ON HIGHER ORDER STABLE IMPLICIT METHODS FOR SOLVING PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS

By Richard S. Varga

Reprinted from Journal of Mathematics and Physics
Vol. XL, No. 3, October, 1961
Printed in U.S.A.

Reprinted from JOURNAL OF MATHEMATICS AND PHYSICS
Vol. 40, No. 3, October, 1961
Printed in U.S.A.

ON HIGHER ORDER STABLE IMPLICIT METHODS FOR SOLVING PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS

BY RICHARD S. VARGA

1. Introduction. The numerical solutions of self-adjoint parabolic partial differential equations, such as

$$\frac{\partial u(x,t)}{\partial t} = \left\{\frac{\partial}{\partial x}\, p(x)\, \frac{\partial u}{\partial x}\right\} - \sigma(x)u(x,t) + s(x), \qquad a < x < b, \quad t > 0, \tag{1.1}$$

in one space variable where

$$u(x, 0+) = g(x), \qquad a \leqq x \leqq b, \tag{1.2}$$

and

$$u(a,t) = \alpha, \qquad u(b,t) = \beta, \qquad t > 0, \tag{1.3}$$

have been considered by many authors (e.g. [5], [12], [17], [19]), and concepts such as the *stability* and *consistency* of discrete approximations, and *convergence* of the discrete approximations to the solution of (1.1) have been previously introduced. However, while most previous papers on this topic simultaneously discretized *both* space and time variables and analysed by means of Fourier series the properties of the resulting system of linear equations, our approach, like that of [8], is to discretize first only the space variables, leaving the time variable continuous, and to analyse the resulting system of ordinary differential equations by matrix methods, avoiding Fourier series arguments.

As a result of this approach, we shall introduce new numerical methods which are unconditionally stable, and show that well known numerical methods for solving (1.1), viz. the explicit method, the backward time implicit method [14], and the Crank-Nicolson implicit method [19], can be generated (Theorem 2) from particular Padé rational approximations $r_{p,q}(z)$ of e^{-z}. These results are applicable to a large class of parabolic partial differential equations which are approximated on non-uniform spatial meshes.

Considering discrete spatial meshes, we also give error estimates for the difference between the continuous time problem, and the discrete time problem used in actual numerical calculations. This consideration leads to *a priori* estimates of the discrete time mesh in order to achieve a particular accuracy. Finally, because of the monotonicity of the error of Padé rational approximations for e^{-z}, we consider also the best rational approximations for e^{-z} in the Chebyshev sense. The interesting feature of the associated matrix approximations is their accuracy for *any* size time step (§5), and as a result, offer the possibility of obtaining approximations (of limited accuracy) of (1.1) for any time t from a *single* time step.

2. The Continuous Time-Discrete Space Equations. To be explicit, we consider in particular the parabolic partial differential equation of (1.1–3), where we assume that $p(x)$ and $\sigma(x)$ are given continuous functions[1] on $a \leqq x \leqq b$ with

$$p(x) > 0, \qquad \sigma(x) \geqq 0, \qquad a \leqq x \leqq b. \tag{2.1}$$

The given functions $s(x)$ and $g(x)$ are also assumed to be continuous on $a \leqq x \leqq b$. If $\{x_i\}_{i=0}^{n+1}$ is any set of spatial mesh points with $a = x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} = b$, define $h_i = x_i - x_{i-1}$ for $1 \leqq i \leqq n+1$. Upon integrating (1.1) from $x_i = h_i/2$ to $x_i + h_{i+1}/2$, and using central difference approximations as in [20], we obtain, denoting $u(x_i, t)$ by $u_i(t)$,

$$\begin{aligned}\left(\frac{h_i + h_{i+1}}{2}\right)\frac{d}{dt}u_i(t) &= -u_i(t)\left\{\frac{p(x_i + \tfrac{1}{2}h_{i+1})}{h_{i+1}} + \frac{p(x_i - \tfrac{1}{2}h_i)}{h_i} + \sigma_i\left(\frac{h_i + h_{i+1}}{2}\right)\right\} \\ &\quad + \frac{p(x_i + \tfrac{1}{2}h_{i+1})}{h_{i+1}}u_{i+1}(t) + \frac{p(x_i - \tfrac{1}{2}h_i)}{h_i}u_{i-1}(t) + \left(\frac{h_i + h_{i+1}}{2}\right)s_i,\end{aligned} \tag{2.2}$$

for $t > 0$, $1 \leqq i \leqq n$, which, along with $u_i(0) = g_i$, defines our continuous time-discrete space approximation of (1.1). In matrix notation, this becomes

$$D\frac{d\mathbf{u}(t)}{dt} = -A\mathbf{u}(t) + \mathbf{s}, \qquad t \geqq 0, \tag{2.3}$$

where

$$\mathbf{u}(0) = \mathbf{g}. \tag{2.4}$$

The matrix D is a positive real diagonal $n \times n$ matrix, and, from (2.1) and (2.2), A is a real symmetric and positive definite $n \times n$ matrix. More precisely, A is a tridiagonal $n \times n$ *Stieltjes matrix* [3]. It should be pointed out that the boundary conditions of (1.3) are implicit in the vector $\mathbf{s}$ of (2.3).

In what is to follow, we need only assume that D is a positive real diagonal $n \times n$ matrix, and that A is a positive definite Hermitian $n \times n$ matrix. Because of this, our results apply to parabolic partial differential equations in more space variables, such as

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} = \operatorname{div}\{p(\mathbf{x})\operatorname{grad} u(\mathbf{x}, t)\} - \sigma(\mathbf{x})u(\mathbf{x}, t) + s(\mathbf{x}), \tag{2.5}$$

for which there are spatial discretization [20] of the form (2.3) satisfying these assumptions.

We now normalize the system of ordinary differential equations of (2.3). If

[1] The derivation [20], based on integration, of the spatial difference equations of this section shows that the case where $p(x)$ and $\sigma(x)$ are piecewise continuous in $0 \leqq x \leqq 1$ can be treated similarly. See also [8] and [15]

$D^{\frac{1}{2}}$ is the positive real diagonal $n \times n$ matrix whose square is D, let

$$\mathbf{v}(t) \equiv D^{\frac{1}{2}}\mathbf{u}(t), \qquad D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \equiv B. \tag{2.6}$$

Then, (2.4) can be written as

$$\frac{d\mathbf{v}(t)}{dt} = -B\mathbf{v}(t) + \mathbf{\Sigma}, \qquad t \geqq 0 \tag{2.7}$$

where $\mathbf{\Sigma} = D^{-\frac{1}{2}}\mathbf{s}$, and

$$\mathbf{v}(0) = D^{+\frac{1}{2}}\mathbf{g} \equiv \overset{*}{\mathbf{g}}. \tag{2.8}$$

From the definition of the $n \times n$ matrix B in (2.6), we see that B is also a positive definite Hermitian matrix. Moreover, there is a 1-1 correspondence between the vectors $\mathbf{u}(t)$ and $\mathbf{v}(t)$.

That the ordinary matrix differential equation of (2.7) is *stable*[2] [2] from the point of view of differential equations follows immediately from the positive definite Hermitian nature of B, and the solution of (2.7), subject to the boundary condition (2.8) is given explicitly by

$$\mathbf{v}(t) = e^{-tB}\overset{*}{\mathbf{g}} + (I - e^{-tB})B^{-1}\mathbf{\Sigma}, \qquad t \geqq 0, \tag{2.9}$$

or equivalently,

$$\mathbf{v}(t_0 + \Delta t) = e^{-\Delta t B}\mathbf{v}(t_0) + (I - e^{-\Delta t B})B^{-1}\mathbf{\Sigma} \tag{2.10}$$

where $t_0 \geqq 0$, $\Delta t \geqq 0$, the matrix e^{-tB} being defined by the convergent matrix series $I - tB + \frac{1}{2}t^2B^2 - \cdots$.

It is now interesting to show the relationship between well known numerical methods for approximating (1.1) and the continuous time-discrete space approximation of (2.7). In terms of (2.7), these well known approximations are

$$\textit{Explicit:} \qquad \left(\frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t}\right) = -B\mathbf{v}(t) + \mathbf{\Sigma}; \tag{2.11}$$

$$\textit{Backward Implicit:} \qquad \left(\frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t}\right) = -B\mathbf{v}(t + \Delta t) + \mathbf{\Sigma}; \tag{2.11'}$$

$$\textit{Crank-Nicolson:} \left(\frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t}\right) = -\frac{B}{2}(\mathbf{v}(t + \Delta t) + \mathbf{v}(t)) + \mathbf{\Sigma}. \tag{2.11*}$$

Solving for $\mathbf{v}(t + \Delta t)$ in terms of $\mathbf{v}(t)$ and $\mathbf{\Sigma}(t)$, these expressions reduce to

$$\textit{Explicit:} \qquad \mathbf{v}(t + \Delta t) = (I - \Delta tB)\mathbf{v}(t) + \Delta t\mathbf{\Sigma}; \tag{2.12}$$

$$\textit{Backward Implicit:} \quad \mathbf{v}(t + \Delta t) = (I + \Delta tB)^{-1}\mathbf{v}(t) + \Delta t(I + \Delta tB)^{-1}\mathbf{\Sigma}; \tag{2.12'}$$

$$\textit{Crank-Nicolson:} \quad \mathbf{v}(t + \Delta t) = (I + \tfrac{1}{2}\Delta tB)^{-1}(I - \tfrac{1}{2}\Delta tB)\mathbf{v}(t) + \Delta t(I + \tfrac{1}{2}\Delta tB)^{-1}\mathbf{\Sigma}. \tag{2.12*}$$

Comparing with (2.10), it is clear that these three methods each give rise to a particular rational approximation for the exponential matrix $e^{-\Delta tB}$.

[2] See also [8].

3. The Padé Table for e^{-z}. The Padé table [18, 22] of a function $f(z)$ analytic in a region containing the point $z = 0$ is a double entry table, such that the rational approximation for $f(z)$,

$$f(z) = \frac{n_{p,q}(z)}{d_{p,q}(z)} + 0(|z|^r), \qquad |z| \to 0, \tag{3.1}$$

where $n_{p,q}(z)$ and $d_{p,q}(z)$ are polynomials of degree q and p, respectively, gives the best approximation (highest power r) for $f(z)$ in the neighborhood of the origin. Except for a multiplicative factor, for fixed non-negative integers p and q the polynomials $n_{p,q}(z)$ and $d_{p,q}(z)$ are uniquely determined, and $r = p + q + 1$. For the function $f(z) = e^{-z}$, the entries of the Padé table are given explicitly [11, 18] by

$$n_{p,q}(z) = \sum_{k=0}^{q} \frac{(p+q-k)!q!}{(p+q)!k!(q-k)!}(-z)^k, \tag{3.2}$$

and

$$d_{p,q}(z) = \sum_{k=0}^{p} \frac{(p+q-k)!p!}{(p+q)!k!(p-k)!}z^k, \tag{3.2'}$$

and if

$$e^{-z} \equiv \frac{n_{p,q}(z)}{d_{p,q}(z)} + R_{p,q}(z), \tag{3.3}$$

then [18]

$$R_{p,q}(z) = \frac{(-1)^{q+1}z^{p+q+1}}{(p+q)!d_{p,q}(z)} \int_0^1 e^{-z(1-u)}u^q(1-u)^p\,du. \tag{3.4}$$

For $z \geqq 0$, we can apply the mean value theorem for integrals to the expression above, and it follows [11] that

$$R_{p,q}(z) = \frac{(-1)^{q+1}p!q!z^{p+q+1}e^{-\theta}}{(p+q)!(p+q+1)!d_{p,q}(z)}, \qquad z \geqq \theta \geqq 0. \tag{3.5}$$

The function e^{-z} is uniformly bounded by unity for all $z \geqq 0$. It is natural to ask the same question of the Padé approximations for e^{-z}.

Lemma 1. For arbitrary p and $q \geqq 0$, $d_{p,q}(z) \geqq 1$ for all $z \geqq 0$.

Proof. The coefficients of the polynomial $d_{p,q}(z)$ are all positive real numbers, with constant coefficient unity for all $p, q \geqq 0$. Thus, if $z \geqq 0$, then $d_{p,q}(z) \geqq 1$.

Lemma 2. $\left|\frac{n_{p,q}(z)}{d_{p,q}(z)}\right| \leqq 1$ for all $z \geqq 0$ if and only if $p \geqq q$.

Proof. Clearly, $\left|\frac{n_{p,q}(z)}{d_{p,q}(z)}\right| = 0(z^{q-p})$ for $z \to +\infty$ if $q \geqq p$. Thus, the ratio being bounded for all $z \geqq 0$ implies that $p \geqq q$. Conversely, if $p \geqq q$, then $|d_{p,q}(z)| \geqq |n_{p,q}(z)|$ if

$$\frac{(p+q-k)!p!}{(p+q)!k!(p-k)!} \geqq \frac{(p+q-k)!q!}{(p+q)!k!(q-k)!}, \qquad 0 \leqq k \leqq q,$$

which is obviously true.

We remark that the error $R_{p,q}(z)$ of the Padé approximation (3.3) for e^{-z} is, from (3.5) and Lemma 1, of one sign for all $z \geqq 0$ for all non-negative integers p and q.

We now consider formal Padé approximations[3] for the exponential matrix e^{-tB} of the form

$$e^{-tB} \doteq [d_{p,q}(tB)]^{-1} \cdot [n_{p,q}(tB)] \equiv M_{p,q}(tB) \tag{3.6}$$

stemming naturally from the approximations of (3.3). We make the usual definition [5].

Definition 1. Let $M(t)$ be an $n \times n$ matrix whose entries are functions of the parameter t. Then $M(t_0)$ is *stable* if and only if all the eigenvalues $\lambda_i(t_0)$ of $M(t_0)$ satisfy $|\lambda_i(t_0)| \leqq 1$.

For the $n \times n$ positive definite Hermitian matrix B, let $\rho(B)$ denote the *spectral radius* [10] of B, i.e. $\rho(B) = \max_{1 \leqq i \leqq n} |\lambda_i|$, where the λ_i's are the (real and positive) eigenvalues of B. Moreover, let

$$\tau_{p,q} \equiv \sup \left\{ l \geqq 0 \,\middle|\, \left| \frac{n_{p,q}(z)}{d_{p,q}(z)} \right| \leqq 1 \quad \text{for} \quad 0 \leqq z \leqq l \right\}. \tag{3.7}$$

Theorem 1. Let B be an $n \times n$ positive definite Hermitian matrix, and let

$$M_{p,q}(\Delta tB) = [d_{p,q}(\Delta tB)]^{-1}[n_{p,q}(\Delta tB)]. \tag{3.8}$$

Then, $M_{p,q}(\Delta tB)$ is stable for all Δt with $0 \leqq \Delta t \leqq \Lambda$ if and only if

$$\Lambda \leqq \tau_{p,q}/\rho(B). \tag{3.9}$$

Proof. If the eigenvalues of B are λ_i, then the eigenvalues of $d_{p,q}(\Delta tB)$ are obviously $d_{p,q}(\Delta t\lambda_i)$, and since Δt is non-negative and each λ_i is real and positive, Lemma 1 show that the matrix $d_{p,q}(\Delta tB)$ is non-singular for any p and q, and any non-negative Δt. Thus the matrix $M_{p,q}(\Delta tB)$ is defined for all $\Delta t \geqq 0$ and all non-negative integers p and q.

The eigenvalues of $M_{p,q}(\Delta tB)$ are $n_{p,q}(\Delta t\lambda_i)/d_{p,q}(\Delta t\lambda_i)$. It is clear that if $0 \leqq \Delta t \leqq \tau_{p,q}/\rho(B)$, the definition of $\tau_{p,q}$ shows that $M_{p,q}(\Delta tB)$ is stable for Δt in this range. Conversely, if $M_{p,q}(\Delta tB)$ is stable for all Δt with $0 \leqq \Delta t \leqq \Lambda$, then as $\rho(B)$ is an eigenvalue of B,

$$\left| \frac{n_{p,q}(\Delta t\rho(B))}{d_{p,q}(\Delta t\rho(B))} \right| \leqq 1$$

for all such Δt, and thus, from (3.7), $\Delta t\rho(B) \leqq \Lambda\rho(B) \leqq \tau_{p,q}$, which completes the proof.

From Lemma 2, we see that $\tau_{p,q} = +\infty$ if and only if $p \geqq q$, which gives us

Corollary 1. Let B be an $n \times n$ positive definite Hermitian matrix. Then, $M_{p,q}(tB)$ is stable for all $t \geqq 0$ if and only if $p \geqq q$.

We explicitly list now particular entries from the Padé table for e^{-z}.

[3] See also [13] for rational approximations of other operators useful in numerical analysis.

$$(3.10)\qquad \frac{n_{0,1}(z)}{d_{0,1}(z)} = 1 - z; \qquad \frac{n_{1,0}(z)}{d_{1,0}(z)} = \frac{1}{1+z} \qquad \frac{n_{1,1}(z)}{d_{1,1}(z)} = \frac{1 - \frac{1}{2}z}{1 + \frac{1}{2}z}.$$

If we now form, using (3.6) and (3.10), the matrices $M_{0,1}(\Delta tB)$, $M_{1,0}(\Delta tB)$, and $M_{1,1}(\Delta tB)$ as particular approximations to $e^{-\Delta tB}$, and these approximations are substituted respectively in (2.10), we obtain precisely the equations (2.12)–(2.12*). Thus, with Corollary 1, we have

Theorem 2. Let B be an $n \times n$ positive definite Hermitian matrix. Then, the Padé exponential approximations $M_{0,1}(\Delta tB)$, and $M_{1,0}(\Delta tB)$, and $M_{1,1}(\Delta tB)$ generate respectively from (2.10) the explicit, the backward implicit, and the Crank-Nicolson methods of (2.12)–(2.12*). Moreover, the latter two methods are stable for all $\Delta t \geqq 0$.

Of course, it is well known (e.g. [8]) that these latter two methods are unconditionally stable. But it now follows that any Padé exponential approximation $M_{p,q}(\Delta tB)$ with $p \geqq q$ also gives rise to an unconditionally stable method. It is interesting that the Padé exponential approximation $M_{0,4}(\Delta tB)$ generates, in a like manner, the fourth-order Runge-Kutta method for approximating (2.7).

We now consider the cases where $0 \leqq p < q$. It is first of all not difficult to algebraically determine $\tau_{p,q}$, and simply applying the Hadamard-Gerschgorin Lemma to the matrix B gives an upper bound estimate of $\rho(B)$ which, when inserted in (3.9), gives a conservative estimate for a stability interval for Δt. To illustrate this, consider the explicit method of (2.12), which corresponds to $M_{0,1}(\Delta tB)$. It is clear that $\tau_{0,1} = 2$, so that the explicit method has an interval of stability $0 \leqq \Delta t \leqq 2/\rho(B)$. If we select a uniform spatial mesh $h = \Delta x$, in the special case of the heat equation, $p(x) \equiv 1$, $\sigma(x) \equiv 0$ in (1.1), the matrix B obtained through normalizing, by means of (2.6), the matrix A of (2.2), is given by

$$(3.11)\qquad B = \frac{1}{(\Delta x)^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix},$$

and thus, by the Hadamard-Gerschgorin Lemma, $\rho(B) \leqq 4/(\Delta x)^2$. Therefore, from (3.9), it is interesting, but not surprising, that we obtain the well known [4, 17] stability interval

$$(3.12)\qquad 0 \leqq \frac{\Delta t}{(\Delta x)^2} \leqq \frac{1}{2}$$

for the explicit method, in this special case. Notice that this was obtained without the use of Fourier series, and that the analysis above is equally valid for non-uniform spatial meshes.

4. Error Analysis for the Padé Approximations. We first write (2.10) in the equivalent form

$$\mathbf{v}(t_0 + \Delta t) = B^{-1}\mathbf{\Sigma} + e^{-\Delta t B}\{\mathbf{v}(t_0) - B^{-1}\mathbf{\Sigma}\}. \tag{4.1}$$

With our definition (3.6) of the matrix Padé approximations $M_{p,q}(\Delta t B)$ of $e^{-\Delta t B}$, we correspondingly define

$$\mathbf{w}_{p,q}((j+1)\Delta t) = B^{-1}\mathbf{\Sigma} + M_{p,q}(\Delta t B)\{\mathbf{w}_{p,q}(j\Delta t) - B^{-1}\mathbf{\Sigma}\}, \tag{4.2}$$

for $j \geqq 0$, $\Delta t > 0$, as discrete Padé approximations of the solution vector $\mathbf{v}((j+1)\Delta t)$, where

$$\mathbf{w}_{p,q}(0) \equiv \overset{*}{\mathbf{g}} = \mathbf{v}(0). \tag{4.2'}$$

In this form (4.2), it follows that

$$\mathbf{w}_{p,q}(j\Delta t) = B^{-1}\mathbf{\Sigma} + M^{j}_{p,q}(\Delta t B)\{\overset{*}{\mathbf{g}} - B^{-1}\mathbf{\Sigma}\}, \qquad j \geqq 1. \tag{4.3}$$

In order to carry out an error analysis between the continuous time-discrete space solution (4.1) and the Padé discrete time-discrete space solution (4.3), we define now the quantity

$$r^{(m)}_{p,q}(z) = \sup_{0 \leqq x \leqq z} | e^{-x} - M^{m}_{p,q}(x/m) | \tag{4.4}$$

for all non-negative integers p and q, and positive integers m. It is easy to see, using (3.5), that for fixed p, q, and z, $r^{(m)}_{p,q}(z) \to 0$ as $m \to +\infty$.

If $\|\mathbf{x}\| \equiv (\sum_{i=1}^{n} | x_i |^2)^{\frac{1}{2}}$ denotes the euclidean norm of the vector $\mathbf{x}$ whose components are x_i, $1 \leqq i \leqq n$, then the spectral norm [10] $\|C\|$ of an arbitrary $n \times n$ complex matrix C is defined by

$$\| C \| \equiv \sup_{\mathbf{x} \neq 0} \frac{\| C\mathbf{x} \|}{\| \mathbf{x} \|}. \tag{4.5}$$

Thus,

$$\|C\mathbf{x}\| \leqq \|C\| \cdot \|\mathbf{x}\| \tag{4.5'}$$

for any vector $\mathbf{x}$. It is moreover known [10] that if C is Hermitian (or normal), then $\|C\| = \rho(B)$.

Lemma 3. Let B be an $n \times n$ positive definite Hermitian matrix. For any vector $\mathbf{x}$,

$$\|(e^{-tB} - M^{m}_{p,q}(tB/m))\mathbf{x}\| \leqq r^{(m)}_{p,q}(t\rho(B)) \cdot \|\mathbf{x}\|, \tag{4.6}$$

for all non-negative integers p, q, all positive integers m, and all $t \geqq 0$.

Proof. From the discussion above, it is only necessary to show that

$$\| e^{-tB} - M^{m}_{p,q}(tB/m) \| \leqq r^{(m)}_{p,q}(t\rho(B)).$$

But as the matrix $e^{-tB} - M^{m}_{p,q}(tB/m)$ is Hermitian,

$$\left\| e^{-tB} - M^{m}_{p,q}(tB/m) \right\| = \max_{1 \leqq i \leqq n} \left| e^{-t\lambda_i} - M^{m}_{p,q}\left(\frac{t\lambda_i}{m}\right) \right| \leqq r^{(m)}_{p,q}(t\rho(B)),$$

which completes the proof.

From (4.1) and (4.3), we have

$$\mathbf{v}(m\Delta t) - \mathbf{w}_{p,q}(m\Delta t) = \{e^{-m\Delta t B} - M^m_{p,q}(\Delta t B)\}(\overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma}),$$

and as an immediate consequence of (4.5′) and Lemma 3, we obtain

Theorem 3. Let $\mathbf{v}(t)$ be the solution of the matrix differential equation (2.7), subject to the initial condition of (2.8), and let $\mathbf{w}_{p,q}(m\Delta t)$ be defined from (4.2) and (4.2′). Then, for all non-negative integers p and q, all positive integers m, and all $\Delta t > 0$,

$$\| \mathbf{v}(m\Delta t) - \mathbf{w}_{p,q}(m\Delta t) \| \leqq r^{(m)}_{p,q}(m\Delta t \rho(B)) \cdot \| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \|. \tag{4.7}$$

To illustrate the usefulness of higher order implicit methods, consider the particular problem (3.11) with a uniform spatial mesh $\Delta x = 0.1$, and suppose that we seek approximations (cf. [6, p. 138]) to $\mathbf{v}(T)$, where $T = 10^4$. With $\rho(B)$ bounded above in this case by $4/(\Delta x)^2 = 400$, we list estimates for the least positive integer $m_{p,q} = T/\Delta t$ for which the coefficient of $\| \overset{*}{\mathbf{g}} - \mathbf{h} \|$ in (4.6) is less than or equal to 0.0073, for various choices of non-negative integers p and q:

$$m_{0,1} \doteq 2.0 \times 10^6, \qquad m_{1,1} \doteq 2.2 \times 10^3, \qquad m_{2,2} \doteq 1.2 \times 10^3. \tag{4.8}$$

In general, it is now clear how *a priori* estimates of $m_{p,q}$, the total number of time steps, can be determined to insure a particular accuracy k between $\mathbf{v}(T)$ and $\mathbf{w}_{p,q}(T)$, where T is given. If $\tilde{\rho}$ is an estimate of the spectral radius $\rho(B)$ of the matrix B, we algebraically determine the least positive solution m of

$$r^m_{p,q}(T\tilde{\rho}) \leqq \frac{k}{\| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \|}, \tag{4.9}$$

and set $\Delta t = T/m$.

While stability considerations imply that $p \geqq q$ in the Padé matrix approximations $M_{p,q}(tB)$ of (3.8) for e^{-tB}, practical considerations imply that one would examine only the diagonal entries $p = q$ of the Padé table. This follows from the fact that forming the matrix polynomial $d_{p,q}(tB)$ of degree p requires explicitly the matrices B^ν, $q \leqq \nu \leqq p$ which could thus also be used in the formation of $n_{p,q}(tB)$ if q were increased to p. Moreover, the matrix inversion of $d_{p,q}(tB)$ usually outweighs in computational effort the formation of $n_{p,q}(tB)$, even when $p = q$.

The diagonal Padé matrix approximations $M_{p,p}(tB)$ can also be derived as approximants of the known continued fraction expansion [22, p. 348] for e^{-z}:

$$e^{-z} = \cfrac{1}{1 + \cfrac{z}{1 - \cfrac{z}{2 + \cfrac{z}{3 - \cfrac{z}{2 + \cfrac{z}{5 - \cfrac{z}{2 - \ddots}}}}}}}, \tag{4.10}$$

and these diagonal Padé matrix approximations can also be derived from repeated differentiations of (2.7), coupled with matrix Taylor series expansions based on central differences.

We shall show in the next section why the particular constant $k = 0.0073$ was chosen in the above discussion.

5. Chebyshev Rational Approximations. We had remarked in §3 that the error $R_{p,q}(z)$ of the Padé approximation for e^{-z} is of one sign for all $z \geqq 0$. While the asymptotic behavior of $R_{p,q}(z)$ in the neighborhood of $z = 0$ allows us, from Theorem 3, to obtain arbitrarily high accuracy between the continuous time and discrete time models, these Padé rational approximations nevertheless are not the best rational approximations for e^{-z} in the Chebyshev sense [1]. For a fixed non-negative integer p, consider

$$H_p \equiv \max_{z \geqq 0} \left| e^{-z} - \frac{n_p(z)}{d_p(z)} \right|, \tag{5.1}$$

where $n_p(z)$ and $d_p(z)$ are any real polynomials of degree p in z, and where $d_p(z) \not\equiv 0$ for all $z \geqq 0$. Now let

$$\hat{H}_p = \min H_p \tag{5.2}$$

where the min is taken over all such real polynomials $n_p(z)$ and $d_p(z)$. The celebrated theorem of Chebyshev (see [1]) states that the rational function $Q_p(z) = n_p(z)/d_p(z)$ which minimizes H_p is uniquely determined (assuming that $Q_p(z)$ is irreducible), and is characterized by the property that $e^{-z} - Q_p(z)$, with alternate change of signs, takes on the value $\hat{H}_p$ not less than $2p + 2$ times. Because of this required oscillation of the error, the diagonal Padé approximations for e^{-z} of §3 are evidently different from the Chebyshev rational approximations $Q_p(z)$ for e^{-z}. We now exhibit some of the rational functions $Q_p(z)$ which best approximate e^{-z} in the Chebyshev sense.[4]

$$\begin{aligned} Q_1(z) &= \frac{1 - 0.108\,196\,z}{0.937\,355 + 1.618\,932\,z}; \qquad \hat{H}_1 = 0.066\,83 \\ Q_2(z) &= \frac{1 - 0.189\,729\,z + 0.004\,242\,15\,z^2}{1.007\,413 + 0.674\,264\,z + 0.576\,492\,z^2}; \qquad \hat{H}_2 = .0073\,59. \end{aligned} \tag{5.3}$$

The particular choice of $k = 0.0073$ in §4 is now clear.

For an error analysis based on these Chebyshev rational approximations, let

$$\mathbf{w}_p(t) \equiv B^{-1}\boldsymbol{\Sigma} + Q_p(tB)\{\overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma}\}, \quad t \geqq 0, \tag{5.4}$$

be the Chebyshev rational approximation of (4.1) corresponding to a *single* time step of length t, where

$$\mathbf{w}_p(0) \equiv \overset{*}{\mathbf{g}} = \mathbf{v}(0), \tag{5.4'}$$

and

[4] The author wishes to express his thanks to Mr. Fred Chapman of the Case Institute of Technology Computing Center for obtaining these results on the IBM 650.

$$Q_p(tB) \equiv [d_p(tB)]^{-1} \cdot [n_p(tB)]. \tag{5.5}$$

Theorem 4. Let $\mathbf{v}(t)$ be the solution of the matrix differential equation of (2.7), subject to the initial condition of (2.8), and let $\mathbf{w}_p(t)$ be defined from (5.3). Then, for *all* $t \geqq 0$

$$\| \mathbf{v}(t) - \mathbf{w}_p(t) \| \leqq \hat{H}_p \cdot \| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \| . \tag{5.6}$$

Proof. From (4.5′),

$$\| \mathbf{v}(t) - \mathbf{w}_p(t) \| \leqq \| e^{-tB} - Q_p(tB) \| \cdot \| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \| .$$

Using the fact that the matrix $e^{-tB} - Q_p(tB)$ is Hermitian, then

$$\| e^{-tB} - Q_p(tB) \| = \max_{1 \leqq i \leqq n} \left| e^{-t\lambda_i} - \frac{n_p(t\lambda_i)}{d_p(t\lambda_i)} \right| \leqq \hat{H}_p ,$$

which completes the proof.

Thus, in comparison with the results in (4.7), it would appear that we can obtain *in one step* approximately the same accuracy with the Chebyshev rational approximation for $p = 2$, as, say, that given by the Crank-Nicolson method, with approximately 10^3 time steps. This requires some explanation, however. If the quantity $\| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \|$ is very large, so that $(.0073) \| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \|$ is itself large, it means that the Chebyshev rational approximation method $p = q = 2$ with one step could give rise to approximations $\mathbf{w}_p(t)$ of $\mathbf{v}(t)$ which have unacceptably large deviations from $\mathbf{v}(t)$. In this case, use of the Padé approximations is indicated, as $r_{p,q}^{(m)}(z)$ can be made as small as possible by choosing m sufficiently large. On the other hand, if $\| \overset{*}{\mathbf{g}} - B^{-1}\boldsymbol{\Sigma} \|$ is small, potentially large savings in digital computer time seem possible by using Chebyshev rational approximations of e^{-tB} for approximating the solution of (2.7).

6. Applications. In the numerical solution of parabolic partial differential equations in one space variable, the inversion of linear polynomials in the matrix B, corresponding to tridiagonal matrices, is carried out directly in practical applications of methods such as the backward implicit method, and the Crank-Nicolson method. But as quadratic polynomials in the matrix B are in this case only five-diagonal matrices, such direct inversions are still quite efficient. Hence, higher order methods, based either on the Padé or Chebyshev rational approximations of e^{-z}, can be used without great difficulty in one space dimension.

In two spatial dimensions, the inversion of linear polynomials in the matrix B is not generally carried out directly in large problems, but rather an iterative technique, such as the Young-Frankel successive overrelaxation method [23, 7], is used in effect to invert such polynomials. These iterative methods can also be applied to higher order methods. For example, the case $p = q = 2$, which requires the inversion of quadratic polynomials in the matrix B, can be rigorously treated by use of the iterative S2LOR [21], as well as the extension of the cyclic Chebyshev semi-iterative method to this case [9]. The methods which are contained in

this paper easily give rise to theoretical generalizations in several directions. For example, the treatment of time-varying forcing functions, as well as a certain class of hyperbolic partial differential equations, can also be examined from this point of view. We shall however leave such theoretical extensions and numerical results to a subsequent paper.

BIBLIOGRAPHY

1. N. I. Achieser, *Theory of Approximation*, translated by Charles J. Hyman, Frederick Ungar Publishing Co., New York, 1956.
2. R. E. Bellman, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.
3. G. Birkhoff and R. S. Varga, Reactor criticability and non-negative matrices, *J. Soc. Indust. Appl. Math.*, **6** (1958): 354–377.
4. R. Courant, K. Friedrichs and H. Levy, Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.* **100** (1928): 32–74.
5. J. Douglas, On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations, *J. Soc. Indust. Appl. Math.*, **4** (1956): 20–37.
6. E. C. DuFort and S. P. Frankel, Stability conditions in the numerical treatment of parabolic differential equations, *Math. Tables Aids to Comp.* **7** (1953): 135–152.
7. S. P. Frankel, Convergence rates of iterative treatments of partial differential equations, *Math. Tables Aids Comp.* **4** (1950): 65–75.
8. J. Franklin, Numerical stability in digital and analog computation for diffusion problems, *J. Math. and Phys.* **37** (1959): 305–315.
9. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods*, Numerische Mathematik **3** (1961): 147–156, 157–168.
10. A. S. Householder, The approximate solution of matrix problems, *J. Assoc. Comp. Mach.* **5** (1958): 155–169.
11. P. M. Hummel and C. L. Seebeck, A generalization of Taylor's Theorem, *Amer. Math. Monthly* **56** (1949): 243–247.
12. M. L. Juncosa and D. Young, On the convergence of a solution of a difference equation to a solution of the equation of diffusion, *Proc. Amer. Math. Soc.* **5** (1954): 168–174.
13. Z. Kopal, "Operational methods in numerical analysis based on rational approximations", *On Numerical Approximation*, University of Wisconsin Press, Madison, 1959: 25–43.
14. P. Laasonen, Über eine Methode zur Lösung der Wärmeleitungsgleichung, *Acta Math.* **81** (1949): 309–317.
15. M. Lotkin, The numerical integration of heat conduction equations, *J. Math. and Phys.* **37** (1958): 178–187.
16. Y. L. Luke, Rational approximations to the exponential function, *J. Assoc. Comp. Mach.* **4** (1957): 24–29.
17. G. G. O'Brien, M. A. Hyman, and S. Kaplan, A study of the numerical solution of partial differential equations, *J. Math. and Phys.* **29** (1951): 223–251.
18. H. Padé, *Sur la représentation approchée d'une fonction par des fractions rationnelles*, Thesis, Ann. de l'Éc. Nor., (3), **9** (1892).
19. R. D. Richtmyer, *Difference Methods for Initial-Value Problems*, Interscience Publishers, Inc., New York, 1957.
20. R. S. Varga, Numerical solution of the two-group diffusion equations in x-y geometry, *IRE Trans. on Nuclear Science* NS **4** (1957): 52–62.
21. R. S. Varga, "Factorization and normalized iterative methods", *Boundary Problems in Differential Equations*, University of Wisconsin, Madison, (1959): 121–142.

22. H. S. WALL, *Analytic Theory of Continued Fractions*, D. van Nostrand Company, Inc., Princeton, 1948.
23. D. YOUNG, Iterative methods for solving partial difference equations of elliptic type, *Trans. Amer. Math. Soc.* **76** (1954): 92–111.

CASE INSTITUTE OF TECHNOLOGY

(Received December 15, 1960)