TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 277, Number 2, June 1983

THE OPTIMAL ACCURACY OF DIFFERENCE SCHEMES

BY

ARIEH ISERLES AND GILBERT STRANG[1]

ABSTRACT. We consider difference approximations to the model hyperbolic equation $u_t = u_x$ which compute each new value $U(x, t + \Delta t)$ as a combination of the known values $U(x - r\Delta x, t), \dots, U(x + s\Delta x, \Delta t)$. For such schemes we find the optimal order of accuracy: stability is possible for small $\Delta t/\Delta x$ if and only if $p \leq \min\{r + s, 2r + 2, 2s\}$. A similar bound is established for implicit methods. In this case the most accurate schemes are based on Padé approximations $P(z)/Q(z)$ to z^λ near $z = 1$, and we find an expression for the difference $|Q|^2 - |P|^2$; this allows us to test the von Neumann condition $|P/Q| \leq 1$. We also determine the number of zeros of Q in the unit circle, which decides whether the implicit part is uniformly invertible.

1. Introduction. This note returns to a theme which continues to appear in the construction of finite difference equations, the conflict between accuracy and stability. In some cases there is a severe limit on the order of accuracy, a limit which we reach in practice and would like to exceed. This occurs in both ordinary and partial difference methods, and we mention three examples:

(1) An A-stable multistep method for $u_t = \lambda u$ cannot have accuracy $p > 2$ (Dahlquist [**2**]).

(2) A method for $u_t = u_x$ with nonnegative coefficients cannot have accuracy $p > 1$ (Lax [**12**]; also in the Russian literature and in [**20**]).

(3) A stable one-sided method for $u_t = u_x$—in which $u(x, 0)$ can influence the approximation $U(x', t)$ only if $x \geq x'$—cannot have accuracy $p \geq 2$ (Strang [**15**], Engquist and Osher [**6**], Iserles [**9**]).

There are other examples, including one of the first and most celebrated of all, in which higher accuracies are involved; we refer to the bound $p \leq 2[(k + 2)/2]$ for a k-step difference method, in which it is zero-stability and not A-stability that is required (Dahlquist [**3**]). This paper gives the corresponding bounds for hyperbolic equations.

First we comment briefly on the relation between $u_t = \lambda u$ and $u_t = u_x$, illustrated by (1) and (3) above.

With $\lambda = i\omega$, and therefore $u = e^{i\omega t}$, the accuracy of a multistep method is determined by substituting this solution into the difference equation:

$$e^{i\omega\Delta t} = \sum_{j \leq 0} c_j e^{ij\omega\Delta t} + O(\Delta t^{p+1}). \tag{1.1}$$

Received by the editors June 9, 1982.

1980 *Mathematics Subject Classification*. Primary 65M10, 33A30.

[1] The work of the second author was supported by the Army Research Office (DAAG 29-80-K0033) and the National Science Foundation (MCS 81-02371).

By comparison, with initial condition $e^{i\omega x}$ in the equation $u_t = u_x$, accuracy for a one-sided method depends on

$$e^{i\omega\Delta t} = \sum_{j\geqslant 0} c_j e^{ij\omega\Delta x} + O(\Delta x^{p+1}). \tag{1.2}$$

There is the difference between Δt and Δx, and the definitions of stability are not identical, but what is most striking is the shift from $j \leqslant 0$ to $j \geqslant 0$. The former is a problem in extrapolation and the latter in interpolation. The Courant-Friedrichs-Lewy condition about domains of dependence would forbid stability in (1.2) if we switched to $j \leqslant 0$; the true solution is $u = e^{i\omega(x+t)}$, depending as always on the data at points $x + t \geqslant x$. This right side (x, ∞) would be ignored, and the left side $(-\infty, x)$ used exclusively, if we changed to $j \leqslant 0$ in (1.2); stability and convergence could not survive.

One may therefore wonder how A-stability can ever be achieved. It is only because the coefficients c_j are allowed to depend on λ (or ω), which is unthinkable in the partial differential problem; there ω was in the initial condition and outside of our control, whereas it appears directly in the ordinary differential equation $u_t = \lambda u$ and therefore in its difference analogue. In the trapezoidal scheme we have

$$\frac{u_{n+1} - u_n}{\Delta t} = \frac{(\lambda u)_{n+1} + (\lambda u)_n}{2}, \quad \text{or} \quad u_{n+1} = \frac{2 + \lambda\Delta t}{2 - \lambda\Delta t} u_n,$$

and this coefficient c_0 is a rational function of λ. It gives $p = 2$ with A-stability ($|c_0| \leqslant 1$ for $\operatorname{Re}\lambda \leqslant 0$), and no more complicated multistep formula can do better.[2]

This digression will return us to the real subject of the paper, if we ask the corresponding question for hyperbolic equations. It is this: If instead of one-sided schemes we allow the use of r values to the left of the origin and s values to the right, what order of accuracy can be combined with stability? If $r = 0$, we are back to question (3), and the Lax-Wendroff scheme is optimal; it achieves $p = 2$. (This is the uncentered scheme with $s = 2$; the normal centering with $r = s = 1$ is also stable with $p = 2$, and works better.) If $r > 0$ then we are using points outside the domain of dependence of $u_t = u_x$, on the left where no characteristics emerge instead of on the right, intending to improve the accuracy and still retain stability. The goal is to decide what is possible. There is a special choice of coefficients $c_{-r}, \ldots, c_s$ which gives the maximal accuracy $p = r + s$, but it will not always be stable.

We concentrate first on explicit schemes for $u_t = u_x$:

$$U(x, t + \Delta t) = \sum_{-r}^{s} c_j U(x + j\Delta x, t). \tag{1.3}$$

The coefficients c_j depend on the ratio $\mu = \Delta t/\Delta x$, and the von Neumann condition is the test for stability: We let $U(x, 0) = e^{i\omega x}$, so that $U(x, \Delta t) = (\Sigma c_j e^{ij\omega\Delta x})e^{i\omega x} = ae^{i\omega x}$, and this *amplification factor a* cannot exceed unity. Otherwise, since $U(x, n\Delta t) = a^n U(x, 0)$, the solution will explode. With $\theta = \omega\Delta x$, stability therefore means

$$|a(e^{i\theta})| = |\sum c_j e^{ij\theta}| \leqslant 1 \quad \text{for all real } \theta. \tag{1.4}$$

[2] The denominator and numerator are kept linear in λ—otherwise Dahlquist's barrier $p = 2$ is raised [**19**].

To determine the accuracy, we compare this same factor with the corresponding factor $e^{i\omega\Delta t} = e^{i\mu\theta}$ which appears at time Δt in the true solution $u = e^{i\omega(x+t)}$. Accuracy of order p means agreement through the term in θ^p:

$$\sum c_j e^{ij\theta} = e^{i\mu\theta} + O(\theta^{p+1}). \tag{1.5}$$

Finally, we agree to stay within the range $\Delta t < \Delta x$, or $0 < \mu < 1$. Nothing is sacrificed; if an equation (1.3) with coefficients c'_j is stable for $\mu' = \frac{3}{2}$, say, then by shifting each coefficient to the left $(c_j = c'_{j+1})$ we are back to $\mu = \frac{1}{2}$. Multiplying throughout by $e^{-i\theta}$ brings no change in accuracy or stability:

$$\sum_{-r}^{s} c'_j e^{ij\theta} \approx e^{i\mu'\theta} \quad \text{if and only if} \quad \sum_{-(r+1)}^{s-1} c_j e^{ij\theta} \approx e^{i\mu\theta},$$

$$\left|\sum c'_j e^{ij\theta}\right| \leqslant 1 \quad \text{if and only if} \quad \left|\sum c_j e^{ij\theta}\right| \leqslant 1.$$

Of course infinite accuracy is achieved for $\Delta x = \Delta t$ in the special case $U(x, t + \Delta t) = U(x + \Delta x, t)$, which is identically satisfied by the general solution $u(x, t) = u_0(x + t)$—but this cannot extend from the model equation $u_t = u_x$ to the real problem of general hyperbolic systems.

The problem is to reconcile (1.4) and (1.5): stability and accuracy. For $s = 0$ it is impossible by the Courant-Friedrichs-Lewy condition, and $p = 0$ is optimal. For $r = 0$, the maximal is $p = 2$. For general r and s, we prove

THEOREM 1. *The maximum order of accuracy of a stable scheme is*

$$p = \min\{r + s, 2r + 2, 2s\}. \tag{1.6}$$

There are two steps in the proof. The first is to show that higher accuracy is not possible. Here we rely on the theory of order-stars. This set of ideas was introduced in [**19**] and it led directly to the proof of two outstanding conjectures: the first Ehle conjecture [**5**] on the A-acceptability of Padé approximations to the exponential function and the Daniel-Moore multiderivative form [**4**] of the Dahlquist barrier. Since then it has played a similar role for rational interpolation of e^x [**8**]. The contribution of order-stars is mostly negative, and completely crucial. They demonstrate that a particular distribution of zeros or poles cannot occur; we make a similar use of them here. However we need an extension of the original theory, to carry out the application to hyperbolic difference schemes.

The other step is to show that the accuracy given in (1.6) can be achieved by a stable scheme. In an equivalent but slightly disguised form, this is already known. The second author showed earlier [**14**] that the maximally accurate schemes, those with $p = r + s$, are stable for $0 < \mu < 1$ provided that $s = r$, $s = r + 1$, or $s = r + 2$. (Instability for other combinations was not proved; that is one of the contributions of order-stars.)

It follows that the value of p in the theorem can be stably attained for any r and s:

(i) If $s \leqslant r$ then $\min\{r + s, 2r + 2, 2s\} = 2s$, and this accuracy is achieved with $c_j = 0$ for $j < -s$ and with the most accurate coefficients (given in [**14**]) from $-s$ to s.

(ii) If $s = r + 1$ then the most accurate scheme is stable, and

$$\min\{r + s, 2r + 2, 2s\} = r + s = p.$$

(iii) If $s > r + 1$ then $\min\{r + s, 2r + 2, 2s\} = 2r + 2$, and this accuracy is achieved with $c_j = 0$ for $j > r + 2$ and with the most accurate coefficients from $-r$ to $r + 2$.

Theorem 1 implies that no other choice of these coefficients c_j can increase the accuracy without destroying stability. The easy choice $c_j = 0$ outside the central range is in this sense optimal.

There are two important variations on the difference approximation (1.3), each raising its own problems of accuracy and stability. Both approximate the same hyperbolic equation $u_t = u_x$, and the first is *semidiscrete*:

$$\frac{\partial U}{\partial t}(x, t) = \sum_{-r}^{s} g_j U(x + j\Delta x, t)/\Delta x. \tag{1.7}$$

Again the test solution is $e^{i\omega(x+t)}$. Starting from $U(x, 0) = e^{i\omega x}$, we find

$$U(x, t) = \exp\left(\sum g_j e^{ij\omega\Delta x} t/\Delta x\right) e^{i\omega x}.$$

Comparing the exponents, the order of accuracy is decided by

$$\sum g_j e^{ij\omega\Delta x} t/\Delta x \approx i\omega t, \quad \text{or} \quad \sum g_j e^{ij\theta} = i\theta + O(\theta^{p+1}). \tag{1.8}$$

Stability requires the exponential to stay bounded, and therefore the exponent itself must have nonpositive real part:

$$\operatorname{Re} \sum_{-r}^{s} g_j e^{ij\theta} \leqslant 0 \quad \text{for all } \theta. \tag{1.9}$$

For this problem, the balance between accuracy and stability is exactly the same. The value of p in Theorem 1 is also the optimal order of accuracy for semidiscrete approximations [**9**] and it is achieved as before by the special schemes with $s = r$, $r + 1$, or $r + 2$. (The bias to the right of the origin stems from the choice of the model $u_t = u_x$; for $u_t = -u_x$, with characteristics coming from the left, this bias is reversed.) The proof that higher accuracy implies instability was the first success in applying order-stars to hyperbolic equations.

Remark. The upper limit $p = 2$ in the one-sided semidiscrete case $r = 0$ was established in [**6**]. In fact it is a consequence of the earlier theorem for fully discrete approximations [**15**]: If $p = 3$ were possible in (1.8), and $g(e^{i\theta}) = \Sigma g_j e^{ij\theta}$ with $\operatorname{Re} g \leqslant 0$, then

$$\sum_{0} c_j e^{ij\theta} = 1 + \mu g + \frac{1}{2}(\mu g)^2 + \frac{1}{6}(\mu g)^3 \tag{1.10}$$

would represent a one-sided difference method that combined $p = 3$ with stability for small μ. (The modulus of (1.10) is bounded by unity for small μ if $\operatorname{Re} g \leqslant 0$.) This contradicts [**15**], and therefore the semidiscrete result for $r = 0$ follows from the fully discrete case. We note that [**14**] considered $u_t = -u_x$ with $r = 1$; the bound $p \leqslant 2$ proved there applies to $u_t = u_x$ with $s = 1$ (by reversing the sign of x) or with $r = 0$ (by shifting the scheme through one meshpoint).

Theorem 1 retains this natural correspondence between s and $r + 1$, and goes beyond the original case $p \leqslant 2$.

A similar transition is also possible for much more general difference methods: *The* "derivative" *of a fully discrete scheme yields a semidiscrete scheme of the same accuracy*. The coefficients are $g_j = dc_j/d\mu$, evaluated at $\mu = 0$, and from (1.5) we have

$$\sum_{-r}^{s} \frac{dc_j}{d\mu}\bigg|_{\mu=0} e^{ij\theta} = i\theta + O(\theta^{p+1}). \tag{1.11}$$

Given that the original scheme was stable, with coefficients reducing to the usual $c_j = \delta_{0j}$ at $\mu = 0$, the semidiscrete scheme is also stable:

$$\operatorname{Re} \sum g_j e^{ij\theta} = \lim_{\mu\to 0} \frac{1}{\mu} \operatorname{Re}\left(\sum c_j e^{ij\theta} - 1\right) \leqslant 0.$$

The stability of the special semidiscrete methods with $s = r$, $r + 1$, or $r + 2$ is therefore confirmed both by the explicit computation of $\Sigma g_j e^{ij\theta}$ in **[9]** and by this relationship to the maximally accurate schemes in **[14]**.

In the general case, this same step from fully discrete to semidiscrete allows an immediate proof of Theorem 1. The order of accuracy is not reduced in the transition, and stability is not lost; therefore the upper limit on p which is known in the semidiscrete case applies also to the discrete equation. The limit in (1.6) cannot be exceeded, since otherwise it would be exceeded by the semidiscrete "derivative," and that is impossible by **[9]**.

The other modification of the difference equation (1.3) is to make it *implicit*:

$$\sum_{-R}^{S} b_j U(x + j\Delta x, t + \Delta t) = \sum_{-r}^{s} c_j U(x + j\Delta x, t). \tag{1.12}$$

This presents new problems, or rather the same problems with new difficulties. There is little change in the test for accuracy:

$$a(e^{i\theta}) = \frac{\Sigma c_j e^{ij\theta}}{\Sigma b_j e^{ij\theta}} = e^{i\mu\theta} + O(\theta^{p+1}). \tag{1.13}$$

For stability the test case $u(x, 0) = e^{i\omega x}$ shows that von Neumann's condition is still necessary:

$$|a(e^{i\theta})| \leqslant 1 \quad \text{for all } \theta. \tag{1.14}$$

Because the difference equation is implicit, there is also a further requirement. Its role is to locate the "center" of the scheme, since the explicit and implicit parts could be shifted in unison without affecting (1.13) or (1.14); the numerator and denominator would be multiplied by the same power of $e^{i\theta}$. On the whole line $-\infty < x < \infty$, the effect would be undetectable. On a half line, however, or on the interval $0 \leqslant x \leqslant 1$, the shift makes a difference. We may think of the implicit operator as a Toeplitz matrix, with the entry b_0 along its main diagonal and the other b_j on adjacent diagonals. For the difference equation (1.12) to be correctly posed, this matrix B must have a bounded inverse (or more precisely, the matrices $B_{\Delta x}^{-1}$ should be uniformly bounded). Suppose, for example, that only one of the coefficients b_j is nonzero. Then that coefficient must be b_0; otherwise we have a multiple of the shift operator and invertibility is lost. The extra requirement can be expressed in terms of the winding number of $\Sigma b_j e^{ij\theta}$ (it must be zero). Or, since there are finitely many terms in the sum, it is a condition on the poles of $a(z)$: the polynomial

$$(1.15)\quad Q(z) = z^R \sum_{-R}^{S} b_j z^j \text{ must have } R \text{ zeros in } |z| < 1 \text{ and } S \text{ zeros in } |z| > 1.$$

This "pole condition" makes B uniformly invertible, and together with the von Neumann condition it is necessary and sufficient for stability in l_2 [**16, 17**].

§2 extends these stability conditions to semidiscrete approximations. The argument depends on a "Wiener-Hopf factorization" of B into triangular Toeplitz matrices, in the opposite order $B = UL$ from the factors in Gaussian elimination. Then we find, for implicit equations, the analogue of Theorem 1. The limit on accuracy depends not only on the number of mesh values that appear in (1.12), but also on their balance: if the equation is stable, then

$$(1.16)\qquad p \leqslant \min\{r + s + R + S, 2(r + R + 1), 2(s + S)\}.$$

This is Theorem 2, proved by a generalization of order-stars.

The remaining problem is to construct implicit approximations that are accurate and stable. We look for coefficients that combine (1.13)–(1.15), and are therefore subject to the limitations on p given above.

It is natural to look first at the most accurate choice, for which $p = r + s + R + S$. This requirement determines the coefficients uniquely, and it must correspond to a *Padé approximation*

$$(1.17)\quad \frac{\Sigma_{-r}^{s} c_j e^{ij\theta}}{\Sigma_{-R}^{S} b_j e^{ij\theta}} \approx e^{i\mu\theta} \quad \text{or} \quad \frac{P(z)}{Q(z)} = \frac{z^r \Sigma c_j z^j}{z^R \Sigma b_j z^j} = z^\lambda + O(|z-1|^{p+1}),$$

where $\lambda = r - R + \mu$. This is the Padé approximation to z^λ at $z = 1$, a polynomial $P_{m/n}$ of degree $m = r + s$ divided by a polynomial $Q_{m/n}$ of degree $n = R + S$. By good fortune these two polynomials were computed in [**10**]; they are limits of hypergeometric functions, we can ask whether they satisfy the von Neumann condition $|P| \leqslant |Q|$ and the pole condition (1.15).

This question is answered, in part, in the second half of the paper. In §3 we find a formula for the difference $D = |Q(e^{i\theta})|^2 - |P(e^{i\theta})|^2$, and we transform Q into a generalized Jacobi polynomial. Therefore the von Neumann condition becomes $D \geqslant 0$, and the pole condition depends on the zeros of $P_n^{(\alpha,\beta)}(z)$. In the symmetric case $m = n$ these tests are comparatively easy to apply; we have $D \equiv 0$ and we can reach $\alpha = \beta$. The classical theory of orthogonal polynomials yields (in Theorem 4A) the only stable possibilities: $r = S$ and $s = R$, or $r = S - 1$ and $s = R + 1$. The nearly symmetric cases $m - n = \pm 1$ are also resolved in §4, using Markoff's theorem on the monotonicity of the zeros with respect to changes in the weight function. Then §5 decides the stability problem for a large class of Padé schemes with general m and n—for which orthogonality is lost (it requires α, $\beta > -1$) and the von Neumann condition is much more difficult.[3] For arbitrary r, s, R, and S, the complete answer remains unknown.

2. The upper limits on accuracy. This section is devoted to the dark side of the theory of difference equations. Their accuracy is restricted, in the hyperbolic case, not only by the number of coefficients in the equation

$$(2.1)\qquad \sum_{-R}^{S} b_j U(x+j\Delta x, t+\Delta t) = \sum_{-r}^{s} c_j U(x, t+\Delta t),$$

but also by the balance between $r+R$ and $s+S$ and by the need for stability. The limit on accuracy is given below. It implies that high accuracy near boundaries cannot be achieved by unbalancing the equation and staying inside the domain. Therefore the alternative that is already adopted in practice is the right one—to introduce extra "boundary equations" for values of U outside the domain. This complicates the stability theory (von Neumann is assisted by Kreiss) but it makes possible a more centered and more stable method.

Our limits show that centering is necessary, for high accuracy with stability:

THEOREM 2. *The accuracy of a stable implicit difference scheme for $u_t = u_x$ is limited by*

$$p \leqslant \min\{r+s+R+S, 2(r+R+1), 2(s+S)\}.$$

Stability will mean that there is a range $0 \leqslant \mu < \mu_0$ of Courant numbers $\mu = \Delta t/\Delta x$ in which the method satisfies von Neumann's condition (and also, in the implicit case, the condition on the poles of $Q(z)$). We suppose that the coefficients b_j and c_j depend analytically (probably polynomially) on μ.

We note immediately that the first limitation $p \leqslant r+s+R+S$ is well known. Equality is achieved only by the Padé schemes, which choose coefficients by matching as many powers of θ as possible in the expansion

$$\sum c_j e^{ij\theta} = e^{i\mu\theta}\left(\sum b_j e^{ij\theta}\right).$$

Equivalently, the ratio $A(z) = P(z)/Q(z)$ is the $r+s/R+S$ Padé approximation to $z^{\mu+r-R}$, and this approximation is known to be "normal" [**1**]; an extra order of accuracy never occurs for $0 < \mu < 1$. We study the stability of these implicit methods, which generalize the explicit and maximally accurate "Lagrange" methods of [**14**], in the following sections.

This section admits any choice of coefficients, and the bounds $p \leqslant 2(r+R+1)$ and $p \leqslant 2(s+S)$ will be proved by the techniques of order-stars.[4] As a preliminary simplification, we derive from the given fully discrete scheme its associated semidiscrete form. If the original scheme is explicit, then this step is the one described in the

[3] The second author may perhaps add a personal confession: he thought it would be impossible. It was the first author who succeeded, leading to a remarkable extension of the results known earlier. But it is still not understood why, in one application after another, stability is achieved along precisely three diagonals.

[4] The case $s = S = 1$ was given first (with the change $x \to -x$) in [**16**].

introduction: the coefficients are $g_j = dc_j/d\mu$ evaluated at $\mu = 0$. In that case the limit on p was known for semidiscrete schemes and Theorem 1 (the bound for discrete explicit methods) was immediate. In the implicit case the corresponding limits on accuracy are not yet established, and that represents our chief task.

In analogy with $c_j \to \delta_{j0}$ in the explicit case, we assume that as $\mu \to 0$ the map in (2.1) from $U(t)$ to $U(t + \Delta t)$ approaches the identity. In other words, we require that $\Sigma b_j(0)e^{ij\theta} = \Sigma c_j(0)e^{ij\theta}$, or

$$\frac{P(z,0)}{Q(z,0)} = \lim_{\mu\to 0} \frac{z^r \Sigma c_j(\mu)z^j}{z^R \Sigma b_j(\mu)z^j} = z^{r-R}. \tag{2.2}$$

From the pole condition for stability, the denominator Q cannot vanish for $|z| = 1$ and $0 \leqslant \mu < \mu_0$.

The semidiscrete scheme derived from (2.1) is also implicit, of the form

$$\sum_{-R^*}^{S^*} f_j \frac{dU}{dt}(x + j\Delta x, t) = \sum_{-r^*}^{s^*} g_j U(x + j\Delta x, t). \tag{2.3}$$

To determine its coefficients, we can begin with

$$\frac{\Sigma c_j(\mu)e^{ij\theta}}{\Sigma b_j(\mu)e^{ij\theta}} = e^{i\mu\theta} + O(\theta^{p+1}),$$

and differentiate with respect to μ. Evaluated at $\mu = 0$, this becomes an approximation to $i\theta$ of order $p^* \geqslant p$:

$$h(e^{i\theta}) = \frac{\Sigma g_j e^{ij\theta}}{\Sigma f_j e^{ij\theta}} = i\theta + O(\theta^{p^*+1}). \tag{2.4}$$

We note that

$$h = \frac{d}{d\mu}\left(\frac{c}{b}\right)\bigg|_{\mu=0} = \frac{c_\mu - b_\mu}{b}\bigg|_{\mu=0} = \frac{c_\mu - b_\mu}{c}\bigg|_{\mu=0}$$

because $b = c$ at $\mu = 0$. Therefore the powers of $e^{i\theta}$ in the numerator extend at most from $-r^*$ to s^*, and in the denominator from $-R^*$ to S^*, with

$$r^* = \max\{r, R\}, \quad R^* = \min\{r, R\}, \quad s^* = \max\{s, S\}, \quad S^* = \min\{s, S\}.$$

This semidiscrete limit can be derived in a different but equivalent way: we start with the rational function $A(z, \mu) = P/Q$, and construct

$$H(z) = \frac{G(z)}{F(z)} = \frac{d}{d\mu}\ln A\bigg|_{\mu=0}. \tag{2.5}$$

From the assumption (2.2) that $A = z^{r-R}$ at $\mu = 0$, this becomes

$$H(z) = \frac{z^R P_\mu(z,0) - z^r Q_\mu(z,0)}{z^r Q(z,0)} = \frac{z^R P_\mu(z,0) - z^r Q_\mu(z,0)}{z^R P(z,0)}.$$

Again the degrees of numerator and denominator are $r^* + s^*$ and $R^* + S^*$. And since $A(z)$ approximated $z^{r-R+\mu}$, the logarithmic derivative yields

$$H(z) = \ln z + O(|z-1|^{p^*+1}), \qquad p^* \geqslant p. \tag{2.6}$$

To prove that stability is inherited along with accuracy, we need to establish the conditions for an implicit semidiscrete scheme to be stable:

LEMMA 2.1. *An equation of the form* (2.3) *is stable if and only if it satisfies*:
(1) *The von Neumann condition*: $\operatorname{Re} h(e^{i\theta}) \leq 0$ *for all* θ.
(2) *The pole condition*: $F(z)$ *has* R^* *zeros in* $|z| < 1$ *and* S^* *zeros in* $|z| > 1$.

It follows that (2.3) is stable if (2.1) is stable: the zeros of $F(z)$ are the zeros of $Q(z,0)$, apart from zeros at the origin which correct for the difference in degree, and one von Neumann condition implies the other:

$$\operatorname{Re} h(e^{i\theta}) = \lim \frac{1}{\mu}\left(\operatorname{Re}\frac{\Sigma c_j e^{ij\theta}}{\Sigma b_j e^{ij\theta}} - 1\right) \leq 0.$$

We sketch the proof of Lemma 2.1, taking this opportunity to explain the relation of the pole condition to the Wiener-Hopf factorization referred to in the introduction. Effectively, it converts an implicit method with finitely many coefficients f_j and g_j into an explicit method with infinitely many. On the whole line $-\infty < x < \infty$, this step does not require Wiener-Hopf; the ratio $H = G/F$ can be expanded in a Laurent series (powers of z and z^{-1}) and the coefficients are those of a large explicit scheme. Its stability hinges entirely on the von Neumann condition $\operatorname{Re} h(e^{i\theta}) \leq 0$, as demonstrated by Fourier analysis.

What changes on a semi-infinite or finite interval is the problem of uniform invertibility of the implicit part. It is no longer sufficient to require only that the denominator be nonzero for $|z| = 1$. We remarked in the introduction that the implicit operator is a Toeplitz matrix: the equation (2.3) in vector form is

$$(2.7)\qquad F\frac{dV}{dt} = GV, \quad \text{with } F_{ij} = f_{j-i},\ G_{ij} = g_{j-i}, \text{ and } V_j(t) = U(j\Delta x, t).$$

Our problem is to show that the pole condition makes F uniformly invertible and that $\operatorname{Re} h \leq 0$ gives $F^{-1}G$ a bounded exponential (and conversely). Then the solution remains bounded and the difference scheme is stable.

On a half line $0 \leq x < \infty$ the matrices F and G are infinite and the Wiener-Hopf technique applies directly; it factors F into a product UL of upper and lower triangular matrices which are themselves Toeplitz. We begin by factoring the associated polynomial $F(z)$, whose coefficients f_j lie on the diagonals of the matrix F:

$$F(z) = z^{R^*}\sum f_j z^j = c(z - z_1)(z - z_2)\cdots(z - z_{R^*+S^*}) = U(z)L(z),$$

where the S^* factors corresponding to the largest roots go into $U(z)$, and the other R^* factors into $L(z)$. The constant c is the coefficient f_{S^*} of the leading term, and can go into U. Then the polynomial $U(z) = \Sigma u_j z^j$ corresponds to an upper triangular Toeplitz matrix with $U_{ij} = u_{j-i}$, and $L(z) = \Sigma l_j z^j$ corresponds to a lower triangular Toeplitz matrix with $L_{ij} = l_{j-i+R^*}$. (It is here that we compensate for the factor z^{R^*} in $R(z)$.) The Wiener-Hopf method depends entirely on the properties of this correspondence:

(i) $F(z) = U(z)L(z)$ implies that $F = UL$.

(ii) U is invertible if and only if the S^* roots of $U(z)$ satisfy $|z_i| > 1$, and U^{-1} is again Toeplitz.

(iii) L is invertible if and only if the R^* roots of $L(z)$ satisfy $|z_i| < 1$, and L^{-1} is again Toeplitz.

(iv) F is invertible if and only if both factors are invertible.

(v) $F^{-1}G = L^{-1}U^{-1}G$ is similar to $U^{-1}GL^{-1}$, which is a Toeplitz matrix.

These properties are not difficult to establish, but neither are they automatic; a product T_1T_2 of Toeplitz matrices is not normally Toeplitz. It is, if T_1 is upper triangular or T_2 is lower triangular, and that explains the last step: the matrix $H = U^{-1}GL^{-1}$ is constant down each diagonal, and those constants are the coefficients in the expansion of $h = \Sigma g_j e^{ij\theta}/\Sigma f_j e^{ij\theta}$ in powers of $e^{i\theta}$. Therefore the correspondence between functions and matrices has one further property:

(vi) If $\operatorname{Re} h(e^{i\theta}) \leqslant 0$ for all θ then $H + H^T$ is negative semidefinite.

Now we can prove the stability of $F\,dV/dt = GV$. If the pole condition holds, properties (i)–(v) permit us to introduce the new unknown $W(t) = LV(t)$, and the equation becomes

$$FL^{-1}\frac{dW}{dt} = GL^{-1}W, \quad \text{or} \quad \frac{dW}{dt} = HW.$$

But then the von Neumann condition, by (vi), implies that this equation is dissipative: $(d/dt)(W, W) = ((H + H^T)W, W) \leqslant 0$. Therefore the original problem was stable:

$$\|W(t)\| \leqslant \|W(0)\| \quad \text{or} \quad \|V(t)\| \leqslant \text{constant}\|V(0)\|.$$

The constant is the condition number $\|L\|\,\|L^{-1}\|$ of L.

For the necessity of conditions (1) and (2) we look first at (iv) above; on $0 \leqslant x < \infty$ the matrix F is invertible only if the pole condition holds. Then stability requires $\operatorname{Re} h(e^{i\theta}) \leqslant 0$. If this fails for some θ, we take as initial data for W the exponential $W_j = e^{ij\theta}$. The solution to $W' = HW$ is $W(t) = e^{tH}W(0)$, or $W_j(t) = \exp(ij\theta + th)$, and with $\operatorname{Re} h > 0$ this explodes as $t \to \infty$. Therefore $W' = HW$, $FV' = GV$, and the original implicit equation (2.3), are all unstable.

For the extension to a finite interval $0 \leqslant x \leqslant 1$, on which the right boundary $x = 1$ interferes with the factorization into Toeplitz matrices, we refer to [**17**]. A "finite section" Toeplitz matrix may be invertible without the pole condition—a typical example would have $f_0 = 1$ on the main diagonal and $f_1 = 2$ on the adjacent diagonal—but uniform invertibility will not hold as $\Delta x \to 0$. The sequence of matrices of increasing size behaves like the corresponding infinite Toeplitz matrix. If the pole condition holds, the triangular factors approach U and L, and the von Neumann condition leads to stability; this completes the lemma.

We come now to Theorem 2, which requires us to show that the pole condition and the von Neumann condition imply bounds on p^* and therefore on p. The argument is carried out by the construction of order-stars and the verification of their properties, but it is technical and we begin with examples.

The order-star itself generalizes the one introduced in [**19**]. Instead of considering $H(z)$ on the halfplane $\operatorname{Re} z < 0$, we study the function $\sigma(z) = H(e^z) - z$ on the strip $\mathcal{S} = \{|\operatorname{Im} z| \leqslant \pi\}$. The essential properties of H will be reflected in the sets

$$\mathcal{A}^* = \{z \in \mathcal{S}\colon \operatorname{Re} \sigma(z) > 0\}, \quad \mathcal{D}^* = \{z \in \mathcal{S}\colon \operatorname{Re} \sigma(z) < 0\}.$$

We call $\mathcal{A}^*$ and $\mathcal{D}^*$ the *order-star* and the *dual order-star* of H. Their connected components are called $\mathcal{A}_0^*$-regions or $\mathcal{A}_\infty^*$-regions (and $\mathcal{D}_0^*$-regions or $\mathcal{D}_\infty^*$-regions) according to whether they are bounded or unbounded.

We give four examples, choosing in each case the scheme with the maximum order of accuracy p:

(a) $r = s = R = S = 1$; $p = 4$:

$$P(z,\mu) = \tfrac{1}{12}\left[(1-\mu)(2-\mu) + 2(4-\mu^2)z + (1+\mu)(2+\mu)z^2\right],$$
$$Q(z,\mu) = \tfrac{1}{12}\left[(1+\mu)(2+\mu) + 2(4-\mu^2)z + (1-\mu)(2-\mu)z^2\right].$$

It is easily seen that $|P| \equiv |Q|$. For every $0 \leqslant \mu < 1$, Q has one zero in the unit circle; hence the method is stable. Its semidiscrete derivative is similarly balanced at the extreme edge of stability:

$$H(z) = \frac{G(z)}{F(z)} = \frac{-1 + z^2}{\frac{1}{3} + \frac{4}{3}z + \frac{1}{3}z^2} \quad \text{and} \quad \operatorname{Re} H \equiv 0.$$

By Lemma 2.1 this is also stable; its order-star is illustrated in Figure 1(a).

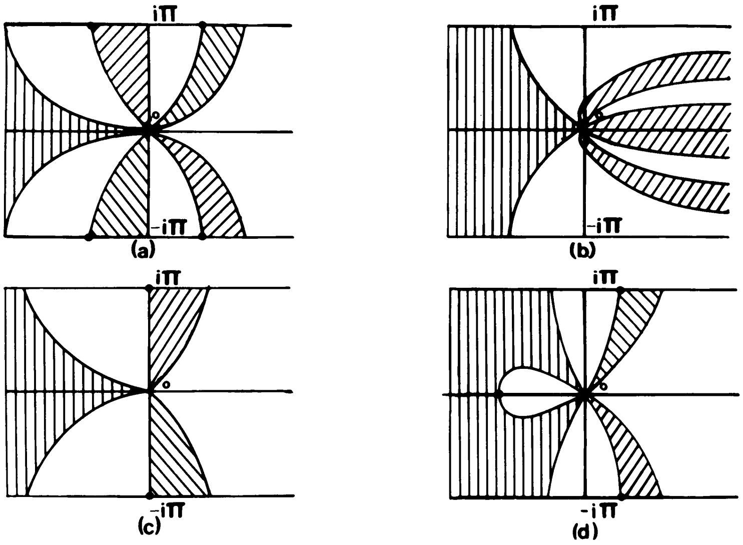

FIGURE 1. Order-stars with $\mathcal{A}^*$ shaded and the poles indicated

(b) $r = R = S = 0, s = 3, p = 3$:

$$P(z,\mu) = \tfrac{1}{6}\left[(1-\mu)(2-\mu)(3-\mu) + 3\mu(2-\mu)(3-\mu)z - 3\mu(2-\mu)(3-\mu)z^2 + \mu(1-\mu)(2-\mu)z^3\right],$$
$$Q(z,\mu) = 1.$$

Hence

$$|Q|^2 - |P|^2 = -\tfrac{1}{3}\mu(1-\mu)(2-\mu)(3-\mu)(1-\cos\theta)^2$$
$$+2/9\mu(1-\mu)^2(2-\mu)^2(3-\mu)(1-\cos\theta)^3$$

and so $|Q| < |P|$ for $\mu \to 0$, $\theta \to 0$, and the scheme is unstable. Moreover

$$H(z) = -\tfrac{11}{6} + 3z - \tfrac{3}{2}z^2 + \tfrac{1}{3}z^3 \quad \text{and} \quad \operatorname{Re} H = \tfrac{1}{3}(1-\cos\theta)^2(4\cos\theta - 1) > 0$$

for $\theta \to 0$. The semidiscrete scheme is unstable; see Figure 1(b) for its order-star.

(c) $r = R = 1, s = S = 0, p = 2$:

$$P(z,\mu) = \tfrac{1}{2}[(1-\mu) + (1+\mu)z],$$
$$Q(z,\mu) = \tfrac{1}{2}[(1+\mu) + (1-\mu)z].$$

It is easy to see that $|P| \equiv |Q|$. However, the single zero of Q is $(\mu+1)/(\mu-1)$, outside the unit circle for $0 < \mu < 1$, and the pole condition is violated. This zero tends to -1 when $\mu \to 0$, and so also the semidiscrete scheme $H(z) = G(z)/F(z) = (-1+z)/(\frac{1}{2}+\frac{1}{2}z)$ is unstable; the pole condition now fails for F.

(d) $r = 0, s = S = R = 1, p = 3$:

$$P(z,\mu) = \tfrac{1}{3}[(2-\mu) + (1+\mu)z],$$
$$Q(z,\mu) = \tfrac{1}{6}[\mu(1+\mu) + 2(2-\mu)(1+\mu)z + (2-\mu)(1-\mu)z^2].$$

In this case

$$|Q|^2 - |P|^2 = \tfrac{1}{9}(2-\mu)(1-\mu)\mu(1+\mu)(1-\cos\theta)^2 \geqslant 0$$

for $0 \leqslant \mu \leqslant 1$, and Q has one zero inside and one outside the unit circle. Hence the discrete scheme is stable. Furthermore,

$$H(z) = \frac{-\frac{1}{2} - 2z + \frac{5}{2}z^2}{2z + z^2}$$

and

$$\operatorname{Re} H(e^{i\theta}) = -(1-\cos\theta)^2/(5 + 4\cos\theta) \leqslant 0.$$

The zeros of F are -2 and 0, and therefore the associated semidiscrete scheme is also stable.

Inspection of Figure 1 suggests that the geometry of each order-star is linked to the stability of the scheme and to its order of accuracy.

The following four lemmas establish this connection. They are completely parallel to Propositions 4.1–4.4 in Iserles [**9**].

LEMMA 2.2. *The scheme* (2.3) *is stable if and only if* $\mathcal{A}^* \cap [-i\pi, i\pi]$ *is empty and the function* σ *has* R^* *poles in* $\mathcal{S}^- = \mathcal{S} \cap \{\operatorname{Re} z < 0\}$ *and* S^* *poles in* $\mathcal{S}^+ = \mathcal{S} \cap \{\operatorname{Re} z > 0\}$.

LEMMA 2.3. *The equation* (2.3) *is of order* p^* *only if for* $z \to 0$, $\mathcal{A}^*$ *consists of* $p^* + 1$ *sectors of angle* $\pi/(p^*+1)$, *separated by* $p^* + 1$ *corresponding sectors of* $\mathcal{D}^*$.

LEMMA 2.4. (a) *If* $s \neq S$ *then for* $\operatorname{Re} z \gg 0$, *the line segment* $\mathcal{A}^* \cap [\operatorname{Re} z - i\pi, \operatorname{Re} z + i\pi]$ *is composed of* $|s - S|$ *intervals separated by intervals of* $\mathcal{D}^*$, *all of length asymptotically* $\pi/|s - S|$. *If* $s = S$ *then* $[\operatorname{Re} z - i\pi, \operatorname{Re} z + i\pi]$ *belongs to* $\mathcal{D}^*$ *for* $\operatorname{Re} z \gg 0$.

(b) *If* $r \neq R$ *then for* $\operatorname{Re} z \ll 0$, *the line segment* $\mathcal{A}^* \cap [\operatorname{Re} z - i\pi, \operatorname{Re} z + i\pi]$ *is composed of* $|r - R|$ *intervals separated by intervals of* $\mathcal{D}^*$, *all of length asymptotically* $\pi/|r - R|$. *If* $r = R$ *then* $[\operatorname{Re} z - i\pi, \operatorname{Re} z + i\pi]$ *belongs to* $\mathcal{A}^*$ *for* $\operatorname{Re} z \ll 0$.

Note that in this count we identify $x + i\pi$ with $x - i\pi$. In Figure 1(b) there are three sectors of $\mathcal{D}^*$ which tend to infinity, one of which is bisected by $\pm i\pi + \mathbf{R}$.

LEMMA 2.5. *Every zero of the denominator* F *lies on the boundary of* $\mathcal{A}$. *Furthermore each* $\mathcal{A}_0^*$*-region and* $\mathcal{D}_0^*$*-region has at least one zero of* F *on its boundary.*

From these lemmas we show that stability imposes restrictions on the order p^*.

LEMMA 2.6. *If* (2.3) *is stable then* $p^* \leqslant 2(r + R + 1)$.

PROOF. We bound from above the number of sectors of $\mathcal{A}^*$ which may reach the origin from $\mathcal{S}^-$. By Lemma 2.4 there are at most $|r - R|$ $\mathcal{A}_\infty^*$-regions in $\mathcal{S}^-$. With the sole exception of an $\mathcal{A}_\infty^*$-region which is bisected by $\pm i\pi + \mathbf{R}$, such a region reaches the origin more than once only if it encircles $\mathcal{D}_0^*$-regions which reach the origin. By Lemma 2.5 these $\mathcal{D}_0^*$-regions have zeros of F on their boundaries. Those zeros must lie inside $\mathcal{S}^-$, because these $\mathcal{D}_0^*$-regions are separated from $\pm i\pi + \mathbf{R}$ by the appropriate $\mathcal{A}_\infty^*$-regions.

The origin can also be reached in $\mathcal{S}^-$ by sectors of $\mathcal{A}_0^*$-regions. By Lemma 2.5 every boundary of such a region contains a zero of F. Because of stability, there are R^* zeros of F in $\mathcal{S}^-$. Furthermore, by Lemma 2.2, no sector of $\mathcal{A}^*$ may approach the origin through the imaginary axis.

Let M^- and M^+ denote the number of sectors of $\mathcal{A}^*$ which reach the origin in $\mathcal{S}^-$ and in $\mathcal{S}^+$. By Lemmas 2.2 and 2.3

$$p^* = M^+ + M^- - 1; \qquad M^+ - 1 \leqslant M^- \leqslant M^+ + 1. \tag{2.8}$$

The largest possible value of M^- is attained when all the zeros of F in $\mathcal{S}^-$ lie on $\pm i\pi + \mathbf{R}$, because then each zero accounts for two sectors of $\mathcal{A}_0^*$-regions which adjoin the origin. Thus $M^- \leqslant |r - R| + 1 + 2R^* = r + R + 1$; in this case $|r - R| + 1$ sectors from $\mathcal{A}_\infty^*$-regions and $2R^*$ sectors of $\mathcal{A}_0^*$-regions reach the origin in $\mathcal{S}^-$. Hence (2.8) yields the limit $p^* \leqslant 2(r + R + 1)$.

A similar argument will give $p^* \leqslant 2(s + S + 1)$. However it is not optimal; a more careful analysis will strengthen this bound. The original differential equation $u_t = u_x$ is not symmetric in x—$u(x, t)$ is determined by the initial value $u_0(x + t)$, so that signals travel from right to left—and we have assumed $\mu = \Delta t/\Delta x < 1$. Therefore we may expect that the number of points $r + 1$ to the left of the first mesh interval corresponds to the number s to its right, and similarly for $R + 1$ and S.

Assume that $M^+ \geqslant s + S + 1$. By repeating the proof of the last lemma for $\mathcal{S}^+$, instead of $\mathcal{S}^-$, it is readily seen that this may happen only if two of the sectors of $\mathcal{A}^*$ which adjoin the origin in $\mathcal{S}^+$ belong to a single $\mathcal{A}_\infty^*$-region, which is bisected by $\pm i\pi + \mathbf{R}$. Hence there is some $y_0 > 0$ such that, for every $x \geqslant y_0$, $\pm i\pi + x \in \mathcal{A}$ or, by definition, $\operatorname{Re}\{H(-e^x) - x\} = \operatorname{Re} \sigma(\pm i\pi + x) > 0$.

It is evident from (2.6) that $f^* \neq 0$ exists such that $H(z) = f^* z^{|s-S|}(1 + O(\frac{1}{z}))$. Hence $M^+ \geqslant s + S + 1$ implies that $(-1)^{s-S} f^* > 0$.

LEMMA 2.7. *If* $(-1)^{s-S}H(z) < 0$ *for* $z \gg 0$ *and* (2.3) *is stable, then* $p^* \leqslant 2(s + S)$.

PROOF. $(-1)^{s-S}H(z) < 0$ for $z \gg 0$ implies $(-1)^{s-S}f^* < 0$. Hence $M^+ \leqslant s + S$. Then stability and (2.8) imply the lemma.

Suppose now that $p^* > 2(s + S)$. Then $M^+ \geqslant s + S + 1$ and $(-1)^{s-S}f^* > 0$. We know from [9] that for every $q \geqslant 0$ the explicit semidiscrete scheme

(2.9)

$$\frac{dU}{dt}(x, t) = \frac{1}{\Delta x}\sum_{i=1}^{q} \frac{(q!)^2}{(q-i)!\,(q+i)!}\,\frac{(-1)^{i+1}}{i}\big(U(x + i\Delta x, t) - U(x - i\Delta x, t)\big)$$

is stable and of order $2q$. Let

$$\Sigma_q(z) = z^{-q}\sum_{i=1}^{q} \frac{(q!)^2}{(q-i)!\,(q+i)!}\,\frac{(-1)^{i+1}}{i}\left(z^{i+q} - z^{i}\right).$$

Then Σ_q is the function H which corresponds to the method (2.9). We now set, for $0 \leqslant \alpha \leqslant 1$,

$$H_\alpha(z) = \alpha H(z) + (1 - \alpha)\Sigma_{s+S}(z) = \frac{G_\alpha(z)}{F_\alpha(z)}.$$

Therefore G_α is of degree at most $s + S + r^* + s^*$, and F_α is of degree $s + S + r^* + S^*$. Furthermore $F_\alpha = z^{s+S}F$ has exactly $s + S + r^*$ zeros in the unit circle and S^* zeros outside. Combined with

$$\operatorname{Re} H_\alpha(e^{i\theta}) = \alpha \operatorname{Re} H(e^{i\theta}) + (1 - \alpha)\operatorname{Re}\Sigma_{s+S}(e^{i\theta}) \leqslant 0,$$

this implies that H_α represents for every $\alpha \in [0, 1]$ a stable method.

Because $(-1)^{s-S}f^* > 0$, there exists $\alpha_0 \in (0, 1)$ at which the degree of G_α is one less than that given above. Therefore the order-star of H_α at α_0 has only $|s - S| - 1$ sectors of $\mathcal{A}^*$ and $|s - S| - 1$ sectors of $\mathcal{D}^*$ tending to infinity in $\mathcal{S}^+$. The order p^* of this method is $2(s + S)$.

Note that G_α and F_α have larger degrees than G and F, but F_α and F have the same number of zeros in $|z| > 1$; the number of sectors of $\mathcal{A}^*$ which tend to infinity in $\mathcal{S}^+$ is actually diminished at $\alpha = \alpha_0$.

If $(-1)^{s-S-1}H_\alpha < 0$ for $\alpha = \alpha_0$ and $z \gg 0$, then by an argument similar to the proof of Lemma 2.7, $p_\alpha^* \leqslant 2(s + S - 1)$. Hence, because $p_\alpha^* = 2(s + S)$, necessarily $(-1)^{s-S-1}H_\alpha > 0$.

By induction we define a sequence of methods $H_k(z)$, $0 \leqslant k \leqslant |s - S|$, such that $H_k(z) = G_k(z)/F_k(z)$ and

(i) F_k has exactly S^* zeros in $|z| > 1$,

(ii) the order-star of H_k has $|s - S| - k$ sectors of $\mathcal{A}^*$ tending to infinity in $\mathcal{S}^+$,

(iii) H_k represents a stable scheme of order $p_k^* \geqslant 2(s + S - k) + 1$,

(iv) $H_0 = H$ and $H_1 = H_\alpha$ at $\alpha = \alpha_0$.

Let us consider $H_{|s-S|}$. In the order-star of this scheme there are no sectors of $\mathcal{A}^*$ tending to infinity in $\mathcal{S}^+$. Hence stability implies that $M^+ \leqslant 2S^*$ and, by (2.8), the order may not exceed $4S^*$. But $p_{|s-S|}^* \geqslant 4S^* + 1$, which is a contradiction. Thus the hypothesis that $p^* \geqslant 2(s+S)+1$ leads to contradiction, and we have

LEMMA 2.8. *If* $(-1)^{s-S}H(z) > 0$ *for* $z \gg 0$, *stability implies* $p^* \leqslant 2(s+S)$.

Lemmas 2.6–2.8 complete the proof of Theorem 2.

3. The Padé schemes. This secton begins the proof that certain special difference approximations which attain the maximal accuracy $p = r+s+R+S$ are at the same time stable. We emphasize that most choices of these integers leave no chance for such a proof; for this value of p, Theorem 2 implies instability unless

$$(3.1)\quad r+s+R+S \leqslant 2(r+R+1) \quad\text{and}\quad r+s+R+S \leqslant 2(s+S).$$

In other words, the only chance for stability occurs when the scheme is sufficiently centered:

$$(3.2)\qquad r+R \leqslant s+S \leqslant r+R+2.$$

When stability is proved, it will hold over the full interval $0 < \mu = \Delta t/\Delta x < 1$.

The coefficients $b_j(\mu)$ and $c_j(\mu)$ in the difference equation are determined by the requirement of maximal accuracy,

$$\frac{\Sigma c_j e^{ij\theta}}{\Sigma b_j e^{ij\theta}} = e^{i\mu\theta} + O(\theta^{r+s+R+S+1}).$$

In terms of the polynomials $P = z^r \Sigma c_j z^j$ and $Q = z^R \Sigma b_j z^j$, this means that

$$(3.3)\qquad A = \frac{P}{Q} = z^\lambda + O(|z-1|^{m+n+1}),$$

where $m = r+s$ is the degree of P, $n = R+S$ is the degree of Q, and $\lambda = r - R + \mu$. Therefore A is the $[m/n]$ Padé approximation to z^λ at $z = 1$; it is the unique rational function of the given degree that satisfies (3.3). It is natural to call the associated difference approximation a *Padé scheme*. The rest of this paper is devoted to the stability problem for Padé schemes.

We recall the two requirements for stability:

(1) The *von Neumann condition*: $|P/Q| \leqslant 1$ for $|z| = 1$.

(2) The *pole condition*: Q has R zeros in $|z| < 1$ and S zeros in $|z| > 1$.

This section prepares for the verification of these two conditions. First, we compute explicitly the difference $|Q|^2 - |P|^2$, which must be nonnegative, and then we identify Q as the Möbius transform of a generalized Jacobi polynomial. We denote $|Q(e^{i\theta}, \lambda)|^2 - |P(e^{i\theta}, \lambda)|^2$ by D, and $2(1-\cos\theta)$ by X.

THEOREM 3. *For a Padé scheme with* $m = r+s < n = R+S$,

$$(3.4)\quad D = (-1)^n \frac{m!\,n!}{[(m+n)!]^2} \sum_{k=[(n+m)/2]+1}^{n} \frac{(n+M-k)!\,(k-m-1)!}{k!\,(2k-n-m-1)!\,(n-k)!} \cdot (-n-\lambda)_k(-m+\lambda)_k X^k.$$

For $m > n$ the signs of D and λ, and the integers m and n, are reversed:

$$(3.4') \quad D = (-1)^{m+1} \frac{m!n!}{[(m+n)!]^2} \sum_{[(n+m)/2]+1}^{m} \frac{(n+m-k)!\,(k-n-1)!}{k!\,(2k-n-m-1)!\,(m-k)!} \cdot (-n-\lambda)_k(-m+\lambda)_k X^k.$$

For $m = n$ the difference is $D = |Q|^2 - |P|^2 \equiv 0$.

The pole condition is satisfied if and only if the Jacobi polynomial $P_r^{(\alpha,\beta)}(z)$, with $\alpha = r - R + \mu$ and $\beta = s - S - \mu$, has R zeros in the right halfplane $\operatorname{Re} z > 0$ *and S zeros in* $\operatorname{Re} z < 0$.

We recall that $(y)_0 = 1$ and $(y)_k = y(y+1)\cdots(y+k-1)$.

The proof of (3.4) begins by connecting the polynomials P and Q to the hypergeometric function

$$_2F_1(a, b; d; 1-z) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{k!\,(d)_k}(1-z)^k.$$

The link is given by an identity of Euler [**13**, p. 60]:

$$(3.5) \qquad _2F_1(d-a, d-b; d; 1-z) = z^{a+b-d}\,{}_2F_1(a, b; d; 1-z).$$

We set $a = -n$, $b = \lambda - m + \varepsilon$, and $d = -n - m + \varepsilon$, avoiding by means of ε the inadmissible values $d = 0, -1, -2, \ldots$. Then $a + b - d = \lambda$, and Iserles [**10**] computed the limits of both series in (3.5) as $\varepsilon \to 0_+$. The one on the right is a polynomial, because the factor $(a)_k = (-n)_k$ vanishes after the term with $k = n$. The sum on the left side of (3.5) omits in the limit of powers $k = m+1, \ldots, m+n$, leaving

$$(3.6) \quad \sum_{k=0}^{m} \frac{(n+m-k)!m!\,(-n-\lambda)_k}{(n+m)!k!\,(m-k)!}(1-z)^k - z^{\lambda} \sum_{k=0}^{n} \frac{(n+m-k)!n!\,(\lambda-m)_k}{(n+m)!k!\,(n-k)!}(1-z)^k = O(|1-z|^{n+m+1}).$$

These sums must therefore be the polynomials P and Q in the Padé approximation to z^{λ}. Because the second made no contribution to the error term, it gives a direct representation of Q:

$$(3.7) \qquad Q_{m/n}(z, \lambda) = \lim_{\varepsilon \to 0} {}_2F_1(-n, \lambda - m + \varepsilon; -n - m + \varepsilon; 1 - z).$$

Then P must appear in a similar form if we change the sign of λ; (3.6) becomes an $[m/n]$ Padé approximation to $z^{-\lambda}$, so that multiplying through by z^{λ} and exchanging m and n isolates P as

$$(3.8) \qquad P_{m/n}(z, \lambda) = Q_{n/m}(z, -\lambda).$$

Now we turn to $|Q|^2$. There is a theorem of Burchnall and Chaundy which must have been invented for exactly this calculation [7, p. 83]: they proved that

$$_2F_1(a, b; d; z_1)\,{}_2F_1(a, b; d, z_2) = \sum_{0}^{\infty} \frac{(a)_k(b)_k(d-a)_k(d-b)_k}{k!\,(d)_k(d)_{2k}}(z_1z_2)^k\,{}_2F_1(a+k, b+k; d+2k; z_1 + z_2 - z_1z_2).$$

In our case $z_1 = 1 - e^{i\theta}$ and $z_2 = \bar{z}_1 = 1 - e^{-i\theta}$. Thus $z_1 + z_2 = z_1 z_2 = X = 2(1 - \cos\theta)$, and the hypergeometric function on the right side is identically 1. Therefore

$$(3.9) \qquad |Q|^2 = \lim_{\varepsilon\to 0} \sum_0^\infty \frac{(-n)_k(\lambda - m + \varepsilon)_k(-m+\varepsilon)_k(-n-\lambda)_k}{k!\,(-n-m+\varepsilon)_k(\,-n-m+\varepsilon)_{2k}} X^k.$$

For $m > n$ the limiting process is trivial; the term $(-n)_k$ is $(-1)^k n!/(n-k)!$ for $k \leqslant n$, and vanishes for $k > n$. To the other terms in (3.9) we apply the same identity, with vanishing ε, and with $C = m!n![(m+n)!]^{-2}$ we get

$$(3.10) \quad |Q|^2 = C\sum_0^n (-1)^k \frac{(n+m-k)!\,(n+m-2k)!}{k!\,(n-k)!\,(m-k)!}(-n-\lambda)_k(-m+\lambda)_k X^k.$$

The case $m < n$ is more delicate and we need it for P. The term $(-m+\varepsilon)_k$ in the numerator approaches zero for $k > m$, but then for $k > [(n+m)/2]$ so does $(-n-m+\varepsilon)_{2k}$ in the denominator. In the latter case their ratio gives

$$\lim_{\varepsilon\to 0}\frac{(-m+\varepsilon)_k}{(-n-m+\varepsilon)_{2k}} = (-1)^n\frac{m!\,(k-m-1)!}{(n+m)!\,(2k-n-m+1)!}.$$

Therefore (3.9) for $m < n$ yields
(3.11)

$$|Q|^2 = C\left[\sum_{k=0}^{m}(-1)^k\frac{(n+m-k)!\,(n+m-2k)!}{k!\,(n-k)!\,(m-k)!}(-n-\lambda)_k(-m+\lambda)_k X^k\right.$$
$$\left. +(-1)^n \sum_{k=[(n+m)/2]+1}^{n}\frac{(n+m-k)!(k-m-1)!}{k!(2k-n-m+1)!(n-k)!}(-n-\lambda)_k(-m+\lambda)_k X^k\right].$$

Finally we use the identity $P_{m/n}(z,\lambda) = Q_{n/m}(z,-\lambda)$ to reach a similar expression for $|P|^2$. For $m = n$ the result is identical to (3.10), and $D = |Q|^2 - |P|^2$ is zero. For $m < n$ we look at (3.11) for $|Q|^2$ and at (3.10) for $|P|^2$. In the latter, reversing m and n as well as the sign of λ leaves $(-n-\lambda)_k(-m+\lambda)_k$ invariant; therefore the first sum in (3.11) exactly cancels the new (3.10). This leaves the second sum in (3.11) as the difference D, and this is our formula (3.4).

The second part of the theorem, concerning the poles of Q, comes from identifying the hypergeometric form (3.7) as a multiple of a Jacobi polynomial. The two are connected by the identity [**13**, p. 255]

$$P_n^{(\alpha,\beta)}(w) = \frac{(1+\alpha+\beta)_{2n}}{n!\,(1+\alpha+\beta)_n}\left(\frac{w+1}{2}\right)^n {}_2F_1\left(-n,-\beta-n;-\alpha-\beta-2n;\frac{2}{w+1}\right).$$

We set $\alpha = \lambda$ and $\beta = m - n - \lambda - \varepsilon$; as $\varepsilon \to 0$ the ratio approaches

$$(3.12) \qquad \lim\frac{(1+\alpha+\beta)_{2n}}{(1+\alpha+\beta)_n} = \frac{(m+n)!}{m!}$$

and the hypergeometric function approaches Q. Then with $w = (1+z)/(1-z)$, or $2/(w+1) = 1 - z$, the identity becomes

$$(3.13) \qquad Q_{m/n}(z,\lambda) = \frac{m!n!}{(m+n)!}(1-z)^n P_n^{(\lambda,m-n-\lambda)}\left(\frac{1+z}{1-z}\right).$$

Finally we can make contact with the poles. They are the zeros of the generalized Jacobi polynomial $P_n^{(\alpha,\beta)}(w)$, with indices $\alpha = \lambda = r - R + \mu$ and $\beta = m - n - \lambda = s - S - \mu$. The argument is $w = (1 + z)/(1 - z)$, the Möbius transform, so that the zeros of Q in $|z| < 1$ correspond to zeros of $P_n^{(\alpha,\beta)}$ in $\operatorname{Re} w > 0$. There must be R such zeros, if the implicit part of the difference equation is to be uniformly invertible, and the other $n - R = S$ zeros must have negative real part. (We note that $Q(1, \lambda) \neq 0$ from the exact order $m + n$ of the Padé approximation.) Therefore Theorem 3 is proved.

For $\alpha, \beta > -1$ the Jacobi polynomials are orthogonal over $[-1, 1]$ with respect to $(1 - x)^\alpha(1 + x)^\beta$, and their zeros are in this interval. For indices below -1, which will occur as the implicit part uses more meshpoints, the zeros may be complex—but without ruling out the possibility of stability.

Remark. For the explicit case $n = 0$, these Padé schemes become the "methods of maximum accuracy" studied by the second author in **[14]**. The stability established there should be confirmed by our formula (3.4′); with $\lambda = r + \mu$ and $X = 2(1 - \cos\theta)$ it becomes

$$D = 1 - |P|^2 = \frac{(-1)^{m+1}}{m!} \sum_{[m/2]+1}^{m} \frac{(r + \mu - m)_k(-r - \mu)_k}{k(2k - m - 1)!} X^k. \tag{3.14}$$

Suppose first that the scheme is centered: $r = s$. Then $m = 2r$, and every term in the sum is negative for $0 < \mu < 1$. (The factors in the numerator have opposite sign for all $k > r$.) Therefore the sign $(-1)^{m+1}$ makes D positive, and verifies the von Neumann condition $|P| \leqslant 1$.

If $-1 < \mu < 0$, then (3.14) remains positive (D is an even function of μ). Or, to follow our system more faithfully, we shift the scheme and verify stability for $0 < \mu < 1$ after s is increased and r is decreased by 1. In the remaining stable case $s = r + 1$, m is odd and the terms in the sum are positive.

The proof that no other choice is stable comes more easily from Theorem 1 than from (3.14). For methods of maximum accuracy p is $r + s$, and then Theorem 1 gives $r \leqslant s \leqslant r + 2$ as in (3.2). Therefore the stability limits found in **[14]** were the largest possible; if for example the scheme based on $U(x - r\Delta x), \ldots, U(x + r\Delta x)$ were stable on an interval outside $-1 \leqslant \mu = \Delta t/\Delta x \leqslant 1$, we could shift the method to move this new interval into $0 < \mu < 1$. The accuracy would still be $r + s$, as noted just before Theorem 1, but then that theorem rules out stability. After a time step of twenty years, the stable cases remain stable and the others are finally excluded.

4. The cases $m = n$ and $m - n = \pm 1$. In this section we characterize all the stable Padé schemes in which the number of points at the two time levels differs at most by one. If $m = n$ then $D = 0$ by Theorem 3, and the Padé method is stable if and only if Q has the correct number of zeros in the unit circle. For $m - n = \pm 1$, there is only a single term in D and we can determine its sign; it is again the pole condition which presents the difficulty.

For $m = n$, Theorem 2 drastically restricts the candidates r, s, R, and S for stability. In what follows we will determine directly all the stable combinations without resorting to order-stars. First we need a lemma on the behavior of rational approximations (not necessarily the Padé P and Q) to z^λ. This lemma is close in spirit and in proof to the Maximal Interpolation Theorem (Iserles [**11**]).

LEMMA 4.1. *Let $B(z, \lambda) = P/Q$ belong to $\pi_{m/n}[z]$. Then for every noninteger value $\lambda < m + 1$, the equation $B(x, \lambda) = x^\lambda$ has at most $n + m + 1$ real roots, counted with their multiplicity.*

PROOF. Every root is also a zero of $\psi(x, \lambda) = P(x, \lambda) - x^\lambda Q(x, \lambda)$, of at most the same multiplicity. Hence it is sufficient to show that ψ has at most $n + m + 1$ real zeros.

By repeated differentiation of ψ with respect to x we find

$$\frac{\partial^k \psi}{\partial x^k} = P_k - x^{\lambda - k} Q_k,$$

where $P_k \in \pi_{m-k}[x]$, $Q_k \in \pi_n[x]$. In particular

$$\frac{\partial^{m+1}\psi}{\partial x^{m+1}} = -x^{\lambda - m - 1} Q_{m+1}.$$

If Q_{m+1} were identically zero, then $Q_{m+1} = \lambda Q_m + x\partial Q_m/\partial x$ would imply $Q_m = Cx^{-\lambda}$. This is impossible for $C \neq 0$, because λ is not an integer while Q_m is a polynomial. Thus Q_m will be identically zero, and by induction we obtain $Q = Q_0 \equiv 0$ which is impossible. Hence Q_{m+1} is not identically zero.

The zeros of $\partial^{m+1}\psi/\partial x^{m+1}$ coincide with the zeros of Q_{m+1}, because $\lambda < m + 1$, and Q_{m+1} has degree n. We apply the Rolle theorem $m + 1$ times and obtain $n + m + 1$ as an upper bound on the number of zeros of ψ. This completes the proof.

LEMMA 4.2. *For Padé, if $D \geqslant 0$ for every $0 \leqslant \theta \leqslant 2\pi$ and $0 < \mu < 1$, then $Q \neq 0$ over the same intervals. Therefore stability for $\mu \to 0$ implies stability for $0 < \mu < 1$.*

PROOF. If $Q = 0$ at θ_0, μ_0, then $D = -|P|^2 = 0$ and so $\exp(i\theta_0)$ is also a zero of P. Hence P and Q have a nontrivial (linear or quadratic) common factor. After reduction by this factor we obtain a rational function of lower degree that violates Lemma 4.1 at the value $\lambda = r - R + \mu_0$.

Thus with $D \geqslant 0$, the zeros of Q cannot touch (or cross) the unit circle for $0 < \mu < 1$. After the Möbius transform from the circle to the imaginary axis $i\mathbf{R}$, we let $\xi_+(\mu)$, $\xi_-(\mu)$, and $\xi_0(\mu)$ denote the number of zeros in $\operatorname{Re} z > 0$, $\operatorname{Re} z < 0$, and on $i\mathbf{R}$. Then it is sufficient in verifying the pole condition to know these numbers at $\mu = 0$ and (in case $\xi_0(0) > 0$) at $\mu = 1$.

We begin with the case $m = n$, for which $|P| = |Q|$ and stability depends on the zeros of the Jacobi polynomial $P_n^{(r-R+\mu, R-r+\mu)}$. To find those zeros for $\mu = 0$, we begin with the identity [**18**, p. 64]

$$P_n^{(-k,\alpha)}(z) = \frac{(-n-\alpha)_k}{(-n)_k}\left(\frac{z-1}{2}\right)^k P_{n-k}^{(k,\alpha)}(z), \qquad 0 \leqslant k \leqslant n. \tag{4.1}$$

Furthermore, because $P_n^{(\alpha,\beta)}(z) = (-1)^n P_n^{(\beta,\alpha)}(-z)$,

$$P_n^{(\alpha,-k)}(z) = \frac{(-n-\alpha)_k}{(-n)_k}\left(\frac{z+1}{2}\right)^k P_{n-k}^{(\alpha,k)}(z); \tag{4.2}$$

$$P_n^{(-K,-k)}(z) = \frac{(-n+k)_K(-n)_k}{(-n)_K(-n+K)_k}\left(\frac{z-1}{2}\right)^K\left(\frac{z+1}{2}\right)^k P_{n-K-k}^{(K,k)}(z).$$

Let $y_+ = \max\{y, 0\}$. When $\mu = 0$ we obtain from (4.1) and (4.2), with $r + s = R + S = n$ and $r - R = y$,

$$P_n^{(r-R,R-r)}(z) = \frac{(r+S)!\,(s+R)!}{[(R+S)!]^2}\left(\frac{z+1}{2}\right)^{y_+}\left(\frac{z-1}{2}\right)^{(-y)_+} P_{n-|y|}^{(|y|,|y|)}(z).$$

For every $\alpha > -1$, $P_n^{(\alpha,\alpha)}(z)$ is an ultraspherical polynomial whose zeros are in $(-1, 1)$ and symmetric with respect to the origin. Therefore

$$\xi_-(0) = (r-R)_+ + \left[\tfrac{1}{2}(R+S-|r-R|)\right] = \left[\tfrac{1}{2}(r+S)\right];$$
$$\xi_+(0) = (R-r)_+ + \left[\tfrac{1}{2}(R+S-|r-R|)\right] = \left[\tfrac{1}{2}(s+R)\right];$$
$$\xi_0(0) = R + S - \xi_-(0) - \xi_+(0).$$

The value of $\xi_0(0)$ may not exceed one, because the zeros of the orthogonal polynomial $P_{n-|y|}^{(|y|,|y|)}$ are simple. Hence the pole condition requires $S - 1 \leqslant \xi_-(0) \leqslant S$ and $R - 1 \leqslant \xi_+(0) \leqslant R$. There are three choices of r and s, for any given R and S, which satisfy these inequalities:

$$\begin{array}{llll} r = S-1, & s = R+1 \Rightarrow \xi_-(0) = S-1, & \xi_+(0) = R, & \xi_0(0) = 1; \\ r = S, & s = R \Rightarrow \xi_-(0) = S, & \xi_+(0) = R, & \xi_0(0) = 0; \\ r = S+1, & s = R-1 \Rightarrow \xi_-(0) = S, & \xi_+(0) = R-1, & \xi_0(0) = 1. \end{array} \tag{4.3}$$

We now examine the zeros when $\mu = 1$. By (4.1) and (4.2)

$$P_n^{(y+1,-y-1)}(z) = \frac{(r+S+1)!\,(s+R-1)!}{[(R+S)!]^2}\left(\frac{z+1}{2}\right)^{(y+1)} + \left(\frac{z-1}{2}\right)^{(-y-1)} + P_{n-|y+1|}^{(|y+1|,|y+1|)}(z).$$

By repeating the analysis given for $\mu = 0$, we obtain $\xi_-(1) = [\frac{1}{2}(r + S + 1)]$, $\xi_+(1) = [\frac{1}{2}(s + R - 1)]$. Once again, three choices of r and s satisfy the inequalities $S - 1 \leqslant \xi_-(1) \leqslant S$, $R - 1 \leqslant \xi_+(1) \leqslant R$, namely

$$\begin{array}{llll} r = S-2, & s = R+2 \Rightarrow \xi_-(1) = S-1, & \xi_+(1) = R, & \xi_0(1) = 1; \\ r = S-1, & s = R+1 \Rightarrow \xi_-(1) = S, & \xi_+(1) = R, & \xi_0(1) = 0; \\ r = S, & s = R \Rightarrow \xi_-(1) = S, & \xi_+(1) = R-1, & \xi_0(1) = 1. \end{array} \tag{4.4}$$

By comparing (4.3) and (4.4) we obtain all the stable Padé methods with $r + s = R + S$, or $m = n$:

THEOREM 4A. *For every given R and S there are exactly two choices of r and s, with $r + s = R + S$, which give stable methods*: $r = S$, $s = R$ *and* $r = S - 1$, $s = R + 1$. *Both choices yield stability for every* $0 < \mu < 1$.

EXAMPLES. $r = s = R = S = 1$ gives

$$P = \tfrac{1}{12}((1-\mu)(2-\mu) + 2(2+\mu)(2-\mu)z + (1+\mu)(2+\mu)z^2),$$
$$Q = \tfrac{1}{12}((1+\mu)(2+\mu) + 2(2-\mu)(2+\mu)z + (1-\mu)(2-\mu)z^2).$$

Hence $|P| = |Q|$ and $D = 0$. The zeros of Q are

$$\frac{-(2-\mu) \pm \sqrt{3(2+\mu)/(2-\mu)}}{1-\mu}.$$

When $\mu \to 0$ they tend to $-2 \pm \sqrt{3}$. By contrast, $r = 0$, $s = 2$, $R = S = 1$ leads to

$$P = \tfrac{1}{12}((2-\mu)(3-\mu) + 2(1+\mu)(3-\mu)z + \mu(1+\mu)z^2),$$
$$Q = \tfrac{1}{12}(\mu(1+\mu) + 2(1+\mu)(3-\mu)z + (2-\mu)(3-\mu)z^2).$$

Again $|P| = |Q|$, but now one of the zeros approaches -1 as $\mu \to 0$.

These examples are typical: the symmetric case $r = S$, $s = R$ remains completely safe as $\mu \to 0$, while the choice $r = S - 1$, $s = R + 1$ approaches weak instability.

Before we investigate the Padé schemes with $m - n = \pm 1$, we need to study further the zeros of Jacobi polynomials. We recall the Markoff Theorem [**18**, p. 115]: Let $\{\varphi_n(x;\alpha)\}_{n=0}^{\infty}$ be a set of polynomials, orthogonal with respect to the continuous weight function $w(x;\alpha) > 0$ in the interval (a, b). Let $x_1^{(n)}(\alpha) < x_2^{(n)}(\alpha) < \cdots < x_n^{(n)}(\alpha)$ denote the zeros of φ_n. If $\partial \ln w(x;\alpha)/\partial\alpha$ is an increasing function of x in (a, b) and if the integrals $\int_a^b x^\nu \partial w/\partial\alpha\, dx$ are uniformly convergent for every $\nu \geqslant 0$, then $dx_k^{(n)}(\alpha)/d\alpha > 0$ for $1 \leqslant k \leqslant n$.

In our case $w = (1-x)^{k+\mu}(1+x)^{k-\mu}$, for $k = 0, 1 \cdots$. Hence $\partial \ln w(x;\mu)/\partial\mu = \ln(1-x) - \ln(1+x)$, a decreasing function in $(-1, 1)$. Therefore $dx_k^{(n)}(\mu)/d\mu < 0$ for all n zeros of $P_n^{(k+\mu,k-\mu)}$.

Let $\mu = 0$. Then $P_n^{(k,k)}$ is an ultraspherical polynomial [**13**, p. 276] and so $\xi_+(0) = \xi_-(0) = [n/2]$, $\xi_0(0) = [(n+1)/2] - [n/2]$.

LEMMA 4.3. *Suppose that* $r - R = s - S = k \geqslant 0$ *and that* $D(e^{i\theta}, r - R + \mu) \geqslant 0$ *for every* θ *in* $[0, 2\pi]$ *and* μ *in* $(0, 1)$. *Then*

$$\xi_+(0) = \xi_-(0) = [n/2]; \quad \xi_0(0) = [(n+1)/2] - [n/2];$$
$$\xi_+(\mu) = [n/2] \quad \textit{and} \quad \xi_-(\mu) = [(n+1)/2] \quad \textit{for}\ 0 < \mu < 1.$$

PROOF. Lemma 4.2 led to $\xi_0(\mu) = 0$ for every $\mu \in (0, 1)$. Hence, because of the monotonicity of the zeros as functions of μ, $\xi_-(\mu) = \xi_-(0) + \xi_0(0)$, $\xi_+(\mu) = \xi_+(0)$. This gives the desired result.

We turn to the zeros of $P_n^{(\gamma+1,\gamma)}$ for $\gamma > 1$. We set $\alpha = \beta = \gamma$ and $\alpha = \beta = \gamma + 1$, respectively, in the identities [**13**, p. 265]

$$(1+z)P_n^{(\alpha,\beta+1)}(z) + (1-z)P_n^{(\alpha+1,\beta)}(z) = 2P_n^{(\alpha,\beta)}(z),$$
$$P_n^{(\alpha,\beta-1)}(z) - P_n^{(\alpha-1,\beta)}(z) = P_{n-1}^{(\alpha,\beta)}(z).$$

This yields

$$P_n^{(\gamma,\gamma+1)}(0) + P_n^{(\gamma+1,\gamma)}(0) = 2P_n^{(\gamma,\gamma)}(0), \tag{4.5}$$
$$-P_n^{(\gamma,\gamma+1)}(0) + P_n^{(\gamma+1,\gamma)}(0) = P_{n-1}^{(\gamma+1,\gamma+1)}(0).$$

Adding, we find

$$(4.6)\qquad P_n^{(\gamma+1,\gamma)}(0) = P_n^{(\gamma,\gamma)}(0) + \tfrac{1}{2}P_{n-1}^{(\gamma+1,\gamma+1)}(0) \neq 0,$$

because if n is even then $P_n^{(\gamma,\gamma)}(0) \neq 0$, $P_{n-1}^{(\gamma+1,\gamma+1)}(0) = 0$, while if n is odd then $P_n^{(\gamma,\gamma)}(0) = 0$, $P_{n-1}^{(\gamma+1,\gamma+1)}(0) \neq 0$.

All the zeros of $P_n^{(\gamma+1,\gamma)}$ are real. Hence, because (4.6) implies $\xi_0(\gamma) = 0$, $\xi_+(\gamma)$ and $\xi_-(\gamma)$ must be constant for $\gamma > -1$. Consequently, they can be evaluated at the particular choice $\gamma = -\frac{1}{2}$. To examine the zeros of $P_n^{(1/2,-1/2)}$, we identify the Padé method which corresponds to this polynomial by virtue of Theorem 3.

If n is even we set $r = s = R = S$, to obtain $D \equiv 0$. For odd n we set $r = R = (n-1)/2$, $s = S = (n+1)/2$, and again $D \equiv 0$. We now apply Lemma 4.3 to the polynomial $P_n^{(1/2,-1/2)}$, to deduce $\xi_+ = [n/2]$, $\xi_- = [(n+1)/2]$.

This analysis, together with $P_n^{(\alpha,\beta)}(z) = (-1)^n P_n^{(\beta,\alpha)}(-z)$, proves the following lemma.

LEMMA 4.4. *Let* $\gamma > -1$. *Then* (a) $P_n^{(\gamma+1,\gamma)}(z)$ *has* $[n/2]$ *zeros in* $(0,1)$ *and* $[(n+1)/2]$ *zeros in* $(-1,0)$; (b) $P_n^{(\gamma,\gamma+1)}(z)$ *has* $[(n+1)/2]$ *zeros in* $(0,1)$ *and* $[n/2]$ *zeros in* $(-1,0)$.

We can now examine the stability of the Padé schemes with $r + s = R + S + 1$. Theorem 3 gives, after a straightforward calculation,

$$(4.7)\qquad D = (-1)^{r+S}\frac{[(R+S)!]^2}{(2R+2S+1)!}(\mu)_{r+S+1}(-\mu)_{s+R}X^{r+s}.$$

Therefore a necessary condition for stability is that $r + S$ is even. We proceed to investigate the zeros of $P_{R+S}^{(r-R,R-r+1)}(z)$. Let $r > R$. Then, by (4.2), this polynomial has $r - R + 1$ zeros at -1 and its other zeros coincide with the zeros of $P_{s+R}^{(r-R,r-R-1)}$. Hence, by Lemma 4.4(a),

$$\xi_+ = \left[\frac{s+R}{2}\right],\quad \xi_- = \left[\frac{r+S}{2}\right],\quad \xi_0 = 0.$$

Now let $r \leqslant R$. Then by (4.1), $P_{R+S}^{(r-R,R-r+1)}$ has $R - r$ zeros at $+1$ and otherwise its zeros are identical to the zeros of $P_{r+S}^{(R-r,R-r+1)}$; Lemma 4.4(b) gives

$$\xi_+ = \left[\frac{s+R}{2}\right],\quad \xi_- = \left[\frac{r+S}{2}\right],\quad \xi_0 = 0.$$

The requirements $\xi_+ = R$, $\xi_- = S$, $r + S$ even, leave the single choice $s = R + 1$, $r = S$. This satisfies the von Neumann condition $D \geqslant 0$, by (4.7), and the pole condition by Lemma 4.2. Therefore we have proved

THEOREM 4B. *Let* $r + s = R + S + 1$. *Then the only stable Padé method, for every* $0 < \mu < 1$, *is given by* $s = R + 1$, $r = S$. *Any other choice is unstable for all* $0 < \mu < 1$.

By proceeding in the same way with $P_{R+S}^{(r-R,R-r-1)}$ we reach

THEOREM 4C. *Let* $r + s + 1 = R + S$. *Then the only stable Padé method, for every* $0 < \mu < 1$, *is given by* $s = R$, $r = S - 1$. *Any other choice is unstable for all* $0 < \mu < 1$.

Note that each stable choice falls within the limits (3.2) set by Theorem 2. However, not every choice which satisfies these constraints leads to stability.

5. The case $R = S$ and $m \geqslant n$. We know that the most accurate explicit schemes, the Padé schemes with $R = S = 0$, are stable for $r \leqslant s \leqslant r + 2$ and not otherwise. Thus every case admitted as possible by Theorem 1 is actually stable. We recall from (3.2) that for more general Padé schemes the corresponding limitation (a direct consequence of Theorem 2) was

$$r + R \leqslant s + S \leqslant r + R + 2. \tag{5.1}$$

Our goal is to establish that for any choice with $R = S$ and $m \geqslant n$, (5.1) is exactly the condition for stability. These are the centered, maximally accurate, and "mostly explicit" methods, and they are stable over the interval $0 < \mu < 1$.

THEOREM 5. *For the Padé schemes with $R = S$ and $m \geqslant n$, the inequalities* (5.1) *are sufficient as well as necessary for stability.*

PROOF. Each term in the decisive quantity $D = |Q|^2 - |P|^2$, according to Theorem 3, is a positive multiple of the corresponding term

$$\begin{aligned}\Omega_k(\mu) &= (-1)^{r+s+1}(-n-\lambda)_k(-m+\lambda)_k\\ &= (-1)^{r+s+1}(-r-S-\mu)_k(-s-R+\mu)_k.\end{aligned}$$

We rewrite the last expressions as

$$\begin{aligned}(-r-S-\mu)_k &= (-r-S-\mu)(-r-S-\mu+1)\cdots(-r-S+k-1-\mu)\\ &= (-1)^{r+S+1}[\mu(\mu+1)\cdots(\mu+R+S)][(1-\mu)\cdots(k-r-S-1-\mu)]\\ &= (-1)^{r+S+1}(\mu)_{r+S+1}(1-\mu)_{k-r-S-1}\end{aligned}$$

and as

$$(-s-R+\mu)_k = (-1)^{s+R}(\mu)_{k-s-R}(1-\mu)_{s+R}.$$

Therefore

$$\Omega_k(\mu) = (-1)^{R+S}(\mu)_{r+S+1}(1-\mu)_{k-r-S+1}(\mu)_{k-s-R}(1-\mu)_{s+R}. \tag{5.2}$$

Now we can verify that $D \geqslant 0$ for $R = S \leqslant r \leqslant s \leqslant r + 2$. In this case the first index k to appear in the expression for D is

$$k = \left[\frac{r+s+R+S}{2}\right] + 1 = \max(s+R, r+S+1).$$

But Ω_k in (5.2) is positive for this and all larger k, for any $0 < \mu < 1$. Therefore $D \geqslant 0$, and the von Neumann condition is verified.

There remains only the pole condition: the Padé methods under consideration are stable if and only if the Jacobi polynomials $P_{2R}^{(r-R+\mu,s-R-\mu)}$ have R zeros in C^- and in C^+. We will show that at $\mu = 0$ we have $\xi_+ = \xi_- = R$, $\xi_0 = 0$, for all three of the cases $s = r$, $r + 1$, and $r + 2$. Applying Lemma 4.2, that completes the proof of stability for $0 < \mu < 1$.

If $s = r$ we have the ultraspherical polynomial $P_{2R}^{(r-R,r-R)}$, symmetric about the origin. Hence indeed $\xi_+(0) = \xi_-(0) = R$.

If $s = r + 1$ then the polynomial is $P_{2R}^{(r-R,r-R+1)}$, and Lemma 4.4(b) gives at once $\xi_+(0) = \xi_-(0) = R$.

Finally, $s = r + 2$ yields $P_{2R}^{(r-R,r-R+2)}$. Let us examine, in general, the zeros of $P_{2R}^{(\gamma-1,\gamma+1)}$ as a function of $\gamma > 0$. By setting $n = 2R$, $\alpha = \gamma$, $\beta = \gamma + 1$, $z = 0$, in

$$P_n^{(\alpha,\beta-1)}(z) - P_n^{(\alpha-1,\beta)}(z) = P_{n-1}^{(\alpha,\beta)}(z)$$

we obtain

$$P_{2R}^{(\gamma-1,\gamma+1)}(0) = P_{2R}^{(\gamma,\gamma)}(0) - P_{2R-1}^{(\gamma,\gamma+1)}(0). \tag{5.3}$$

But because of (4.6),

$$P_{2R-1}^{(\gamma,\gamma+1)}(0) = P_{2R-1}^{(\gamma,\gamma)}(0) + \tfrac{1}{2}P_{2(R-1)}^{(\gamma+1,\gamma+1)}(0)$$

and, by symmetry, $P_{2R-1}^{(\gamma,\gamma)}(0) = 0$. A substitution in (5.3) yields.

$$P_{2R}^{(\gamma-1,\gamma+1)}(0) = P_{2R}^{(\gamma,\gamma)}(0) - \tfrac{1}{2}P_{2(R-1)}^{(\gamma+1,\gamma+1)}(0). \tag{5.4}$$

This quantity is not zero; the link to Gegenbauer polynomials [**13**, pp. 277–278] gives

$$P_{2R}^{(\gamma,\gamma)}(0) = \frac{(\gamma+1)_{2R}}{(2\gamma+1)_{2R}}C_{2R}^{\gamma+1/2}(0) = (-1)^R\frac{(\gamma+1)_{2R}(\gamma+1/2)_{2R}}{(2\gamma+1)_{2R}(2R)!},$$

and the final term in (5.4) reinforces this sign.

Therefore, with $\mu \to 1$ in Lemma 4.3, $P_{2R}^{(\gamma-1,\gamma+1)}$ has R zeros in $(-1, 0)$ and R zeros in $(0, 1)$ for every positive integer γ. We take $\gamma = r - R + 1$, and so $\xi_+(0) = \xi_-(0) = R$ for $P_{2R}^{(r-R,r-R+2)}$. Thus the cases $R = S \leqslant r \leqslant s \leqslant r + 2$ are stable.

The hypotheses $R = S$ and $m \geqslant n$ admit only one other possibility that satisfies the necessary conditions (5.1). It is $r = S - 1$, $s = R + 1$, whose stability was established in Theorem 4A. Therefore the proof of Theorem 5 is complete; for these Padé approximations the necessary conditions for stability are also sufficient.

There is one more class for which we can prove the same result. It is the case $r = s$ and $n \geqslant m$, of centered and "mostly implicit" schemes; stability is confirmed exactly when $R \leqslant S \leqslant R + 2$, by studying the reciprocal of P/Q and reversing the sign of μ.

The most general case, in which r, s, R, and S are entirely arbitrary, is still too delicate even to conjecture the right conditions for stability. Our proofs depended on the hypergeometric identity for $|Q|^2 - |P|^2$ in Theorem 3, and on the verification that every term in that sum was positive. In general this is too much to expect. The identity will remain correct and fundamental, but with terms of opposite sign the von Neumann condition becomes a close thing. Nevertheless we are optimistic about the possibility of a complete solution.

References

1. G. A. Baker, *Essentials of Padé approximants*, Academic Press, New York, 1975.

2. G. Dahlquist, *A special stability problem for linear multistep methods*, BIT **3** (1963), 27–43.

3. ______, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. **4** (1956), 33–53.

4. J. W. Daniel and R. E. Moore, *Computation and theory in ordinary differential equations*, Freeman, San Francisco, 1970.

5. B. L. Ehle, *A-stable methods and Padé approximations to the exponential*, Siam J. Numer. Anal. **4** (1973), 571–580.

6. B. Engquist and S. Osher, *One-sided difference approximations for non-linear conservation laws*, Math. Comp. **36** (1981), 321–352.

7. A. Erdelyi, et al., *Higher transcendental functions*. I, McGraw-Hill, New York, 1953.

8. A. Iserles and M. J. D. Powell, *On the A-acceptability of rational approximations that interpolate the exponential function*, IMA J. Numer. Anal. **1** (1981), 241–251.

9. A. Iserles, *Order stars and a saturation theorem for first order hyperbolics*, IMA J. Numer. Anal. (to appear).

10. ______, *A note on Padé approximations and generalized hypergeometric functions*, BIT **19** (1979), 543–545.

11. ______, *On the generalized Padé approximations to the exponential function*, Siam J. Numer. Anal. **16** (1979), 631–636.

12. P. D. Lax, *On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients*, Comm. Pure Appl. Math. **14** (1961), 497–520.

13. E. D. Rainville, *Special functions*, Macmillan, New York, 1967.

14. G. Strang, *Trigonometric polynomials and difference methods of maximum accuracy*, J. Math. Phys. **41** (1962), 147–154.

15. ______, *Accurate partial difference methods*. II. *Non-linear problems*, Numer. Math. **6** (1964), 37–46.

16. ______, *Wiener-Hopf difference equations*, J. Math. Mech. **13** (1964), 37–46.

17. ______, *Implicit difference methods for initial-boundary value problems*, J. Math. Anal. Appl. **16** (1966), 188–198.

18. G. Szegö, *Orthogonal polynomials*, Amer. Math. Soc. Colloq. Publ., vol. 23, Amer. Math. Soc., Providence, R.I., 1939.

19. G. E. Wanner, E. Hairer and S. P. Nørsett, *Order stars and stability theorems*, BIT **18** (1978), 475–489.

20. O. B. Widlund, *On Lax's theorem on Friedrichs type finite difference schemes*, Comm. Pure Appl. Math. **24** (1971), 117–123.

DEPARTMENT OF MATHEMATICS, KINGS' COLLEGE, CAMBRIDGE, ENGLAND

DEPARTMENT OF MATHEMATICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139