

8. Matrix Methods for Parabolic Partial Differential Equations

8.1 Semi-Discrete Approximation

Many of the problems of physics and engineering that require numerical approximations are special cases of the following second-order linear parabolic differential equation:

$$(8.1) \quad \phi(\mathbf{x})u_t(\mathbf{x};t) = \sum_{i=1}^n (K_i(\mathbf{x})u_{x_i})_{x_i} + \sum_{i=1}^n G_i(\mathbf{x})u_{x_i} - \sigma(\mathbf{x})u(\mathbf{x};t) + S(\mathbf{x};t), \quad \mathbf{x} \in R, t > 0,$$

where R is a given finite (connected) region in Euclidean n -dimensional space, with (external) boundary conditions

$$(8.2) \quad \alpha(\mathbf{x})u(\mathbf{x};t) + \beta(\mathbf{x})\frac{\partial u(\mathbf{x};t)}{\partial n} = \gamma(\mathbf{x}), \quad \mathbf{x} \in \Gamma, t > 0,$$

where Γ is the external boundary of R . Characteristic of such problems is the additional *initial condition*

$$(8.3) \quad u(\mathbf{x};0) = g(\mathbf{x}), \quad \mathbf{x} \in R.$$

We assume for simplicity that the given functions $\phi, K_i, G_i, \sigma, S$, and g are continuous¹ in \bar{R} , the closure of R , and satisfy the following conditions:

$$(8.4) \quad \begin{array}{l} 1. \phi, K_i \text{ are positive in } \bar{R}, \quad 1 \leq i \leq n, \\ 2. \sigma \text{ is nonnegative in } \bar{R}. \end{array}$$

For the external boundary condition of (8.2), we assume that the functions α, β , and γ are piecewise continuous on Γ , and, as in Chap. 6, satisfy

$$(8.5) \quad \alpha(\mathbf{x}) \geq 0, \quad \beta(\mathbf{x}) \geq 0, \quad \alpha(\mathbf{x}) + \beta(\mathbf{x}) > 0, \quad \mathbf{x} \in \Gamma.$$

These assumptions cover important physics and engineering problems. For example, in reactor physics, the time-dependent density of neutrons $u(\mathbf{x};t)$ of

¹ It will be clear from the discussion that follows that problems with internal interfaces, where the functions $\phi, K_i, G_i, \sigma, S$, and g are only *piecewise* continuous in \bar{R} , can similarly be treated.

a particular average energy in a reactor satisfies, in a diffusion approximation, the equation

$$\frac{1}{v}u_t(\mathbf{x};t) = \sum_{i=1}^n (K(\mathbf{x})u_{x_i})_{x_i} - \sigma(\mathbf{x})u(\mathbf{x},t) + S(\mathbf{x};t),$$

which physically represents a conservation of neutrons. Here, v is the average velocity of these neutrons, $K(\mathbf{x})$ is the *diffusion coefficient*, and $\sigma(\mathbf{x})$ is the *total removal cross section*.² In petroleum engineering, the flow of a compressible fluid in a homogeneous porous medium is given by

$$\phi u_t(\mathbf{x};t) = \nabla^2 u(\mathbf{x};t) = \sum_{i=1}^n u_{x_i x_i}(\mathbf{x};t),$$

where ϕ depends on the density, compressibility, and viscosity of the fluid, as well as the permeability and porosity of the medium.³

More widely known, however, is the *heat or diffusion equation*. To be specific, the temperature $u(x,t)$, in an infinite thin rod, satisfies

$$(8.6) \quad u_t(x,t) = Ku_{xx}(x,t), \quad -\infty < x < +\infty, t > 0,$$

where the constant $K > 0$ is the *diffusivity*⁴ of the rod. For an initial condition, we are given the initial temperature distribution

$$(8.7) \quad u(x,0) = g(x), \quad -\infty < x < \infty,$$

where $g(x)$ is continuous and uniformly bounded on $-\infty < x < \infty$. It is well known that the solution of this problem is *unique* and is explicitly given by⁵

$$(8.8) \quad u(x,t) = \frac{1}{\sqrt{4\pi Kt}} \int_{-\infty}^{+\infty} \exp\left(\frac{-(x-x')^2}{4Kt}\right) g(x') dx', \quad -\infty < x < +\infty, t > 0.$$

We can also express this solution in the form

$$(8.9) \quad u(x,t + \Delta t) = \frac{1}{\sqrt{4\pi K\Delta t}} \int_{-\infty}^{+\infty} \exp\left(\frac{-(x-x')^2}{4K\Delta t}\right) u(x',t) dx',$$

where $t \geq 0, \Delta t > 0$. The point of this last representation is that $u(x,t + \Delta t)$ is linked (through positive weights) to $u(x',t)$ for *all* x' . Hence, we might

² See Glasstone and Edlund (1952), p. 291.

³ See Muskat (1937), p. 627. In actual petroleum problems, ϕ can be flow-dependent, which makes this partial differential equation nonlinear.

⁴ See, for example, Carslaw and Jaeger (1959), p. 9.

⁵ See Carslaw and Jaeger (1959), p. 35. For a discussion of existence and uniqueness for the more general problem of (8.1)-(8.3), see, for example, Bernstein (1950) and Hadamard (1952).

expect in this case that matrix equations arising from finite difference approximations of (8.6) would preserve some form of this particular coupling, i.e.,

$$u_m^{n+1} := u(m\Delta x, (n+1)\Delta t)$$

would be coupled through *nonnegative* coefficients to each u_j^n . Clearly, this suggests that the Perron-Frobenius theory of nonnegative matrices would again be useful in finite difference approximations of (8.1). Indeed, one of the main objectives of this chapter is to show that this property of nonnegative couplings, as noted for the infinite thin rod, is a general characteristic of discrete approximations to (8.1)-(8.3) and gives a further association of the Perron-Frobenius theory of nonnegative matrices to the numerical solution of partial differential equations.

To obtain finite difference approximations of the general initial value problem (8.1)-(8.3), we discretize first only the spatial variables, leaving the time variable *continuous*. In this **semi-discrete** form,⁶ matrix properties of the resulting system of ordinary differential equations are studied. Then, when the time variable is finally discretized, the concept of **stability** is introduced. In this way, stability, or lack thereof, is seen to be a property of matrix approximations only. Again, our primary concern in this chapter is with the analysis of matrix methods, rather than with questions of convergence of such finite difference approximations to the solution of (8.1)-(8.3).

Returning to the parabolic partial differential equation of (8.1) with boundary conditions given in (8.2), we now assume, for simplicity, that the number of spatial variables is $n = 2$, although it will be clear that the results to be obtained extend to the general case. Letting R_h denote a general two-dimensional (nonuniform) Cartesian mesh region which approximates \bar{R} , we derive spatial equations, as in Sect. 6.3, by integrating the differential equation (8.1) at each mesh (x_i, y_j) over its corresponding two-dimensional mesh rectangle $r_{i,j}$. For the integration of the left-side of (8.1), we make the approximation⁷

$$(8.10) \quad \int_{r_{i,j}} \int \phi(x, y) \frac{\partial u}{\partial t} dx dy \doteq \frac{du(x_i, y_i; t)}{dt} \int_{r_{i,j}} \int \phi(x, y) dx dy.$$

The five-point approximation to the right side of (8.1), by means of integration, is carried out as in Sect. 6.3, and we thus obtain the *ordinary* matrix differential equation

$$(8.11) \quad C \frac{d\mathbf{u}(t)}{dt} = -A\mathbf{u}(t) + \mathbf{s}(t) + \tilde{\tau}(t), \quad t > 0,$$

⁶ These approximations are also called *semi-explicit*. See Lees (1961).

⁷ Since ϕ is a known function in \bar{R} , we can *in principle* exactly evaluate the integral of ϕ over each rectangular mesh region $r_{i,j}$ of \bar{R}_i .

where C and A are $n \times n$ real matrices with time-independent entries,⁸ and

$$(8.12) \quad \mathbf{u}(0) = \mathbf{g}.$$

Note that the integration method of Sect. 6.3 directly utilizes the boundary conditions of (8.2), which are incorporated in the vector $\mathbf{s}(t)$ and the matrix A , and that the vector \mathbf{g} of (8.12) is some integral average on R_h of the function $g(x, y)$ defined on \bar{R} . The solution of (8.11)-(8.12) is called the **semi-discrete approximation** of the solution of (8.1)-(8.2), in that the spatial variables have been discretized, while t , the time variable, remains continuous.

For properties of the matrices C and A , we have from (8.10) that C is by construction a real diagonal matrix whose diagonal entries are integrals of the function ϕ . But from (8.4), C is then a positive diagonal matrix. For the real matrix A , we have, by virtue of the five-point approximation, that there are at most five nonzero entries in any row of A . Also, using the hypotheses of (8.4) and (8.5), we have, for *all sufficiently fine* Cartesian mesh regions R_h , that A is irreducibly diagonally dominant⁹ with positive diagonal entries and nonpositive off-diagonal entries. From Corollary 3.20 and Definition 3.22, it is thus an irreducible M -matrix. Multiplying on the left by C^{-1} in (8.11), we have

$$(8.13) \quad \frac{d\mathbf{u}(t)}{dt} = -C^{-1}A\mathbf{u}(t) + C^{-1}\mathbf{s}(t) + \boldsymbol{\tau}(t), \quad t > 0,$$

where $\boldsymbol{\tau}(t) := C^{-1}\bar{\boldsymbol{\tau}}(t)$. With the above properties for the matrices C and A , it follows that $C^{-1}A$ is also an *irreducible M -matrix* for sufficiently fine mesh regions R_h .

To give a particular estimate for $\boldsymbol{\tau}(t)$, consider the following special case of (8.1):

$$u_t(x, y; t) = u_{xx}(x, y; t) + u_{yy}(x, y; t), \quad (x, y) \in R, t > 0,$$

in a rectangle R . If the Cartesian mesh R_h which approximates \bar{R} is uniform, i.e., $\Delta x = \Delta y = h$ and $\beta := 0$ in (8.2), it can be verified, by means of Taylor's series expansions, that for all $0 \leq t \leq T$,

$$(8.14) \quad \tau_i(t) = O(h^2)$$

if the fourth derivatives u_{xxxx} and u_{yyyy} are continuous and bounded¹⁰ in \bar{R} for $0 \leq t \leq T$.

⁸ See Exer. 7 for cases where the matrices A and C have time-independent entries even when the functions α, β , and γ of (8.2) are time-dependent.

⁹ This requires the additional hypothesis that $\sigma(\mathbf{x}) := 0$ in \bar{R} and $\alpha(\mathbf{x}) := 0$ on Γ cannot simultaneously occur.

¹⁰ For example, see Douglas (1961b) for more general results concerning discretization errors.

To solve the ordinary matrix differential equation of (8.13), we make use of the exponential of a matrix:

$$\exp(M) := I + M + \frac{M^2}{2!} + \cdots,$$

which is convergent¹¹ for any $n \times n$ matrix M . With this definition, it follows that¹²

$$(8.15) \quad \mathbf{u}(t) = \exp(-tC^{-1}A)\mathbf{u}(0) + \exp(-tC^{-1}A) \cdot \int_0^t \exp(\lambda C^{-1}A)[C^{-1}\mathbf{s}(\lambda) + \boldsymbol{\tau}(\lambda)]d\lambda, \quad t \geq 0,$$

with $\mathbf{u}(0) := \mathbf{g}$, is the solution of (8.11)-(8.12).

For convenience of exposition of the remaining sections, we now assume that the source term $S(x, y; t)$ of (8.1) is *time-independent*. Because of this, the vector \mathbf{s} of (8.11) is also time-independent, and the solution of (8.11) and (8.12) takes the simpler form

$$(8.16) \quad \mathbf{u}(t) = A^{-1}\mathbf{s} + \exp(-tC^{-1}A)\{\mathbf{u}(0) - A^{-1}\mathbf{s}\} + \exp(-tC^{-1}A) \int_0^t \exp(\lambda C^{-1}A)\boldsymbol{\tau}(\lambda)d\lambda, \quad t \geq 0.$$

Neglecting the term involving the vector $\boldsymbol{\tau}(t)$, the previous equation gives rise to the vector approximation

$$(8.17) \quad \mathbf{v}(t) = A^{-1}\mathbf{s} + \exp(-tC^{-1}A)\{\mathbf{v}(0) - A^{-1}\mathbf{s}\},$$

where

$$(8.18) \quad \mathbf{v}(0) := \mathbf{g}.$$

Note that $\mathbf{v}(t)$ of (8.17)-(8.18) is just the solution of

$$(8.19) \quad C \frac{d\mathbf{v}(t)}{dt} = -A\mathbf{v}(t) + \mathbf{s}, \text{ for } t > 0, \text{ with } \mathbf{v}(0) = \mathbf{g}.$$

Using (8.16) and (8.17), it follows that

$$(8.20) \quad \mathbf{u}(t) - \mathbf{v}(t) = \int_0^t \exp[-(t-\lambda)C^{-1}A]\boldsymbol{\tau}(\lambda)d\lambda, \quad t \geq 0,$$

which can be used to estimate $\|\mathbf{u}(t) - \mathbf{v}(t)\|$. (See Exer. 6.)

By analogy with the physical problem, the expression of (8.17) suggests that the terms of the right side respectively correspond to the **steady-state solution** of (8.19) and a **transient term**. The matrix properties of $\exp(-tC^{-1}A)$, which correctly establish this, will be developed in the next section.

¹¹ See Exer. 1 of Sect. 3.5.

¹² By $\int_\alpha^\beta \mathbf{g}(x)dx$, we mean a vector with components $\int_\alpha^\beta g_i(x)dx$.

Exercises

1. If C and D are both $n \times n$ matrices, show that

$$\exp(tC + tD) = \exp(tC) \cdot \exp(tD), \quad \text{for all } t > 0,$$

if and only if $CD = DC$. As a consequence, if A is the matrix of (1.9) and B is the matrix of (1.10), conclude that

$$\exp(-tA) = \exp(-t) \cdot \exp(tB), \quad \text{for all } t.$$

2. Let A be any $n \times n$ complex matrix. Prove that

a. $\exp(A) \cdot \exp(-A) = I$. Thus, $\exp(A)$ is always nonsingular.

b. $\|\exp(A)\| \leq \exp(\|A\|)$.

3. If $[A, B] := AB - BA$, show for small values of t that

$$\begin{aligned} & \exp\{t(C + D)\} - \exp(tC) \cdot \exp(tD) \\ &= \frac{t^2}{2}[D, C] + \frac{t^3}{6}\{[CD, C] + [D^2, C] + [D, C^2] + [D, CD]\} + O(t^4). \end{aligned}$$

4. Verify that (8.15) is the solution of (8.11) and (8.12).

5. If the matrix $C^{-1}A$ is Hermitian and positive definite, using (8.17) show that

$$\|\mathbf{v}(t)\| \leq \|A^{-1}\mathbf{s}\| + \|\mathbf{v}(0) - A^{-1}\mathbf{s}\| \quad \text{for all } t \geq 0.$$

For the vector norms $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_\infty$ of Exer. 1, Sect. 1.3, show that $\|\mathbf{v}(t)\|_1$ and $\|\mathbf{v}(t)\|_\infty$ are also *uniformly bounded* for all $t \geq 0$.

6. If the matrix $C^{-1}A$ is Hermitian and positive definite, and if $\|\tau(t)\| \leq M(\Delta x)^2$ for all $0 \leq t \leq T$, show from (8.20) that

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq TM(\Delta x)^2, \quad 0 \leq t \leq T.$$

7. Suppose that the functions α, β and γ of the boundary condition (8.2) satisfy (8.5), but are *time-dependent*. Show that the matrices A and C , derived in the manner of this section, still have time-independent entries, if

a. whenever $\alpha(\mathbf{x}; t) > 0$, then $\frac{\beta(\mathbf{x}; t)}{\alpha(\mathbf{x}; t)}$ is time-independent;

b. whenever $\beta(\mathbf{x}; t) > 0$, then $\frac{\alpha(\mathbf{x}; t)}{\beta(\mathbf{x}; t)}$ is time-independent.

8. To show that choices other than diagonal matrices are possible for the matrix C of (8.11), consider the problem

$$\frac{\partial u(x, t)}{\partial t} = \frac{\partial^2 u(x, t)}{\partial x^2}, \quad 0 < x < 1, \quad t > 0,$$

where $u(0; t) = \alpha_1$, $u(1; t) = \alpha_2$ and $u(x; 0) = g(x)$, for $0 < x < 1$, with α_1 and α_2 scalars. Letting $x_i = ih$ where $h = 1/(N+1)$ for $0 \leq i \leq N+1$ and $u(x_i; t) := u_i(t)$, show that

$$\begin{aligned} \frac{d}{dt} \left[\frac{1}{12} \{10u_i(t) + u_{i-1}(t) \quad u_{i+1}(t)\} \right] \\ = \frac{u_{i-1}(t) - 2u_i(t) + u_{i+1}(t)}{h^2} + \tilde{\tau}_i(t), \end{aligned}$$

$1 \leq i \leq N$, where $\tilde{\tau}_i(t) = O(h^4)$ as $h \rightarrow 0$ for $0 \leq t \leq T$ if $\partial^6 u(x, t)/\partial x^6$ is continuous and bounded in $0 \leq x \leq 1, 0 \leq t \leq T$. In this case, the matrix C is tridiagonal.

9. Let $B := \begin{bmatrix} O & F \\ F^* & O \end{bmatrix}$. Show that

$$\exp(tB) = \left[\begin{array}{c|c} \cosh(t\sqrt{FF^*}) & \left(\frac{\sinh(t\sqrt{FF^*})}{t\sqrt{FF^*}} \right) tF \\ \hline \left(\frac{\sinh(t\sqrt{tF^*F})}{t\sqrt{F^*F}} \right) tF^* & \cosh(t\sqrt{F^*F}) \end{array} \right].$$

(Note that if $\cosh(\sqrt{z}) := \sum_{k=0}^{\infty} z^k/(2k)!$ and if $\sinh(\sqrt{z})/\sqrt{z} := \sum_{k=0}^{\infty} z^k/(2k+1)!$, then $\cosh(\sqrt{z})$ and $\sinh(\sqrt{z})/\sqrt{z}$ are *entire functions* of z , i.e., they are analytic functions of z in all of \mathbb{C} . Hence, from Exer. 1, Sect. 3.5, the matrices $\cosh \sqrt{A}$ and $\sinh(\sqrt{A})/\sqrt{A}$ are defined for *any* square complex matrix A .)

8.2 Essentially Positive Matrices

The exponential matrix $\exp(-tC^{-1}A)$, which resulted from the semi-discrete approximations of the previous section, has interesting matrix properties. We begin with

Definition 8.1. A real $n \times n$ matrix $Q = [q_{i,j}]$ is **essentially positive** if $q_{i,j} \geq 0$ for all $i \neq j$, $1 \leq i, j \leq n$, and Q is irreducible.

With this definition, it is clear that Q is an essentially positive matrix if and only if $(Q + sI)$ is a nonnegative, irreducible, and primitive matrix for all sufficiently large $s > 0$. Similarly, we have

Theorem 8.2. *A matrix Q is essentially positive if and only if $\exp(tQ) > O$ for all $t > 0$.*

Proof. If the matrix $\exp(tQ) = I + tQ + t^2Q^2/2! + \cdots$ has only positive entries for all $t > 0$, then Q is evidently irreducible. Otherwise, Q , and all its powers are reducible, which implies that $\exp(tQ)$ would have some zero entries. If there exists a $q_{i,j} < 0$ for some $i \neq j$, then clearly $\exp(tQ) = I + tQ + \cdots$ has a negative entry for all sufficiently small $t > 0$. Thus, $\exp(tQ) > O$ implies that Q is essentially positive. On the other hand, assume that Q is essentially positive. Then $Q + sI$ is nonnegative, irreducible, and primitive for all sufficiently large $s > 0$. Since the powers of $Q + sI$ are nonnegative and all sufficiently high powers of $Q + sI$ are positive by Frobenius' Theorem 2.18, then $\exp(Q + sI) > O$. But it is easily verified¹³ that

$$\exp(tQ) = \exp(-st)\exp(t(Q + sI)) > O,$$

which completes the proof. ■

As an immediate application of this result, note that as the matrix $C^{-1}A$ of (8.13) is, by construction, an irreducible M -matrix for all sufficiently fine mesh regions R_h , then $Q := -C^{-1}A$ is necessarily essentially positive. Thus, the semi-discrete approximation of (8.17), which can be expressed as

$$(8.21) \quad \mathbf{v}(t) = \exp(-tC^{-1}A)\mathbf{v}(0) + \{I - \exp(-tC^{-1}A)\}A^{-1}\mathbf{s},$$

couples (through positive coefficients), for sufficiently fine mesh regions R_h , each component of $\mathbf{v}(t)$ to every component of $\mathbf{v}(0)$ for all $t > 0$, and is the matrix analogue of the behavior noted for the infinite thin rod.

Closely related to the Perron-Frobenius Theorem 2.7 is

Theorem 8.3. *Let Q be an essentially positive $n \times n$ matrix. Then, Q has a real eigenvalue $\zeta(Q)$ such that*

1. *To $\zeta(Q)$ there corresponds an eigenvector $\mathbf{x} > \mathbf{0}$.*
2. *If α is any other eigenvalue of Q , then $\operatorname{Re} \alpha < \zeta(Q)$.*
3. *$\zeta(Q)$ increases when any element of Q increases.*
4. *$\zeta(Q)$ is a simple eigenvalue of Q .*

¹³ See Exer. 1 of Sect. 8.1.

Proof. If Q is essentially positive, then the matrix $T := Q + sI$ is nonnegative, irreducible, and primitive for some real $s > 0$. Thus, from the Perron-Frobenius Theorem 2.7, there exists a vector $\mathbf{x} > \mathbf{0}$ such that $T\mathbf{x} = \rho(T)\mathbf{x}$. But this implies that

$$Q\mathbf{x} = (\rho(T) - s)\mathbf{x} =: \zeta(Q)\mathbf{x}.$$

The other conclusions of this theorem follow similarly from the Perron-Frobenius Theorem 2.7. ■

Physically, using the infinite thin rod of Sect. 8.1 as an example, we know that the temperature of the rod is bounded for all $t \geq 0$ and that there is a steady-state temperature of the rod, i.e., $\lim_{t \rightarrow \infty} u(x; t)$ exists for all $-\infty < x < +\infty$. Analogously, we ask if the semi-discrete approximations of Sect. 8.1 possess these basic properties. In order to answer this, we first determine the general asymptotic behavior of $\exp(tQ)$ for essentially positive matrices Q , in terms of norms. With the explicit matrices of (3.37), the following lemma is readily established, along the lines of the proof of Lemma 3.2.

Lemma 8.4. *Let J be the upper bi-diagonal $p \times p$ complex matrix of the form*

$$(8.22) \quad J = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda \end{bmatrix}.$$

Then,

$$(8.23) \quad \|\exp(tJ)\| \sim \frac{t^{p-1}}{(p-1)!} \exp(t \operatorname{Re} \lambda), \quad t \rightarrow +\infty.$$

Let Q be an $n \times n$ essentially positive matrix. If S is the nonsingular $n \times n$ matrix such that $Q = SJS^{-1}$, where J is the Jordan normal form of Q , then we apply Lemma 8.4 to the diagonal blocks of the matrix J . Since Q is essentially positive, the unique eigenvalue of Q with the largest real part, $\zeta(Q)$, is simple from Theorem 8.3, which in turn implies that the corresponding diagonal submatrix of J is 1×1 . Then, as

$$\exp(tQ) = S \exp(tJ) S^{-1}$$

holds, and as (8.23) applies to each block diagonal of $\exp(tJ)$, we have the result of

Theorem 8.5. *Let Q be an $n \times n$ essentially positive matrix. If $\zeta(Q)$ is the eigenvalue of Theorem 8.3, then*

$$(8.24) \quad \|\exp(tQ)\| \sim K \exp(t\zeta(Q)), \quad t \rightarrow +\infty,$$

where K is a positive constant, independent of t .

As $\zeta(Q)$ in (8.24) then dictates the asymptotic behavior of $\|\exp(tQ)\|$ for large t when Q is essentially positive, we accordingly make¹⁴

Definition 8.6. If Q is an essentially positive $n \times n$ matrix, then Q is **super-critical**, **critical**, or **subcritical** if its real eigenvalue $\zeta(Q)$, from Theorem 8.3, is respectively positive, zero, or negative.

Consider now the nonhomogeneous ordinary matrix differential equation

$$(8.25) \quad \frac{d\mathbf{v}(t)}{dt} = Q\mathbf{v}(t) + \mathbf{r},$$

where $\mathbf{v}(0)$ in \mathbb{R}^n is the given initial vector condition, and \mathbf{r} in \mathbb{R}^n is a given time-independent vector. If Q is nonsingular, then the unique solution of (8.25), satisfying the initial vector condition, is

$$(8.26) \quad \mathbf{v}(t) = -Q^{-1}\mathbf{r} + \exp(tQ) \cdot \{\mathbf{v}(0) + Q^{-1}\mathbf{r}\}, \quad t \geq 0.$$

Theorem 8.7. Let the $n \times n$ matrix Q of (8.25) be essentially positive and nonsingular. If Q is supercritical, then for certain initial vectors $\mathbf{v}(0)$, the solution $\mathbf{v}(t)$ of (8.25) satisfies

$$(8.27) \quad \lim_{t \rightarrow +\infty} \|\mathbf{v}(t)\| = +\infty.$$

If Q is subcritical, then the solution vector $\mathbf{v}(t)$ of (8.25) is uniformly bounded in norm for all $t \geq 0$, and satisfies

$$(8.28) \quad \lim_{t \rightarrow +\infty} \mathbf{v}(t) = -Q^{-1}\mathbf{r}.$$

Proof. Since Q is by hypothesis nonsingular, then Q has no zero eigenvalues. Thus, Q is either supercritical or subcritical. The results then follow from Theorem 8.5 and (8.26). Note that the vector $-Q^{-1}\mathbf{r}$, as Q is nonsingular, is just the solution of (8.25) with the derivative term set to zero. ■

With the previous theorem, we arrive at the final result in this section, which connects the theory in this section to the ordinary matrix differential equation of (8.19).

Corollary 8.8. If C is an $n \times n$ positive diagonal matrix, and A is an irreducible $n \times n$ M -matrix, then the unique vector solution of

¹⁴ These terms are the outgrowth of investigations concerning mathematical models for nuclear reactor theory. See, for example, Birkhoff and Varga (1958).

$$(8.29) \quad C \frac{d\mathbf{v}(t)}{dt} = -A\mathbf{v} + \mathbf{s}, \quad t \geq 0,$$

subject to the initial vector condition $\mathbf{v}(0)$, is uniformly bounded in norm for all $t \geq 0$, and satisfies

$$(8.30) \quad \lim_{t \rightarrow +\infty} \mathbf{v}(t) = A^{-1}\mathbf{s}.$$

Proof. Let $Q := -C^{-1}A$. Then Q is a nonsingular essentially positive matrix. Using Theorem 8.3, let $Q\mathbf{x} = \zeta(Q)\mathbf{x}$ where $\mathbf{x} > \mathbf{0}$. It follows that

$$A^{-1}C\mathbf{x} = \left(\frac{-1}{\zeta(Q)} \right) \mathbf{x}.$$

But by the hypotheses, $A^{-1}C$ is a positive matrix, so that $\zeta(Q)$ is necessarily a negative real number, which proves that Q is subcritical. The remainder then follows from Theorem 8.7. ■

We have thus shown that for sufficiently fine mesh regions R_h , the semi-discrete approximations (8.29), as introduced in Sect. 8.1, possess solutions which are *uniformly bounded* in norm for all $t \geq 0$, and possess finite steady-state solutions, as suggested by the example of the infinite thin rod. Furthermore, the semi-discrete approximations for the partial differential equation of (8.1) possess the positive couplings noted for this physical example.

Exercises

1. A real $n \times n$ matrix $Q = [q_{i,j}]$ is *essentially nonnegative* if $q_{i,j} \geq 0$ for all $i \neq j$, $1 \leq i, j \leq n$. Prove that Q is essentially nonnegative if and only if $\exp(tQ) \geq O$ for all $t \geq 0$.
2. Prove that Q is essentially positive if and only if $-Q + sI$ is an irreducible M -matrix for some real s .
3. Complete the proof of Theorem 8.3.
4. Prove Lemma 8.4.
5. Let Q be an arbitrary $n \times n$ complex matrix with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, where $\operatorname{Re} \lambda_1 \leq \operatorname{Re} \lambda_2 \leq \dots \leq \operatorname{Re} \lambda_n$. Extending Lemma 8.4, find a necessary and sufficient condition that $\|\exp(tQ)\|$ be bounded for all $t \geq 0$.

6. Let Q be an essentially positive $n \times n$ matrix, and let \mathbf{y} be the positive eigenvector of $\exp(tQ^T)$, where $\|\mathbf{y}\| = 1$. If $\mathbf{v}(t)$ is the solution of

$$\frac{d\mathbf{v}(t)}{dt} = Q\mathbf{v}(t),$$

where $\mathbf{v}(0) = \mathbf{g}$ is the initial vector condition, show that

$$\mathbf{y}^T \mathbf{v}(t) = \exp(t\zeta(Q)) \mathbf{y}^T \mathbf{g}$$

for all $t \geq 0$.

- *7. Let Q be an essentially positive $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, where $\operatorname{Re} \lambda_1 \leq \operatorname{Re} \lambda_2 \leq \dots < \operatorname{Re} \lambda_n = \zeta(Q)$, and let the initial vector \mathbf{g} of Exer. 6 have positive components. If \mathbf{y} is the positive eigenvector of $\exp(tQ^T)$, show that the solution of

$$\frac{d\mathbf{v}(t)}{dt} = Q\mathbf{v}(t)$$

can be expressed as

$$\mathbf{v}(t) = \tilde{K} \exp(t\zeta(Q)) \mathbf{x} + O(\exp(t\mu)), \quad t \rightarrow +\infty,$$

where \mathbf{x} is the positive vector of Theorem 8.3, and where

$$\zeta(Q) > \mu > \max_{1 \leq j < n} \operatorname{Re} \lambda_j, \text{ and } \tilde{K} := \mathbf{y}^T \mathbf{g} / \mathbf{y}^T \mathbf{x} > 0$$

(Birkhoff and Varga (1958)).

8. For Theorem 8.5, show that the result of (8.24) is valid for the matrix norms $\|A\|_1$ and $\|A\|_\infty$, defined in Exer. 2 of Sect. 1.3. Characterize the constant K in each case, as well as the case of (8.24).
9. Let Q be any $n \times n$ complex matrix such that $\|\exp(tQ)\|$ is uniformly bounded for all $t \geq 0$. Show that there exists a positive scalar K such that

$$\|(Q - zI)^{-1}\| \leq \frac{K}{\operatorname{Re} z}$$

for any complex z with $\operatorname{Re} z > 0$ (Kreiss (1959a)).

10. Let Q be an essentially positive $n \times n$ matrix. If $\zeta(Q)$ is the eigenvalue of Q characterized in Theorem 8.3, show that $\zeta(Q)$ has a min-max representation analogous to that of Theorem 2.9:

$$\max_{\mathbf{x} \in P^*} \left\{ \min_{1 \leq i \leq n} \frac{\sum_{j=1}^n a_{i,j} x_j}{x_i} \right\} = \zeta(Q) = \min_{\mathbf{x} \in P^*} \left\{ \max_{1 \leq i \leq n} \frac{\sum_{j=1}^n a_{i,j} x_j}{x_i} \right\},$$

where P^* is the hyperoctant of vectors $\mathbf{x} > \mathbf{0}$. (See (Beckenbach and Bellman (1961), p. 84).)

8.3 Matrix Approximations for $\exp(-tS)$

The exponentials of matrices, introduced in this chapter to solve the semi-discrete approximations of Sect. 8.1, had the attractive features of uniform boundedness in norm, as well as the positive couplings from one time step to the next. Unfortunately, the direct determination of these exponentials of matrices, relative to even large computers, seems impractical in two- or three-dimensional problems¹⁵ because of the enormous storage problem created by the positivity of these matrices. To illustrate this, suppose that a two-dimensional problem (8.1) is approximated on a mesh consisting of 50 subdivisions in each coordinate direction. To completely specify the matrix A of (8.11) would require (using symmetry) only 7500 nonzero coefficients. To specify the matrix $\exp(-tC^{-1}A)$, on the other hand, would require approximately $3 \cdot 10^6$ nonzero coefficients.

We now consider matrix approximations of $\exp(-tS)$, where $S := C^{-1}A$, from the construction of Sect. 8.1, is necessarily a sparse matrix. First, we shall consider some well-known numerical methods for solving parabolic partial differential equations. As we shall see, we can consider these methods as either matrix approximations for the matrix $\exp(-\Delta tS)$ in

$$(8.31) \quad \mathbf{v}(t_0 + \Delta t) = \exp(-\Delta tS)\mathbf{v}(t_0) + \{I - \exp(-\Delta tS)\}A^{-1}\mathbf{s},$$

or discrete approximations in time of the ordinary matrix differential equation

$$(8.32) \quad C \frac{d\mathbf{v}(t)}{dt} = -A\mathbf{v}(t) + \mathbf{s}.$$

¹⁵ For one-space variable problems, it may not be so impractical.

The Forward Difference Method.

The matrix approximation

$$(8.33) \quad \exp(-\Delta t S) \doteq I - \Delta t S, \quad S := C^{-1}A,$$

obtained by taking the first two terms of the expansion for $\exp(-\Delta t S)$, when substituted in (8.31), results in

$$(8.34) \quad \mathbf{w}(t_0 + \Delta t) = (I - \Delta t S)\mathbf{w}(t_0) + \Delta t C^{-1}\mathbf{s}.$$

By rearranging, this can be written as

$$(8.35) \quad C \left\{ \frac{\mathbf{w}(t_0 + \Delta t) - \mathbf{w}(t_0)}{\Delta t} \right\} = -A\mathbf{w}(t_0) + \mathbf{s}.$$

This last equation then appears as a discrete approximation of (8.32). Starting with the given initial vector $\mathbf{w}(0)$, one can explicitly step ahead, finding successively $\mathbf{w}(\Delta t)$, $\mathbf{w}(2\Delta t)$, etc. This well-known method is appropriately called the **forward difference** or **explicit method**. Note that the time increments need *not* be constant. From (8.35), we observe that only the positive diagonal matrix C must be inverted to carry out this method.

The Backward Difference Implicit Method.

Consider now the matrix approximation

$$(8.36) \quad \exp(-\Delta t S) \doteq (I + \Delta t S)^{-1}, \quad \Delta t \geq 0.$$

Since $S = C^{-1}A$ is an irreducible M -matrix, all eigenvalues of S have their real parts positive.¹⁶ Thus, $I + \Delta t S$ is nonsingular for all $\Delta t \geq 0$. For Δt sufficiently small, we can write

$$(I + \Delta t S)^{-1} = I - \Delta t S + (\Delta t S)^2 - \dots,$$

which shows that the matrix approximation of (8.36) also agrees through linear terms with $\exp(-\Delta t S)$. Substituting in (8.31) gives

$$(8.37) \quad (I + \Delta t S)\mathbf{w}(t_0 + \Delta t) = \mathbf{w}(t_0) + \mathbf{s},$$

which can be written equivalently as

$$(8.38) \quad C \left\{ \frac{\mathbf{w}(t_0 + \Delta t) - \mathbf{w}(t_0)}{\Delta t} \right\} = -A\mathbf{w}(t_0 + \Delta t) + \mathbf{s}.$$

The name **backward difference method** for this procedure stems from the observation that one can explicitly use (8.37) to step *backward* in

¹⁶ See Exer. 4 of Sect. 3.5.

time to calculate $\mathbf{w}(t_0)$ from $\mathbf{w}(t_0 + \Delta t)$. Note that this would involve only matrix multiplications and vector additions. Used, however, as a numerical procedure to calculate $\mathbf{w}(t_0 + \Delta t)$ from $\mathbf{w}(t_0)$, (8.37) shows this method to be implicit, since it requires the solution of a matrix problem.

The Central Difference Implicit Method.

With S an irreducible M -matrix, we can also form the approximation

$$(8.39) \quad \exp(-\Delta t S) \doteq \left(I + \frac{\Delta t}{2} S\right)^{-1} \left(I - \frac{\Delta t}{2} S\right), \quad \Delta t \geq 0,$$

which for small Δt gives an expansion

$$\left(I + \frac{\Delta t}{2} S\right)^{-1} \left(I - \frac{\Delta t}{2} S\right) = I - \Delta t S + \frac{(\Delta t S)^2}{2} - \frac{(\Delta t S)^3}{4} + \dots$$

This now agrees through *quadratic* terms with the expansion for $\exp(-\Delta t S)$. Substituting in (8.31), we have

$$(8.40) \quad \left(I + \frac{\Delta t}{2} S\right) \mathbf{w}(t_0 + \Delta t) = \left(I - \frac{\Delta t}{2} S\right) \mathbf{w}(t_0) + \Delta t C^{-1} \mathbf{s},$$

which can be written as

$$(8.41) \quad C \left\{ \frac{\mathbf{w}(t_0 + \Delta t) - \mathbf{w}(t_0)}{\Delta t} \right\} = -\frac{A}{2} \{\mathbf{w}(t_0 + \Delta t) + \mathbf{w}(t_0)\} + \mathbf{s}.$$

It is evident that the above approximation in (8.41), unlike (8.38) and (8.35), is a *central difference approximation* to (8.32) for $t = t_0 + \Delta t/2$. From (8.40), we see that this method for generating $\mathbf{w}(t_0 + \Delta t)$ from $\mathbf{w}(t_0)$ is also implicit. Crank and Nicolson (1947) first considered this approximation in the numerical solution of the one-dimensional heat equation (8.1), and this is now known in the literature as the **Crank-Nicolson method**. In this case, the matrix $S = C^{-1}A$ is a nonsingular real tridiagonal matrix, and the matrix equation (8.40) can be efficiently solved by Gaussian elimination, as described in Sect. 6.4. Note that the same is true of the numerical solution (8.37) for the backward difference method.

The three methods just described are among the best known and most widely used numerical methods for approximating the solution of the parabolic partial differential equation of (8.1).

If the matrix $S = C^{-1}A$ is an irreducible M -matrix, then $-S$ is a subcritical essentially positive matrix. From Theorem 8.5, the matrix norms $\|\exp(-\Delta t S)\|$ are *uniformly bounded* for all $\Delta t \geq 0$, and this implies (Corollary 8.8) that the vectors $\mathbf{v}(t)$ of (8.29) are uniformly bounded in norm for all $t \geq 0$. The different approximations for $\exp(-\Delta t S)$ that we have just considered do *not* all share this boundedness property. For example, the forward difference approximation $(I - \Delta t S)$ has its spectral norm bounded below by

$$\|(I - \Delta t S)\| \geq \rho(I - \Delta t S) = \max_{1 \leq i \leq n} |1 - \Delta t \lambda_i|,$$

where the λ_i 's, the eigenvalues of S , satisfy $0 < \alpha \leq \operatorname{Re} \lambda_i \leq \beta$. It is evident that $\rho(I - \Delta t S)$ is *not* uniformly bounded for all $\Delta t \geq 0$.

We now concentrate on those approximations for $\exp(-\Delta t S)$ which have their spectral radii less than unity for certain choices of $\Delta t \geq 0$. The reason that these approximations are of interest lies simply in the following: Let $T(\Delta t)$ be any matrix approximation of $\exp(-\Delta t S)$. Then, we would approximate the vector $\mathbf{v}(t_0 + \Delta t)$ of (8.31) by

$$(8.42) \quad \mathbf{w}(t_0 + \Delta t) = T(\Delta t) \cdot \{\mathbf{w}(t_0) - A^{-1}\mathbf{s}\} + A^{-1}\mathbf{s}, \quad \Delta t > 0.$$

Starting with $t_0 = 0$, we arrive by induction at

$$(8.43) \quad \mathbf{w}(m\Delta t) = [T(\Delta t)]^m \{\mathbf{w}(0) - A^{-1}\mathbf{s}\} + A^{-1}\mathbf{s}, \quad m \geq 0.$$

If $\rho(T(\Delta t)) \geq 1$, we see that the sequence of vectors $\mathbf{w}(m\Delta t)$ will *not* in general be bounded in norm for all choices of $\mathbf{w}(0)$. Actually, what we are considering here is the topic of **stability**. It has been recognized for some time that, although in practice one wishes to take longer time steps (Δt large) in order to arrive at an approximation $\mathbf{w}(t)$ for the solution $u(\mathbf{x}; t)$ of (8.1) with as *few* arithmetic computations as possible, one must restrict the size of $\Delta t = t/m$ in order to insure that $\mathbf{w}(m\Delta t)$ is a reasonable approximation to $u(\mathbf{x}; t)$. This brings us to

Definition 8.9. The matrix $T(t)$ is **stable** for $0 \leq t \leq t_0$ if $\rho(T(t)) \leq 1$ for this interval. It is **unconditionally stable** if $\rho(T(t)) < 1$ for all $t > 0$.

Note that the definition of stability is *independent* of the closeness of approximation of $T(t)$ to $\exp(-tS)$.

For stability intervals for the approximations considered, we have

Theorem 8.10. Let S be an $n \times n$ matrix whose eigenvalues λ_i satisfy $0 < \alpha \leq \operatorname{Re} \lambda_i \leq \beta$, $1 \leq i \leq n$. Then, the forward difference matrix approximation $I - \Delta t S$ is stable for

$$(8.44) \quad 0 \leq \Delta t \leq \min_{1 \leq i \leq n} \left\{ \frac{2 \operatorname{Re} \lambda_i}{|\lambda_i|^2} \right\}.$$

On the other hand, the matrix approximations

$$(I + \Delta t S)^{-1} \text{ and } \left(I + \frac{\Delta t}{2} S \right)^{-1} \cdot \left(I - \frac{\Delta t}{2} S \right)$$

for the backward difference and Crank-Nicolson matrix approximations are *unconditionally stable*.

Proof. The result of (8.44) follows by direct computation. For the backward difference matrix approximation $(I + \Delta t S)^{-1}$, the eigenvalues of this matrix are $(1 + \Delta t \lambda_i)^{-1}$. Since the real part of $(1 + \Delta t \lambda_i)$ is greater than unity for *all* $\Delta t \geq 0$, then this matrix approximation is unconditionally stable. The proof for the Crank-Nicolson matrix approximation follows similarly. (See Exer. 7.) ■

One might ask how the approximations of $\exp(-\Delta t S)$, which we have discussed, as well as other approximations, are generated. This is most simply described in terms of **Padé rational approximations**¹⁷ from classical complex analysis. To begin, let $f(z)$ be any analytic function in the neighborhood of the origin:

$$(8.45) \quad f(z) = a_0 + a_1 z + a_2 z^2 + \cdots.$$

We consider approximations to $f(z)$ defined by

$$f(z) \doteq \frac{n_{p,q}(z)}{d_{p,q}(z)},$$

with $n_{p,q}(z)$ and $d_{p,q}(z)$ respectively polynomials of degree q and p in z , and we assume that $d_{p,q}(0) \neq 0$. We now select, for *each* pair of nonnegative integers p and q , those polynomials $n_{p,q}(z)$ and $d_{p,q}(z)$ such that the Taylor's series expansion of $n_{p,q}(z)/d_{p,q}(z)$ about the origin agrees with *as many* leading terms of $f(z)$ of (8.45), as is possible. Since the ratio $n_{p,q}(z)/d_{p,q}(z)$ contains $p + q + 1$ *essential* unknown coefficients, it is evident that the expression

$$(8.46) \quad d_{p,q}(z)f(z) - n_{p,q}(z) = O(|z|^{p+q+1}), \quad |z| \rightarrow 0,$$

gives rise to $p + q + 1$ linear equations in these essential unknowns, whose solution determines these unknown coefficients. In this way, one can generate the double-entry *Padé table* for $f(z)$. It can be verified¹⁸ that the first few entries of the Padé table for $f(z) = \exp(-z)$ are given by:

¹⁷ Due to Padé (1892). See, for example, Wall (1948), Chap. XX.

¹⁸ See Exer. 3.

	$q = 0$	$q = 1$	$q = 2$
$p = 0$	1	$1 - z$	$1 - z + \frac{z^2}{2}$
$\exp(-z) : p = 1$	$\frac{1}{1 + z}$	$\frac{2 - z}{2 + z}$	$\frac{6 - 4z + z^2}{6 + 2z}$
$p = 2$	$\frac{1}{1 + z + \frac{z^2}{2}}$	$\frac{6 - 2z}{6 + 4z + z^2}$	$\frac{12 - 6z + z^2}{12 + 6z + z^2}$

These rational approximations for $\exp(-z)$ generate matrix approximations of $\exp(-\Delta t S)$, in an obvious way. *Formally*, we merely replace the variable z by the matrix $\Delta t S$, and we define

$$(8.47) \quad E_{p,q}(\Delta t S) := [d_{p,q}(\Delta t S)]^{-1} [n_{p,q}(\Delta t S)]$$

to be the (p, q) -Padé matrix approximation of $\exp(-\Delta t S)$. From the entries of the Padé table for $\exp(-z)$, we see that the matrix approximations for $\exp(-\Delta t S)$ corresponding to the forward difference, backward difference, and Crank-Nicolson methods are *exactly* the Padé matrix approximations $E_{0,1}(\Delta t S)$, $E_{1,0}(\Delta t S)$, and $E_{1,1}(\Delta t S)$, respectively.

Our purpose in introducing these Padé approximations is threefold. First, we see that the matrix approximations associated with the forward difference, backward difference, and Crank-Nicolson methods are special cases of Padé matrix approximations of $\exp(-\Delta t S)$. Second, these Padé approximations generate many other useful matrix approximations of $\exp(-\Delta t S)$. Although it is true that such higher-order Padé approximations have been less used in practical applications, it is entirely possible that they may be of use on larger computers where higher-order implicit methods could be feasible. Third, the direct association of these matrix approximations with the classical analysis topic of rational approximation of analytic functions gives a powerful tool for the analysis of stability of these approximations. For example, it can be shown¹⁹ that if the eigenvalues of S are positive real numbers, then the Padé matrix approximation $E_{p,q}(\Delta t S)$ in (8.47) is *unconditionally stable* if and only if $p \geq q$.

Another important concept, due to Lax and Richtmyer (1956), on approximations of the matrix $\exp(-\Delta t S)$ is given by

¹⁹ See Varga (1961).

Definition 8.11. The matrix $T(t)$ is a **consistent approximation** of $\exp(-tS)$ if $T(t)$ has a matrix power series development about $t = 0$ that agrees through at least linear terms with the expansion of $\exp(-tS)$.

Note that all Padé matrix approximations for $p + q > 0$ are by definition consistent approximations of $\exp(-tS)$. (See also Exer. 4.)

Finally, we end this section with a result further linking the Perron-Frobenius theory of nonnegative matrices to the numerical solution of the parabolic problem (8.1).

Theorem 8.12. *Let $T(t)$ be any consistent matrix approximation of $\exp(-tS)$ where S is an irreducible M -matrix. If $T(t) \geq O$ for $0 \leq t \leq t_0$, where $t_0 > 0$, then there exists a $t_1 > 0$ such that $T(t)$ is primitive (i.e., $T(t)$ is nonnegative, irreducible, and noncyclic) for $0 < t < t_1$.*

Proof. Since $T(t)$ is a consistent approximation of $\exp(-tS)$, then

$$T(t) = I - tS + O(t^2), \quad \text{as } t \rightarrow 0.$$

Since, by hypothesis, $T(t) \geq O$ for positive t sufficiently small, then as S is an irreducible M -matrix, $T(t)$ is nonnegative, irreducible, with positive diagonal entries for positive t sufficiently small. But from Lemma 2.17, $T(t)$ is evidently primitive for some interval in t , completing the proof. ■

The physical, as well as mathematical, significance of this result is that, from (8.43), we have

$$(8.48) \quad \mathbf{w}(m\Delta t) = [T(\Delta t)]^m \{\mathbf{w}(0) - A^{-1}\mathbf{s}\} + A^{-1}\mathbf{s}, \quad m \geq 0.$$

Thus, for all $\Delta t > 0$ sufficiently small and m sufficiently large, each component $w_i(m\Delta t)$ is coupled through positive entries to *every* component $w_j(0)$, $1 \leq j \leq n$, so that the *primitivity* of the matrix approximation $T(\Delta t)$ of $\exp(-\Delta tS)$ is the *natural* matrix analogue of the positive coupling noted for the heat equation.

Exercises

1. Consider the heat equation of (8.6) for the finite interval $0 < x < 1$, with boundary conditions $u(0, t) := \mu_1$, $u(1, t) := \mu_2$, where μ_1 and μ_2 are positive constants. Using a uniform mesh $\Delta x = 1/(n+1)$ and setting $x_i = i\Delta x$, $0 \leq i \leq n+1$, show first that the $n \times n$ matrix $S = C^{-1}A$ is given explicitly by

$$S = \frac{K}{(\Delta x)^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Show next that the forward difference method approximation $(I - \Delta t S)$ of $\exp(-\Delta t S)$ is stable for

$$0 \leq \Delta t \leq \frac{(\Delta x)^2}{2K}.$$

(*Hint:* Use Corollary 1.12 to estimate $\rho(I - \Delta t S)$.) (See Courant, Friedrichs, and Lewy (1928).)

2. Consider the heat equation

$$\frac{\partial u}{\partial t} = K \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}$$

for the unit n -dimensional hypercube, $0 < x_i < 1$, $1 \leq i \leq n$, where $u(\mathbf{x}, t)$ is specified for $\mathbf{x} \in \Gamma$, the boundary of this hypercube. Using a uniform mesh $\Delta x_i = 1/(N+1)$ in each coordinate direction, show that the forward difference method approximation $I - \Delta t S$ of $\exp(-\Delta t S)$ is stable for

$$0 \leq \Delta t \leq \frac{(\Delta x)^2}{2nK}.$$

(*Hint:* Generalize the result of the previous exercise.)

- *3. Show that the (p, q) entry of the Padé table for $\exp(-z)$ is determined by

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

and

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

(Padé (1892)) and (Hummel and Seebeck (1949)).

4. Show that $\exp(-z) - n_{p,q}(z)/d_{p,q}(z) \sim c_{p,q} z^{p+q+1}$ as $|z| \rightarrow 0$, and determine $c_{p,q}$.

- * 5. Show that $\exp(-z) - n_{p,q}(z)/d_{p,q}(z)$ is of one sign for all $z \geq 0$ (Hummel and Seebeck (1949)).
- * 6. It is known (Wall (1948), p. 348) that $\exp(-z)$ has the continued fraction expansion

$$\exp(-z) = \frac{1}{1 + z \cfrac{1}{1 - z \cfrac{2}{2 + z \cfrac{3}{3 - z \cfrac{2}{2 + z \cfrac{5}{5 - z \cfrac{2}{2 + \dots}}}}}}}}$$

Prove that the successive approximants from this continued fraction, i.e.,

$$1, \frac{1}{1+z}, \frac{1}{1+z \cfrac{1}{1-z/2}}, \dots,$$

are particular entries of the Padé table for $\exp(-z)$.

7. Let S be any $n \times n$ complex matrix with eigenvalues λ_i satisfying $\operatorname{Re} \lambda_i > 0$. Show that the matrix $(I + (\Delta t/2)S)^{-1}(I - (\Delta t/2)S)$ is an unconditionally stable and consistent approximation of $\exp(-\Delta t S)$, for $\Delta t > 0$.
8. Let S be an irreducible M -matrix. Prove that the Padé matrix approximations $E_{0,1}(\Delta t S)$, $E_{1,0}(\Delta t S)$, and $E_{1,1}(\Delta t S)$ of $\exp(-\Delta t S)$ of (8.47) are *nonnegative* matrices for certain intervals in $\Delta t \geq 0$. What are these intervals? In particular, show, for the matrix S of Exer. 1, that the interval for which $I - \Delta t S$ is nonnegative is *exactly* the same as the stability interval for $I - \Delta t S$.
9. Let $T(\Delta t)$ be a *strictly* stable approximation (i.e., $\rho(T(\Delta t)) < 1$) of $\exp(-\Delta t S)$. If $\mathbf{w}(m\Delta t)$ is given by (8.48), show that there exists a positive constant M such that

$$\|\mathbf{w}(m\Delta t)\| \leq \|A^{-1}\mathbf{s}\| + M\|\mathbf{w}(0) - A^{-1}\mathbf{s}\|,$$

for all $m \geq 0$. Show that this inequality is also true for the vector norms $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_\infty$ of Exer. 1, Sect. 1.3. If S is the matrix of Exer. 1 of this section, show that the constant M can be chosen to be unity for $0 \leq \Delta t \leq (\Delta x)^2/2K$, for all these vector norms.

8.4 Relationship with Iterative Methods for Solving Elliptic Difference Equations

The form of equation (8.42), i.e.,

$$(8.49) \quad \mathbf{w}((n+1)\Delta t) = T(\Delta t)\mathbf{w}(n\Delta t) + \{I - T(\Delta t)\}A^{-1}\mathbf{s}, \quad \Delta t > 0,$$

where $\mathbf{w}(0) = \mathbf{v}(0) = \mathbf{g}$, suggests that matrix approximations of $\exp(-\Delta t S)$ induce abstract iterative methods for solving the matrix problem

$$(8.50) \quad A\mathbf{w} = \mathbf{s}.$$

In fact, if $\rho(T(\Delta t)) < 1$, the solution of (8.50) is obviously also the solution of the matrix equation

$$\mathbf{w} = T(\Delta t)\mathbf{w} + (I - T(\Delta t))A^{-1}\mathbf{s}.$$

Inserting iteration exponents in the above equation produces the convergent iterative method

$$(8.51) \quad \mathbf{w}^{(n+1)} = T(\Delta t)\mathbf{w}^{(n)} + (I - T(\Delta t))A^{-1}\mathbf{s},$$

which is identical with (8.49) for $\mathbf{w}^{(n)} := \mathbf{w}((n)\Delta t)$, and the convergence of the iterative method in (8.51) results in

$$(8.52) \quad \lim_{n \rightarrow \infty} \mathbf{w}^{(n)} = A^{-1}\mathbf{s}.$$

The relationship of the iterative method of (8.49) or (8.51) with the solution of the matrix differential equation

$$(8.53) \quad \frac{d\mathbf{v}(t)}{dt} = -C^{-1}A\mathbf{v}(t) + C^{-1}\mathbf{s}, \quad t \geq 0,$$

where $\mathbf{v}(0) = \mathbf{g}$, becomes clearer if we recall that, when $C^{-1}A$ is an irreducible M -matrix, the solution vector $\mathbf{v}(t)$ of (8.53) tends to the steady-state solution, $A^{-1}\mathbf{s}$, of (8.53). Thus, we have established that any stable approximation $T(\Delta t)$ of $\exp(-\Delta t C^{-1}A)$ with $\rho(T(\Delta t)) < 1$ gives rise to a convergent iterative method (8.51), whose solution is the steady-state solution of (8.53).

What is of interest to us now is essentially the converse of the above observation. For all the iterative methods previously described in this book,

we shall now show that we can regard these iterative methods specifically in terms of discrete approximations to parabolic differential equations, so that the actual process of iteration can be viewed as simply marking progress in time to the steady-state solution of an equation of the form (8.53). That such a relationship exists is intuitively very plausible, and is in reality an old idea. In order to bring out this relationship, we are first reminded that in an iterative procedure, such as in (8.51), the initial vector $\mathbf{w}^{(0)}$ is some vector approximation of the solution of $A\mathbf{x} = \mathbf{s}$. Although this initial approximation corresponds to the initial condition for a parabolic partial differential equation, it is, however, *not* prescribed in contrast to the matrix differential equation of (8.11)-(8.12). Next, the solution of (8.50) is *independent* of the matrix C of (8.53). This gives us the additional freedom of *choosing*, to our advantage, any nonsingular matrix C which might lead to more rapidly convergent iterative methods.

The Point Jacobi Iterative Method.

We first express the $n \times n$ matrix A in the form

$$(8.54) \quad A = D - E - F,$$

where D is a positive diagonal matrix and E and F are respectively strictly lower and strictly upper triangular matrices. Setting $C := D$, consider the consistent Padé matrix approximation $E_{0,1}(\Delta t S)$ for $\exp(-\Delta t S)$ where $S = C^{-1}A$:

$$(8.55) \quad \exp(-\Delta t S) \doteq I - \Delta t S = I - \Delta t D^{-1}(D - E - F) =: T_1(\Delta t).$$

With $L := D^{-1}E$ and $U := D^{-1}F$, the matrix $T_1(\Delta t)$ of (8.55) gives rise to the iterative procedure

$$(8.56) \quad \mathbf{w}((n+1)\Delta t) = \{(1 - \Delta t)I + \Delta t(L + U)\}\mathbf{w}(n\Delta t) + \Delta t D^{-1}\mathbf{s}.$$

Clearly, the choice $\Delta t = 1$ gives us the familiar *point Jacobi iterative method*. To extend this to the block Jacobi method, we partition the matrix A as in (3.68), and now let the matrix C be defined as the block diagonal matrix D of (3.69). Again, the choice of $\Delta t = 1$ in the approximation of (8.55) gives the associated block Jacobi iterative method.

To carry the analogy further, we now use a sequence of time steps Δt_i , $1 \leq i \leq m$, in (8.56). With $B := L + U$, this defines an iterative procedure whose error vectors $\epsilon^{(j)}$ satisfy

$$(8.57) \quad \epsilon^{(j)} = \prod_{i=1}^j \{(1 - \Delta t_i)I + \Delta t_i B\} \epsilon^{(0)}, \quad 1 \leq j \leq m.$$

It is possible to select the Δt_i 's such that

$$\prod_{i=1}^m \{(1 - \Delta t_i)I + \Delta t_i B\}$$

is the polynomial $\tilde{p}_m(B)$ of Sect. 5.1. In other words, we can generate the Chebyshev semi-iterative method with respect to the Jacobi method simply by taking proper nonuniform time increments Δt_i .

The Point Successive Overrelaxation Iterative Method.

With $A = D - E - F$ and $C := D$, consider the matrix approximation

$$(8.58) \quad \exp(-\Delta t S) \doteq (I - \Delta t L)^{-1} \{\Delta t U + (1 - \Delta t)I\} =: T_2(\Delta t).$$

For $\Delta t > 0$ sufficiently small, the expansion of $T_2(\Delta t)$ agrees through linear terms with the expansion of $\exp(-\Delta t S)$. Thus, $T_2(\Delta t)$ is a *consistent* approximation of $\exp(-\Delta t S)$. The matrix $T_2(\Delta t)$ gives rise to the iterative method

$$(8.59) \quad \mathbf{w}((n+1)\Delta t) = (I - \Delta t L)^{-1} \{\Delta t U + (1 - \Delta t)I\} \mathbf{w}(n\Delta t) + \Delta t (I - \Delta t L)^{-1} D^{-1} \mathbf{s}.$$

It is clear why this is called the *point successive overrelaxation iterative method*, since the matrix $T_2(\Delta t)$ is precisely the point successive overrelaxation matrix $\mathcal{L}_{\Delta t}$ of (3.20). Note that the relaxation factor ω corresponds *exactly* to the time increment Δt . Again, the application of block techniques to successive overrelaxation iterative methods can be similarly realized in terms of consistent approximations for an exponential matrix.

Again, let $A = D - E - F$ and $C := D$. Now, we suppose that $S = C^{-1}A = I - B$ is such that B is weakly cyclic of index 2 and has the form

$$B = \begin{bmatrix} O & H \\ H^* & O \end{bmatrix}.$$

If we take the consistent Padé matrix approximation $E_{0,2}(\Delta t S)$, then

$$(8.60) \quad \exp(-\Delta t S) \doteq I - \Delta t S + \frac{(\Delta t)^2}{2} S^2 =: \hat{T}(\Delta t),$$

where

$$(8.61) \quad \hat{T}_2(t) = \left[\begin{array}{c|c} \left(1 - t + \frac{t^2}{2}\right) I + \frac{t^2}{2} H H^* & tH - t^2 H \\ \hline tH^* - t^2 H^* & \left(1 - t + \frac{t^2}{2}\right) I + \frac{t^2}{2} H^* H \end{array} \right],$$

and we immediately see that the choice $t = 1$ is such that $\hat{T}_2(1)$ is *completely reducible* (see Theorem 2.19). This complete reducibility is the basis for the cyclic reduction iterative methods of Sect. 5.4.

The Peaceman-Rachford Iterative Method

The iterative methods described above have all been linked with Padé matrix approximations $p = 0$ and $q \geq 1$, which are *explicit*, and not in general unconditionally stable. For example, if A is Hermitian and positive definite, then from the Ostrowski-Reich Theorem 3.12, the matrix approximation $T_2(\Delta t)$ of (8.58) is stable only for the finite interval $0 \leq \Delta t \leq 2$. The Peaceman-Rachford variant of the alternating-direction implicit methods, however, is related to the implicit (and unconditionally stable) Padé matrix approximation $E_{1,1}(\Delta t S)$, which we have called the *Crank-Nicolson method*. Recalling from Sect. 7.1 that

$$(8.62) \quad A = H + V + \Sigma = H_1 + V_1,$$

where

$$(8.63) \quad H_1 := H + \frac{1}{2}\Sigma, \quad V_1 := V + \frac{1}{2}\Sigma,$$

we now consider matrix approximations for $\exp(-\Delta t C^{-1}A)$ where $C := I$. From (8.62), we have

$$(8.64) \quad \begin{aligned} \exp(-\Delta t A) &= \exp(-\Delta t [H_1 + V_1]) \\ &\doteq \exp(-\Delta t H_1) \cdot \exp(-\Delta t V_1) \end{aligned}$$

where equality is *valid* if H_1 and V_1 commute. The nature of the matrices H_1 and V_1 was such that each was, after suitable permutations, the direct sum of tridiagonal matrices. The Crank-Nicolson Padé matrix approximation $E_{1,1}(\Delta t H_1)$ of the factor $\exp(-\Delta t H_1)$ is

$$(8.65) \quad \exp(-\Delta t H_1) \doteq \left(I + \frac{\Delta t}{2} H_1 \right)^{-1} \left(I - \frac{\Delta t}{2} H_1 \right).$$

Thus, making the same approximation for $\exp(-\Delta t V_1)$, we arrive at

$$(8.66) \quad \begin{aligned} \exp(-\Delta t A) &\doteq \left(I + \frac{\Delta t}{2} H_1 \right)^{-1} \left(I - \frac{\Delta t}{2} H_1 \right) \left(I + \frac{\Delta t}{2} V_1 \right)^{-1} \\ &\quad \cdot \left(I - \frac{\Delta t}{2} V_1 \right). \end{aligned}$$

Permuting the factors, which of course is valid when H_1 and V_1 commute, we then have

$$(8.67) \quad \exp(-\Delta t A) \doteq \left(I + \frac{\Delta t}{2} V_1\right)^{-1} \left(I - \frac{\Delta t}{2} H_1\right) \left(I + \frac{\Delta t}{2} H_1\right)^{-1} \\ \cdot \left(I - \frac{\Delta t}{2} V_1\right) =: T_3(\Delta t).$$

Since $T_3(\Delta t)$ can also be written for $\Delta t > 0$ as

$$(8.68) \quad T_3(\Delta t) = \left(V_1 + \frac{2}{\Delta t} I\right)^{-1} \left(\frac{2}{\Delta t} I - H_1\right) \left(H_1 + \frac{2}{\Delta t} I\right)^{-1} \\ \cdot \left(\frac{2}{\Delta t} I - V_1\right),$$

we see then that $T_3(\Delta t)$ is the Peaceman-Rachford matrix of (7.13). Note that the acceleration parameter r for the Peaceman-Rachford iterative method is given by

$$(8.69) \quad r = \frac{2}{\Delta t}.$$

Several items are of interest here. First, it can be verified that the matrix $T_3(\Delta t)$ is a consistent approximation of $\exp(-\Delta t A)$. Moreover, if Δt is small, then the expression of $T_3(\Delta t)$ agrees through *quadratic terms* with the expansion of $\exp(-\Delta t A)$. Second, we know from Theorem 8.10 that the Crank-Nicolson Padé matrix approximation $E_{1,1}(\Delta t A)$ is unconditionally stable when A has its eigenvalues λ_i satisfying $\operatorname{Re} \lambda_i > 0$. It is not surprising then that the interval in r for which the Peaceman-Rachford matrix for a single parameter is convergent is $(0, +\infty)$. (See Theorem 7.1).

Using the idea that different choices of the diagonal matrix C in (8.53) can be useful in determining properties of iterative methods, it is now interesting to describe the Wachspress-Habetler variant (see Sect. 7.4) of the Peaceman-Rachford iterative method in terms of this approach. Let F be the positive diagonal matrix, corresponding to $\phi \equiv 1$ in (8.1). In other words, the positive diagonal entries of F are from Sect. 8.1 *exactly* the areas of the mesh rectangles $r_{i,j}$ of Sect. 8.1. Thus, from

$$(8.70) \quad F \frac{d\mathbf{v}(t)}{dt} = -(H_1 + V_1)\mathbf{v}(t) + \mathbf{s} = -A\mathbf{v}(t) + \mathbf{s},$$

where $\mathbf{v}(0)$ will be our initial approximation to $A^{-1}\mathbf{s}$, we consider matrix approximations to $\exp(-\Delta t F^{-1}A)$ of the form

$$(8.71) \quad \exp(-\Delta t F^{-1}A) \doteq \left(F + \frac{\Delta t}{2} V_1\right)^{-1} \left(F - \frac{\Delta t}{2} H_1\right) \\ \cdot \left(F + \frac{\Delta t}{2} H_1\right)^{-1} \left(F - \frac{\Delta t}{2} V_1\right) := T_4(\Delta t),$$

which can be shown to be a consistent approximation to $\exp(-\Delta t F^{-1}A)$. The matrix $T_4(\Delta t)$ turns out, of course, to be the Wachspress-Habetler variant (7.81)-(7.82) of the Peaceman-Rachford iterative method, corresponding to the acceleration parameter $r = 2/\Delta t$. As previously mentioned, this choice of the diagonal matrix $C = F$, which appears to be a very natural choice in solving (8.1) with $\phi \equiv 1$, allows one to rigorously apply the commutative theory (Sect. 7.2) for alternating-direction implicit methods to the numerical solution of the Dirichlet problem in a rectangle with *nonuniform* mesh spacings.

The Symmetric Successive Overrelaxation Iterative Method.

We now assume that the matrix $A = D - E - E^*$ is Hermitian and positive definite, that the associated matrix D is Hermitian and positive definite, and that $D - \omega E$ is nonsingular for $0 \leq \omega \leq 2$. Setting $C := D$, consider the matrix approximation

$$(8.72) \quad \exp(-tC^{-1}A) \doteq \left(D - \frac{t}{2}E^*\right)^{-1} \left(\frac{t}{2}E + \left(1 - \frac{t}{2}\right)D\right) \\ \cdot \left(D - \frac{t}{2}E\right)^{-1} \left(\frac{t}{2}E^* + \left(1 - \frac{t}{2}\right)D\right) =: T_5(t).$$

Although $T_5(t)$ can be verified to be a consistent approximation of $\exp(-tC^{-1}A)$, what interests us here is that $T_5(t)$ can be expressed as

$$(8.73) \quad T_5(t) = \left(I - \frac{t}{2}U\right)^{-1} \left(\frac{t}{2}L + \left(1 - \frac{t}{2}\right)I\right) \left(I - \frac{t}{2}L\right)^{-1} \\ \cdot \left(\frac{t}{2}U + \left(1 - \frac{t}{2}\right)I\right),$$

where $L = D^{-1}E$ and $U = D^{-1}E^*$. Since the inner two bracketed terms commute and can be interchanged, then $T_5(t)$ is similar to

$$\tilde{T}_5(t) := \left(I - \frac{t}{2}U\right) T_5(t) \left(I - \frac{t}{2}U\right)^{-1},$$

which can be written as

$$(8.74) \quad \tilde{T}_5(t) = \left(I - \frac{t}{2}L\right)^{-1} \left(\frac{t}{2}L + \left(1 - \frac{t}{2}\right)I\right) \\ \cdot \left(\frac{t}{2}U + \left(1 - \frac{t}{2}\right)I\right) \left(I - \frac{t}{2}U\right)^{-1},$$

which can be expressed equivalently as

$$\begin{aligned}
 \tilde{T}_5(t) &= \left(D - \frac{t}{2}E\right)^{-1} \left(\frac{t}{2}E + \left(1 - \frac{t}{2}\right)D\right) \\
 (8.75) \quad &\cdot \left(\frac{t}{2}E^* + \left(1 - \frac{t}{2}\right)D\right) \left(D - \frac{t}{2}E^*\right)^{-1} \\
 &= Q^*Q,
 \end{aligned}$$

where

$$(8.76) \quad Q := \left(\frac{t}{2}E^* + \left(1 - \frac{t}{2}\right)D\right) \left(D - \frac{t}{2}E^*\right)^{-1}.$$

As $\tilde{T}_5(t) = Q^*Q$, then $\tilde{T}_5(t)$ is a nonnegative definite Hermitian matrix, which proves that the eigenvalues of $T_5(t)$ are nonnegative real numbers. Moreover, in a similar manner it can be shown that the eigenvalues of $T_5(t)$ are less than unity for $0 < t/2 < 2$. (See Exer. 6.)

For the iterative method resulting from the matrix approximation $T_5(t)$ of $\exp(-tC^{-1}A)$, this can be written as a two-step method:

$$\begin{aligned}
 \mathbf{w}^{(n+1/2)} &= \left(I - \frac{t}{2}L\right)^{-1} \left\{ \frac{t}{2}U + \left(1 - \frac{t}{2}\right)I \right\} \mathbf{w}^{(n)} \\
 &\quad + \frac{t}{2} \left(I - \frac{t}{2}L\right)^{-1} D^{-1}\mathbf{s}, \\
 (8.77) \quad \mathbf{w}^{(n+1)} &= \left(I - \frac{t}{2}U\right)^{-1} \left\{ \frac{t}{2}L + \left(1 - \frac{t}{2}\right)I \right\} \mathbf{w}^{(n+1/2)} \\
 &\quad + \frac{t}{2} \left(I - \frac{t}{2}U\right)^{-1} D^{-1}\mathbf{s},
 \end{aligned}$$

which is called the **symmetric successive overrelaxation iterative method** and has been considered by Aitken (1950), Sheldon (1955), and Habetler and Wachspress (1961). This iterative method corresponds to sweeping the mesh in one direction, and then reversing the direction for another sweep of the mesh.²⁰ Sheldon (1955) observed that, since the resulting iteration matrix $T_5(t)$ has nonnegative real eigenvalues less than unity for $0 < t/2 < 2$, one can rigorously apply the Chebyshev semi-iterative method of Sect. 5.1 to (8.77) to accelerate convergence further. In effect then, one first selects the optimum value of t which gives the smallest spectral radius for the matrix $T_5(t)$ and then generates a three-term Chebyshev recurrence relation for the vectors $\mathbf{w}^{(n)}$.

As a final result in this chapter, we recall that in our discussions of the successive overrelaxation iterative method and the Peaceman-Rachford variant of the alternating-direction implicit iterative methods, each of the associated iteration matrices was shown to be *primitive* for certain choices of acceleration parameters.²¹ However, these isolated and seemingly independent results

²⁰ For this reason, this method is sometimes called the *to-fro* or *forward-backward* iterative method.

²¹ See Exer. 4 of Sect. 3.6 and Theorem 7.15 of Sect. 7.3.

are merely corollaries to Theorem 8.12, as the associated matrix approximations of (8.58) and (8.67) are nonnegative consistent approximations to an irreducible M -matrix in each case.

In summary, the main purpose of this section was to establish a correspondence between the iterative methods previously described and matrix methods for approximating the solution of parabolic partial differential equations. Although it is true that all these iterative methods could be adapted for use in approximating the solution of parabolic partial differential equations, the variants of the alternating-direction implicit method are more widely used in practical parabolic problems, mainly because of their inherent unconditional stability.

Exercises

1. Show that one can select time increments $\Delta t_i, 1 \leq i \leq m$, such that the matrix of (8.57),

$$\prod_{i=1}^m \{(1 - \Delta t_i)I + \Delta t_i B\},$$

is precisely the polynomial $\tilde{p}_m(B)$ of (5.21).

2. Let $A = D - E - F$ be an $n \times n$ matrix where D is a positive diagonal matrix and E and F are respectively strictly lower and strictly upper triangular matrices. For the splittings

$$\begin{aligned} M_1 &= D, & N_1 &= E + F, \\ M_2 &= D - E, & N_2 &= F, \\ M_3 &= \frac{1}{\omega}(D - \omega E), & N_3 &= \frac{(1 - \omega)}{\omega}D + F, \quad \omega \neq 0, \end{aligned}$$

let $C_i := M_i, i = 1, 2, 3$, and consider respectively the Padé matrix approximation $E_{0,1}(\Delta t C_i^{-1} A)$ for $\exp(-\Delta t C_i^{-1} A)$. Show for $\Delta t = 1$ that these Padé approximations exactly correspond to the point Jacobi, point Gauss-Seidel, and point successive overrelaxation iterative methods, respectively.

3. Verify that the matrices $T_i(\Delta t), 2 \leq i \leq 5$, of (8.58), (8.67), (8.71), and (8.72) are each consistent approximations of $\exp(-\Delta t S)$. Next, for small Δt show that the expansion of $T_3(\Delta t)$ in (8.67) agrees through *quadratic* terms of $\exp(-\Delta t A)$ without assuming that H_1 and V_1 commute.

4. Let $A := I - B$, where

$$B = \left[\begin{array}{c|c} O & H \\ \hline H^* & O \end{array} \right]$$

is convergent, and set $C := I$. Show that the optimum value of $t/2 = \omega$ which minimizes the spectral radius of $T_5(t)$ of (8.72) is $\omega = 1$ (Kahan (1958)).

- *5. Consider two iterative methods for solving the matrix problem $Ax = k$, where $A := I - B$ and B is the convergent matrix of the previous Exer. 4. The first iterative method is defined by applying the Chebyshev semi-iterative method of Sect. 5.1 to the optimized basic iteration matrix $T_5(t)$ of (8.72) with $t/2 = \omega = 1$ and $C := I$. The second iterative method is defined by applying the Chebyshev semi-iterative method to the basic cyclic reduction iterative method of Sect. 5.4. Show that the second method is iteratively faster for m iterations than the first method, for every $m > 1$.
- *6. Let $A = D - E - E^*$ and D be $n \times n$ Hermitian and positive definite matrices. Prove that the eigenvalues λ_i of $T_5(t)$ of (8.72) satisfy $0 < \lambda_i < 1$ for all $1 \leq i \leq n$ when $0 < t/2 < 2$. (*Hint*: Apply Stein's result of Exer. 6, Sect. 1.3) (Habetler and Wachspress (1961)).

8.5 Chebyshev Rational Approximations for $\exp(-tS)$

The Padé rational approximations of e^{-z} , which led to Padé rational matrix approximations of $\exp(-tS)$ in Section 8.3, are defined as best *local* rational approximations of e^{-z} at $z = 0$, i.e. (cf. (8.46)),

$$(8.78) \quad e^{-z} - \frac{n_{p,q}(z)}{d_{p,q}(z)} = O(|z|^{p+q+1}), \quad \text{as } |z| \rightarrow 0,$$

for each pair (p, q) of nonnegative integers. But, results from complex approximation theory provide us with a different rational approximation which also has applications to the numerical solution of semi-discrete parabolic matrix equations.

Specifically, consider the following problem of Chebyshev rational approximations to e^{-x} on the *infinite* interval $[0, +\infty)$. For a nonnegative integer m , let π_m denote the set of all real polynomials of degree at most m , and, for a pair (n, m) of nonnegative integers, let $\pi_{n,m}$ analogously denote the set of all real rational functions

$$(8.79) \quad r_{n,m}(x) = p(x)/q(x), \text{ where } p \in \pi_n \text{ and } q \in \pi_m.$$

We then define

$$(8.80) \quad \lambda_{n,m} := \inf_{r_{n,m} \in \pi_{n,m}} \|e^{-x} - r_{n,m}(x)\|_{L_\infty[0,+\infty)},$$

where $\lambda_{n,m}$ is called the constant of best uniform rational Chebyshev approximation of e^{-x} on $[0, +\infty)$. It is obvious from (8.80) that $\lambda_{n,m}$ is finite if and only if $0 \leq n \leq m$, so we assume, henceforth, that $0 \leq n \leq m$. With the usual compactness considerations, it can then be shown (cf. Achieser (1956), p. 55) that, after dividing out possible common factors, there is a unique $\hat{r}_{n,m}(x) = \hat{p}_{n,m}(x)/\hat{q}_{n,m}(x)$ in $\pi_{n,m}$ where $\hat{q}_{n,m}(x) > 0$ on $[0, +\infty)$, such that

$$(8.81) \quad \lambda_{n,m} = \|e^{-x} - \hat{r}_{n,m}(x)\|_{L_\infty[0,+\infty)}.$$

It is evident from the definition in (8.80) that

$$(8.82) \quad 0 < \lambda_{m,m} \leq \lambda_{m-1,m} \leq \cdots \leq \lambda_{0,m} \text{ for any nonnegative integer } m.$$

To give some history for this problem, Cody, Meinardus and Varga (1969) showed that for $s_m(x) := \sum_{k=0}^m x^k/k!$, the familiar m -th partial sum of e^x , there holds

$$(8.83) \quad \lim_{m \rightarrow \infty} \left\{ \|e^{-x} - 1/s_m(x)\|_{L_\infty[0,+\infty)} \right\}^{1/m} \leq \frac{1}{2},$$

i.e., **geometric convergence** is obtained. But as $1/s_m(x)$ is an element of $\pi_{0,m}$, and not necessarily the best element $\hat{r}_{0,m}(x)$ in this case from (8.81), it follows from (8.82) that, for any sequence $\{n(m)\}_{m=0}^\infty$ of nonnegative integers with $0 \leq n(m) \leq m$, there holds

$$(8.84) \quad \overline{\lim}_{m \rightarrow \infty} (\lambda_{n(m),m})^{1/m} \leq \frac{1}{2}.$$

The result of (8.84), on the geometric convergence of the Chebyshev constants for e^{-x} on $[0, +\infty)$, gave rise to further interesting numerical and theoretical results. As will be indicated later, work considerations, in applications of the Chebyshev rational approximations of e^{-x} on $[0, +\infty)$ to the numerical solution of the semi-discrete matrix approximations of parabolic partial differential equations in Section 8.1, always lead to the choice $n(m) = m$ for all m , so that interest then focused on the specific Chebyshev constants $\lambda_{m,m}$. From (8.82), this is the smallest such constant from the set $\{\lambda_{j,m}\}_{j=0}^m$. (The numerical and theoretical work on the behavior of $\{\lambda_{m,m}\}_{m=0}^\infty$ is covered in detail in Varga (1990), Chap. 2.)

Based on numerical and theoretical results up to that time, the following conjecture was made in Saff and Varga (1977):

$$(8.85) \quad \text{Conjecture : } \lim_{m \rightarrow \infty} \lambda_{m,m}^{1/m} = \frac{1}{9}.$$

It turns out that the above conjecture is *false*; the correct result, due to Gonchar and Rakhmanov (1987), is

$$(8.86) \quad \lim_{m \rightarrow \infty} \lambda_{m,m}^{1/m} = \Lambda$$

where

$$(8.87) \quad \Lambda = \frac{1}{9.28902 \dots}.$$

Hence, their constant is actually *smaller* than the conjectured constant $1/9$ of (8.85), so that the geometric convergence is *faster* than was conjectured.

It is interesting to briefly describe the analytical result of Gonchar and Rakhmanov (1987). Using deep potential-theoretic techniques in complex approximation theory, they established the following beautiful result.

Theorem 8.13. *The number Λ of (8.86) can be characterized in the following number-theoretic way. Define*

$$(8.88) \quad f(z) := \sum_{j=1}^{\infty} a_j z^j,$$

where

$$(8.89) \quad a_j := \left| \sum_{d|j} (-1)^d d \right| \quad (j = 1, 2, \dots),$$

so that $f(z)$ is a real analytic function in $|z| < 1$, which is strictly increasing on the interval $[0, 1)$. Then, Λ is the unique positive root of the equation

$$(8.90) \quad f(\Lambda) = \frac{1}{8}.$$

It is interesting to mention A. P. Magnus wrote in late 1986 to A. A. Gonchar that the constant Λ of (8.87) is also the unique solution (less than unity) of the equation

$$\sum_{n=0}^{\infty} (2n+1)^2 (-\Lambda)^{n(n+1)/2} = 0,$$

and that this same constant Λ appeared *exactly one hundred years earlier* in Halphen (1886), who had computed Λ to six significant digits. Halphen had arrived at the above equation in his studies of theta functions. It seems appropriate to call Λ the “**Halphen constant**”!

In calculations carried out in Carpenter, Ruttan, and Varga (1984), using the second Remez algorithm, the numbers $\{\lambda_{m,m}\}_{m=0}^{30}$ of (8.80) were determined to approximately 200 decimal digits, along with explicit determinations of the associated rational functions $\{\hat{r}_{m,m}(x)\}_{m=0}^{30}$ of (8.81). For example, these calculations give that

$$\lambda_{4,4} = 8.652406 \cdot 10^{-5}, \lambda_{5,5} = 9.345713 \cdot 10^{-6}, \text{ and } \lambda_{6,6} = 1.008454 \cdot 10^{-6},$$

so that, in particular from (8.81), we have

$$|e^{-x} - \hat{r}_{4,4}(x)| \leq 8.652407 \cdot 10^{-5} \quad \text{for all } x \geq 0.$$

For applications of the above approximation theoretic results, from (8.32), consider the simplified matrix problem

$$(8.91) \quad \frac{d\mathbf{v}(t)}{dt} = -A\mathbf{v}(t) + \mathbf{s}, \text{ for } t > 0, \text{ with } \mathbf{v}(0) = \mathbf{g},$$

where we assume A is a real symmetric and positive definite $n \times n$ matrix, and that \mathbf{s} is a vector, independent of t . The solution of (8.91) is then

$$(8.92) \quad \mathbf{v}(t) = A^{-1}\mathbf{s} + \exp(-tA) \cdot \{\mathbf{g} - A^{-1}\mathbf{s}\} \quad (t \geq 0).$$

Choosing the rational approximation $\hat{r}_{m,m}(x)$ of e^{-x} in (8.81), and replacing the variable x by the matrix tA , our vector approximation for $\mathbf{v}(t)$ is then defined as

$$(8.93) \quad \mathbf{w}_m(t) := A^{-1}\mathbf{s} + \hat{r}_{m,m}(tA) \cdot \{\mathbf{g} - A^{-1}\mathbf{s}\} \quad (t \geq 0),$$

i.e., with $\hat{r}_{m,m}(x) = \hat{p}_{m,m}(x)/\hat{q}_{m,m}(x)$, this becomes

$$(8.94) \quad \mathbf{w}_m(t) = A^{-1}\mathbf{s} + (\hat{q}_{m,m}(tA))^{-1}(\hat{p}_{m,m}(tA) \cdot \{\mathbf{g} - A^{-1}\mathbf{s}\}) \quad (t \geq 0),$$

which defines the m -th **matrix Chebyshev rational approximation** of the solution of (8.91). (At this point, we could have chosen $\hat{r}_{n,m}(x)$, with $0 \leq n < m$. But as the work of inverting the matrix $\hat{q}_{n,m}(tA)$, a polynomial of degree m in A , is dominant and far exceeds the work in multiplying the vector $\{\mathbf{g} - A^{-1}\mathbf{s}\}$ by the matrix $\hat{p}_{n,m}(tA)$, and as $\lambda_{m,m} \leq \lambda_{n,m}$ from (8.82), then choosing $n = m$ is indicated.) Applying norms to the difference $\mathbf{v}(t) - \mathbf{w}_m(t)$, from (8.92) and (8.94), we have

$$(8.95) \quad \|\mathbf{v}(t) - \mathbf{w}_m(t)\| \leq \|\exp(-tA) - \hat{r}_{m,m}(tA)\| \cdot \|\mathbf{g} - A^{-1}\mathbf{s}\| \quad (t \geq 0).$$

Because A is a real symmetric and positive definite matrix, it follows from Corollary 1.8 that, with $\sigma(A)$ denoting the spectrum of A ,

$$\|\exp(-tA) - \hat{r}_{m,m}(tA)\| = \max_{\lambda_i \in \sigma(A)} |e^{-t\lambda_i} - \hat{r}_{m,m}(t\lambda_i)| \leq \lambda_{m,m},$$

the last inequality following from (8.81) and the fact that $0 \leq t\lambda_i < \infty$ for all eigenvalues λ_i of A , all $t \geq 0$. Hence, from (8.95), we have

$$(8.96) \quad \|\mathbf{v}(t) - \mathbf{w}_m(t)\| \leq \lambda_{m,m} \cdot \|\mathbf{g} - A^{-1}\mathbf{s}\| \quad \text{for all } t \geq 0.$$

Because the above uniform error bound involves these small constants $\lambda_{m,m}$, these Chebyshev rational approximations were used in Reusch, et. al (1988) and Gallopoulos and Saad (1992) in numerically solving discretizations of time-dependent parabolic problems. It should be pointed out that these matrix Chebyshev rational approximations do not directly apply to *nonlinear* forms of the parabolic equation (8.1), i.e., when the coefficients K_i , G_i , σ , and S depend *nonlinearly* on the sought-solution $u(\mathbf{x}, t)$.

Bibliography and Discussion

- 8.1 It is interesting that various forms of semi-discrete approximations have been considered independently by several authors. Perhaps the latest comprehensive use of such semi-discrete approximations appears in the excellent book of Thomée (1997), while the first such approach was taken by Hartree and Womersley (1937), who, in numerically approximating the solution of the one-dimensional heat equation (8.6), discretized the *time* variable but left the space variable continuous. The main reason for this approach was to solve the resulting system of ordinary differential equations by means of a differential analyzer. This general semi-discrete approach, of course, is closely related to the Russian “method of lines,” where an elliptic partial differential equation in two space variables is approximated by a system of ordinary differential equations. See Faddeeva (1949) and Kantorovich and Krylov (1958), p. 321. Franklin (1959b), Lees (1961), and Varga (1961) all consider approximations to parabolic partial differential equations in which the spatial variables are discretized but the time variable is kept continuous.

The explicit representation of the solution (8.15) of the particular ordinary matrix differential equation of (8.13), using exponentials of a square matrix, is a special case of results to be found in Bellman (1953), p. 12, and is further connected to the abstract theory of semigroups of operators, as the matrix A is the *infinitesimal generator* for the semigroup of matrix operators $\exp(tA)$. For the theory of semigroups, see Phillips (1961) and Hille and Phillips (1957). Bellman (1953) also gives explicit representations for the solution of the more general differential equation

$$\frac{d\mathbf{v}(t)}{dt} = -A(t)\mathbf{v}(t) + \mathbf{s}(t),$$

where the matrix $A(t)$ has time-dependent entries. Because of this, a corresponding discussion of semi-discrete approximations to the partial differential equation (8.1) in which the coefficients $\phi(\mathbf{x}; t)$, $K_i(\mathbf{x}; t)$, etc., are functions of time can also be given.

Results similar to those of Exers. 5 and 6, in which bounds for the norm of the approximate solution $\mathbf{v}(t)$ as well as bounds for the norm of $(\mathbf{u}(t) - \mathbf{v}(t))$ are obtained, are sometimes derived from the *maximum principle* for such parabolic problems. See Douglas (1961b), Keller (1960b), and Lees (1960).

- 8.2 The terms *essentially positive*, *subcritical*, *critical*, and *supercritical* were introduced by Birkhoff and Varga (1958) in the study of the numerical solution of the time-dependent multigroup diffusion equations of reactor physics, and the results of this section are largely drawn from this reference. For generalizations to more general operators, see Birkhoff (1961) and Habetler and Martino (1961), and references given therein.

Essentially positive matrices, as defined in this section, are also called *input-output* matrices and *Leontieff matrices* in problems arising from economic considerations. For a detailed discussion of such matrices and their economic applications, see Karlin (1959b), and references cited there. See also Beckenbach and Bellman (1961), (pp. 83 and 94), where the result of Theorem 8.3, for example, is given in a slightly weaker form.

From Theorem 8.5, we deduce that the norms $\|\exp(tQ)\|$ of a subcritical essentially positive matrix are uniformly bounded for all $t \geq 0$. More generally, one can deduce, from Lemma 8.4, necessary and sufficient conditions that $\|\exp(tQ)\|$ be bounded for all $t \geq 0$ for a *general* $n \times n$ complex matrix. For such extensions, see Bellman (1953), p. 25, and Kreiss (1959a), as well as Exers. 5 and 9. It should also be stated that the result of Exer. 9 is clearly related to the Hille-Yosida theorem of semigroup theory. See Phillips (1961).

- 8.3 The analyses of the forward difference, backward difference, and Crank-Nicolson methods for numerically approximating the solution of parabolic partial differential equations have received a great deal of attention. For excellent bibliographical coverage, see Todd (1956), Richtmyer (1957), Young (1961), and Douglas (1961b).

The initial works of Courant, Friedrichs, and Lewy (1928), von Neumann (see O'Brien, Hyman, and Kaplan (1951)), Laasonen, P. (1949), Crank and Nicolson (1947), and many subsequent articles, give analyses of these finite difference methods for uniform mesh spacings, and generally one finds in the literature that Fourier series methods are a commonly used tool. Although the goal in this section, following Varga (1961), has been to show the relevance of these basic numerical methods to particular Padé matrix approximations of $\exp(-tA)$, the stability analysis of these Padé approximations in Theorem 8.10 for M -matrices is applicable under fairly wide circumstances, since non-homogeneous, nonrectangular problems with internal interfaces and non-uniform spatial meshes do not require special treatment. See also Householder (1958) for a related use of norms in studying stability criteria.

Not all known methods for approximating solutions of parabolic partial differential equations can be obtained from Padé matrix approximations of $\exp(-tA)$. For example, multilevel approximations such as the unconditionally stable method of DuFort and Frankel (1953) and the unconditionally *unstable* method of Richardson (1910) require special treatment. See, for example, Douglas (1961b), Todd (1956), and Richtmyer (1957).

Padé rational approximations are only a particular type of rational approximation of $\exp(-tA)$. Franklin (1959b) considers consistent explicit polynomial approximations of high order in order to maximize the stability interval in Δt . Varga (1961) considers Chebyshev rational approximations of $\exp(-tA)$ that are aimed at obtaining useful vector approximations in just *one* time step.

It is interesting to note that several definitions of stability exist in the literature. For example, O'Brien, Hyman, and Kaplan (1951) relate stability to growth of rounding errors, where Lax and Richtmyer (1956) define an operator $C(\Delta t)$ to be stable if for some $\tau > 0$, the set of norms $\|C^m(\Delta t)\|$ is uniformly bounded for $0 < \Delta t \leq \tau, 0 \leq m\Delta t \leq T$. However, if the size of $m\Delta t$ is unrestricted, a supposition of interest in problems admitting a finite steady-state solution, this definition of Lax and Richtmyer reduces to that given in Definition 8.9. See also Richtmyer (1957) and Esch (1960).

Because we have omitted the entire topic of *convergence* of the difference approximations to the solution of the differential equation, the basic results of Lax and Richtmyer (1956) and Douglas (1956), which in essence show the *equivalence* of the concepts of stability and convergence, are not discussed here. These analyses consider the combined errors due to spatial as well as time discretizations. In this regard, the article of John (1952) is basic. See also Lees (1959), Strang (1960), and Kreiss (1959b).

The matrix result of Theorem 8.12, linking the Perron-Frobenius theory of nonnegative matrices to the solutions of discrete approximations to certain parabolic partial differential equations, is apparently new.

It must also be mentioned that our treatment in Sect. 8.3 does *not* cover the parabolic partial differential equations for which the coefficients of the differential equation are functions of time, or functions of the unknown solution, nor does it cover coupled systems of parabolic partial differential equations. For the numerical solution of such problems, we recommend the survey article by Douglas (1961b).

- 8.4 Many authors, including Frankel (1950) and Garabedian (1956), have either indicated or made use of the heuristic analogy between iterative methods for solving elliptic differential equations and numerical methods for solving parabolic partial differential equations. For the alternating-direction implicit iterative methods of Peaceman and Rachford (1955) and Douglas and Rachford (1956), this association is quite clear. The conclusion in this section that all such iterative methods can be obtained as consistent approximations of exponential matrices seems to be new, but other approaches have also produced interesting results. Notably, Garabedian (1956) has linked the successive overrelaxation iterative method with *hyperbolic* partial differential equations, and this will be described in Chap. 9.

The application of the results of Theorem 8.12, further tying in the Perron-Frobenius theory of nonnegative primitive matrices, also appears to be new.

The *symmetric* successive overrelaxation iterative method is compared with the successive overrelaxation iterative method by Habetler and Wachspress (1961), where it is shown that the symmetric successive overrelaxation iterative method is superior in problems with limited application.

- 8.5 The excursion in this section into Chebyshev rational approximations of e^{-x} on $[0, +\infty)$ is perhaps of interest in that it couples results from complex approximation theory to the numerical solution of certain parabolic problems. But, this investigation has also opened up related research in complex approximation theory. For example, are there real continuous functions, other than $f(x) = e^x$, defined on $[0, +\infty)$, such that $1/f(x)$ admits geometric convergence by reciprocals of polynomials on $[0, +\infty)$, i.e., there exist polynomials $\{p_m(x)\}_{m=0}^\infty$, with $p_m \in \pi_m$ for all $m \geq 0$, and a real number $q > 1$, such that

$$(8.97) \quad \overline{\lim}_{m \rightarrow \infty} \left\{ \left\| \frac{1}{p_m} - \frac{1}{f} \right\|_{L_\infty[0, +\infty)} \right\}^{1/m} \leq \frac{1}{q} < 1?$$

This has been studied in Meinardus, Reddy, Taylor and Varga (1972), where necessary and nearly equivalent sufficient conditions for this have been found. (See also Andrievskii, Blatt and Kovacheva (1997).) It turns out that if (8.97) is valid, then $f(x)$, defined on $[0, +\infty)$, has an extension to an entire function $F(z)$, where $F(x) = f(x)$ for all $x \geq 0$, with $F(z)$ having finite exponential growth. See also Varga (1990), Chap. 2.



<http://www.springer.com/978-3-642-05154-8>

Matrix Iterative Analysis

Varga, R.S.

2000, X, 358 p., Softcover

ISBN: 978-3-642-05154-8