot expect (nor do we wish) to exclude such solutions from the numerical work.

19. The Du Fort-Frankel Equations

Du Fort and Frankel [12] have approximated the diffusion equation

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2} \qquad (\sigma = \text{constant} > 0) \tag{41}$$

by the difference equation

$$(42)\quad u^{n+1}(x)-u^{n-1}(x)=\frac{2\sigma\Delta t}{(\Delta x)^2}\left[u^n(x+\Delta x)-u^{n+1}(x)-u^{n-1}(x)+u^n(x-\Delta x)\right].$$

This system is of interest for two reasons, the first having to do with consistency and the second with stability. It is readily verified that the consistency condition of Section 5 is satisfied if and only if $\Delta t/\Delta x \to 0$ as $\Delta t,\ \Delta x \to 0$. In fact, if $\Delta t/\Delta x \to \mu$ where μ is a constant, it is clear that the difference equation (42) approximates the differential equation

$$\frac{\partial u}{\partial t}=\sigma\frac{\partial^2 u}{\partial x^2}-\mu^2\frac{\partial^2 u}{\partial t^2}$$

rather than (41). Just how large values of $\Delta t/\Delta x$ can be tolerated in practice is of course not settled by our argument.

To write (42) in the form required by our general theory, we must introduce another dependent variable, say $\phi(x)$, as follows:

$$u^{n+1}(x)=\phi^n(x)+\frac{2\sigma\Delta t}{(\Delta x)^2}\left[u^n(x+\Delta x)-u^{n+1}(x)-\phi^n(x)+u^n(x-\Delta x)\right],$$
$$\phi^{n+1}(x)=u^n(x).$$

The amplification matrix is

$$(43)\qquad G(\Delta t,\Delta x,k)=\begin{pmatrix}\dfrac{2\gamma}{1+\gamma}\cos k\Delta x & \dfrac{1-\gamma}{1+\gamma}\\ 1 & 0\end{pmatrix}$$

where $\gamma=\dfrac{2\sigma\Delta t}{(\Delta x)^2}$. The characteristic values of G are

$$(44)\qquad \lambda=\frac{\gamma\cos k\Delta x\pm\sqrt{1-\gamma^2\sin^2 k\Delta x}}{1+\gamma},$$

and it is readily found that a) the Von Neumann condition is always satisfied, and in fact $R_1(\Delta t)=1$, b) for any fixed value of γ, $R_2(\Delta t)=$ constant >1 so that condition (25) of Section 14 is of no use, c) the determinant Δ of the normalized eigenvectors of G vanishes when $\gamma\sin k\,\Delta x=1$, so that the condition (Theorem 2) of Section 15 is of no use, d) the criterion of Section 16 is satisfied because the lesser of the roots (44) is always bounded, in absolute value, by $\left|\dfrac{\gamma-1}{\gamma+1}\right|^{1/2}$, so that the Von Neumann condition is again sufficient as well as necessary for stability.

Therefore, the Du Fort-Frankel equations are always stable, but the time increment must be limited on account of the consistency condition.

20. Positive Operators

In this section we present a sufficient condition, due to Friedrichs [4], for the stability of certain difference schemes. The schemes considered are explicit two level schemes for vector-valued unknowns; i.e., the value of the approximate solution at time $t + h$ and position x is expressed as a linear combination of its computed values at the time t:

$$\mathbf{u}(t + \Delta t) = \sum_r c_r \mathbf{u}(\mathbf{x} + \Delta t \mathbf{d}_r) = C(\Delta t)\mathbf{u}. \tag{45}$$

In Friedrichs' theory the displacement vectors $\mathbf{d}_r$ (finite in number) need not lie on a lattice. The coefficient matrices c_r are functions of $\mathbf{x}$, and are required to satisfy the following conditions:

i) $\sum_r c_r(\mathbf{x}) = I$ (the identity matrix),

ii) $c_r(\mathbf{x})$ is symmetric and positive definite,

iii) $c_r(\mathbf{x})$ is a Lipschitz continuous function of the vector variable $\mathbf{x}$.

Conclusion: The difference scheme (45) is stable.

We reproduce Friedrichs' proof, and show that the norm of $C(\Delta t)$ with respect to the L^2 norm over the whole space, is bounded by

$$| C(\Delta t) | \leqq 1 + A\Delta t, \tag{46}$$

where the constant A depends on the Lipschitz constant of the c_r, and on the number of coefficients. Since stability means the uniform boundedness of $| C^n(\Delta t) |$, $n\Delta t \leqq T$, and $| C^n(\Delta t) | \leqq | C(\Delta t) |^n$, the estimate (46) implies stability.

To estimate $|| C ||$, Friedrichs uses this characterization of the norm of an operator:

$$|| C || = \text{Sup } (\mathbf{u}, C\mathbf{v}),$$
$$|| \mathbf{u} || = || \mathbf{v} || = 1,$$

where the bracket denotes the L^2 scalar product over the period parallelogram:

$$(\mathbf{u}, C\mathbf{v}) = \int \sum \mathbf{u}'(\mathbf{x}) c_r(\mathbf{x}) \mathbf{v}(\mathbf{x} + \Delta t \mathbf{d}_r) d\mathbf{x}. \tag{47}$$

It follows from this characterization of the norm of C that any upper bound for $(\mathbf{u}, C\mathbf{v})$ valid for all vectors $\mathbf{u}$ and $\mathbf{v}$ of unit length is an upper bound for $|| C ||$. We shall find an upper bound for $(\mathbf{u}, C\mathbf{v})$ from (47). Since the matrices c were assumed to be positive, we can apply the Schwarz inequality to the terms $\mathbf{u}'c\mathbf{v}$ in the integrand. We get, after throwing in the inequality about the arithmetic and the geometric mean,

$$\mathbf{u}'c\mathbf{v} \leqq \tfrac{1}{2}\mathbf{u}'c\mathbf{u} + \tfrac{1}{2}\mathbf{v}'c\mathbf{v}.$$

Substituting this into the integrand in (47) we have the following inequality:

$$(48)\quad (\mathbf{u}, C\mathbf{v}) \leqq \tfrac{1}{2}\sum \int \mathbf{u}'(\mathbf{x})c_r(\mathbf{x})\mathbf{u}(\mathbf{x}) + \tfrac{1}{2}\sum \int \mathbf{v}'(\mathbf{x} + \Delta t\mathbf{d}_r)c_r(\mathbf{x})\mathbf{v}'(\mathbf{x} + \Delta t\mathbf{d}_r).$$

The first term on the right in (48) is, on account of the requirement that $\Sigma\, c_r(\mathbf{x})$ is the identity, just $\frac{1}{2}(\mathbf{u}, \mathbf{u})$ which is $\frac{1}{2}$, since $\mathbf{u}$ has unit norm. In the second group of terms introduce $\mathbf{x}' = \mathbf{x} + \Delta t\mathbf{d}_r$ as new independent variable; we obtain

$$\tfrac{1}{2}\sum \int \mathbf{v}'(\mathbf{x}')c_r(\mathbf{x}' - \Delta t\mathbf{d}_r)\mathbf{v}(\mathbf{x}').$$

If in the above expression we replace $c_r(\mathbf{x}' - \Delta t\mathbf{d}_r)$ by $c_r(\mathbf{x}')$, the error committed is at most a constant times Δt, on account of the assumed Lipschitz continuity of the coefficients c. Imagine such a substitution performed, and treat the resulting group of terms the same way as the first group of terms. This way we find that the value of the second group of terms is at most $1/2 +$ const. Δt, and have the desired $1 +$ const. Δt estimate for the whole expression (48).

Such symmetric positive difference operators come up in difference approximations to solutions of symmetric hyperbolic systems, i.e., equations of the form

$$(49)\qquad \mathbf{u}_t + a_k\mathbf{u}_{x^k} + b\mathbf{u} = 0,$$

where the coefficients a_k are symmetric matrices. A majority of the equations of mathematical physics which describe reversible phenomena are of this form; the general theory of such equations has been developed by Friedrichs (loc. cit), where he gives a recipe for associating a positive symmetric operator to each symmetric hyperbolic operator. We give here such a recipe:

Take for the displacement vectors $\mathbf{d}_r$ the $2d$ vectors

$$\mathbf{d}_r = \pm\ (0,\ 0, \cdots \lambda_r\ ,\ 0, \cdots, 0), \qquad r = 1, \cdots, d.$$

Here the λ_r are arbitrary constants, the side lengths of a rectangular lattice in $\mathbf{x}$-space. Replace the $\mathbf{x}$-space derivative $\mathbf{u}_{x^k}$ in the differential equation (49) by centered difference quotients between $\mathbf{x} + \Delta t\mathbf{d}_k$ and $\mathbf{x} - \Delta t\mathbf{d}_k$, and the time derivative by the forward difference quotient $\mathbf{u}(\mathbf{x}, t + \Delta t) - \bar{\mathbf{u}}(\mathbf{x}, t)$ where $\bar{\mathbf{u}}$ is the weighted average $\frac{1}{2}\alpha_k\mathbf{u}(\mathbf{x} \pm h\mathbf{d}_k\ , t)$, the sum of the weights α being one. The resulting difference equation can be solved for $\mathbf{u}(\mathbf{x}, t + \Delta t)$:

$$u(x,\ t + \Delta t) = Cu(x,\ t) = \sum c_{\pm r}u(x \pm \lambda_r\Delta t),$$

where[3]

$$c_{\pm r} = \tfrac{1}{2}\{\alpha_r I \mp \lambda_r^{-1} a_r\}.$$

[3] For the sake of simplicity we have taken b, the coefficient of $\mathbf{u}$ in (49), to be zero.

Clearly, if the α_r are fixed positive constants, the coefficients $c_{\pm r}$ can be made positive definite by takıng λ_r large enough. Of course in practice it is the space mesh that stays constant and Δt is made small enough.

Bibliography

[1] Courant, R., Friedrichs, K. O., and Lewy, H., *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann., Vol. 100, 1928, pp. 32–74.

[2] Courant, R., Isaacson, E., and Rees, M., *On the solution of nonlinear hyperbolic differential equations by finite differences*, Comm. Pure Appl. Math., Vol. 5, 1952, pp. 243–255.

[3] Eddy, R. P., *Stability in the numerical solution of initial value problems in partial differential equations*, Naval Ordnance Laboratory Memorandum 10232.

[4] Friedrichs, K. O., *Symmetric hyperbolic linear differential equations*, Comm. Pure Appl. Math., Vol. 7, 1954, pp. 345–392.

[5] Kantorovitch, L. V., *Functional analysis and applied mathematics*, Uspehi Matem. Nauk, Vol. 3, 1948, p. 89; Bureau of Standards Report 1509, 1952.

[6] Keller, J. B., and Lax, P. D., *The initial and mixed initial and boundary value problems for hyperbolic systems*, Los Alamos Report No. 1210, 1951.

[7] Laasonen, *Über eine Methode zur Lösung der Wärmeleitungsgleichung*, Acta Math.,Vol. 81, 1949, pp. 309–317.

[8] O'Brien, G. G., Hyman, M. A., and Kaplan, S., *A study of the numerical solution of partial differential equations*, J. Math. Physics, Vol. 29, 1951, pp. 223–251.

[9] Lewy, H., *On the convergence of solutions of difference equations*, in *Studies and Essays*, Courant Anniversary Volume, Interscience Publishers, New York, 1948, pp.211–214.

[10] Von Neumann, J., and Richtmyer, R. D., *A method for the numerical calculation of hydrodynamic shocks*, J. Appl. Physics, Vol. 21, 1950, pp. 232–237.

[11] Thomas, L. H., *Stability of partial differential equations*, Symposium on Theoretical Compressible Flow, U. S. Naval Ordnance Laboratory, 1949, NOLR 1132, 1950.

[12] Du Fort, E. C., and Frankel, S. P., *Stability conditions in the numerical treatment of parabolic differential equations*, Math. Tables and Other Aids to Computation, Vol. 7, 1953, pp. 135–152.

[13] John, F., *On integration of parabolic equations by difference methods*, Comm. Pure Appl. Math., Vol. 5, 1952, pp. 155–211.

[14] Lax, P. D., *Difference approximation to solutions of linear differential equations – an operator theoretical approach*, Symposium on Partial Differential Equations, Berkeley, Summer 1955; Report of the University of Kansas Mathematics Department (to appear).

Received October, 1955.