Chapter 4.
Accuracy, Stability, and Convergence

Mighty oaks from little acorns grow!
— *ANONYMOUS*

The problem of stability is pervasive in the numerical solution of partial differential equations. In the absence of computational experience, one would hardly be likely to guess that instability was an issue at all,* yet it is a dominant consideration in almost every computation. Its impact is visible in the nature of algorithms all across this subject—most basically, in the central importance of linear algebra, since stability so often necessitates the use of implicit or semi-implicit formulas whose implementation involves large systems of discrete equations.

The relationship between stability and convergence was hinted at by Courant, Friedrichs, and Lewy in the 1920's, identified more clearly by von Neumann in the 1940's, and brought into organized form by Lax and Richtmyer in the 1950's—the Lax Equivalence Theorem. After presenting an example, we shall begin with the latter, and then relate it to the CFL and von Neumann conditions. After that we discuss the important problem of determining stability of the method of lines. For problems that lead to normal matrices, it is enough to make sure that the spectra of the spatial discretization operators lie within a distance $O(k)$ of the stability region of the time-stepping formula, but if the matrices are not normal, one has to consider pseudospectra instead.

The essential issues of this chapter are the same as those that came up for ordinary differential equations in Sections 1.5–1.7. For partial differential equations, however, the details are more complicated, and more interesting.

In addition to Richtmyer and Morton, a good reference on the material of this chapter is V. Thomée, "Stability theory for partial difference operators," *SIAM Review 11* (1969), 152–195.

---

*In particular, L. F. Richardson, the originator of finite-difference methods for partial differential equations, did not discover instability; see his book *Weather Prediction by Numerical Processes*, Cambridge University Press, 1922 (!), reprinted by Dover in 1965.

4.1. An example

Consider the model partial differential equation

$$u_t = u_x, \qquad x \in \mathbb{R}, \quad t \geq 0 \tag{4.1.1}$$

together with initial data

$$u(x,0) = \begin{cases} \cos^2 x & |x| \leq \frac{\pi}{2}, \\ 0 & |x| \geq \frac{\pi}{2}. \end{cases} \tag{4.1.2}$$

Let us solve this initial-value problem numerically by the leap frog formula (3.2.2), with space and time steps

$$h = 0.04\,\pi, \qquad k = \lambda h,$$

where λ is a constant known as the **mesh ratio.** Thus the leap frog formula takes the form

$$v_j^{n+1} = v_j^{n-1} + \lambda(v_{j+1}^n - v_{j-1}^n), \tag{4.1.3}$$

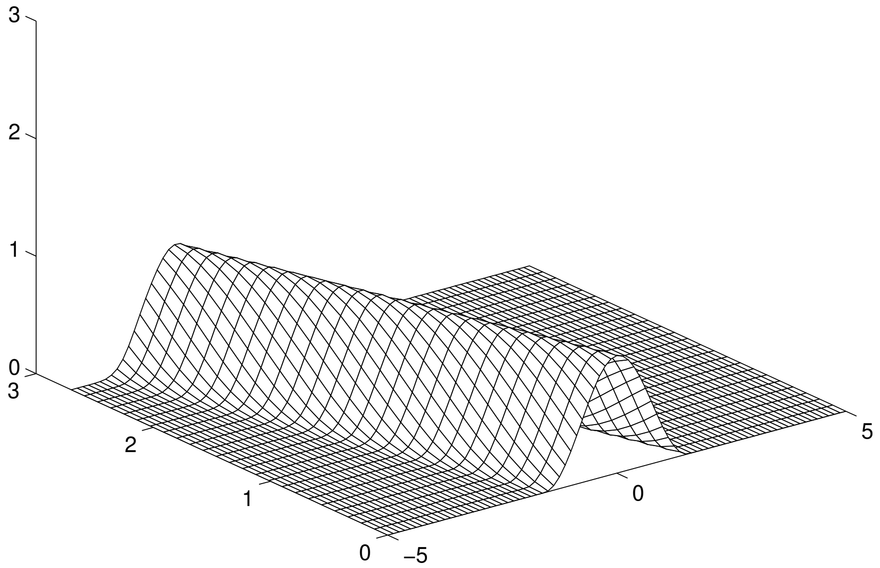
with the bump in the initial function represented by 25 grid points. The starting values at $t = 0$ and k will both be taken from the exact solution $u(x,t) = u(x+t,0)$.

Figure 4.1.1 shows computed results with $\lambda = 0.9$ and $\lambda = 1.1$, and they are dramatically different. For $\lambda < 1$ the leap frog formula is stable, generating a left-propagating wave as expected. For $\lambda > 1$ it is unstable. The errors introduced at each step are not much bigger than before, but they grow exponentially in subsequent time steps until the wave solution is obliterated by a sawtooth oscillation with 4 points per wavelength. This rapid blow-up of a sawtooth mode is typical of unstable finite difference formulas.

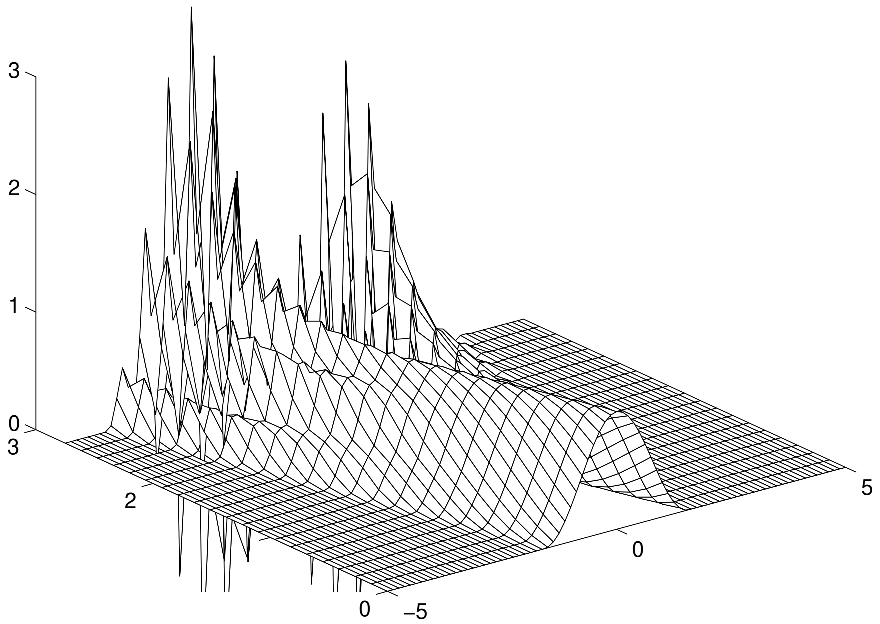
Although rounding errors can excite an instability, more often it is discretization errors that do so, and this particular experiment is quite typical in this respect. Figure 4.1.1 would have looked the same even if the computation had been carried out in exact arithmetic.

This chapter is devoted to understanding instability phenomena in a general way. Let us briefly mention how each of the sections to follow relates to the particular example of Figure 4.1.1.

First, §4.2 presents the celebrated Lax Equivalence Theorem: a consistent finite difference formula is convergent if and only if it is stable. Our example



(a) $\lambda = 0.9$



(b) $\lambda = 1.1$

Figure 4.1.1. Stable and unstable leap frog approximations to $u_t = u_x$.

(a) largest value $2i\lambda\sin\xi h$ *(b)* corresponding amplification factors z

Figure 4.1.2. Right-hand sides and corresponding amplification factors z of (4.1.6). The circles correspond to Figure 4.1.1a and the crosses to Figure 4.1.1b.

is consistent for any λ, but stable only for $\lambda < 1$. Thus as Figure 4.1.1 suggests, the numerical results would converge to the correct solution as $h,k \to 0$ in case *(a)*, but not in case *(b)*.

Next, §4.3 presents the CFL condition: a finite difference formula can be stable only if its numerical domain of dependence is at least as large as the mathematical domain of dependence. In the space-time grid of Figure 4.1.1*(b)*, information travels under the leap frog model at speed at most $(1.1)^{-1}$, which is less than the propagation speed 1 for the PDE itself. Thus *something* had to go wrong in that computation.

Section 4.4 presents the von Neumann approach to stability: Fourier analysis and amplification factors. This is the workhorse of stability analysis, and the foundations were given already in §§3.5,3.6. For our leap frog model (4.1.3), inserting the trial solution

$$v_j^n = z^n e^{i\xi x_j}, \qquad \xi \in \mathbb{R} \tag{4.1.4}$$

leads to the equation

$$z = z^{-1} + \lambda(e^{i\xi h} - e^{-i\xi h}), \tag{4.1.5}$$

that is,

$$z - z^{-1} = 2i\,\lambda\sin\xi h. \tag{4.1.6}$$

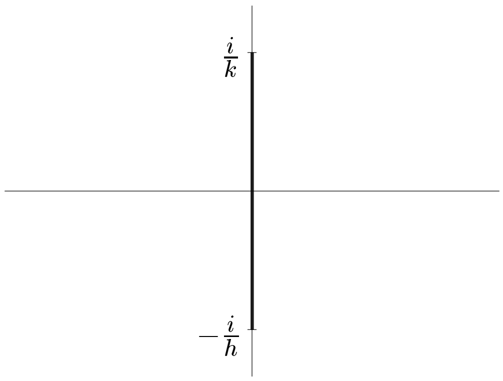
This is a quadratic equation in z with two complex roots (in general). As ξ varies, the right-hand side ranges over the complex interval $[-2i\lambda, 2i\lambda]$. For $|\lambda| \le 1$ this interval is a subset of $[-2i, 2i]$, and so the roots z lie in symmetric positions on the unit circle $|z|=1$ (Figure 4.1.2). For $|\lambda|>1$, on the other hand, some values of ξ lead to right-hand side values not contained in $[-2i,2i]$, and the roots z then move off the unit circle—one inside and one outside. The root z outside the circle amounts to an "amplification factor" greater than 1, and causes instability. The largest z occurs for $\sin\xi h=\pm 1$, which explains why the instability of Figure 4.1.1*(b)* had 4 points per wavelength.

Finally, §§4.5,4.6 discuss stability analysis via stability regions for problems in the form of the method of lines. Our leap frog example can be interpreted as the midpoint rule in time coupled with the centered difference operator δ_0 in space. Figure 4.1.3*(a)* shows the stability region in the a-plane (not the $\bar{k}=ka$ plane) for the midpoint rule, repeated from Figure 1.7.1: it is the complex open interval $(-i/k,i/k)$ on the imaginary axis. Figure 4.1.3*(b)* shows the eigenvalues* of δ_0—its eigenfunctions are the functions $v_j=e^{i\xi x_j}$, $\xi\in\mathbb{R}$, with corresponding eigenvalues

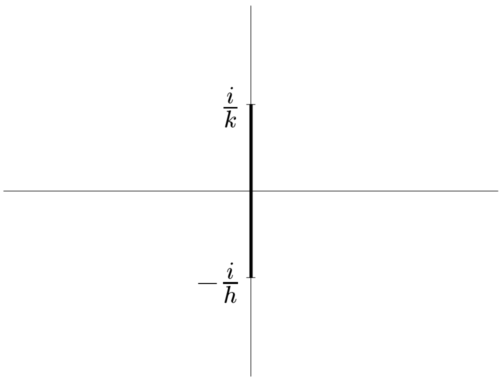
$$\frac{1}{2h}(e^{i\xi h}-e^{-i\xi h}) = \frac{i}{h}\sin\xi h. \tag{4.1.7}$$

Thus the eigenvalues cover the complex interval $[-i/h,i/h]$. For absolute stability of the system of ODEs that arise in the method of lines, these eigenvalues must lie in the stability region, leading once again to the condition $h^{-1}<k^{-1}$, that is, $k<h$. Section 4.6 shows that this condition relates to true stability as well as to absolute stability.

*We are being careless with the term "eigenvalue." Since the functions $v_j=e^{i\xi x_j}$ do not belong to ℓ_h^2, they are not true eigenfunctions, and the proper term for the quantities (4.1.7) is "spectral values." However, this technicality is of little importance for our purposes, and goes away when one considers problems on a bounded interval or switches to the ℓ_h^∞ norm, so we shall ignore it and speak of "eigenvalues" anyway.



(a) Stability region for midpoint rule



(b) Eigenvalues of δ_0

Figure 4.1.3. Absolute stability analysis of Example 4.1.1.

EXERCISES

▷ *4.1.1.* Determine the unstable mode that dominates the behavior of Figure 4.1.1(b). In particular, what is the factor by which the unstable solution is amplified from one step to the next?

4.2. The Lax Equivalence Theorem

[This section is rather sketchy at the moment, omitting some essential points of rigor as well as explanation, especially in two areas: the application of the operator A on a dense subspace rather than the whole space $\mathcal{B}$, and the relationship between continuous and discrete norms. For a more precise discussion see Richtmyer and Morton.]

The essential idea of the Lax Equivalence Theorem is this: for consistent *linear* finite difference models, stability is a necessary and sufficient condition for convergence. This is an analog of the Dahlquist Equivalence Theorem for ordinary differential equations (Theorem 1.10), except that the latter is valid for nonlinear problems too.

Aside from the assumption of linearity, the formulation of the Lax Equivalence Theorem is very general. Let $\mathcal{B}$ be a Banach space (a complete normed vector space) with norm denoted by $\|\cdot\|$. In applications of interest here, each element of $\mathcal{B}$ will be a function of one or more space variables x. Let $A:\mathcal{B}\to\mathcal{B}$ be a linear operator on this space. Here, A will be a differential operator. We are given the **initial value problem**

$$u_t(t)=Au(t), \quad 0\le t\le T, \qquad u(0)=u_0, \tag{4.2.1}$$

where A is fixed but u_0 may range over all elements of $\mathcal{B}$. [Actually, A only has to be defined on a dense subset of $\mathcal{B}$.] This initial value problem is assumed to be **well-posed**, which means that a unique solution $u(t)$ exists for any initial data u_0 and $u(t)$ depends continuously upon the initial data.

EXAMPLE 4.2.1. Suppose we wish to solve $u_t=u_x$ for $x\in(-\infty,\infty)$, $t\ge 0$, with initial data $f(x)$, and we wish to look for solutions in the space L^2. In this case each $u(t)$ in (4.2.1) is a function of x, namely $u(x,t)$. (Technically we should not use the same symbol u in both places.) The Banach space $\mathcal{B}$ is L^2, A is the first-order differentiation operator ∂_x, and $u_0=f$.

EXAMPLE 4.2.2. Suppose we wish to solve $u_t=u_{xx}$ for $x\in[-1,1]$, $t\ge 0$, with initial data $f(x)$ and boundary conditions $u(-1,t)=u(1,t)=0$. Again, each $u(t)$ is a function of x. Now an appropriate Banach space might be $C_0[-1,1]$, the set of continuous functions of $x\in[-1,1]$ with value 0 at $x=\pm 1$, together with the supremum norm.

Abstractly, $u(t)$ is nothing more than an element in a Banach space $\mathcal{B}$, and this leaves room for applications to a wide variety of problems in differential equations. As in the examples above, $\mathcal{B}$ might be L^2 and A might be ∂_x or ∂_x^2, and homogeneous boundary conditions could be included by restricting $\mathcal{B}$ appropriately. More generally, $u(t)$ could be an vector-valued function of multiple space variables.

The next step is to define a general finite difference formula. In the abstract setting, this is a family of bounded linear operators

$$S_k:\mathcal{B}\to\mathcal{B}, \tag{4.2.2}$$

where the subscript k indicates that the coefficients of the finite difference formula depend on the time step. We advance from one step to the next by a single application of S_k:

$$v^{n+1} = S_k v^n, \qquad \text{hence} \qquad v^n = S_k^n v^0, \tag{4.2.3}$$

where S_k^n abbreviates $(S_k)^n$. (Watch out for the usual confusion of notation: the n in v^n is a superscript, while in S_k^n it is an exponent.) For simplicity, but no more essential reason, we are assuming that the problem (4.2.1) and hence S_k have no explicit dependence on t. But S_k does potentially depend on k, and this is an important point. On the other hand it does not explicitly depend on the space step h, for we adopt the following rule:

h is a fixed function $h(k)$ of k.

For example, we might have $h = k/\lambda$ (λ constant) or $h = \sqrt{k/\sigma}$ (σ constant). If there are several space dimensions, each may have its own function $h_j(k)$. More generally, what we really need is grid(k), not $h(k)$; there is no need at all for the grid to be regular in the space dimensions.

EXAMPLE 4.2.3. *Lower-order terms.* For the UW model of $u_t = u_x$, the discrete solution operator is defined by $S_k v_j^n = v_j^n + \lambda(v_{j+1}^n - v_j^n)$, and if λ is held constant as $k \to 0$, this formula happens to be independent of k. The natural extension of UW to $u_t = u_x + u$, on the other hand, is $S_k v_j^n = v_j^n + \lambda(v_{j+1}^n - v_j^n) + k v_j^n$, and here there is an explicit dependence on k. This kind of k-dependence appears whenever the operator A involves derivatives of different orders.

Implicit or multistep finite difference formulas are not excluded by this formulation. As explained in §3.5, an implicit formula may still define a bounded operator S_k on an appropriate space such as ℓ_h^2, and a multistep formula can be reduced to an equivalent one-step formula by the introduction of a vector $w^n = (v^n, \ldots, v^{n+1-s})$.

Let us now be a bit more systematic in summarizing how the setup for the Lax Equivalence Theorem does or does not handle the various complications that make real problems differ from $u_t = u_x$ and $u_t = u_{xx}$.

- *Nonlinearity*
  The restriction here is essential: the Lax-Richtmyer theory does not handle nonlinear problems. (However, see various more recent papers by Sanz-Serna and others.)

- *Multiple space dimensions*
- *Implicit finite difference formulas*
  Both of these are included as part of the standard formulation.

- *Time-varying coefficients*
  Initial-value problems with time-varying coefficients are not covered in the description given here or in Richtmyer and Morton, but this restriction is not essential. The theory can be straightforwardly extended to such problems.

- *Boundary conditions*
- *Space-varying coefficients*
- *Lower-order terms*
  All of these are included as part of the standard formulation, and they have in common the property that they all lead to finite-difference approximations S_k that depend on k, as illustrated in Example 4.2.3 above.

- *Systems of equations*
- *Higher-order initial-value problems*
- *Multistep finite difference formulas*
  These are covered by the theory, if we make use of the usual device of reducing a one-step vector finite difference approximation to a first-order initial-value problem.

As in Chapter 1, we begin a statement of the Lax-Richtmyer theory by defining the order of accuracy and consistency of a finite difference formula.

> $\{S_k\}$ *has* **order of accuracy** p *if*
>
> $$\|u(t+k)-S_k u(t)\| = O(k^{p+1}) \qquad \text{as } k\to 0 \tag{4.2.4}$$
>
> *for any* $t\in[0,T]$*, where* $u(t)$ *is any sufficiently smooth solution to the initial-value problem (4.2.1). It is* **consistent** *if it has order of accuracy* $p>0$.

There are differences between this definition and the definition of order of accuracy for linear multistep formulas in §1.3. Here, the finite difference formula is applied not to an arbitrary function u, but to a solution of the initial value problem. In practice, however, one still calculates order of accuracy by substituting formal Taylor expansions and determining up to what order the terms cancel (Exercise 4.2.1).

Another difference is that in the case of linear multistep formulas for ordinary differential equations, the order of accuracy was always an integer, and so consistency amounted to $p \ge 1$. Here, non-integral orders of accuracy are possible, although they are uncommon in practice.

EXAMPLE 4.2.4. *Non-integral orders of accuracy.* The finite difference approximation to $u_t = u_x$,

$$S_k v_j^n = v_j^n + \lambda(v_{j+1}^n - v_j^n) + k^{p+1}, \qquad \lambda = \text{constant} \tag{4.2.5}$$

is a (contrived) example with order of accuracy p, if p is any constant in the range $[0,1]$. A slightly less contrived example with order of accuracy p is

$$S_k v_j^n = v_j^n + \frac{k}{2h}(v_{j+1}^n - v_{j-1}^n), \quad \text{with} \quad h = k^{p/2} \tag{4.2.6}$$

for any $p \in [0,2]$.

As with ordinary differential equations, a finite difference formula for a partial differential equation is defined to be convergent if and only if it converges to the correct solution as $k \to 0$ for arbitrary initial data:

> $\{S_k\}$ *is* **convergent** *if*
>
> $$\lim_{\substack{k\to 0\\ nk=t}} \|S_k^n u(0) - u(t)\| = 0 \tag{4.2.7}$$
>
> *for any $t \in [0,T]$, where $u(t)$ is the solution to the initial-value problem (4.2.1) for any initial data u_0.*

Note that there is a big change in this definition from the definition of convergence for linear multistep formulas in §1.5. There, a fixed formula had to apply successfully to any differential equation and initial data. Here, the differential equation is fixed and only the initial data vary.

The definition of stability is slightly changed from the ordinary differential equation case, because of the dependence on k:

> $\{S_k\}$ *is* **stable** *if for some $C > 0$,*
>
> $$\|S_k^n\| \le C \tag{4.2.8}$$
>
> *for all n and k such that $0 \le nk \le T$.*

This bound on the operator norms $\|S_k^n\|$ is equivalent to

$$\|v^n\| = \|S_k^n v^0\| \le C\|v^0\|$$

for all $v^0 \in \mathcal{B}$ and $0 \le nk \le T$.

Here is the Lax Equivalence Theorem (compare Theorem 1.10):

LAX EQUIVALENCE THEOREM

Theorem 4.1. *Let $\{S_k\}$ be a consistent approximation to a well-posed linear initial-value problem (4.2.1). Then $\{S_k\}$ is convergent if and only if it is stable.*

Proof. [Not yet written] ▮

The following analog to Theorem 1.11 establishes that stable discrete formulas have the expected rate of convergence.

GLOBAL ACCURACY

Theorem 4.2. *Let a convergent approximation method of order of accuracy p be applied to a well-posed initial-value problem (4.2.1) [with some additional smoothness assumptions...]. Then the computed solution satisfies*

$$\|v(t) - u(t)\| = O(k^p) \qquad \text{as } k \to 0 \tag{4.2.9}$$

uniformly for all $t \in [0,T]$.

Proof. [Not yet written] ▮

A number of remarks should be made about the developments of this section.

• The definitions of convergence and of consistency make reference to the initial-value problem (4.2.1), but the definition of stability does not. One can ask whether a finite difference formula is stable or unstable without having any knowledge of what partial differential equation, if any, it approximates.

• No assumption has been made that the initial-value problem is hyperbolic or parabolic, so long as it is well-posed. Indeed, as far as the theory is concerned, the initial-value problem may not involve a differential operator or an x variable at all. (There are some useful applications of the Lax Equivalence Theorem of this kind, one of which involves the so-called Trotter product formula of mathematical physics.)

• As in §1.4, we are dealing with the limit in which t is fixed and $k \to 0$. The situation for $t \to \infty$ with k fixed will be discussed in §4.5.

• The definition of the finite difference formula (4.2.2) depends upon the mesh function $h = h(k)$. Consequently, whether the formula is stable or not may depend on $h(k)$ too. Examples are given in §4.4.

• As in §1.4, it is quite possible for an unstable finite difference formula to give convergent results for some initial data. For example, this might happen if the initial data were particularly smooth. But the definition of convergence requires good results for all possible initial data.

• Relatedly, the theory assumes exact arithmetic: discretization errors are included, but not rounding errors. The justification for this omission is that the same phenomena of stability and instability govern the propagation of both kinds of errors, so that in most situations a prediction based on discretization errors alone will be realistic. On the other hand, the fact that rounding errors occur in practice is one motivation for requiring convergence for all initial data. The initial data prescribed mathematically for a particular computation might be smooth, but the rounding errors superimposed on them will not be.

• Consistency, convergence, and stability are all defined in terms of a norm $\|\cdot\|$, and it must be the same norm in each case.

• It is quite possible for a finite-difference model to be stable in one norm and unstable in others; see §5.5. This may sound like a defect in the theory, but in such cases the instability is usually so weak that the behavior is more or less stable in practice.

We close this section by mentioning a general theorem that follows from the definition of stability:

PERTURBATIONS OF A STABLE FAMILY
Theorem 4.3. *Let $\{S_k\}$ be a stable family of operators, and let T_k be a family of operators satisfying $\|T_k\| = O(k)$ as $k \to 0$. Then $\{S_k + T_k\}$ is also a stable family.*

Proof. [not yet written; see Richtmyer & Morton, §3.9.]

Theorem 4.3 has an important corollary that generalizes Example 4.2.3:

LOWER ORDER TERMS
Theorem 4.4. *Let $\{S_k\}$ be a consistent finite difference approximation to a well-posed linear initial-value problem (4.2.1) in which A is a differential operator acting on one or more space variables. The stability of $\{S_k\}$ is determined only by the terms that relate to spatial derivatives.*

Proof. (sketch) If the finite difference approximation is consistent, then lower-order terms modify the finite difference formula only by terms of order $O(k)$, so by Theorem 4.3 they do not affect stability. ∎

EXAMPLE 4.2.4. If $\{S_k\}$ is a stable finite difference approximation of $u_t = u_x$ on $(-\infty,\infty)$, then the approximation remains stable if additional terms are added so that it becomes consistent with $u_t = u_x + f(x,u)$, for any function $f(x,u)$. The same is true for the equation $u_t = u_{xx}$. A consistent change from $u_t = u_{xx}$ to $u_t = u_{xx} + f(x,u,u_x)$, on the other hand, might destroy stability.

References:

- Chapters 4 and 5 of R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems,* Wiley, 1967.

- P. D. Lax and R. D. Richtmyer, "Survey of the stability of linear finite difference equations," *Comm. Pure Appl. Math. 9* (1956), 267–293.

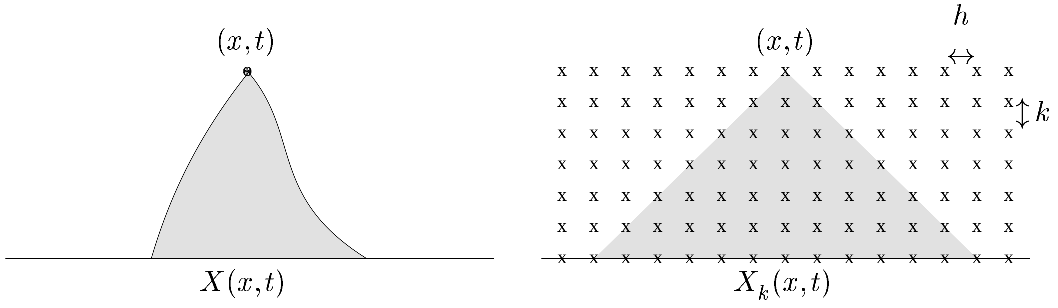
EXERCISES

▷ *4.2.1. Order of accuracy of the Lax-Wendroff formula.* Consider the Lax-Wendroff model of $u_t = u_x$ with $k/h = \lambda =$ constant. In analogy to the developments of §1.3, insert a formal power series for $u(x_{j+\Delta j}, t_{n+\Delta n})$ to obtain a formula for the leading-order nonzero term of the discretization error. Verify that the order of accuracy is 2.

4.3. The CFL condition

In 1928 Richard Courant, Kurt Friedrichs, and Hans Lewy, of the University of Göttingen in Germany, published a famous paper entitled "On the partial difference equations of mathematical physics."* This paper was written long before the invention of digital computers, and its purpose in investigating finite difference approximations was to apply them to prove existence of solutions to partial differential equations. But the "CFL" paper laid the theoretical foundations for practical finite difference computations, too, and in particular, it identified a fundamental necessary condition for convergence of any numerical scheme that has subsequently come to be known as the **CFL condition.**

What Courant, Friedrichs, and Lewy pointed out was that a great deal can be learned by considering the **domains of dependence** of a partial differential equation and of its discrete approximation. As suggested in Figure 4.3.1a, consider an initial-value problem for a partial differential equation, and let (x,t) be some point with $t>0$. (Despite the picture, the spatial grid need not be regular, or one-dimensional.) The **mathematical domain of dependence** of $u(x,t)$, denoted by $X(x,t)$, is the set of all points in space where the initial data at $t=0$ may have some effect on the solution $u(x,t)$.†



(a) (b)

Figure 4.3.1. Mathematical and numerical domains of dependence.

For example, for $u_t = u_{xx}$ or any other parabolic partial differential equation in one space dimension, $X(x,t)$ will be the entire real axis, because under a parabolic equation, information travels infinitely fast. The magnitude of the influence of far-away data may decay exponentially with distance, but in the definition of the domain of dependence it matters only whether this influence is zero or nonzero. The same conclusion holds for the Schrödinger equation $u_t = iu_{xx}$.

On the other hand for $u_t = u_x$, $u_t = u_x + u$, $u_{tt} = u_{xx}$, or any other hyperbolic partial differential equation or system of equations, including nonlinear equations such as $u_t = (\frac{1}{2}u^2)_x$,

*In German: "Über die partiellen Differenzengleichungen der mathematischen Physik," *Math. Ann. 100* (1928), 32–74. An English translation appeared much later in *IBM Journal 11* (1967), 215–234.

†More precisely, for a problem in d space dimensions, $X(x,t)$ is the intersection of all closed sets $E \subseteq \mathbb{R}^d$ with the property that the data on $\mathbb{R}^d \backslash E$ has no effect on $u(x,t)$.

$X(x,t)$ is finite for each x and t. The reason is that in hyperbolic problems, information travels at a finite speed. Figure 4.3.1a suggests a problem of this sort, since the domain of dependence shown there is finite. For the model problem $u_t = u_x$, $X(x,t)$ is the single point $\{x+t\}$, but the more typical situation for hyperbolic problems is that the domain of dependence covers a bounded range of values of x. In one-dimensional problems the curves that bound this range, as in Figure 4.3.1a, are the **characteristic curves** for the partial differential equation, and these are straight lines in simple examples but usually more general curves in problems containing variable coefficients or nonlinearity.

A numerical approximation also has a domain of dependence, and this is suggested in Figure 4.3.1b. With an implicit finite difference formula, each value v_j^n depends on all the values at one or more earlier steps, and the domain of dependence is unbounded. On the other hand with an explicit formula, v_j^n depends on only a finite range of values at previous steps. For any fixed k and h, the domain of dependence will then fan out in a triangle that goes backwards in time. The triangle will be symmetrical for a three-point formula like Lax-Wendroff or leap frog, symmetrical and twice as wide for a five-point formula like fourth-order leap frog, asymmetrical for a one-sided formula like upwind, and so on.

The numerical domain of dependence for a fixed value k, denoted by $X_k(x,t)$, is defined to be the set of points x_j whose initial data v_j^0 enter into the computation of $v(x,t)$. For each time step k, this set is discrete, but what really matters is the limit $k \to 0$. We define the limiting **numerical domain of dependence**, $X_0(x,t)$, to be the set of all limit points of the sets $X_k(x,t)$ as $k \to 0$.* This will be a closed subset of the spatial domain—typically an interval if there is one space variable, or a parallelepiped if there are several.

From the point of view of domain of dependence, there are three general classes of discrete approximations. In the case of an implicit finite difference model, $X_k(x,t)$ is unbounded for each k, and therefore, provided only that the spatial grid becomes finer everywhere as $k \to 0$, $X_0(x,t)$ will be the entire spatial domain. (Spectral methods are also essentially of this type: although the time stepping may be explicit, their stencils cover the entire spatial domain, and so $X_k(x,t)$ is unbounded for each k.)

At the other extreme is the case of an explicit finite difference formula with a spatial grid that scales in proportion to k, so that $X_k(x,t)$ and $X_0(x,t)$ are bounded sets for each x and t. In the particular case of a regular grid in space with mesh ratio $k/h = \lambda =$ constant, their size is proportional to the width of the stencil and inversely proportional to λ. The latter statement should be obvious from Figure 4.3.1b—if λ is cut in half, it will take twice as many time steps to get to (x,t) for a fixed h, and so the width of the triangle will double.

Between these two situations lies the case of an explicit finite difference formula whose spatial grid is refined more slowly than the time step as $k \to 0$—for example, a regular grid with $k = o(h)$. Here, $X_k(x,t)$ is bounded for each k, but $X_0(x,t)$ is unbounded. This in-between situation shows that even an explicit formula can have an unbounded domain of dependence in the limiting sense.

The observation made by Courant, Friedrichs, and Lewy is beautifully simple: a numerical approximation cannot converge for arbitrary initial data unless it takes all of the necessary data into account. "Taking the data into account" means the following:

*More precisely, for a problem in d space dimensions, $X_0(x,t)$ is the set of all points $s \in \mathbb{R}^d$ every open neighborhood of which contains a point of $X_k(x,t)$ for all sufficiently small k.

THE CFL CONDITION. For each (x,t), the mathematical domain of dependence is contained in the numerical domain of dependence:

$$X(x,t) \subseteq X_0(x,t).$$

Here is their conclusion:

CFL THEOREM

Theorem 4.5. *The CFL condition is a necessary condition for the convergence of a numerical approximation of a partial differential equation, linear or nonlinear.*

The justification of Theorem 4.5 is so obvious that we shall not attempt to state or prove it more formally. But a few words are in order to clarify what the terms mean. First, the theorem is certainly valid for the particular case of the linear initial-value problem (4.2.1) with the definition of convergence (4.2.7) provided in the last section. In particular, unlike the von Neumann condition of the next section, it holds for any norm $\|\cdot\|$ and any partial differential equation, including problems with boundary conditions, variable coefficients, or nonlinearity. But one thing that cannot be changed is that Theorem 4.5 must always be interpreted in terms of a definition of convergence that involves arbitrary initial data. The reason is that for special initial data, a numerical method might converge even though it had a seemingly inadequate domain of dependence. The classic example of this is the situation in which the initial data are so smooth that they form an analytic function. Since an analytic function is determined globally by its behavior near any point, a finite difference model might sample "too little" initial data in an analytic case and still converge to the correct solution—at least in the absence of rounding errors.

A priori, the CFL condition is a necessary condition for *convergence.* But for linear problems, the Lax Equivalence Theorem asserts that convergence is equivalent to stability. From Theorems 4.1 and 4.5 we therefore obtain:

CFL CONDITION AND STABILITY

Theorem 4.6. *Let $\{S_k\}$ be a consistent approximation to a well-posed linear initial-value problem (4.2.1). Then the CFL condition is a necessary condition for the stability of $\{S_k\}$.*

Unlike Theorem 4.5, Theorem 4.6 is valid only if its meaning is restricted to the linear formulations of the last section. The problem of stability of finite difference models had not yet been identified when the CFL paper appeared in 1928, but Theorem 4.6 is the form in which the CFL result is now generally remembered. The reason is that the connection between convergence and stability is so universally recognized now that one habitually thinks of stability as the essential matter to be worried about.

The reader may have noticed a rather strange feature of Theorem 4.6. In the last section it was emphasized that the stability of $\{S_k\}$ has nothing to do with what initial-value problem, if any, it approximates. Yet Theorem 4.6 states a stability criterion based on an initial-value problem. This is not a logical inconsistency, for nowhere has it been claimed that the theorem is applicable to every finite difference model $\{S_k\}$; an initial-value problem is brought into the act only when a consistency condition happens to be satisfied. For more on this point, see Exercise 4.4.4.

Before giving examples, we shall state a fundamental consequence of Theorem 4.6:

EXPLICIT MODELS OF PARABOLIC PROBLEMS
Theorem 4.7. *If an explicit finite difference approximation of a parabolic initial-value problem is convergent, then the time and space steps must satisfy* $k = o(h)$ *as* $k \to 0$.

Proof. This assertion follows from Theorem 4.6 and the fact that an explicit finite difference model with $k \neq o(h)$ has a bounded numerical domain of dependence $X_0(x,t)$ for each (x,t), whereas the mathematical domain of dependence $X(x,t)$ is unbounded. ∎

The impact of Theorem 4.7 is far-reaching: parabolic problems must always be solved by implicit formulas, or by explicit formulas with small step sizes. This would make them generally more difficult to treat than hyperbolic problems, were it not that hyperbolic problems tend to feature shock waves and other strongly nonlinear phenomena—a different source of difficulty that evens the score somewhat.

In computations involving more complicated equations with both convective and diffusive terms, such as the Navier-Stokes equations of fluid dynamics, the considerations of Theorem 4.7 often lead to numerical methods in which the time iteration is based on a splitting into an explicit substep for the convective terms and an implicit substep for the diffusive terms. See any book on computational fluid dynamics.

EXAMPLE 4.3.1. *Approximations of* $u_t = u_x$. For the equation $u_t = u_x$, all information propagates leftward at speed exactly 1, and so the mathematical domain of dependence for each (x,t) is the single point $X(x,t) = \{x+t\}$. Figure 4.3.2 suggests how this relates to the domain of dependence for various finite difference formulas. As always, the CFL condition is necessary but not sufficient for stability: it can prove a method unstable, but not stable. For the finite difference formulas of Table 3.2.1 with $k/h = \lambda =$ constant, we reach the following conclusions:

$$
\begin{aligned}
\text{LF4:}&\quad \text{unstable for } \lambda > 2;\\
\text{UW, LF, LW, EU}_x\text{, LXF:}&\quad \text{unstable for } \lambda > 1;\\
\text{“Downwind” formula:}&\quad \text{unstable for all } \lambda;\\
\text{BE}_x\text{, CN}_x\text{, BOX}_x\text{:}&\quad \text{no restriction on } \lambda.
\end{aligned}
$$

We shall see in the next section that these conclusions are sharp except in the cases of LF4 and EU_x (and LF, marginally, whose precise condition for instability is $\lambda \geq 1$ rather than $\lambda > 1$.)

EXAMPLE 4.3.2. *Approximations of* $u_t = u_{xx}$. Since $u_t = u_{xx}$ is parabolic, Theorem 4.7 asserts that no consistent explicit finite difference approximation can be stable unless $k = o(h)$ as $k \to 0$. Thus for the finite difference formulas of Table 3.3.2, it implies

$$
\begin{aligned}
\text{EU}_{xx}\text{, LF}_{xx}\text{:}&\quad \text{unstable unless } k = o(h),\\
\text{BE}_{xx}\text{, CN, BOX}_{xx}\text{, CN4:}&\quad \text{no restriction on } h(k).
\end{aligned}
$$

(a) LW or EU, $\lambda < 1$ *(b)* LW or EU, $\lambda > 1$ *(c)* UW, $\lambda < 1$

Figure 4.3.2. The CFL condition and $u_t = u_x$. The dashed line represents the characteristic of the PDE and the solid line represents the stencil of the finite difference formula.

The next section will show that these conclusions for LF_{xx} and EU_{xx} are not sharp. In fact, for example, EU_{xx} is stable only if $k \le \frac{1}{2}h^2$.

The virtue of the CFL condition is that it is extremely easy to apply. Its weakness is that it is necessary but not sufficient for convergence. As a practical matter, the CFL condition often suggests the correct limits on stability, but not always, and therefore it must be supplemented by more careful analysis.

EXERCISES

▷ *4.3.1. Multidimensional wave equation.* Consider the second-order wave equation in d space dimensions:

$$u_{tt} = u_{x_1x_1} + \cdots + u_{x_dx_d}.$$

(a) Write down the d-dimensional analog of the simple second-order finite difference approximation

$$v_j^{n+1} - 2v_j^n + v_j^{n-1} = \lambda^2 (v_{j+1}^n - 2v_j^n + v_{j-1}^n)$$

for a regular grid with space step h in all directions and $\lambda = k/h =$ constant.

(b) What does the CFL condition tell you about values of λ for which this model must be unstable? (*Hint:* if you are in doubt about the speed of propagation of energy under the multidimensional wave equation, consider the fact that the equation is isotropic—i.e., energy propagates in the same manner regardless of direction.) (*Another hint:* be careful!)

4.4. The von Neumann condition for scalar one-step formulas

Von Neumann analysis is the analysis of stability by Fourier methods. In principle this restricts its applicability to a narrow range of linear, constant-coefficient finite difference formulas on regular grids, with errors measured in the ℓ_h^2 norm. In practice, the von Neumann approach has something to say about almost every finite difference model. It is the mainstay of practical stability analysis.

The essential idea is to combine the ideas of §4.2, specialized to the ℓ_h^2 norm, with those of §§3.5,3.6. For one-step scalar models the ensuing results give a complete characterization of stability in terms of the "amplification factors" introduction in §3.5, which for stability must lie within a distance $O(k)$ of the unit disk as $k\to 0$. For multistep or vector problems one works with the associated "amplification matrices" introduced in §3.6, and a complete analysis of stability requires looking not only at the spectra of these matrices, which amount to amplification factors, but also at their resolvents or pseudospectra. We shall treat the one-step scalar case in this section and the general case in the next two sections.

We begin with two easy lemmas. The first is a generalization of the well-known inequality $(1+\epsilon)^{1/\epsilon} < e$:

Lemma 4.4.1. *For any real numbers $a \ge -1$ and $b \ge 0$,*

$$(1+a)^b \le e^{ab}. \tag{4.4.1}$$

Proof. Both sides are nonnegative, so taking the bth root shows that it is enough to prove $1+a \le e^a$. This inequality is trivial (just draw a graph!). ∎

The second lemma deals with arbitrary sets of numbers $s_k \ge 0$, which in our applications will be norms such as $\|S_k\|$:

Lemma 4.4.2. *Let $\{s_k\}$ be a set of nonnegative numbers indexed by $k>0$, and let $T>0$ be fixed. Then $(s_k)^n \le C_1$ for some $C_1 \ge 0$, uniformly for all k and n with $kn \le T$, if and only if $s_k \le 1+O(k)$ as $k \to 0$.*

Proof. If $s_k \le 1+C_2 k$ for each k, then by Lemma 4.4.1, $(s_k)^n \le (1+C_2k)^n \le e^{C_2kn} \le e^{C_2T}$. Conversely, if $(s_k)^n \le C_1$ for all $kn \le T$, then in particular, for each k, $(s_k)^n \le C_1$ for some value of n with $nk > T/2$. This implies $s_k \le C_1^{1/n} = (C_1^{2/T})^{T/2n} \le (C_1^{2/T})^k \le 1+O(k)$. ∎

Now we are ready for von Neumann analysis. Suppose that (4.2.1) is a well-posed linear initial-value problem in which $u(t)$ is a scalar function of x and $\mathcal{B}$ is the space ℓ_h^2, with $\|\cdot\|$ denoting the ℓ_h^2-norm. Suppose also that for each $k>0$, the operator S_k of (4.2.2) denotes an explicit or implicit one-step finite difference formula (3.5.20) with coefficients $\{\alpha_\mu\}$ and $\{\beta_\mu\}$ that are constant except for the possible dependence on k. If S_k is implicit, it is assumed to satisfy the solvability condition (3.5.26), which ensures that it has a bounded amplification factor function $g_k(\xi) = \hat{a}_k(\xi)/\hat{b}_k(\xi)$.

By (4.2.8), the formula is stable if and only if

$$\|S_k^n\| \le C \qquad \text{for } 0 \le nk \le T$$

for some constant C. By Theorem 3.4, this is equivalent to the condition

$$(\|g_k\|_\infty)^n \le C \qquad \text{for } 0 \le nk \le T,$$

or equivalently,

$$|g_k(\xi)|^n \le C \qquad \text{for } 0 \le nk \le T, \tag{4.4.2}$$

where C is a constant independent of ξ. By Lemma 4.4.2, this is equivalent to

$$|g_k(\xi)| \le 1 + O(k) \tag{4.4.3}$$

as $k \to 0$, uniformly in ξ. What this means is that there exists a constant C' such that for all n and k with $0 \le nk \le T$, and all $\xi \in [-\pi/h, \pi/h]$,

$$|g_k(\xi)| \le 1 + C'k. \tag{4.4.4}$$

We have obtained the following complete characterization of stable finite difference formulas in our scalar ℓ_h^2 problem.

VON NEUMANN CONDITION
FOR SCALAR ONE-STEP FINITE DIFFERENCE FORMULAS

Theorem 4.8. *A linear, scalar, constant-coefficient one-step finite difference formula as described above is stable in ℓ_h^2 if and only if the amplification factors $g_k(\xi)$ satisfy*

$$|g_k(\xi)| \le 1 + O(k) \tag{4.4.5}$$

as $k \to 0$, uniformly for all $\xi \in [-\pi/h, \pi/h]$.

With Theorem 4.8 in hand, we are equipped to analyze the stability of many of the finite difference formulas of §3.2.

EXAMPLE 4.4.1. *Upwind formula for $u_t = u_x$.* In (3.5.5) we computed the amplification factor for the upwind formula as

$$g(\xi) = (1-\lambda) + \lambda e^{i\xi h}, \tag{4.4.6}$$

assuming that $\lambda = k/h$ is a constant. For any λ, this formula describes a circle in the complex plane, shown in Figure 4.4.1, as ξ ranges over $[-\pi/h, \pi/h]$. The circle will lie in the closed unit disk, as required by (4.4.5), if and only if $\lambda \le 1$, which is accordingly the stability condition for the upwind formula. This matches the restriction suggested by the CFL condition (Example 4.3.1).

$1-\lambda$ 1 $1-\lambda$ 1

(a) $\lambda < 1$ (stable) *(b)* $\lambda > 1$ (unstable)

Figure 4.4.1. Amplification factors for the upwind model of $u_t = u_x$ (solid curve). The shaded region is the unit disk.

EXAMPLE 4.4.2. *Crank-Nicolson formulas for $u_t = u_x$, $u_t = u_{xx}$, $u_t = iu_{xx}$.* In (3.5.35) we found the amplification factor for the Crank-Nicolson formula to be

$$g(\xi) = \frac{1 - 2\frac{k}{h^2}\sin^2\frac{\xi h}{2}}{1 + 2\frac{k}{h^2}\sin^2\frac{\xi h}{2}}. \tag{4.4.7}$$

Here $|g(\xi)| \le 1$ for all ξ, regardless of k and h. Therefore the Crank-Nicolson formula is stable as $k \to 0$, no matter how k and h are related. (It will be consistent, hence convergent, so long as $h(k) = o(1)$ as $k \to 0$.)

For the Crank-Nicolson model of $u_t = u_x$, the corresponding formula is

$$g(\xi) = \frac{1 + \frac{ik}{2h}\sin\xi h}{1 - \frac{ik}{2h}\sin\xi h}. \tag{4.4.8}$$

Now $|g(\xi)| = 1$ for all ξ, so the formula is again unconditionally stable. The same is true of the Crank-Nicolson model of $u_t = iu_{xx}$, whose amplification factor function is

$$g(\xi) = \frac{1 - 2i\frac{k}{h^2}\sin^2\frac{\xi h}{2}}{1 + 2i\frac{k}{h^2}\sin^2\frac{\xi h}{2}}. \tag{4.4.9}$$

EXAMPLE 4.4.3. *Euler formulas for $u_t = u_x$ and $u_t = u_{xx}$.* The amplification factor for the Euler model of $u_t = u_{xx}$ was given in (3.5.19) as

$$g(\xi) = 1 - 4\frac{k}{h^2}\sin^2\frac{\xi h}{2}. \tag{4.4.10}$$

As illustrated in Figure 4.4.2a, this expression describes the interval $[1 - 4k/h^2, 1]$ in the complex plane as ξ ranges over $[-\pi/h, \pi/h]$. If $\sigma = k/h^2$ is held constant as $k \to 0$, we conclude that the Euler formula is stable if and only if $\sigma \le \frac{1}{2}$. This is a tighter restriction than the one provided by the CFL condition, which requires only $k = o(h)$.

$1-4\sigma$ 1 1

 ik/h

(a) $u_t = u_{xx}$ (b) $u_t = u_x$

Figure 4.4.2. Amplification factors for the Euler models of $u_t = u_{xx}$ and $u_t = u_x$.

The amplification factor for the Euler model of $u_t = u_x$ is

$$g(\xi) = 1 + \frac{ik}{h}\sin\xi h, \tag{4.4.11}$$

which describes a line segment tangent to the unit circle in the complex plane (Figure 4.4.2b). Therefore the largest amplification factor is

$$\sqrt{1+\frac{k^2}{h^2}}.$$

If $\lambda = k/h$ is held fixed as $k \to 0$, then this finite difference formula is therefore unstable regardless of λ. On the other hand the square root will have the desired magnitude $1+O(k)$ if $k^2/h^2 = O(k)$, i.e. $k = O(h^2)$, and this is accordingly the stability condition for arbitrary mesh relationships $h = h(k)$. Thus in principle the Euler formula is usable for hyperbolic equations, but it is not used in practice since there are alternatives that permit larger time steps $k = O(h)$, as well as having higher accuracy.

EXAMPLE 4.4.4. *Lax-Wendroff formula for $u_t = u_x$.* The amplification factor for the Lax-Wendroff formula was given in (3.5.17) as

$$g(\xi) = 1 + i\lambda\sin\xi h - 2\lambda^2\sin^2\frac{\xi h}{2}, \tag{4.4.12}$$

if $\lambda = k/h$ is a constant. Therefore $|g(\xi)|^2$ is

$$|g(\xi)|^2 = (1 - 4\lambda^2\sin^2\frac{\xi h}{2} + 4\lambda^4\sin^4\frac{\xi h}{2}) + \lambda^2\sin^2\xi h.$$

Applying the identity $\sin^2\theta = 4\sin^2\frac{\theta}{2}\cos^2\frac{\theta}{2} = 4(\sin^2\frac{\theta}{2} - \sin^4\frac{\theta}{2})$ to the last term converts this expression to

$$\begin{aligned} |g(\xi)|^2 &= 1 - 4\lambda^2\sin^2\frac{\xi h}{2} + 4\lambda^4\sin^4\frac{\xi h}{2} + 4\lambda^2\sin^2\frac{\xi h}{2} - 4\lambda^2\sin^4\frac{\xi h}{2} \\ &= 1 + 4(\lambda^4 - \lambda^2)\sin^4\frac{\xi h}{2}. \end{aligned} \tag{4.4.13}$$

If λ is fixed, it follows that the Lax-Wendroff formula is stable provided that $\lambda^4 - \lambda^2 \in [-\frac{1}{2}, 0]$. This is true if and only if $\lambda \le 1$, which is accordingly the stability condition.

Tables 4.4.1 and 4.4.2 summarize the orders of accuracy and the stability limits for the finite difference formulas of Tables 3.2.1 and 3.2.2. The results listed for multistep formulas will be justified in §4.6.

Formula	order of accuracy	CFL stability restriction	Exact stability restriction
(EU_x = Euler)	1	$\lambda \le 1$	unstable
(BE_x = Backward Euler)	1	none	none
(CN_x = Crank-Nicolson)	2	none	none
LF = Leap frog	2	$\lambda \le 1$	$\lambda < 1$
BOX_x = Box	2	none	none
LF4 = Fourth-order Leap frog	2	$\lambda \le 2$	$\lambda < 0.728\ldots$*
LXF = Lax-Friedrichs	1	$\lambda \le 1$	$\lambda \le 1$
UW = Upwind	1	$\lambda \le 1$	$\lambda \le 1$
LW = Lax-Wendroff	2	$\lambda \le 1$	$\lambda \le 1$

Table 4.4.1. Orders of accuracy and stability limits for various finite difference approximations to the wave equation $u_t = u_x$, with $\lambda = k/h =$ constant (see Table 3.2.1).

Formula	order of accuracy	CFL stability restriction	Exact stability restriction
EU_{xx} = Euler	1†	none	$\sigma \le \frac{1}{2}$
BE_{xx} = Backward Euler	1	none	none
CN = Crank-Nicolson	2	none	none
(LF_{xx} = Leap frog)	2	none	unstable
BOX_{xx} = Box	2	none	none
CN4 = Fourth-order CN	2	none	none
DF = DuFort-Frankel	2**	none	none**

Table 4.4.2. Orders of accuracy and stability limits for various finite difference approximations to the heat equation $u_t = u_{xx}$, with $\sigma = k/h^2 =$ constant (see Table 3.2.2). (The orders of accuracy are with respect to h, not k as in (4.2.4).)

---

* See Exercise 4.5.3.
† See Exercise 4.4.5.
** See Exercise 4.5.4.

EXERCISES

▷ *4.4.1. Generalized Crank-Nicolson or "theta method."* Let the heat equation $u_t = u_{xx}$ be modeled by the formula

$$v_j^{n+1} = v_j^n + \frac{k(1-\theta)}{h^2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n) + \frac{k\theta}{h^2}(v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}) \qquad (4.4.19)$$

with $0 \le \theta \le 1$. For $\theta = 0,\ \frac{1}{2},\ 1$ this is Euler, Crank-Nicolson, or backward Euler formula, respectively.

(a) Determine the amplification factor function $g(\xi)$.

(b) Suppose $\sigma = k/h^2$ is held constant as $k \to 0$. For which σ and θ is (4.4.19) stable?

(c) Suppose $\lambda = k/h$ is held constant as $k \to 0$. For which λ and θ is (4.4.19) stable?

(d) Your boss asks you to solve a heat conduction problem involving a space interval of length 1 and a time interval of length 1. She wants an answer with errors on the order of some number $\delta \ll 1$. Roughly speaking (order of magnitude in δ), how many floating-point operations will you have to perform if you use $\theta = 0$, $\theta = \frac{1}{2}$, and $\theta = 1$?

▷ *4.4.2. The downwind formula.* The downwind approximation to $u_t = u_x$ is

$$v_j^{n+1} = \mathcal{S}_k v_j^n = v_j^n + \frac{k}{h}(v_j^n - v_{j-1}^n).$$

In this problem, do not assume that k/h is held constant as $k \to 0$; let $h(k)$ be a completely arbitrary function of k.

(a) For what functions $h(k)$, if any, is this formula stable? (Use Theorem 4.8 and be careful!)

(b) For what functions $h(k)$, if any, is it consistent?

(c) For what functions $h(k)$, if any, is it convergent? (Use the Lax Equivalence Theorem.)

(d) How does the result of *(c)* match the prediction by the CFL condition?

▷ *4.4.3. The CFL-stability link.* (This problem was originally worked out by G. Strang in the 1960's.) **Bernstein's inequality** asserts that if

$$g(\xi) = \sum_{\nu=-m}^{m} \alpha_\nu e^{i\nu\xi}$$

for some constants $\{\alpha_\nu\}$, then

$$\|g'\|_\infty \le m\|g\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the maximum over $[-\pi,\pi]$, and g' is the derivative $dg/d\xi$.

Derive from this the CFL condition: if an explicit finite difference formula

$$v_j^{n+1} = \sum_{\nu=-m}^{m} \alpha_\nu v_{j+\nu}^n$$

is stable as $h \to 0$ with $k/h = \lambda =$ constant, and consistent with $u_t = u_x$, then $\lambda \le m$. (*Hint:* what does consistency imply about the behavior of $g(\xi)$ near $\xi = 0$?)

▷ *4.4.4. A thought experiment.* Suppose the Lax-Wendroff model of $u_t = u_x$ is applied on an infinite grid with $h = 0.01$ to compute an approximate solution at $t = 1$. The initial data are

$$u_1(x) = \cos^2 \pi x \qquad \text{or} \qquad u_2(x) = \max\{0, 1 - |x|\},$$

and the time step is $k = \lambda h$ with

$$\lambda_a = \frac{4}{5} \qquad \text{or} \qquad \lambda_b = \frac{4}{3}.$$

Thus we have four problems: 1a, 1b, 2a, 2b. Let E_{1a}, E_{1b}, E_{2a}, E_{2b} denote the corresponding maximum (ℓ_h^∞) errors over the grid at $t = 1$.

One of the four numbers E_{1a}, E_{1b}, E_{2a}, E_{2b} depends significantly on the machine precision, ϵ, and thus deserves a star * in front of it (see §β); the other three do not. Which one? Explain carefully why each of them does or does not depend on ϵ. In the process, give order-of-magnitude predictions for the four numbers, such as $E \approx 10^{-2}$, $E \approx 10^{10}$, $E \approx {}^\star\epsilon 10^{10}$. Explain your reasoning!

If you wish, you are welcome to turn this thought experiment into a computer experiment.

▷ *4.4.5. The Euler formula for the heat equation.* Show that if $\sigma = 1/6$, the order of accuracy of the Euler formula for $u_t = u_{xx}$ increases from 1 to 2. (*Note:* Such bits of good fortune are not always easy to take advantage of in practice, since coefficients and hence effective mesh ratios may vary from point to point.)

▷ *4.4.6. Weak inequalities.* In Tables 4.4.1 and 4.4.2, the stability restrictions for one-step formulas all involve weak ("$\leq$") rather than strong ("$<$") inequalities. Prove that this is a general phenomenon: the stability condition for a formula of the type considered in Theorem 4.8 may be of the form $k \leq f(h)$ but never $k < f(h)$.

4.5. Resolvents, pseudospectra, and the Kreiss matrix theorem

Powers of matrices appear throughout numerical analysis. In this book they arise mainly from equation (4.2.3), which expresses the computed solution at step n for a numerical model of a linear problem in terms of the solution at step 0:

$$v^n = S_k^n v^0. \tag{4.5.1}$$

In this formula S_k represents a bounded operator on a Banach space $\mathcal{B}$, possibly different for each time step $k>0$. Thus we are dealing with a family of operators $\{S_k\}$ indexed by k. According to the theory presented in §4.2, a key question to ask about such a family of operators is whether it is stable, i.e., whether a bound

$$\|S_k^n\| \le C \tag{4.5.2}$$

holds for all n and k with $0 \le nk \le T$.

This question about operators may be reduced to a question about matrices in two distinct ways. The first is by Fourier or von Neumann analysis, applicable in cases where one is dealing with a regular grid, constant coefficients, and the 2-norm. For one-step scalar formulas Fourier analysis reduces S_k to a scalar, the amplification factor $g(\xi)$ treated in the last section. For vector formulas, the subject of this and the next section, we get the more interesting Fourier analogue

$$\widehat{v^n}(\xi) = G_k(\xi)^n\, \widehat{v^0}(\xi). \tag{4.5.3}$$

Here $G_k(\xi)^n$ is the nth power of the amplification matrix, defined in §3.6, and the equivalence of (4.5.1) and (4.5.3) implies that the norm of $G_k(\xi)$ determines the norm of S_k^n:

$$\|S_k^n\|_2 = \sup_{\xi\in[-\pi/h,\pi/h]} \|G_k(\xi)^n\|.$$

The question of stability becomes, are the norms $\|G_k(\xi)^n\|$ bounded by a constant for all $nk \le T$, uniformly with respect to ξ? Note that here the dimension N of $G_k(\xi)$ is fixed, independent of k and h.

For problems involving variable coefficients, variable grids, or translation-dependent discretizations such as Chebyshev spectral methods (Chapter 8), Fourier analysis cannot be applied. Here a second and more straightforward reduction to matrices becomes relevant. One can simply work with S_k itself as a matrix—that is, work in space itself, not Fourier space. If the grid is unbounded, then the dimension of S_k is infinite, but since the grids one computes with are usually bounded, S_k is usually finite in practice. The dimension increases to ∞, however, as $k \to 0$.

In summary, the stability analysis of numerical methods for partial differential equations leads naturally to questions of whether the powers of a family of matrices have uniformly bounded norms, or, as the problem is often put, whether a family of matrices is **power-bounded**. Depending on the circumstances, the matrices in the family may be of fixed dimension or varying dimensions.

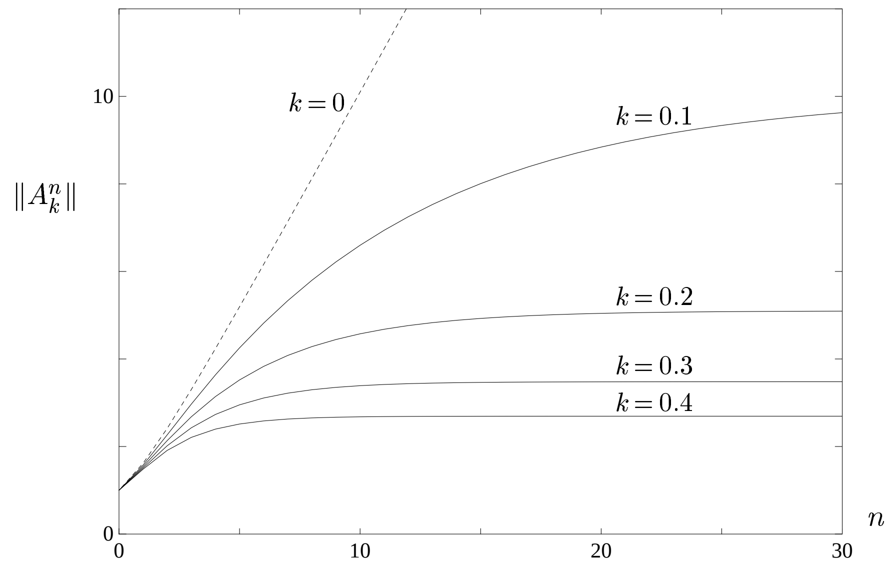


Figure 4.5.1. Norms of powers of the matrix A_k of (4.5.4) for various k. Each A_k for $k>0$ is individually power-bounded, but the family $\{A_k\}_{(k>0)}$ is not power-bounded, since the powers come arbitrarily close to those of the limiting defective matrix A_0 (dashed).

In §1.5 we have already presented an algebraic criterion for power-boundedness of a matrix A. According to the "alternative proof" of Theorem 1.6, the powers $\|A^n\|$ are bounded if and only if the eigenvalues of A lie in the closed unit disk and any eigenvalues on the unit circle are nondefective. For families of matrices, however, matters are not so simple. The eigenvalue condition is still necessary for power-boundedness, but it is not sufficient. To illustrate this, consider the two families of matrices

$$A_k = \begin{pmatrix} 1 & 1 \\ 0 & 1-k \end{pmatrix}, \qquad B_k = \begin{pmatrix} 0 & k^{-1} \\ 0 & 0 \end{pmatrix} \tag{4.5.4}$$

for $k>0$. For each $k>0$, A_k and B_k are individually power-bounded, but neither $\{A_k\}$ nor $\{B_k\}$ is power-bounded as a family. For $\{B_k\}$ the explanation is quite obvious: the upper-right entry diverges to ∞ as $k \to 0$, and thus even the first power $\|B_k\|$ is unbounded as $k \to 0$. For $\{A_k\}$ the explanation is more interesting. As $k \to 0$, these matrices come closer and closer to a defective matrix whose powers increase in norm without bound. Figure 4.5.1 illustrates this effect.

In the next section we shall see that an unstable family much like $\{A_k\}$ arises in the von Neumann analysis of the leap frog discretization of $u_t = u_x$ with mesh ratio $\lambda = 1$. In most applications, however, the structure of the matrices that arise is far from obvious by inspection and one needs a general tool for determining power-boundedness.

What is needed is to consider not just the spectra of the matrices but also their *pseudospectra*. The idea of pseudospectra is as follows. An eigenvalue of a matrix A is a number $z \in \mathbb{C}$ with the property that the **resolvent** matrix $(zI-A)^{-1}$ is undefined. By convention we may write $\|(zI-A)^{-1}\| = \infty$ in this case. A pseudo-eigenvalue of A is a number z where $\|(zI-A)^{-1}\|$, while not necessarily infinite, is very large. Here is the definition:

> *Given $\epsilon > 0$, the number $\lambda \in \mathbb{C}$ is an* **ϵ-pseudo-eigenvalue** *of A if either of the following equivalent conditions is satisfied:*
> *(i)* $\|(\lambda I - A)^{-1}\| \geq \epsilon^{-1}$;
> *(ii)* *λ is an eigenvalue of $A+E$ for some $E \in \mathbb{C}^{N\times N}$ with $\|E\| \leq \epsilon$.*
> *The* **ϵ-pseudospectrum** *of A, denoted by $\Lambda_\epsilon(A)$, is the set of all of its ϵ-pseudo-eigenvalues.*

Condition *(i)* expresses the idea of an ϵ-pseudo-eigenvalue just described. Condition *(ii)* is a mathematical equivalent with quite a different flavor: though an ϵ-pseudo-eigenvalue of A need not be near to any exact eigenvalue of A, it is an exact eigenvalue of some nearby matrix. The equivalence of conditions *(i)* and *(ii)* is easy to prove (Exercise 4.5.1).

If a matrix is **normal**, which means that its eigenvectors are orthogonal,* then the 2-norm of the resolvent is just $\|(zI-A)^{-1}\|_2 = 1/\mathrm{dist}(z,\Lambda(A))$, where $\mathrm{dist}(z,\Lambda(A))$ denotes the distance between the point z and the set $\Lambda(A)$ (Exercise 4.5.2). Therefore for each $\epsilon \geq 0$, $\Lambda_\epsilon(A)$ is equal to the union of the closed balls of radius ϵ about the eigenvalues of A; by condition *(ii)*, the eigenvalues of A are insensitive to perturbations. The interesting cases arise when A is non-normal, where the ϵ-pseudospectra may be much larger and the eigenvalues may be highly sensitive to perturbations. Figure 4.5.2 illustrates this comparison rather mildly by comparing the ϵ-pseudospectra of the two matrices

$$A = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.8 \end{pmatrix}, \qquad \tilde{A} = \begin{pmatrix} 0.9 & 1 \\ 0 & 0.8 \end{pmatrix}, \tag{4.5.5}$$

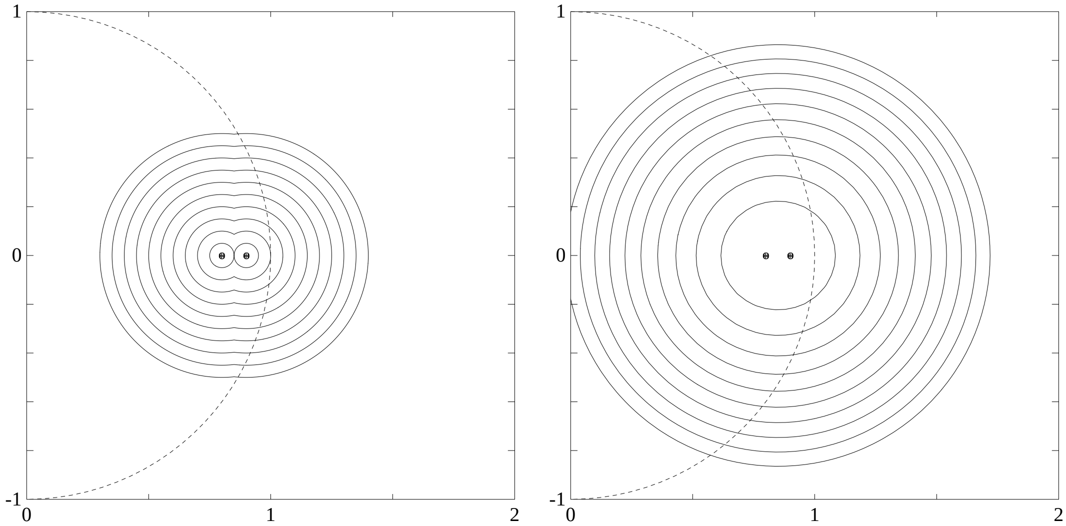
motivated by (4.5.4), for $\epsilon = 0.05, 0.10, 0.15, \ldots, 0.50$. These two matrices both have the spectrum $\{0.8, 0.9\}$, but the pseudospectra of $\tilde{A}$ are considerably larger than those of A.

Roughly speaking, matrices with larger pseudospectra tend to have larger powers, even if the eigenvalues are the same. Figure 4.5.3 illustrates this phenomenon for the case of the two matrices A and $\tilde{A}$ of (4.5.5). Asymptotically as $n \to \infty$, the powers decrease in both cases at the rate $(0.9)^n$ determined by the spectral radius, $\rho(A) = \rho(\tilde{A}) = 0.9$. The transient behavior is noticeably different, however, with the curve for $\|\tilde{A}^n\|$ showing a hump by a factor of about 2.8 centered around $n = 6$ before the eventual decay.

Factors of 2.8 are of little consequence for applications, but then, $\tilde{A}$ is not a very highly non-normal matrix. In more extreme examples the hump in a figure like Figure 4.5.3 may be arbitrarily high. To control it we must have a bound on the pseudospectra, and this is the subject of the Kreiss matrix theorem, first proved in 1962.

---

* More precisely, a matrix is normal if *there exists* a complete set of orthogonal eigenvectors.



(a) A (normal) *(b)* $\tilde{A}$ (non-normal)

Figure 4.5.2. Boundaries of ϵ-pseudospectra of the matrices A and $\tilde{A}$ of (4.5.6) for $\epsilon = 0.05, 0.10, \ldots, 0.50$. The dashed curve is the right half of the unit circle, whose location is important for the Kreiss matrix theorem. The solid dots are the eigenvalues.

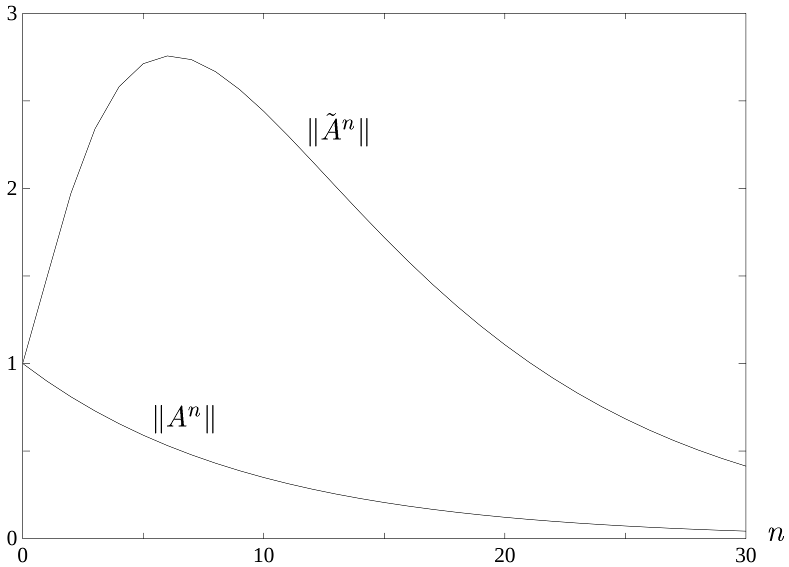


Figure 4.5.3. $\|A^n\|$ and $\|(\tilde{A})^n\|$ for the matrices A and $\tilde{A}$ of (4.5.5).

KREISS MATRIX THEOREM

Theorem 4.10. *Let A be a matrix of dimension N. If*

$$\|A^n\| \le C \qquad \forall n \ge 0 \tag{4.5.6}$$

for some constant C, then

$$|\lambda_\epsilon| \le 1 + C\epsilon \qquad \forall \lambda_\epsilon \in \Lambda_\epsilon(A), \quad \forall \epsilon \ge 0, \tag{4.5.7}$$

or equivalently,

$$\|(zI - A)^{-1}\| \le \frac{C}{|z|-1} \qquad \forall z \in \mathbb{C}, \quad |z| > 1. \tag{4.5.8}$$

Conversely, (4.5.7)-(4.5.8) imply

$$\|A^n\| \le eC \min\{N, n+1\} \qquad \forall n \ge 0. \tag{4.5.9}$$

Theorem 4.10 is stated for an individual matrix, but since the bounds asserted are explicit and quantitative, the same result applies to families of matrices A_ν satisfying a uniform bound $\|A_\nu^n\| \le C$, provided the dimensions N are all the same. If the dimensions N vary unboundedly, then in (4.5.9), only the factor $n+1$ remains meaningful. The inequality (4.5.8) is sometimes called the **Kreiss condition**, and the constant C there is the **Kreiss constant**.

Proof. The equivalence of (4.5.7) and (4.5.8) follows trivially from definition *(i)* of $\Lambda_\epsilon(A)$. The proof that (4.5.6) implies (4.5.8) is also straightforward if one considers the power series expansion

$$(zI - A)^{-1} = z^{-1}I + z^{-2}A + z^{-3}A^2 + \cdots.$$

Thus the real substance of the Kreiss Matrix theorem lies in the assertion that (4.5.8) implies (4.5.9). This is really two assertions: one involving a factor N, the other involving a factor $n+1$.

The starting point for both proofs is to write the matrix A^n as the Cauchy integral

$$A^n = \frac{1}{2\pi i}\int_\Gamma z^n (zI - A)^{-1}\, dz, \tag{4.5.10}$$

where Γ is any curve in the complex plane that encloses the eigenvalues of A. To prove that $\|A^n\|$ satisfies the bound (4.5.9), it is enough to show that $|v^*A^n u|$ satisfies the same bound for any vectors u and v with $\|u\| = \|v\| = 1$. Thus let u and v be arbitrary N-vectors of this kind. Then (4.5.10) implies

$$v^*A^n u = \frac{1}{2\pi i}\int_\Gamma z^n q(z)\, dz, \tag{4.5.11}$$

where $q(z) = v^*(zI - A)^{-1}u$. It can be shown that $q(z)$ is a rational function of order N, i.e., a quotient of two polynomials of degrees at most N, and by (4.5.8), it satisfies

$$|q(z)| \le \frac{C}{|z|-1}. \tag{4.5.12}$$

Let us take Γ to be the circle $\Gamma = \{z\in\mathbb{C}:\ |z| = 1+(n+1)^{-1}\}$, which certainly encloses the eigenvalues of A if (4.5.8) holds. On this circle (4.5.12) implies $|q(z)|\le C(n+1)$. Therefore (4.5.11) yields the bound

$$\begin{aligned}|v^*A^n u| &\le \frac{1}{2\pi}\int_\Gamma |z|^n C(n+1)\,|dz| \\ &\le \frac{1}{2\pi}(1+(n+1)^{-1})^n C(n+1)2\pi(1+(n+1)^{-1}) \\ &\le (1+(n+1)^{-1})^{n+1}C(n+1) \le e\,C(n+1).\end{aligned}$$

In the last of these inequalities we have used Lemma 4.4.1. This proves the part of (4.5.9) involving the factor $n+1$.

The other part of (4.5.9), involving the factor N, is more subtle. Integration by parts of (4.5.11) gives

$$v^*A^n u = \frac{-1}{2\pi i(n+1)}\int_\Gamma z^{n+1}q'(z)\,dz.$$

Using the same contour of integration Γ as before and again the estimate $|z^{n+1}|\le e$, we obtain

$$|v^*A^n u| \le \frac{e}{2\pi(n+1)}\int_\Gamma |q'(z)|\,|dz|.$$

The integral in this formula can be interpreted as the arc length of the image of the circle Γ under the rational function $q(z)$. Now according to a result known as Spijker's Lemma,* if $q(z)$ is a rational function of order N, the arc length of the image of a circle under $q(z)$ can be at most $2\pi N$ times the maximum modulus that $q(z)$ attains on that circle, which in this case is at most $C(n+1)$. Thus we get

$$|v^*A^n u| \le \frac{e}{2\pi(n+1)}\,(2\pi N)C(n+1) = e\,CN.$$

This completes the proof of the Kreiss matrix theorem. ∎

We close this section with a figure to further illustrate the idea of pseudospectra. Consider the 32×32 matrix of the form

$$A = \begin{pmatrix} 0 & 1 & 1 & & & \\ & 0 & 1 & 1 & & \\ & & 0 & 1 & 1 & \\ & & & 0 & 1 & 1 \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix}, \qquad (4.5.13)$$

*Spijker's Lemma was conjectured in 1983 by LeVeque and Trefethen (*BIT* '84) and proved up to a factor of 2. The sharp result was proved by Spijker in 1990 (*BIT* '92). Shortly thereafter it was pointed out by E. Wegert that the heart of the proof is a simple estimate in integral geometry that generalizes the famous "Buffon needle problem" of 1777. See Wegert and Trefethen, "From the Buffon needle problem to the Kreiss matrix theorem," *Amer. Math. Monthly 101* (1994), pp. 132–139.

whose only eigenvalue is $\lambda = 0$. The upper plot of Figure 4.5.4 depicts the boundaries of the ϵ-pseudospectra of this matrix for $\epsilon = 10^{-1}, 10^{-2}, \dots, 10^{-8}$. Even for $\epsilon = 10^{-8}$, the ϵ-pseudospectrum of this matrix is evidently quite a large set, covering a heart-shaped region of the complex plane that extends far from the spectrum $\{0\}$. Thus by condition *(ii)* of the definition of the ϵ-pseudospectrum, it is evident that the eigenvalues of this matrix are exceedingly sensitive to perturbations. The lower plot of Figure 4.5.4 illustrates this fact more directly. It shows a superposition of the eigenvalues of 100 matrices $A+E$, where each E is a random matrix of norm $\|E\|_2 = 10^{-3}$. (The elements of each E are taken as independent, normally distributed complex numbers of mean 0, and then the whole matrix is scaled to achieve this norm.) Thus there are 3200 dots in Figure 4.5.4b, which by definition must lie within in second-largest of the curves in Figure 4.5.4a.

For further illustrations of matrices with interesting pseudospectra, see L. N. Trefethen, "Pseudospectra of matrices," in D. F. Griffiths and G. A. Watson, eds., *Numerical Analysis 1991,* Longman, 1992, pp. 234–266.

EXERCISES

▷ *4.5.1. Equivalent definitions of the pseudospectrum.*

(a) Prove that conditions *(i)* and *(ii)* on p. 175 are equivalent.

(b) Prove that another equivalent condition is
(iii) $\exists u \in \mathbb{C}^n$, $\|u\| = 1$, such that $\|(A-\lambda)u\| \le \epsilon$.
Such a vector u is called an ϵ-*pseudo-eigenvector* of A.

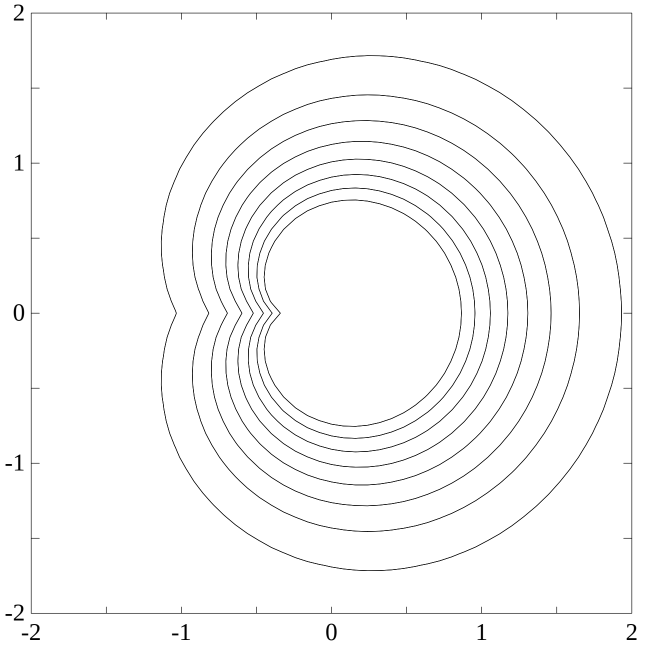
(c) Prove that if $\|\cdot\| = \|\cdot\|_2$, then a further equivalent condition is
(iv) $\sigma_N(\lambda I - A) \le \epsilon$,
where $\sigma_N(\lambda I - A)$ denotes the smallest singular value of $\lambda I - A$.

▷ *4.5.2.* Prove that if A is a normal matrix, then $\|(zI-A)^{-1}\|_2 = 1/\mathrm{dist}(z, \Lambda(A))$ for all $z \in \mathbb{C}$. (*Hint:* if A is normal, then it can be unitarily diagonalized.)

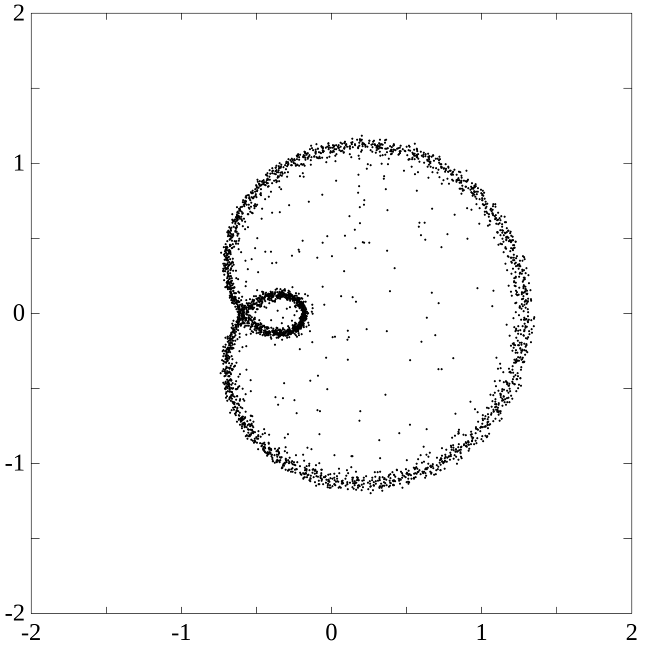
▶ *4.5.3.*

(a) Making use of Figure 4.5.4a, a ruler, and the Kreiss matrix theorem, derive lower and upper bounds as sharp as you can manage for the quantity $\sup_{n\ge 0} \|A^n\|_2$ for the matrix A of (4.5.13).

(b) Find the actual actual number with Matlab. How does it compare with your bounds?



(*a*) Boundaries of $\Lambda_\epsilon(A)$ for $\epsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-8}$.



(*b*) Eigenvalues of 100 randomly perturbed matrices $A+E$, $\|E\|_2 = 10^{-3}$.

Figure 4.5.4. Pseudospectra of the 32×32 matrix A of (4.5.13).

4.6. The von Neumann condition for vector or multistep formulas

Just as §3.6 followed §3.5, Fourier analysis applies to vector or multistep finite difference formulas as well as to the scalar one-step case. The determination of stability becomes more complicated, however, because one must estimate norms of powers of matrices.

Consider a linear, constant-coefficient finite-difference formula on a regular grid, where the dependent variable is an N-vector. By introducing new variables as necessary, we may assume that the formula is written in one-step form $v^{n+1} = S_k v^n$. It may be explicit or implicit, provided that in the implicit case, the solvability condition (3.6.5) is satisfied.

If $\|\cdot\|$ is the 2-norm, then the condition (4.2.8) for stability is equivalent to the condition

$$\|G_k(\xi)^n\|_2 \le C \tag{4.6.1}$$

for all $\xi \in [-\pi/h, \pi/h]$ and n,k with $0 \le nk \le T$. Here $G_k(\xi)$ denotes the amplification matrix, an $N \times N$ function of ξ, as described in §3.6. Stability is thus a question of power-boundedness of a family of $N \times N$ matrices, a family indexed by the two parameters ξ and k. This is just the question that was addressed in the last section.

The simplest estimates of the powers $\|G_k(\xi)^n\|$ are based on the norm $\|G_k(\xi)\|$ or the spectral radius $\rho(G_k(\xi))$, that is, the largest of the moduli of the eigenvalues of $G_k(\xi)$. These two quantities provide a lower and an upper bound on (4.6.1) according to the easily proved inequalities

$$\rho(G_k(\xi))^n \le \|G_k(\xi)^n\| \le \|G_k(\xi)\|^n. \tag{4.6.2}$$

Combining (4.6.1) and (4.6.2) yields:

VON NEUMANN CONDITION
FOR VECTOR FINITE DIFFERENCE FORMULAS

Theorem 4.10. *Let $\{S_k\}$ be a linear, constant-coefficient finite difference formula as described above. Then*

$$(a)\ \rho(G_k(\xi)) \le 1 + O(k) \text{ is necessary for stability, and} \tag{4.6.3}$$

$$(b)\ \|G_k(\xi)\| \le 1 + O(k) \text{ is sufficient for stability.} \tag{4.6.4}$$

Both (a) and (b) are assumed to hold as $k \to 0$, uniformly for all $\xi \in [-\pi/h, \pi/h]$. Condition *(a)* is called the **von Neumann condition.** For the record, let us give it a formal statement:

VON NEUMANN CONDITION. The spectral radius of the amplification matrix satisfies

$$\rho(G_k(\xi)) \le 1 + O(k) \tag{4.6.5}$$

as $k \to 0$, uniformly for all $\xi \in [-\pi/h, \pi/h]$.

In summary, for vector or multistep problems, the von Neumann condition is a statement about eigenvalues of amplification matrices, and it is necessary but not sufficient for stability.

Obviously there is a gap between conditions (4.6.3) and (4.6.4). To eliminate this gap one can apply the Kreiss matrix theorem. For any matrix A and constant $\epsilon \ge 0$, let us define the **ϵ-pseudospectral radius** $\rho_\epsilon(A)$ of A to be the largest of the moduli of its ϵ-pseudo-eigenvalues, that is,

$$\rho_\epsilon(A) = \sup_{\lambda_\epsilon \in \Lambda_\epsilon(A)} |\lambda_\epsilon|. \tag{4.6.6}$$

From condition *(ii)* of the definition of $\Lambda_\epsilon(A)$, it is easily seen that an equivalent definition is

$$\rho_\epsilon(A) = \sup_{\|E\| \le \epsilon} \rho(A+E). \tag{4.6.7}$$

Theorem 4.10 can be restated in terms of the ϵ-pseudospectral radius as follows: a matrix A is power-bounded if and only if

$$\rho_\epsilon(A) \le 1 + O(\epsilon) \tag{4.6.8}$$

as $\epsilon \to 0$. For the purpose of determining stability we need to modify this condition slightly in recognition of the fact that only powers n with $nk \le T$ are of interest. Here is the result:

STABILITY VIA THE KREISS MATRIX THEOREM

Theorem 4.11. *A linear, constant-coefficient finite difference formula* $\{S_k\}$ *is stable in the* 2*-norm if and only if*

$$\rho_\epsilon(G_k(\xi)) \le 1 + O(\epsilon) + O(k) \tag{4.6.9}$$

as $\epsilon \to 0$ *and* $k \to 0$.

The "O" symbols in this theorem are understood to apply uniformly with respect to $\xi \in [-\pi/h, \pi/h]$.

Proof. A more explicit expression of (4.6.9) is

$$|\lambda_\epsilon| \le 1 + C_1\epsilon + C_2 k \tag{4.6.10}$$

for all λ_ϵ in the ϵ-pseudospectrum $\Lambda_\epsilon(G_k(\xi))$, all $\epsilon \ge 0$, all $k > 0$, and all $\xi \in [-\pi/h, \pi/h]$. Equivalently,

$$\|(zI - G(\xi))^{-1}\|_2 \le \frac{C}{|z| - (1 + C_2 k)} \tag{4.6.11}$$

for all $|z| > 1 + C_2 k$. [From here it's easy; to be completed later.] ∎

Theorem 4.11 is a powerful result, giving a necessary and sufficient condition for stability of a wide variety of problems, but it has two limitations. The first is that determining resolvent norms and pseudospectra is not always an easy matter, and that is why it is convenient to have Theorem 4.10 available as well. The second is that not all numerical methods satisfy the rather narrow requirements of constant coefficients and regular unbounded grids that make Fourier analysis applicable. This difficulty will be addressed in the next section.

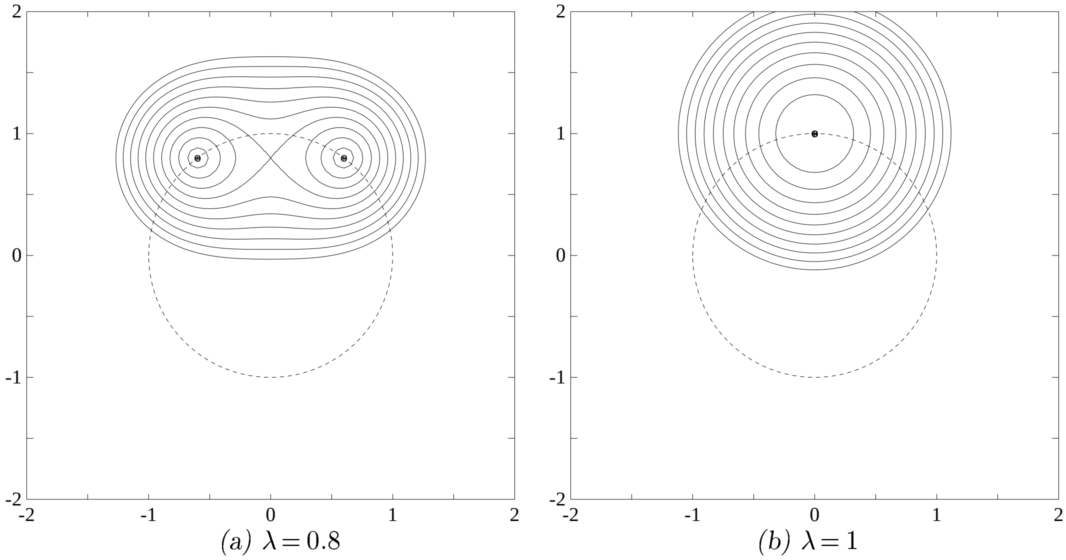


Figure 4.6.1. Boundaries of ϵ-pseudospectra of the leap frog amplification matrix $G(\xi)$ with $\xi = \pi/2$ ($\epsilon = 0.05, 0.10, \ldots, 0.50$). For $\lambda = 1$ there is a defective eigenvalue at $z = i$, the condition $\rho_\epsilon(G(\xi)) \le 1 + O(\epsilon)$ fails, and the formula is unstable.

EXAMPLE 4.6.1. For the leap frog model of $u_t = u_t$, the amplification matrix

$$G(\xi) = \begin{pmatrix} 2i\lambda \sin \xi h & 1 \\ 1 & 0 \end{pmatrix} \tag{4.6.12}$$

was derived in Example 3.6.1. For $\lambda < 1$ this family of matrices is power-bounded, but for $\lambda = 1$, the matrix $G(\pi/2)$ is defective and not power-bounded. The pseudospectra for $G(\pi/2)$ for $\lambda = 0.8$ and $\lambda = 1$ are illustrated and Figure 4.6.1. See Exercise 4.6.2.

In practice, how do people test for stability of finite difference methods? Usually, by computing eigenvalues and checking the von Neumann condition. If it is satisfied, the method is often stable, but sometimes it is not. It is probably accurate to say that when instability occurs, the problem is more often in the boundary conditions or nonlinearities than in the gap between Theorems 4.10 and 4.11 associated with non-normality of the amplification matrices. In Chapter 6 we shall discuss additional tests that can be used to check for instability introduced by boundary conditions.

EXERCISES

▷ *4.6.1. Multidimensional wave equation.* Consider again the second-order wave equation in d space dimensions

$$u_{tt} = u_{x_1x_1} + \cdots + u_{x_dx_d},$$

and the finite difference approximation discussed in Exercise 4.3.1. Use d-dimensional Fourier analysis to determine the stability bound on λ. (You do not have to use matrices and do it rigorously, which would involve an amplification matrix with a defective eigenvalue under certain conditions; just plug in the appropriate Fourier mode solution ("Ansatz") and compute amplification factors. You need not worry about keeping track of strong vs. weak inequalities.) Is it the same as the result of Exercise 4.3.1?

▷ *4.6.2. Stability of leap frog.* Consider the leap frog model of $u_t = u_x$ with $\lambda = k/h =$ constant ≤ 1 (Example 4.6.1 and Figure 4.6.1).

(a) Compute the eigenvalues and spectral radius of $G_k(\xi)$, and verify that the von Neumann condition does not reveal leap frog to be unstable.

(b) Compute the 2-norm $\|G_k(\xi)\|$, and verify that the condition (4.6.4), which would guarantee that leap frog is stable, does not hold.

(c) In fact, leap frog is stable for $\lambda < 1$. Prove this by any means you wish, but be sure that you have shown boundedness of the powers $G_k(\xi)$ uniformly in ξ, not just for each ξ individually. One way to carry out the proof is to argue first that for each fixed ξ, $\|G^n(\xi)\| \le M(\xi)$ for all n for an appropriate function $M(\xi)$, and then argue that $M(\xi)$ must be bounded as a function of ξ. Another method is to compute the eigenvalue decomposition of $G(\xi)$.

▷ *4.6.3. The fourth-order leap frog formula.* In Table 4.4.1 the stability restriction for the fourth-order leap frog formula is listed as $\lambda < 0.728\ldots$. What is this number?

▷ *4.6.4. The DuFort-Frankel formula.* The DuFort-Frankel model of $u_t = u_{xx}$, listed in Table 3.2.2 on p. 120, has some remarkable properties.

(a) Derive an amplification matrix for this formula. (This was done already in Exercise 3.6.1.)

(b) Show that it is unconditionally stable in ℓ_h^2.

(c) What does the result of *(b)*, together with the theorems of this chapter, imply about the consistency of the DuFort-Frankel formula with $u_t = u_{xx}$? Be precise, and state exactly what theorems you have appealed to.

(d) By manipulating Taylor series, derive the precise consistency restriction (involving k and h) for the DuFort-Frankel formula. Is it is same as the result of *(c)*?

4.7. Stability of the method of lines

[This section is not yet written. Most of its results can be found in S. C. Reddy and L. N. Trefethen, "Stability of the method of lines," *Numerische Mathematik 62* (1992), 235–267.]

The Kreiss matrix theorem (Theorem 4.9) asserts that a family of matrices $\{A_\nu\}$ is power-bounded if and only if its ϵ-pseudospectra $\Lambda_\epsilon(A_\nu)$ lie within a distance $O(\epsilon)$ of the unit disk, or equivalently, if and only if its resolvent norms $\|(zI-A_\nu)^{-1}\|$ increase at most inverse-linearly as z approaches the unit disk D from the outside. If the matrices all have a fixed dimension N, then this statement is valid exactly as it stands, and if the dimensions N_ν are variable, one loses a factor $\min\{N,n+1\}$, that is, $O(N)$ or $O(n)$.

The Kreiss matrix theorem has many consequences for stability of numerical methods for time-dependent PDEs. To apply it to this purpose, we first make a small modification so that we can treat stability on a finite interval $[0,T]$, as in §4.2, rather than the infinite interval $[0,\infty)$. All that is involved is to replace $O(\epsilon)$ by $O(\epsilon)+O(k)$, as in Theorem 4.11.

It turns out that there are four particularly important consequences of the Kreiss matrix theorem, which can be arranged in a two-by-two table according to the following two binary choices. Theorem 4.11 was one of these four consequences.

First, one can work either with the operators $\{S_k\}$ as matrices, or with the amplification matrices $\{G_k(\xi)\}$. The latter choice is only possible when Fourier analysis is applicable, i.e., under the usual restrictions of constant coefficients, regular grids, no boundaries, etc. It has the great advantage that the dimensions of the matrices involved are fixed, so there is no factor $O(n)$ or $O(N)$ to worry about, and indeed, $G_k(\xi)$ is often independent of k. When Fourier analysis is inapplicable, however, one always has the option of working directly with the matrices $\{S_k\}$ themselves—in "space space" instead of Fourier space. This is customary for example in the stability analysis of spectral methods on bounded domains.

Thus our first pair of theorems are as follows:

STABILITY

Theorem 4.11 (again). *A linear, constant-coefficient finite difference formula $\{S_k\}$ is stable in the 2-norm if and only if the pseudo-eigenvalues $\lambda_\epsilon \in \Lambda_\epsilon(G_k(\xi))$ of the amplification matrices satisfy*

$$dist(\lambda_\epsilon, D) = O(\epsilon)+O(k) \qquad (4.7.1)$$

STABILITY IN FOURIER SPACE

Theorem 4.12. *A linear finite difference formula $\{S_k\}$ is stable, up to a factor $\min\{N,n+1\}$, if and only if the pseudo-eigenvalues $\lambda_\epsilon \in \Lambda_\epsilon(S_k)$ satisfy*

$$dist(\lambda_\epsilon, D) = O(\epsilon)+O(k) \qquad (4.7.2)$$

In these theorems the order symbols $O(\epsilon)+O(k)$ should be understood to apply uniformly for all k and (where appropriate) ξ as $k\to 0$ and $\xi\to 0$.

As a practical matter, the factor $\min\{N,n+1\}$ is usually not important, because most often the instabilities that cause trouble are exponential. "One derivative of smoothness" in the initial and forcing data for a time-dependent PDE is generally enough to ensure that such a factor will not prevent convergence as the mesh is refined. In a later draft of the book this point will be emphasized in Chapter 4.

The other pair of theorems comes when we deal with the method of lines, discussed previously in §3.3. Following the standard formulation of the Lax-Richtmyer stability theory in Chapter 4, suppose we are given an autonomous linear partial differential equation

$$u_t = \mathcal{L}u, \tag{4.7.3}$$

where $u(t)$ is a function of one or more space variables on a bounded or unbounded domain and $\mathcal{L}$ is a differential operator, independent of t. For each sufficiently small time step $k>0$, let a corresponding finite or infinite spatial grid be defined and let (4.7.3) first be discretized with respect to the space variables only, so that it becomes a system of ordinary differential equations,

$$v_t = L_k v, \tag{4.7.4}$$

where $v(t)$ is a vector of dimension $N_k\le\infty$ and L_k is a matrix or bounded linear operator. With the space discretization determined in this way, let (4.7.4) then be discretized with respect to t by a linear multistep or Runge-Kutta formula with time step k. We assume that the stability region S is bounded by a cusp-free curve. Then one can show:

STABILITY OF THE METHOD OF LINES

Theorem 4.13. *The method of lines discretization described above is stable, up to a factor* $\min\{N,n+1\}$, *if and only if the pseudo-eigenvalues* $\lambda_\epsilon\in\Lambda_\epsilon(kL_k)$ *satisfy*

$$\mathit{dist}(\lambda_\epsilon,S) = O(\epsilon)+O(k) \tag{4.7.5}$$

STABILITY OF THE METHOD OF LINES IN FOURIER SPACE

Theorem 4.14. *A linear, constant-coefficient method of lines discretization as described above is stable in the* 2*-norm if and only if the pseudo-eigenvalues* $\lambda_\epsilon\in\Lambda_\epsilon(k\hat{L}_k(\xi))$ *satisfy*

$$\mathit{dist}(\lambda_\epsilon,S) = O(\epsilon)+O(k) \tag{4.7.6}$$

When the matrices S_k or $G_k(\xi)$ or L_k or $\hat{L}_k(\xi)$ appearing in these theorems are normal, one can simplify the statements by replacing ϵ-pseudo-eigenvalues by eigenvalues and $O(\epsilon)$ by 0. In particular, for a method of lines calculation in which the space discretization matrices L_k are normal, a necessary and sufficient condition for stability is that the eigenvalues of $\{kL_k\}$ lie within a distance $O(k)$ of the stability region S as $k\to 0$.