TRIGONOMETRIC POLYNOMIALS AND DIFFERENCE METHODS OF MAXIMUM ACCURACY*

By Gilbert Strang

1. The fundamental problem in the theory of partial difference approximations to hyperbolic equations is the question of convergence to the true solution through the use of an increasingly fine mesh. We shall examine this question for those explicit methods of greatest interest, the most accurate ones. Happily, our answer is very different in kind from that given by Dahlquist [2] to a similar question for ordinary differential equations. There, the order of accuracy that appears feasible must be reduced by nearly half to achieve stability. In our case, the most exact methods prove to be stable under quite reasonable restrictions on the mesh ratio.

Both the statement and the proof of our chief analytical result (Theorem 1) can be presented without drawing on the possibly unfamiliar vocabulary of the theory discussed above. We therefore do so immediately, hoping the theorem will be found interesting in itself.

Bernstein's inequality for the derivative of a trigonometric polynomial is applied in the latter half of the paper to establish the connection between two standard necessary conditions for convergence. For systems of equations, we reverse the reasoning to obtain a sort of Bernstein theorem (Theorem 6) for the spectral radius when the polynomial's coefficients are matrices.

We owe to Professor Szegö the idea which has very much improved the last step in the proof of Theorem 1, namely to expand a polynomial in terms of its differences. We are also grateful to Professor McCarthy for a good many useful discussions, and to Professor Lax for a copy of his important paper [4] in advance of its publication.

2. Consider the polynomial $p(y, \theta) = \sum_{-N}^{N} a_j(y)e^{ij\theta}$, $N > 0$, defined by either of the properties:

i) $p = e^{i\theta y} + O(\theta^{2N+1})$

ii) p has degree $2N$ in y, $p(r, \theta) = e^{ir\theta}$ for $r = -N, -N + 1, \cdots, N$.

If condition i) is met, equate the coefficients of θ^n to find $\sum_{-n}^{n} j^n a_j = y^n$, $0 \leqq n \leqq 2N$. The coefficients a_j are evidently those needed for Lagrange's interpolation formula at $2N + 1$ equally spaced points, which is exactly the assertion of condition ii). To be specific,

$$a_j = \prod_{k \neq j} \frac{y - k}{j - k}$$

* The author is a NATO Postdoctoral Fellow for 1961–62 at the Mathematical Institute, Oxford, England. This research was supported by the United States Air Force under contract AF-49(638)-42, monitored by the Office of Scientific Research of the Air Research and Development Command.

Similarly, we define $q(y, \theta) = \sum_{-N+1}^{N} b_j(y) e^{i(2j-1)\theta}$ by either of

i) $q = e^{i\theta y} + O(\theta^{2N})$

ii) q has degree $2N - 1$ in y, $q(r, \theta) = e^{ir\theta}$ for $r = \pm 1, \pm 3, \cdots, \pm 2N - 1$.

Theorem 1. If θ is real and $-1 \leqq y \leqq 1$, then $|p| \leqq 1$ and $|q| \leqq 1$.

PROOF. We first write out explicitly $p = c + is$, where $c(y,\theta) = \Sigma a_j(y) \cos j\theta$, $s(y, \theta) = \Sigma a_j(y) \sin j\theta$, to see that

a) c is even and s is odd in y (using $a_j(-y) = a_{-j}(y)$) so that in particular s has degree $<2N$,

b) $|p|^2 = c^2 + s^2$ is even in θ and y, with period 2π in θ, and $p(y, 0) = 1$, so it will be sufficient to show that for $0 \leqq y \leqq 1$ and $0 \leqq \theta \leqq \pi$, we have $(\partial/\partial\theta)(c^2 + s^2) \leqq 0$.

The key step is the fact that c and s share the following property of $\cos \theta y$ and $\sin \theta y$:

$$\frac{\partial c}{\partial \theta} = -ys, \quad \text{or} \quad -\Sigma j a_j \sin j\theta = -y \Sigma a_j \sin j\theta. \tag{1}$$

Both sides have degree $\leqq 2N$ in y, and there is agreement at $y = -N, \cdots, N$, so they are identically equal. Therefore

$$\frac{\partial}{\partial \theta}(c^2 + s^2) = 2\left(c \frac{\partial c}{\partial \theta} + s \frac{\partial s}{\partial \theta}\right) = -2s\left(yc - \frac{\partial s}{\partial \theta}\right). \tag{2}$$

We take the factors on the right separately; first we show

$$yc - \frac{\partial s}{\partial \theta} = \Sigma(y - j) a_j \cos j\theta = \frac{(-1)^N}{2N!}\left(2 \sin \frac{\theta}{2}\right)^{2N} \prod_{-N}^{N} (y - j). \tag{3}$$

Since both sides vanish at $y = -N, \cdots, N$, we need only verify that the coefficients of y^{2N+1} are equal. On the left we have

$$\begin{aligned} \sum_{-N}^{N} \frac{\cos j\theta}{\prod_{k \neq j}(j - k)} &= \sum_{-N}^{N} \frac{(-1)^{N-j} \cos j\theta}{N - j!\, N + j!} = \operatorname{Re} \sum_{-N}^{N} \frac{(-1)^{N-j} e^{ij\theta}}{2N!} \binom{2N}{N + j} \\ &= \operatorname{Re} \frac{(e^{i\theta/2} - e^{-i\theta/2})^{2N}}{2N!} = \frac{(-1)^N}{2N!}\left(2 \sin \frac{\theta}{2}\right)^{2N}. \end{aligned} \tag{4}$$

We remark that another way to establish the θ-dependence is to use $c = \cos \theta y + O(\theta^{2N+2})$, $s = \sin \theta y + O(\theta^{2N+1})$, so that $yc - \partial s/\partial\theta = O(\theta^{2N})$. In the same way, we get $\partial c/\partial\theta + ys = O(\theta^{2N+1})$, which yields an alternative proof of (1).

The final step is to find an expression for the other factor in (2), namely s. We first calculate, for $-N \leqq t \leqq t + k \leqq N$,

$$\Delta^k p(t) = p(t + k) - \binom{k}{1} p(t + k - 1) + \cdots = e^{i\theta t}(e^{i\theta} - 1)^k. \tag{5}$$

Now substituting in the Gauss forward formula for an odd polynomial of degree $<2N$ we get

$$
\begin{aligned}
s(y) &= \sum_1^N \binom{y+k-1}{2k-1} \Delta^{2k-1} s(-k+1) \\
&= \operatorname{Im} \sum_1^N \binom{y+k-1}{2k-1} e^{i\theta(1-k)}(e^{i\theta}-1)^{2k-1} \qquad (6) \\
&= \sin\theta \sum_1^N \frac{y(1^2-y^2)\cdots((k-1)^2-y^2)}{(2k-1)!}\left(2\sin\frac{\theta}{2}\right)^{2k-2}
\end{aligned}
$$

Since (3) and (6) are non-negative if $0 \leqq y \leqq 1$ and $0 \leqq \theta \leqq \pi$, we have from (2) that $(\partial/\partial\theta)(c^2 + s^2) \leqq 0$.

The proof that $|q| \leqq 1$ follows the same line; this time $|q|^2 = C^2 + S^2$ has period π in θ, and to show that $(\partial/\partial\theta)(C^2 + S^2) \leqq 0$ if $-1 \leqq y \leqq 1$ and $0 \leqq \theta \leqq \pi/2$, we derive the following formulas:

$$\frac{\partial S}{\partial\theta} = yC \tag{7}$$

$$\frac{\partial(C^2+S^2)}{\partial\theta} = 2C\left(Sy + \frac{\partial C}{\partial\theta}\right) \tag{8}$$

$$Sy + \frac{\partial C}{\partial\theta} = \frac{(-1)^{N-1}\sin^{2N-1}\theta}{(2N-1)!}\prod_1^N (y^2-(2j-1)^2) \leqq 0 \tag{9}$$

$$C = \cos\theta\left[1 + \sum_1^{N-1} \frac{(1^2-y^2)(3^2-y^2)\cdots((2k-1)^2-y^2)}{2k!}\sin^{2k}\theta\right] \geqq 0. \tag{10}$$

The applications to follow require that we make a note of two details:

i) For $0 < y < 1, 0 < \theta < \pi$, we have $(\partial/\partial\theta)(c^2 + s^2) < 0$; therefore $|p| < 1$ if $0 < |y| < 1$ and $\theta \not\equiv 0 \pmod{2\pi}$.

ii) $|p|^2 = 1 - \dfrac{4N+2}{(2N+2)!}y^2(1^2-y^2)\cdots(N^2-y^2)\theta^{2N+2} + O(\theta^{2N+4})$ for θ near 0.

3. We consider now certain difference analogues of a hyperbolic differential equation of the form

$$\frac{\partial u}{\partial t} = G(x,t)\frac{\partial u}{\partial x}, \qquad u(x,0) = f(x), \quad -\infty < x < \infty. \tag{11}$$

G is a matrix with real eigenvalues $\lambda_i(x, t)$, and u represents a vector variable. An approximation $U_n(x) \sim u(x, n\Delta)$ is provided by the system

$$U_{n+1}(x) = \textstyle\sum_{j=-N}^{N} A_j(x, t, \Delta)\, U_n(x + j\Delta), \qquad U_0(x) = f(x) \tag{12}$$

if we choose the matrices A_j appropriately.

Suppose first that G is constant. Substituting a solution u of (11) into (12) and equating coefficients of Δ^n in a Taylor expansion gives

$$\Sigma j^n A_j = G^n \qquad n = 0, 1, \cdots \tag{13}$$

Clearly the only way to satisfy (13) for $n \leqq 2N$ is to take $A_j = a_j(G)$, where a_j is the Lagrange coefficient from Theorem 1. Of course this formal local accuracy, of maximum order, is insufficient *per se* to assure the convergence of U to u, which we measure in the norm

$$|U - u|^2 = \sup_{n\Delta \leqq 1} \int_{-\infty}^{\infty} |U_n(x) - u(x, n\Delta)|^2 \, dx. \tag{14}$$

Theorem 2. If G is similar to a diagonal matrix, and every eigenvalue satisfies $|\lambda_i| \leqq 1$, then for the most accurate difference method $|U - u| \to 0$ with Δ for all f in L_2.

For the proof, we repeat in summary an argument suggested by von Neumann and developed in detail in [5] for our present case of constant coefficients. Denote by $S(\Delta)$ the linear operator on L_2 defined by the difference equation (12), that is

$$U_{n+1}(x) = S(\Delta) U_n(x).$$

Under a Fourier transformation of L_2, taking $U_n(x)$ into $\hat{U}_n(\xi)$, the operator $S(\Delta)$ is unitarily equivalent to multiplication by the *amplification matrix*

$$\hat{S}(\xi, \Delta) = \Sigma A_j e^{ij\xi\Delta} = p(G, \xi\Delta). \tag{15}$$

Now diagonalize G, say $T^{-1}GT = \operatorname{diag}(\lambda_1, \cdots, \lambda_k)$. Then $T^{-1}\hat{S}^n T = \operatorname{diag}(p(\lambda_1^n, \xi\Delta), \cdots, p(\lambda_k^n, \xi\Delta))$. Since all $|\lambda_i| \leqq 1$, we may use Theorem 1 to obtain $\| T^{-1}\hat{S}^n T \| \leqq 1$, or $\| S^n \| = \| \hat{S}^n \| \leqq \| T \| \, \| T^{-1} \|$.

Thus the powers of S are uniformly bounded; this is called *stability*, and invoking the Equivalence Theorem of [5], we have established Theorem 2.

Theorem 3. If every eigenvalue of G satisfies $|\lambda_i| \leqq 1$, and the data f is sufficiently differentiable in L_2, then for the most accurate difference method $|U - u| = O(\Delta^{2N})$.

Proof. We again need to estimate $\| \hat{S}^n \|$ for $n\Delta \leqq 1$, where $S = p(G, \xi\Delta)$. Suppose M is a unitary matrix, chosen so that $M^{-1}GM$ is triangular. Then $M^{-1}\hat{S}M = p(M^{-1}GM, \xi\Delta)$ will also be triangular, and the diagonal entries satisfy $|p(\lambda_i, \xi\Delta)| \leqq 1$ as before. We can easily bound the off-diagonal elements independently of ξ and Δ by the corresponding elements of $\Sigma |a_j| (|M^{-1}GM|)$, taking the modulus of every coefficient of the a_j and every entry in $M^{-1}GM$. Since now $M^{-1}\hat{S}M$ is majorized by a fixed matrix with k-fold eigenvalue 1, we conclude

$$\| S^n \| = \| \hat{S}^n \| = \| (M^{-1}\hat{S}M)^n \| = O(n^{k-1}) = O(\Delta^{1-k}). \tag{16}$$

We called this property *s-stability* in [6], and proved its equivalence to the conclusion we now reach, that $|U - u| = O(\Delta^{2N})$ for sufficiently smooth initial data.

Notice that when G has non-simple eigenvalues, the original differential problem (11) is well-posed only in a norm which takes account of enough derivatives of f, so that some differentiability condition in the theorem was to be expected.

With a powerful assist from [4] we can extend these results to certain cases when $G = G(x, t)$ is a smoothly varying matrix function. To maintain the local

order of accuracy at $2N$ we certainly must vary the coefficients A_j in the difference scheme. If, for example, we take $A_j(x, t, \Delta)$ to be a polynomial of degree $2N$ in Δ, then the sort of formal Taylor expansion of the difference equation described earlier gives the conditions to be satisfied by the A_j to provide maximum accuracy. Our whole concern is with the conditions on $A_j(x, t, 0)$, the *principal part* of $A_j(x, t, \Delta)$. Those conditions are precisely

$$\sum_{-N}^{N} j^n A_j(x, t, 0) = G^n(x, t), \qquad 0 \leqq n \leqq 2N,$$

so we again find $A_j(x, t, 0) = a_j(G(x, t))$.

Theorem 4. If $G(x, t)$ is real and symmetric, with eigenvalues satisfying $0 < |\lambda_i(x, t)| < 1$, then for a most accurate difference method $|U - u| = O(\Delta^{2N})$, provided G and f have a sufficient number of bounded derivatives in L_2.

All that the proof requires is the observation that by Theorem 1 and the notes which follow it, the hypotheses of Theorem 3.1 of [4] are satisfied; more precisely, we verify that for $\hat{S} = p(G, \xi\Delta)$,

i) $\hat{S}$ is Lipschitz continuous in x, uniformly in ξ, if G has a single bounded derivative;

ii) the eigenvalues of $\hat{S}$ satisfy $|p(\lambda_i(x, t), \xi\Delta)| < 1$ for all x, t, if $\xi\Delta \not\equiv 0 \pmod{2\pi}$;

iii) $$\hat{S}^*\hat{S} = I - \frac{4N + 2}{(2N + 2)!} G^2(I - G^2) \cdots (N^2 I - G^2)(\xi\Delta)^{2N+2} + 0[(\xi\Delta)^{2N+4}],$$

and the coefficient of $(\xi\Delta)^{2N+2}$ is negative definite for all x and t.

Lax's theorem asserts that such a method is stable, from which it follows that $|U - u| = O(\Delta^{2N})$ for smooth enough G and f.

It is perhaps noteworthy that the most accurate wholly implicit scheme, namely

$$\Sigma a_j(G)\, U_{n+1}(x - j\Delta) = U_n(x),$$

is completely *unstable* if $0 < |\lambda_i| < 1$ for any λ_i, since the stability condition is here reversed to $|p(\lambda_i, \xi\Delta)| \geqq 1$. This is a rather unexpected consequence of Theorem 1, since in familiar parabolic problems implicit methods are the more stable.

A number of inessential conventions were introduced in our formulation of the difference equation (12). The use of different mesh sizes for space and time would involve only the replacement of the condition $|\lambda_i| \leqq 1$ by $|\lambda_i\, \Delta t/\Delta x| \leqq 1$. Translation is just as easy; if $U_{n+1}(x) = \Sigma A_j U_n(x + x_0 + j\Delta)$, the most accurate methods are stable provided $x_0 - 1 \leqq \lambda_i \leqq x_0 + 1$. It is convenient, but not important, that U should be defined for all x; this point is discussed at some length in [5]. Finally, if U_{n+1} is determined from the values of U_n at an *even* number of equally spaced points, we deal with q instead of p, and the conclusions parallel those of our three previous theorems.

4. Our concern now is with the simplest equation of the form $u_t = Gu_x$, that is, when G is a scalar constant. The difference analogue

$$U_{n+1}(x) = \sum_{j=N_1}^{N_2} B_j U_n(x + j\Delta), \qquad B_{N_1}B_{N_2} \neq 0, \tag{17}$$

is consistent (has a positive order of accuracy) if

$$\Sigma B_j = 1, \qquad \Sigma j B_j = G. \tag{18}$$

In this scalar case, von Neumann's condition

$$| \Sigma B_j e^{ij\theta} | \leqq 1 \quad \text{for all real } \theta \tag{19}$$

is equivalent to stability, and therefore (18) and (19) are necessary and sufficient for convergence. So it is not surprising that we can prove

Theorem 5. The von Neumann necessary condition for the stability of (17), together with the consistency conditions, implies the Courant-Friedrichs-Lewy necessary condition.

PROOF. Consider the polynomial $F(\theta) = e^{-i(N_1+N_2)\theta/2}\Sigma B_j e^{ij\theta}$. The classical Bernstein inequality for trigonometric functions of degree $(N_2 - N_1)/2$ is

$$\sup_\theta \left| \frac{\partial F}{\partial \theta} \right| \leqq \left(\frac{N_2 - N_1}{2} \right) \sup_\theta | F |$$

We evaluate the left side at $\theta = 0$, and use (18) and (19) to obtain

$$\begin{aligned} \left| \sum i\left(j - \frac{N_1 + N_2}{2}\right) B_j \right| &\leqq \frac{N_2 - N_1}{2} \\ \left| G - \frac{N_1 + N_2}{2} \right| &\leqq \frac{N_2 - N_1}{2}, \qquad N_1 \leqq G \leqq N_2. \end{aligned} \tag{20}$$

This result is easily interpreted as the Courant-Friedrichs-Lewy condition [1] that the domain of dependence of $U_n(x)$ should contain that of $u(x, n\Delta)$. On the one hand, $U_n(x)$ depends on U_{n-1} along the segment from $x + N_1\Delta$ to $x + N_2\Delta$, and thus ultimately on the initial data $U_0 = f$ on the segment between $x + nN_1\Delta$ and $x + nN_2\Delta$. Solving the original equation $u_t = Gu_x$, we have $u(x, n\Delta) = f(x + Gn\Delta)$, so that the second domain of dependence is the single point $(x + Gn\Delta, 0)$. Obviously this point lies in the first domain exactly when $N_1 \leqq G \leqq N_2$.

One can obtain the same result with several independent variables, where the only new device needed is a suitable rotation of coordinates. If there are k dependent variables, $k > 1$, the situation is quite different. It is easy to generalize the Bernstein inequality to polynomials $F(\theta) = \sum_{-N}^{N} B_j e^{ij\theta}$ with matrix coefficients:

$$\sup_\theta \left\| \frac{\partial F}{\partial \theta} \right\| \leqq N \sup_\theta \| F \| \tag{21}$$

in any matrix norm induced by a vector norm. Unfortunately, it is with the spectral radius (the maximum modulus of the eigenvalues), and not with any norm, that the Courant-Friedrichs-Lewy condition has to do. The inequality (21) becomes false if the norm is replaced by the spectral radius; the simplest example is

$$F = \begin{pmatrix} 1 & e^{-i\theta} \\ -e^{i\theta} & -1 \end{pmatrix}, \qquad \rho(F) = 0, \qquad \rho\left(\frac{\partial F}{\partial \theta}\right) = 1.$$

However, we may use the theory of difference equations to find in this context a partial analogue of Bernstein's inequality.

Theorem 6. If $F(0) = I$, and $\rho(F(\theta)) \leqq 1$ for all θ, then $\rho[(\partial F/\partial\theta)(0)] \leqq N$.

PROOF. If we define $G = \sum_{-N}^{N} jB_j$, then (17), with $-N_1 = N_2 = N$, is a difference analogue of $u_t = Gu_x$ which satisfies von Neumann's condition $\rho(F(\theta)) \leqq 1$. For each θ, $F(\theta)$ is unitarily similar to an upper triangular matrix $T(\theta)$. As in the proof of Theorem 3, we may construct a single triangular matrix T, say $T_{ij} = \sup |T(\theta)_{ij}|$, which majorizes all the $T(\theta)$. Since $\rho(T) = 1$, $\| T^n \| = O(n^{k-1}) = O(\Delta^{1-k})$. The difference method is thus s-stable, and will converge for sufficiently smooth initial data.

Therefore the Courant-Friedrichs-Lewy condition on the domains of dependence must hold, or we could modify the data outside the domain of dependence of U without destroying differentiability, and discover that the same approximation U converged to a variety of true solutions u. The domain of dependence of $u(x, n\Delta)$ is easy to determine; it is the set of points $(x + \lambda_i n\Delta, 0)$, where the λ_i are the eigenvalues of G. These points lie on the segment from $x - nN\Delta$ to $x + nN\Delta$ provided

$$N \geqq \sup |\lambda_i| = \rho(G) = \rho[(\partial F/\partial\theta)(0)].$$

It should be remarked that for the most accurate methods studied in the previous section, the restriction $|\lambda_i| \leqq 1$ which we were compelled to impose is, for $N > 1$, stricter than the Courant-Friedrichs-Lewy condition would require.

Our final application of the Bernstein theorem is to difference analogues of the (scalar) heat equation

$$u_t = gu_{xx}, \qquad u(x, 0) = f(x), \qquad g > 0. \tag{22}$$

A consistent approximation $U_n(x) \sim u(n, x\Delta t)$ is furnished by

$$U_{n+1}(x) = \sum_{N_1}^{N_2} C_j U_n(x + j\Delta x) \tag{23}$$

under the conditions

$$\Sigma C_j = 1, \qquad \Sigma j C_j = 0, \qquad \Sigma (j\Delta x)^2 C_j = 2g\Delta t. \tag{24}$$

Theorem 7. If the difference method (23) is convergent, then $2g\Delta t \leqq -N_1 N_2 (\Delta x)^2$.

PROOF. Convergence implies, as usual, that $|\Sigma C_j e^{ij\theta}| \leqq 1$ for all θ. Applying the Bernstein theorem twice to $e^{-i(N_1+N_2)\theta/2}\Sigma C_j e^{ij\theta}$, and evaluating the second derivative at $\theta = 0$, we have

$$\left|\Sigma i^2 \left(j - \frac{N_1 + N_2}{2}\right)^2 C_j\right| \leqq \left(\frac{N_2 - N_1}{2}\right)^2$$

$$\left|\frac{2g\Delta t}{(\Delta x)^2} + \left(\frac{N_1 + N_2}{2}\right)^2\right| \leqq \left(\frac{N_2 - N_1}{2}\right)^2 \tag{25}$$

$$2g\Delta t \leqq -N_1 N_2 (\Delta x)^2.$$

In the case of a three-point method in which $-N_1 = N_2 = 1$, this restriction has long been known.

BIBLIOGRAPHY

[1] R. COURANT, K. O. FRIEDRICHS AND H. LEWY, Über die partiellen Differenzengleichungen der mathematischen Physik, Math. Ann., **100** (1928) 32–74.

[2] G. DAHLQUIST, Convergence and stability in the numerical integration of ordinary differential equations, Math. Scandinavica, **4** (1956) 33–53.

[3] P. D. LAX, Difference approximation to solutions of linear differential equations—an operator theoretical approach, Berkeley Symposium on Partial Differential Equations, Report of the University of Kansas, 1957.

[4] P. D. LAX, On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients, Comm. Pure and Appl. Math., **14** (1961).

[5] P. D. LAX AND R. D. RICHTMYER, Survey of the stability of linear finite difference equations, Comm. Pure and Appl. Math., **9** (1956) 267–293.

[6] W. G. STRANG, Difference methods for mixed boundary-value problems, Duke Math. J., **27** (1960) 221–232.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

(Received June 30, 1961)