8

Convergence, Accuracy, and Stability

Whenever we use a numerical method to solve a differential equation, we should be concerned about the accuracy and convergence properties of the method. In practice we must apply the method on some particular discrete grid with a finite number of points, and we wish to ensure that the numerical solution obtained is a sufficiently good approximation to the true solution. For real problems we generally do not have the true solution to compare against, and we must rely on some combination of the following techniques to gain confidence in our numerical results:

- *Validation on test problems.* The method (and particular implementation) should be tested on simpler problems for which the true solution is known, or on problems for which a highly accurate comparison solution can be computed by other means. In some cases experimental results may also be available for comparison.
- *Theoretical analysis of convergence and accuracy.* Ideally one would like to prove that the method being used converges to the correct solution as the grid is refined, and also obtain reasonable error estimates for the numerical error that will be observed on any particular finite grid.

In this chapter we concentrate on the theoretical analysis. Here we consider only the Cauchy problem on the unbounded spatial domain, since the introduction of boundary conditions leads to a whole new set of difficulties in analyzing the methods. We will generally assume that the initial data has *compact support*, meaning that it is nonzero only over some bounded region. Then the solution to a hyperbolic problem (which has finite propagation speeds) will have compact support for all time, and so the integrals over the whole real line that appear below really reduce to finite intervals and we don't need to worry about issues concerning the behavior at infinity.

8.1 Convergence

In order to talk about accuracy or convergence, we first need a way to quantify the error. We are trying to approximate a function of space and time, and there are many possible ways to measure the magnitude of the error. In one space dimension we have an approximation Q_i^n at each point on space–time grid, or in each grid cell when using a finite volume method. For comparison we will let q_i^n represent the exact value we are hoping to approximate well.

For a finite difference method we would probably choose the pointwise value

$$q_i^n = q(x_i, t_n), \tag{8.1}$$

while for a finite volume method we might instead want to compare Q_i^n with

$$q_i^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x, t_n)\, dx. \tag{8.2}$$

If the function $q(x, t)$ is sufficiently smooth, then the pointwise value (8.1) evaluated at the cell center x_i agrees with the cell average (8.2) to $\mathcal{O}(\Delta x^2)$, and so for the methods considered in this book (which are generally at most second-order accurate), comparison with the pointwise value can be used even for finite volume methods and is often simpler.

To discuss convergence we must first pick some finite time T over which we wish to compute. We expect errors generally to grow with time, and so it would be unreasonable to expect that any finite grid would be capable of yielding good solutions at arbitrarily large times. Note that as we refine the grid, the number of time steps to reach time T will grow like $T/\Delta t$ and go to infinity (in the limit that must be considered in convergence theory), and so even in this case we must deal with an unbounded number of time steps. We will use N to indicate the time level corresponding to time $T = N\Delta t$. The *global error* at this time will be denoted by

$$E^N = Q^N - q^N,$$

and we wish to obtain bounds on this grid function as the grid is refined.

To simplify notation we will generally assume that Δt and Δx are related in a fixed manner as we refine the grid. For hyperbolic problems it is reasonable to assume that the ratio $\Delta t/\Delta x$ is fixed, for example. Then we can speak of letting $\Delta t \to 0$ to refine the grid, and speak of convergence with *order s* if the errors vanish like $\mathcal{O}(\Delta t^s)$ or as $\mathcal{O}(\Delta x^s)$, which are the same thing.

8.1.1 Choice of Norms

To quantify the error, we must choose some norm in which to measure the error at a fixed time. The standard set of norms most commonly used are the p-norms

$$\|E\|_p = \left(\Delta x \sum_{i=-\infty}^{\infty} |E_i|^p \right)^{1/p}. \tag{8.3}$$

These are discrete analogues of the function-space norms

$$\|E\|_p = \left(\int_{-\infty}^{\infty} |E(x)|^p\, dx \right)^{1/p}. \tag{8.4}$$

Note that the factor Δx in (8.3) is very important to give the correct scaling and order of accuracy as the grid is refined.

In particular, the 1-norm (with $p = 1$) is commonly used for conservation laws, since integrals of the solution itself are of particular importance. The 2-norm is often used for

linear problems because of the utility of Fourier analysis in this case (the classical *von Neumann analysis* of linear finite difference methods; see Section 8.3.3). We will use $\|\cdot\|$ without any subscript when we don't wish to specify a particular norm. Note that for a system of m equations, $E \in \mathbb{R}^m$ and the absolute value in (8.3) and (8.4) represents some vector norm on $\mathbb{R}^m$.

We say that the method is convergent at time T in the norm $\|\cdot\|$ if

$$\lim_{\substack{\Delta t \to 0 \\ N\Delta t = T}} \|E^N\| = 0.$$

The method is said to be *accurate of order s* if

$$\|E^N\| = \mathcal{O}(\Delta t^s) \qquad \text{as } \Delta t \to 0. \tag{8.5}$$

Ideally we might hope to have *pointwise convergence* as the grid is refined. This amounts to using the max norm (or ∞-norm) to measure the error:

$$\|E\|_\infty = \max_{-\infty < i < \infty} |E_i|. \tag{8.6}$$

This is the limiting case $p \to \infty$ of (8.3).

If the solution $q(x, t)$ is smooth, then it may be reasonable to expect pointwise convergence. For problems with discontinuous solutions, on the other hand, there will typically always be some smearing at one or more grid points in the neighborhood of the discontinuity. In this case we cannot expect convergence in the max norm, no matter how good the method is. In such cases we generally don't care about pointwise convergence, however. Convergence in the 1-norm is more relevant physically, and we may still hope to obtain this. In general, the rate of convergence observed can depend on what norm is being used, and it is important to use an appropriate norm when measuring convergence. Differences are typically greatest between the max norm and other choices (see Section 8.5 for another example). Except in fairly rare cases, the 1-norm and 2-norm will give similar results, and the choice of norm may depend mostly on which yields an easier mathematical analysis, e.g., the 1-norm for conservation laws and the 2-norm for linear equations.

8.2 One-Step and Local Truncation Errors

It is generally impossible to obtain a simple closed-form expression for the global error after hundreds or thousands of time steps. Instead of trying to obtain the error directly, the approach that is widely used in studying numerical methods for differential equations consists of a two-pronged attack on the problem:

- Study the error introduced in a single time step, showing that the method is *consistent* with the differential equation and introduces a small error in any one step.
- Show that the method is *stable*, so that these local errors do not grow catastrophically and hence a bound on the global error can be obtained in terms of these local errors.

If we can get a bound on the local error in an appropriate sense, then stability can be used to convert this into a bound on the global error that can be used to prove convergence.

Moreover we can generally determine the rate of convergence and perhaps even obtain reasonable error bounds. The *fundamental theorem* of numerical methods for differential equations can then be summarized briefly as

$$\text{consistency} + \text{stability} \iff \text{convergence}. \tag{8.7}$$

This theorem appears in various forms in different contexts, e.g., the Lax equivalence theorem for linear PDEs (Section 8.3.2) or Dahlquist's equivalence theorem for ODEs. The exact form of "stability" needed depends on the type of equation and method. In this section we will study the local error, and in Section 8.3 we turn to the question of stability.

A general explicit numerical method can be written as

$$Q^{n+1} = \mathcal{N}(Q^n),$$

where $\mathcal{N}(\cdot)$ represents the numerical operator mapping the approximate solution at one time step to the approximate solution at the next. The *one-step error* is defined by applying the numerical operator to the true solution (restricted to the grid) at some time and comparing this with the true solution at the next time:

$$\text{one-step error} = \mathcal{N}(q^n) - q^{n+1}. \tag{8.8}$$

Here q^n and q^{n+1} represent the true solution restricted to the grid by (8.1) or (8.2). This gives an indication of how much error is introduced in a single time step by the numerical method. The *local truncation error* is defined by dividing this by Δt:

$$\tau^n = \frac{1}{\Delta t}[\mathcal{N}(q^n) - q^{n+1}]. \tag{8.9}$$

As we will see in Section 8.3, the local truncation error typically gives an indication of the magnitude of the global error, and particularly the order of accuracy, in cases when the method is stable. If the local truncation error is $\mathcal{O}(\Delta x^s)$ as $s \to 0$, then we expect the global error to have this same behavior.

We say that the method is *consistent* with the differential equation if the local truncation error vanishes as $\Delta t \to 0$ for all smooth functions $q(x, t)$ satisfying the differential equation. In this case we expect the method to be convergent, provided it is stable.

The local truncation error is relatively easy to investigate, and for smooth solutions can be well approximated by simple Taylor series expansions. This is illustrated very briefly in the next example.

Example 8.1. Consider the first-order upwind method for the advection equation with $\bar{u} > 0$,

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\bar{u}\big(Q_i^n - Q_{i-1}^n\big).$$

Applying this method to the true solution gives the local truncation error

$$\tau^n = \frac{1}{\Delta t}\Big(q(x_i, t_n) - \frac{\Delta t}{\Delta x}\bar{u}[q(x_i, t_n) - q(x_{i-1}, t_n)] - q(x_i, t_{n+1})\Big). \tag{8.10}$$

We now expand $q(x_{i-1}, t_n)$ and $q(x_i, t_{n+1})$ in Taylor series about (x_i, t_n) and cancel common terms to obtain

$$\tau^n = -[q_t(x_i, t_n) + \bar{u} q_x(x_i, t_n)] + \frac{1}{2}\Delta x\, \bar{u} q_{xx}(x_i, t_n) - \frac{1}{2}\Delta t\, q_{tt}(x_i, t_n) + \cdots. \quad (8.11)$$

The first term in this expression is identically zero, because we assume that q is an exact solution to the advection equation and hence $q_t + \bar{u} q_x = 0$, so we find that

$$\begin{aligned} \text{Upwind:} \qquad \tau^n &= \frac{1}{2}\bar{u}\, \Delta x\, q_{xx}(x_i, t_n) - \frac{1}{2}\Delta t\, q_{tt}(x_i, t_n) + \mathcal{O}(\Delta t^2) \\ &= \frac{1}{2}\bar{u}\, \Delta x\, (1 - \nu)\, q_{xx}(x_i, t_n) + \mathcal{O}(\Delta t^2), \end{aligned} \quad (8.12)$$

where

$$\nu \equiv \bar{u}\, \Delta t / \Delta x \quad (8.13)$$

is the Courant number. The truncation error is dominated by an $\mathcal{O}(\Delta x)$ term, and so the method is first-order accurate.

Similarly, one can compute the local truncation errors for other methods, e.g.,

$$\text{Lax–Friedrichs:} \qquad \tau^n = \frac{1}{2}\left(\frac{\Delta x^2}{\Delta t} - \bar{u}^2 \Delta t\right) q_{xx}(x_i, t_n) + \mathcal{O}(\Delta t^2) \quad (8.14)$$

$$= \frac{1}{2}\bar{u}\, \Delta x\, (1/\nu - \nu) q_{xx}(x_i, t_n) + \mathcal{O}(\Delta t^2), \quad (8.15)$$

$$\text{Lax–Wendroff:} \qquad \tau^n = -\frac{1}{6}\bar{u}(\Delta x)^2(1 - \nu^2) q_{xxx}(x_i, t_n) + \mathcal{O}(\Delta t^3). \quad (8.16)$$

Note that the Lax–Wendroff method is second-order accurate and the dominant term in the truncation error depends on the third derivative of q, whereas the upwind and Lax–Friedrichs methods are both first-order accurate with the dominant error term depending on q_{xx}. The relation between these errors and the diffusive or dispersive nature of the methods is discussed in Section 8.6.

8.3 Stability Theory

In this section we review the basic ideas of stability theory and the derivation of global error bounds from information on the local truncation error. The form of stability bounds discussed here are particularly useful in analyzing linear methods. For nonlinear methods they may be hard to apply, and in Section 12.12 we discuss a different approach to stability theory for such methods.

The essential requirements and importance of stability can be easily seen from the following attempt to bound the global error using a recurrence relation. In time step n suppose we have an approximation Q^n with error E^n, so that

$$Q^n = q^n + E^n.$$

We apply the numerical method to obtain Q^{n+1}:

$$Q^{n+1} = \mathcal{N}(Q^n) = \mathcal{N}(q^n + E^n),$$

and the global error is now

$$\begin{aligned} E^{n+1} &= Q^{n+1} - q^{n+1} \\ &= \mathcal{N}(q^n + E^n) - q^{n+1} \\ &= \mathcal{N}(q^n + E^n) - \mathcal{N}(q^n) + \mathcal{N}(q^n) - q^{n+1} \\ &= [\mathcal{N}(q^n + E^n) - \mathcal{N}(q^n)] + \Delta t\, \tau^n. \end{aligned} \tag{8.17}$$

By introducing $\mathcal{N}(q^n)$ we have written the new global error as the sum of two terms:

- $\mathcal{N}(q^n + E^n) - \mathcal{N}(q^n)$, which measures the effect of the numerical method on the *previous* global error E^n,
- $\Delta t\, \tau^n$, the new one-step error introduced in this time step.

The study of the local truncation error allows us to bound the new one-step error. Stability theory is required to bound the other term, $\mathcal{N}(q^n + E^n) - \mathcal{N}(q^n)$.

8.3.1 Contractive Operators

The numerical solution operator $\mathcal{N}(\cdot)$ is called *contractive* in some norm $\|\cdot\|$ if

$$\|\mathcal{N}(P) - \mathcal{N}(Q)\| \le \|P - Q\| \tag{8.18}$$

for any two grid functions P and Q. If the method is contractive, then it is stable in this norm and we can obtain a bound on the global error from (8.17) very simply using $P = q^n + E^n$ and $Q = q^n$:

$$\begin{aligned} \|E^{n+1}\| &\le \|\mathcal{N}(q^n + E^n) - \mathcal{N}(q^n)\| + \Delta t\, \|\tau^n\| \\ &\le \|E^n\| + \Delta t\, \|\tau^n\|. \end{aligned} \tag{8.19}$$

Applying this recursively gives

$$\|E^N\| \le \|E^0\| + \Delta t \sum_{n=1}^{N-1} \|\tau^n\|.$$

Suppose the local truncation error is bounded by

$$\|\tau\| \equiv \max_{0 \le n \le N} \|\tau^n\|.$$

Then we have

$$\begin{aligned} \|E^N\| &\le \|E^0\| + N\Delta t\, \|\tau\| \\ &\le \|E^0\| + T\|\tau\| \qquad (\text{for } N\Delta t = T). \end{aligned} \tag{8.20}$$

The term $\|E^0\|$ measures the error in the initial data on the grid, and we require that $\|E^0\| \to 0$ as $\Delta t \to 0$ in order to be solving the correct initial-value problem. If the method is consistent, then also $\|\tau\| \to 0$ as $\Delta t \to 0$ and we have proved convergence. Moreover, if $\|\tau\| = \mathcal{O}(\Delta t^s)$, then the global error will have this same behavior as $\Delta t \to 0$ (provided the initial data is sufficiently accurate), and the method has global order of accuracy s.

Actually a somewhat weaker requirement on the operator $\mathcal{N}$ is sufficient for stability. Rather than the contractive property (8.18), it is sufficient to have

$$\|\mathcal{N}(P) - \mathcal{N}(Q)\| \le (1 + \alpha\, \Delta t)\|P - Q\| \tag{8.21}$$

for all P and Q, where α is some constant independent of Δt as $\Delta t \to 0$. (Recall that the one-step operator $\mathcal{N}$ depends on Δt even though we haven't explicitly included this dependence in the notation.) If (8.21) holds then the above proof still goes through with a slight modification. We now have

$$\|E^{n+1}\| \le (1 + \alpha\, \Delta t)\|E^n\| + \Delta t\, \|\tau\|,$$

and so

$$\begin{aligned} \|E^N\| &\le (1 + \alpha\, \Delta t)^N \|E^0\| + \Delta t \sum_{n=1}^{N-1} (1 + \alpha\, \Delta t)^{N-1-n} \|\tau\| \\ &\le e^{\alpha T}(\|E^0\| + T\|\tau\|) \qquad (\text{for } N\Delta t = T). \end{aligned} \tag{8.22}$$

In this case the error may grow exponentially in time, but the key fact is that this growth is bounded independently of the time step Δt. For fixed T we have a bound that goes to zero with Δt. This depends on the fact that any growth in error resulting from the operator $\mathcal{N}$ is at most order $\mathcal{O}(\Delta t)$ in one time step, which is what (8.21) guarantees.

8.3.2 Lax–Richtmyer Stability for Linear Methods

If the operator $\mathcal{N}(\cdot)$ is a *linear* operator, then $\mathcal{N}(q^n + E^n) = \mathcal{N}(q^n) + \mathcal{N}(E^n)$, and so $\mathcal{N}(q^n + E^n) - \mathcal{N}(q^n)$ reduces to simply $\mathcal{N}(E^n)$. In this case, the condition (8.21) reduces simply to requiring that

$$\|\mathcal{N}(E^n)\| \le (1 + \alpha\, \Delta t)\|E^n\| \tag{8.23}$$

for any grid function E^n, which is generally simpler to check. This can also be expressed as a bound on the norm of the linear operator $\mathcal{N}$,

$$\|\mathcal{N}\| \le 1 + \alpha\, \Delta t. \tag{8.24}$$

An even looser version of this stability requirement can be formulated in the linear case. We really only need that, for each time T, there is a constant C such that

$$\|\mathcal{N}^n\| \le C \tag{8.25}$$

for all $n \le N = T/\Delta t$, i.e., the nth power of the operator $\mathcal{N}$ is uniformly bounded up to this time, for then all the terms in (8.22) are uniformly bounded. For linear methods, this form

of stability is generally referred to as *Lax–Richtmyer stability*. The result (8.7) is called the *Lax equivalence theorem* in this context. See [369] for a rigorous proof. Note that if (8.24) holds, then we can take $C = e^{\alpha T}$ in (8.25).

Classical methods such as the first-order upwind or the Lax–Wendroff method for the linear advection equation are all linear methods, and this form of the stability condition can be used. (See Section 8.3.4 for an example.) The high-resolution methods developed in Chapter 6 are *not* linear methods, however, since the limiter function introduces nonlinearity. Proving stability of these methods is more subtle and is discussed briefly in Section 8.3.5 and Chapter 15.

8.3.3 2-Norm Stability and von Neumann Analysis

For linear difference equations, stability analysis is often particularly easy in the 2-norm, since Fourier analysis can then be used to simplify the problem. This is the basis of *von Neumann stability analysis*, which is described more completely in many books on finite difference methods for partial differential equations (e.g., [333] or [427]).

Let Q_I^n $(-\infty < I < \infty)$ represent an arbitrary grid function for the Cauchy problem. In this section we use I for the grid index ($x_I = I\Delta x$) so that $i = \sqrt{-1}$ can be used in the complex exponentials below. We suppose that Q_I^n has finite norm, so that it can be expressed as a Fourier series

$$Q_I^n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{Q}(\xi)\, e^{i\xi I \Delta x}\, d\xi. \tag{8.26}$$

Applying a linear finite difference method to Q_I^n and manipulating the exponentials typically gives an expression of the form

$$Q_I^{n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{Q}(\xi)\, g(\xi, \Delta x, \Delta t)\, e^{i\xi I \Delta x}\, d\xi, \tag{8.27}$$

where $g(\xi, \Delta x, \Delta t)$ is called the *amplification factor* for wave number ξ, since

$$\hat{Q}^{n+1}(\xi) = g(\xi, \Delta x, \Delta t)\, \hat{Q}^n(\xi). \tag{8.28}$$

The 2-norm is now convenient because of *Parseval's relation*, which states that

$$\|Q^n\|_2 = \|\hat{Q}^n\|_2, \tag{8.29}$$

where

$$\|Q^n\|_2 = \left(\Delta x \sum_{I=-\infty}^{\infty} \left|Q_I^n\right|^2 \right)^{1/2} \quad \text{and} \quad \|\hat{Q}^n\|_2 = \left(\int_{-\infty}^{\infty} |\hat{Q}^n(\xi)|^2\, d\xi \right)^{1/2}.$$

To show that the 2-norm of Q^n remains bounded it suffices to show that the 2-norm of $\hat{Q}^n$ does. But whereas all elements of Q^n (as I varies) are coupled together via the difference equations, each element of $\hat{Q}^n$ (as ξ varies) satisfies an equation (8.28) that is decoupled

from all other wave numbers. (The Fourier transform diagonalizes the linear difference operator.) So it suffices to consider an arbitrary single wave number ξ and data of the form

$$Q_I^n = e^{i\xi I\Delta x}. \tag{8.30}$$

From this we can compute $g(\xi, \Delta x, \Delta t)$. Then requiring that $|g(\xi, \Delta x, \Delta t)| \le 1$ for all ξ gives a sufficient condition for stability. In fact it suffices to have $|g(\xi, \Delta x, \Delta t)| \le 1 + \alpha\,\Delta t$ for some constant α independent of ξ.

Example 8.2. Consider the upwind method (4.25) for the advection equation $q_t + \bar{u}q_x = 0$ with $\bar{u} > 0$. Again use $\nu \equiv \bar{u}\,\Delta t/\Delta x$ as shorthand for the Courant number, and write the upwind method (4.25) as

$$\begin{aligned} Q_I^{n+1} &= Q_I^n - \nu\big(Q_I^n - Q_{I-1}^n\big) \\ &= (1-\nu)Q_I^n + \nu Q_{I-1}^n. \end{aligned} \tag{8.31}$$

We will use von Neumann analysis to demonstrate that this method is stable in the 2-norm provided that

$$0 \le \nu \le 1 \tag{8.32}$$

is satisfied, which agrees exactly with the CFL condition for this method (see Section 4.4). Using the data (8.30) yields

$$\begin{aligned} Q_I^{n+1} &= (1-\nu)e^{i\xi I\Delta x} + \nu e^{i\xi(I-1)\Delta x} \\ &= \left[(1-\nu) + \nu e^{-i\xi\Delta x}\right] e^{i\xi I\Delta x} \\ &= g(\xi, \Delta x, \Delta t)\, Q_I^n \end{aligned} \tag{8.33}$$

with

$$g(\xi, \Delta x, \Delta t) = (1-\nu) + \nu e^{-i\xi\Delta x}. \tag{8.34}$$

As ξ varies, $g(\xi, \Delta x, \Delta t)$ lies on a circle of radius ν in the complex plane, centered on the real axis at $1-\nu$. This circle lies entirely inside the unit circle (i.e., $|g| \le 1$ for all ξ) if and only if $0 \le \nu \le 1$, giving the stability limit (8.32) for the upwind method.

8.3.4 1-Norm Stability of the Upwind Method

For conservation laws the 1-norm is often used, particularly for nonlinear problems. We will demonstrate that the upwind method (4.25) considered in the previous example is also stable in the 1-norm under the time-step restriction (8.32). We revert to the usual notation with i as the grid index and write the upwind method (8.31) as

$$Q_i^{n+1} = (1-\nu)Q_i^n + \nu Q_{i-1}^n, \tag{8.35}$$

where ν is again the Courant number (8.13). From this we compute

$$\begin{aligned}\|Q^{n+1}\|_1 &= \Delta x \sum_i \left|Q_i^{n+1}\right| \\ &= \Delta x \sum_i \left|(1-\nu)Q_i^n + \nu Q_{i-1}^n\right| \\ &\le \Delta x \sum_i \left[(1-\nu)\left|Q_i^n\right| + \nu\left|Q_{i-1}^n\right|\right]. \end{aligned} \tag{8.36}$$

In the final step we have used the triangle inequality and then pulled $1-\nu$ and ν outside the absolute values, since these are both positive if (8.32) is satisfied. The sum can be split up into two separate sums, each of which gives $\|Q^n\|_1$, obtaining

$$\|Q^{n+1}\|_1 \le (1-\nu)\|Q^n\|_1 + \nu\|Q^n\|_1 = \|Q^n\|_1.$$

This proves stability in the 1-norm. Note that this only works if (8.32) is satisfied, since we need both $1-\nu$ and ν to be positive.

8.3.5 Total-Variation Stability for Nonlinear Methods

For a nonlinear numerical method, showing that (8.23) holds is generally not sufficient to prove convergence. The stronger contractivity property (8.21) would be sufficient, but is generally difficult to obtain. Even for the linear advection equation, the high-resolution methods of Chapter 6 are nonlinear (since the limiter function depends on the data), and so a different approach to stability must be adopted to prove convergence of these methods.

The total variation introduced in Section 6.7 turns out to be an effective tool for studying stability of nonlinear problems. We make the following definition.

Definition 8.1. *A numerical method is* total-variation bounded *(TVB) if, for any data Q^0 (with $\mathrm{TV}(Q^0) < \infty$) and time T, there is a constant $R > 0$ and a value $\Delta t_0 > 0$ such that*

$$\mathrm{TV}(Q^n) \le R \tag{8.37}$$

for all $n\,\Delta t \le T$ whenever $\Delta t < \Delta t_0$.

This simply requires that we have a uniform bound on the total variation up to time T on all grids sufficiently fine (and hence as $\Delta t \to 0$).

In Section 12.12 we will see how this can be used to prove convergence of the numerical method (using a more subtle argument than the approach taken above for linear problems, based on the compactness of an appropriate function space). For now we just note that, in particular, a method that is TVD (see Section 6.7) is certainly TVB with $R = \mathrm{TV}(Q^0)$ in (8.37) for any T. So the notion of a TVD method, useful in insuring that no spurious oscillations are introduced, is also sufficient to prove convergence. In particular the high-resolution TVD methods introduced in Chapter 6 are all convergent provided the CFL condition is satisfied (since this is required in order to be TVD).

Of course a weaker condition than TVD is sufficient for convergence, since we only need a uniform bound of the form (8.37). For example, a method that satisfies

$$\mathrm{TV}(Q^{n+1}) \le (1 + \alpha\,\Delta t)\mathrm{TV}(Q^n) \tag{8.38}$$

for some constant α independent of Δt (at least for all Δt sufficiently small) will also be TVB.

8.4 Accuracy at Extrema

Examining the results of Figure 6.1 and Figure 6.3 shows that the high-resolution methods developed in Chapter 6 give rather poor accuracy at extrema (local maxima or minima in q), even though the solution is smooth. Figure 8.1(a) gives an indication of why this occurs. All of the limiters discussed in Section 6.9 will give slopes $\sigma = 0$ in cells $i - 2$ and $i - 1$ for this data. This is required in order to prove that the method is truly TVD, as any other choice will give a reconstructed function $\tilde{q}^n(x, t_n)$ with $\mathrm{TV}(\tilde{q}^n(\cdot, t_n)) > \mathrm{TV}(Q^n)$, allowing the possibility that $\mathrm{TV}(Q^{n+1}) > \mathrm{TV}(Q^n)$ with suitable choices of the data and time step. If the solution is smooth near the extremum, then the Lax–Wendroff slope should be close to zero anyway, so this is perhaps not a bad approximation. However, setting the slope to zero will lead to a *clipping* of the solution, and the extreme value will be diminished by $\mathcal{O}(\Delta x^2) = \mathcal{O}(\Delta t^2)$ in each time step. After $T/\Delta t$ time steps this can lead to a global error near extrema that is $\mathcal{O}(\Delta t)$, reducing the method to first-order accuracy in this region. This can be observed in Figure 6.2. Osher and Chakravarthy [351] prove that TVD methods must in fact degenerate to first-order accuracy at extremal points.

Using a better approximation to q_x near extrema, as indicated in Figure 8.1(b), would give a reconstruction that allows smooth peaks to be better represented over time, since the peaks are then reconstructed more accurately from the data. The cost is that the total variation will need to increase slightly at times in order to reconstruct such a peak. But as indicated in Section 8.3.5, the TVD property is not strictly needed for stability. The challenge is to find looser criteria that allow a small increase in the total variation near extremal points while still suppressing oscillations where necessary. This goal has led to several suggestions on

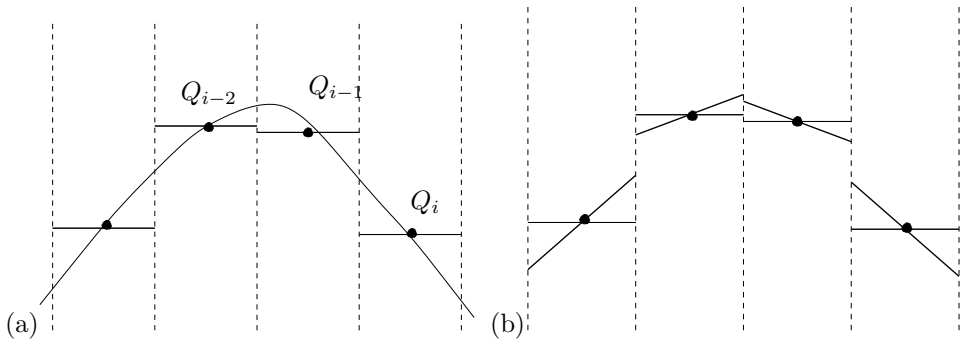

Fig. 8.1. (a) A smooth maximum and cell averages. A TVD slope reconstruction will give clipping of the peak. (b) Better accuracy can be obtained with a reconstruction that is not TVD relative to the cell averages.

other ways to choose the limiter function and criteria other than strictly requiring the TVD property. Shu [407] has developed the theory of *TVB* methods for which convergence can still be proved while better accuracy at extremal points may be obtained. The *essentially nonoscillatory* (ENO) methods are another approach to obtaining high-resolution that often give better than second-order accuracy in smooth regions, including extrema. These are briefly described in Section 10.4.4.

8.5 Order of Accuracy Isn't Everything

The quality of a numerical method is often summarized by a single number, its order of accuracy. This is indeed an important consideration, but it can be a mistake to put too much emphasis on this one attribute. It is not always true that a method with a higher order of accuracy is more accurate on a particular grid or for a particular problem.

Suppose a method has order of accuracy s. Then we expect the error to behave like

$$\|E^N\| = C(\Delta x)^s + \text{higher-order terms} \tag{8.39}$$

as the grid is refined and $\Delta x \to 0$. Here C is some constant that depends on the particular solution being computed (and the time T). The magnitude of the constant C is important as well as the value of s. Also, note that the "higher-order" terms, which depend on higher powers of Δx, are asymptotically negligible as $\Delta x \to 0$, but may in fact be larger than the "dominant" term $C(\Delta x)^s$ on the grid we wish to use in practice.

As a specific example, consider the high-resolution TVD methods developed in Chapter 6 for the scalar advection equation. Because of the nonlinear limiter function, these methods are formally not second-order accurate, even when applied to problems with smooth solutions. The limiter typically leads to a clipping of the solution near extrema, as discussed in Section 8.4.

For discontinuous solutions, as illustrated in Figure 6.1 and Figure 6.2, these methods have clear advantages over the "second-order" Lax–Wendroff or Beam–Warming methods, even though for discontinuous solutions none of these methods exhibit second-order convergence.

But suppose we compare these methods on a problem where the solution is smooth. At least in this case one might think the second-order methods should be better than the high-resolution methods, which have a lower order of accuracy. This is true on a sufficiently fine grid, but may not be at all true on the sort of grids we want to use in practice.

Consider the wave-packet propagation problem illustrated in Figure 6.3. Here the data is smooth and yet the high-resolution method shows a clear advantage over the Lax–Wendroff on the grid shown in this figure. Figure 8.2 shows the results of a mesh refinement study for this particular example. The true and computed solutions are compared on a sequence of grids, and the norm of the error is plotted as a function of Δx. These are shown on a log–log scale because from (8.39) we have

$$\log|E| \approx \log|C| + s\log|\Delta x|, \tag{8.40}$$

so that we expect linear behavior in this plot, with a slope given by the order of accuracy s. Figure 8.2(a) shows errors in the max norm, while Figure 8.2(b) shows the errors in the 1-norm. The Lax–Wendroff method is second-order accurate in both norms; the slope

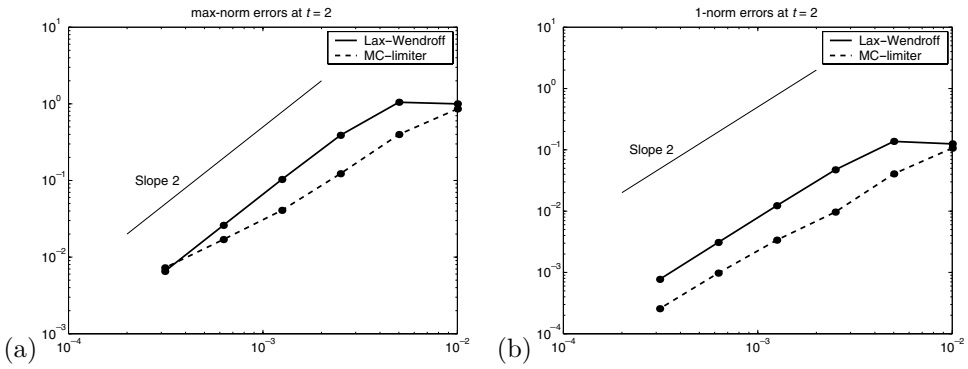

Fig. 8.2. Log–log plot of the error vs. grid size for the Lax–Wendroff and a high-resolution method on the wave-packet problem of Figure 6.3: (a) max-norm errors, (b) 1-norm errors. `[claw/book/chap8/wavepacket]`

of the corresponding curves is about 1.999. Due to the clipping of peaks observed with TVD methods, the error with the high-resolution method is dominated by errors near these few locations, and so the max norm shows larger errors than the 1-norm, which averages over the entire domain. In the max norm the observed order of accuracy is about 1.22, and on sufficiently fine grids the Lax–Wendroff method is superior. However, the crossover point for this particular example is at about $\Delta x = 0.00035$, meaning about 2800 grid cells in the unit interval. This is a much finer grid than one would normally want to use for this problem. Certainly for two- or three-dimensional analogues of this problem it would be unthinkable to use this many grid points in each direction. On the coarser grids one might use in practice, the high-resolution method is superior in spite of its lower order of accuracy.

In the 1-norm the high-resolution method looks even better. In this norm the observed order of accuracy is about 1.92, but the error constant C is about 5 times smaller than what is observed for the Lax–Wendroff method, so that on all the grids tested the error is essentially 5 times smaller. Of course, for very small Δx the Lax–Wendroff method would eventually prove superior, but extrapolating from the results seen in Figure 8.2(b) we find that the crossover point in the 1-norm is around $\Delta x = 10^{-33}$.

Later on we will see other examples where it is wise to look beyond order of accuracy in comparing different methods. For example, in Chapter 17 we will see that a fractional-step method for source terms that is often dismissed as being "only first-order accurate" is in fact essentially identical to second-order accurate methods for many practical purposes, and is often more efficient to use.

8.6 Modified Equations

As discussed in Section 8.2, the local truncation error of a method is determined by seeing how well the true solution of the differential equation satisfies the difference equation. Now we will study a slightly different approach that can be very illuminating in that it reveals much more about the structure and behavior of the numerical solution in addition to the order of the error.

The idea is to ask the following question: Is there a PDE to which our numerical approximation Q_i^n is actually the *exact* solution? Or, less ambitiously, can we at least find an equation that is better satisfied by Q_i^n than the original PDE we were attempting to solve? If so, then studying the behavior of solutions to this PDE should tell us much about how the numerical approximation is behaving. This can be advantageous because it is often easier to study the behavior of solutions of differential equations than of finite-difference formulas.

In fact it is possible to find a PDE that is exactly satisfied by the Q_i^n, by doing Taylor series expansions as we do to compute the local truncation error. However, this PDE will have an infinite number of terms involving higher and higher powers of Δt and Δx. By truncating this series at some point we will obtain a PDE that is simple enough to study and yet gives a good indication of the behavior of the Q_i^n. If the method is accurate to order s, then this equation is generally a modification of the original PDE with new terms of order s, and is called the *modified equation* for the method, or sometimes the *model equation*. Good descriptions of the theory and use of modified equations can be found in Hedstrom [193] or Warming & Hyett [480]. See [61], [114], [126], [170] for some further discussion and other applications of this approach.

8.6.1 The Upwind Method

The derivation of a modified equation is best illustrated with an example. Consider the first-order upwind method for the advection equation $q_t + \bar{u} q_x = 0$ in the case $\bar{u} > 0$,

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\big(Q_i^n - Q_{i-1}^n\big). \tag{8.41}$$

The process of deriving the modified equation is very similar to computing the local truncation error, only now we insert a function $v(x,t)$ into the numerical method instead of the true solution $q(x,t)$. Our goal is to determine a differential equation satisfied by v. We view the method as a finite difference method acting on grid-point values, and v is supposed to be a function that agrees exactly with Q_i^n at the grid points. So, unlike $q(x,t)$, the function $v(x,t)$ satisfies (8.41) exactly:

$$v(x, t+\Delta t) = v(x,t) - \frac{\bar{u}\,\Delta t}{\Delta x}[v(x,t) - v(x-\Delta x, t)].$$

Expanding these terms in Taylor series about (x,t) and simplifying gives

$$\left(v_t + \frac{1}{2}\Delta t\, v_{tt} + \frac{1}{6}(\Delta t)^2 v_{ttt} + \cdots\right) + \bar{u}\left(v_x - \frac{1}{2}\Delta x\, v_{xx} + \frac{1}{6}(\Delta x)^2 v_{xxx} + \cdots\right) = 0.$$

We can rewrite this as

$$v_t + \bar{u} v_x = \frac{1}{2}(\bar{u}\,\Delta x\, v_{xx} - \Delta t\, v_{tt}) - \frac{1}{6}[\bar{u}(\Delta x)^2 v_{xxx} + (\Delta t)^2 v_{ttt}] + \cdots. \tag{8.42}$$

This is the PDE that v satisfies. If we take $\Delta t/\Delta x$ fixed, then the terms on the right-hand side are $\mathcal{O}(\Delta t)$, $\mathcal{O}(\Delta t^2)$, etc., so that for small Δt we can truncate this series to get a PDE that is quite well satisfied by the Q_i^n.

If we drop all the terms on the right-hand side, we just recover the original advection equation. Since we have then dropped terms of $\mathcal{O}(\Delta t)$, we expect that Q_i^n satisfies this equation to $\mathcal{O}(\Delta t)$, as we know to be true, since this upwind method is first-order accurate.

If we keep the $\mathcal{O}(\Delta t)$ terms then we get something more interesting:

$$v_t + \bar{u} v_x = \frac{1}{2}(\bar{u}\,\Delta x\, v_{xx} - \Delta t\, v_{tt}). \tag{8.43}$$

This involves second derivatives in both x and t, but we can derive a slightly different modified equation with the same accuracy by differentiating (8.43) with respect to t to obtain

$$v_{tt} = -\bar{u} v_{xt} + \frac{1}{2}(\bar{u}\,\Delta x\, v_{xxt} - \Delta t\, v_{ttt})$$

and with respect to x to obtain

$$v_{tx} = -\bar{u} v_{xx} + \frac{1}{2}(\bar{u}\,\Delta x\, v_{xxx} - \Delta t\, v_{ttx}).$$

Combining these gives

$$v_{tt} = \bar{u}^2 v_{xx} + \mathcal{O}(\Delta t).$$

Inserting this in (8.43) gives

$$v_t + \bar{u} v_x = \frac{1}{2}(\bar{u}\,\Delta x\, v_{xx} - \bar{u}^2 \Delta t\, v_{xx}) + \mathcal{O}(\Delta t^2).$$

Since we have already decided to drop terms of $\mathcal{O}(\Delta t^2)$, we can drop these terms here also to obtain

$$v_t + \bar{u} v_x = \frac{1}{2}\bar{u}\,\Delta x\,(1 - \nu) v_{xx}, \tag{8.44}$$

where $\nu = \bar{u}\,\Delta t/\Delta x$ is the Courant number. This is now a familiar advection–diffusion equation. The grid values Q_i^n can be viewed as giving a *second-order accurate* approximation to the true solution of this equation (whereas they only give first-order accurate approximations to the true solution of the advection equation).

For higher-order methods this elimination of t-derivatives in terms of x-derivatives can also be done, but must be done carefully and is complicated by the need to include higher-order terms. Warming and Hyett [480] present a general procedure.

The fact that the modified equation for the upwind method is an advection–diffusion equation tells us a great deal about how the numerical solution behaves. Solutions to the advection–diffusion equation translate at the proper speed $\bar{u}$ but also diffuse and are smeared out. This was clearly visible in Figures 6.1 and 6.3, for example.

Note that the diffusion coefficient in (8.43) vanishes in the special case $\bar{u}\,\Delta t = \Delta x$. In this case we already know that the exact solution to the advection equation is recovered by the upwind method; see Figure 4.4 and Exercise 4.2.

Also note that the diffusion coefficient is positive only if $0 < \bar{u}\,\Delta t/\Delta x < 1$. This is precisely the stability limit of the upwind method. If it is violated, then the diffusion coefficient

in the modified equation is negative, giving an ill-posed *backward heat equation* with exponentially growing solutions. Hence we see that some information about stability can also be extracted from the modified equation.

8.6.2 Lax–Wendroff Method

If the same procedure is followed for the Lax–Wendroff method, we find that all $\mathcal{O}(\Delta t)$ terms drop out of the modified equation, as is expected because this method is second-order accurate on the advection equation. The modified equation obtained by retaining the $\mathcal{O}(\Delta t^2)$ term and then replacing time derivatives by spatial derivatives is

$$v_t + \bar{u} v_x = -\frac{1}{6}\bar{u}(\Delta x)^2(1-\nu^2)v_{xxx}. \tag{8.45}$$

The Lax–Wendroff method produces a *third-order* accurate solution to this equation. This equation has a very different character from (8.43). The v_{xxx} term leads to *dispersive* behavior rather than diffusion.

This dispersion is very clearly seen in the wave-packet computation of Figure 6.3, where the Q_i^n computed with the Lax–Wendroff method clearly travels at the wrong speed. Dispersive wave theory predicts that such a packet should travel at the *group velocity*, which for wavenumber ξ in the Lax–Wendroff method is

$$c_g = \bar{u} - \frac{1}{2}\bar{u}(\Delta x)^2(1-\nu^2)\xi^2.$$

See for example [8], [50], [298], [427], [486] for discussions of dispersive equations and group velocities. The utility of this concept in the study of numerical methods has been stressed by Trefethen, in particular in relation to the stability of boundary conditions. A nice summary of some of this theory may be found in Trefethen [458].

The computation shown in Figure 6.3 has $\xi = 80$, $\bar{u} = 1$, $\Delta x = 1/200$, and $\bar{u}\,\Delta t/\Delta x = 0.8$, giving a group velocity of 0.9712 rather than the correct advection speed of 1. At time 10 this predicts the wave packet will be lagging the correct location by a distance of about 0.288, which agrees well with what is seen in the figure.

For data such as that used in Figure 6.1, dispersion means that the high-frequency components of data such as the discontinuity will travel substantially more slowly than the lower-frequency components, since the group velocity is less than $\bar{u}$ for all wave numbers and falls with increasing ξ. As a result the numerical result can be expected to develop a train of oscillations *behind* the peak, with the high wave numbers lagging farthest behind the correct location.

If we retain one more term in the modified equation for the Lax–Wendroff method, we find that the Q_i^n are fourth-order accurate solutions to an equation of the form

$$v_t + \bar{u} v_x = \frac{1}{6}\bar{u}(\Delta x)^2(\nu^2-1)v_{xxx} - \epsilon v_{xxxx}, \tag{8.46}$$

where the ϵ in the fourth-order dissipative term is $\mathcal{O}(\Delta x^3)$ and positive when the stability bound holds. This higher-order dissipation causes the highest wave numbers to be damped, so that there is a limit to the oscillations seen in practice.

Note that the dominant new term in the modified equation corresponds to the dominant term in the local truncation error for each of these methods. Compare (8.45) with (8.16), for example.

8.6.3 Beam–Warming Method

The second-order Beam–Warming method (6.7) has a modified equation similar to that of the Lax–Wendroff method,

$$v_t + \bar{u} v_x = \frac{1}{6}\bar{u}(\Delta x)^2(2 - 3\nu + \nu^2)v_{xxx}. \tag{8.47}$$

In this case the group velocity is greater than $\bar{u}$ for all wave numbers in the case $0 < \nu < 1$, so that the oscillations move ahead of the main hump. This can be observed in Figure 6.1, where $\nu = \bar{u}\,\Delta t/\Delta x = 0.8$ was used. If $1 < \nu < 2$, then the group velocity is less than $\bar{u}$ and the oscillations will fall behind.

8.7 Accuracy Near Discontinuities

In the previous section we derived the modified equation for various numerical methods, a PDE that models the behavior of the numerical solution. This equation was derived using Taylor series expansion, and hence is based on the assumption of smoothness, but it turns out that the modified equation is often a good model even when the true solution of the original hyperbolic problem contains discontinuities. This is because the modified equation typically contains diffusive terms that cause the discontinuities to be immediately smeared out, as also happens with the numerical solution, and so the solution we are studying is smooth and the Taylor series expansion is valid.

Figure 8.3 shows a simple example in which the upwind method has been applied to the scalar advection equation $q_t + \bar{u} q_x = 0$ with discontinuous data $\mathring{q}(x)$ having the value 2 for $x < 0$ and 0 for $x > 0$. The parameters $\bar{u} = 1$, $\Delta x = 0.05$, and $\Delta t = 0.04$ were used giving a Courant number $\nu = 0.8$. The dashed line is the true solution to this equation, $q(x, t) = \mathring{q}(x - \bar{u}t)$. The solid line is the exact solution to the modified equation (8.44). This advection–diffusion equation can be solved exactly to yield

$$v(x, t) = \operatorname{erfc}\left(\frac{x - \bar{u}t}{\sqrt{4\beta t}}\right), \tag{8.48}$$

where

$$\beta = \frac{1}{2}\bar{u}\,\Delta x(1 - \nu) \tag{8.49}$$

is the diffusion coefficient from (8.44), and the *complementary error function* erfc is defined by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-z^2}\, dz. \tag{8.50}$$

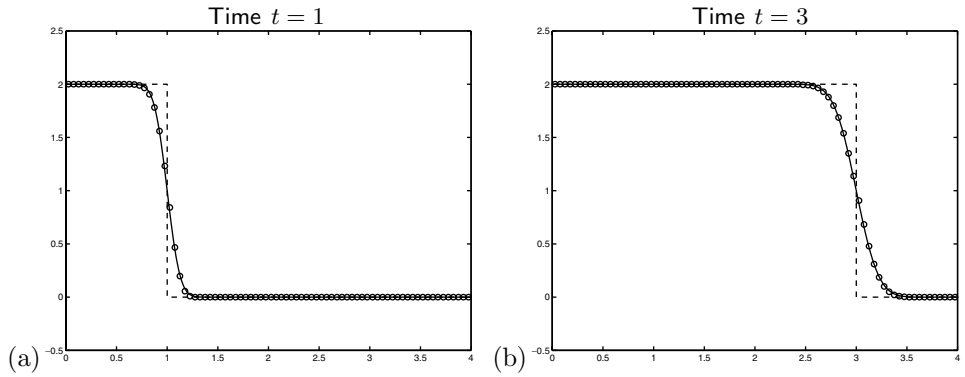

Fig. 8.3. Dashed line: exact solution to the advection equation. Points: numerical solution obtained with the upwind method. Solid line: exact solution to the modified equation (8.44). (a) At time $t = 1$. (b) At time $t = 3$. `[book/chap8/modeqn]`

The numerical solution to the advection equation obtained using the upwind method, marked by the symbols in Figure 8.3, is well approximated by the exact solution to the modified equation.

It follows that we can use the modified equation to give us some insight into the expected accuracy of the upwind method on this problem. Comparing $v(x, t)$ from (8.48) to the true solution $q(x, t) = 2H(\bar{u}t - x)$, it is possible to show that the 1-norm of the difference is

$$\begin{aligned} \|q(\cdot, t) - v(\cdot, t)\|_1 &= 2\int_0^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{4\beta t}}\right) dx \\ &= 2\sqrt{4\beta t}\int_0^\infty \operatorname{erfc}(z)\, dz \\ &= C_1\sqrt{\beta t} \end{aligned} \tag{8.51}$$

for some constant C_1 independent of β and t. Since β is given by (8.49), this gives

$$\|q(\cdot, t) - v(\cdot, t)\|_1 \approx C_2\sqrt{\Delta x\, t} \tag{8.52}$$

as $\Delta x \to 0$ with $\Delta t/\Delta x$ fixed. This indicates that the 1-norm of the error decays only like $(\Delta x)^{1/2}$ even though the method is formally "first-order accurate" based on the local truncation error, which is valid only for smooth solutions.

This informal analysis only gives an indication of the accuracy one might expect from a first-order method on a problem with a discontinuous solution. More detailed error analysis of numerical methods for discontinuous solutions (to nonlinear scalar equations) can be found, for example, in [251], [316], [339], [390], [436], [438].

Exercises

8.1. Consider the centered method (4.19) for the scalar advection equation $q_t + \bar{u}q_x = 0$. Apply von Neumann analysis to show that this method is unstable in the 2-norm for any fixed $\Delta t/\Delta x$.

8.2. Following the proof of Section 8.3.4, show that the upwind method (4.30) is stable provided the CFL condition (4.32) is satisfied.

8.3. Consider the equation

$$q_t + \bar{u} q_x = aq, \qquad q(x,0) = \mathring{q}(x),$$

with solution $q(x,t) = e^{at}\mathring{q}(x - \bar{u}t)$.

(a) Show that the method

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right) + \Delta t\, a Q_i^n$$

is first-order accurate for this equation by computing the local truncation error.

(b) Show that this method is Lax–Richtmyer-stable in the 1-norm provided $|\bar{u}\,\Delta t/\Delta x| \le 1$, by showing that a bound of the form (8.23) holds. Note that when $a > 0$ the numerical solution is growing exponentially in time (as is the true solution) but the method is stable and convergent at any fixed time.

(c) Show that this method is TVB. Is it TVD?

8.4. Show that a method $Q^{n+1} = \mathcal{N}(Q^n)$ that is contractive in the 1-norm (L^1-contractive), so that $\|\mathcal{N}(P) - \mathcal{N}(Q)\|_1 \le \|P - Q\|_1$, must also be TVD, so $\mathrm{TV}(\mathcal{N}(Q)) \le \mathrm{TV}(Q)$. *Hint:* Define the grid function P by $P_i = Q_{i-1}$.

8.5. Prove Harten's Theorem 6.1. Hint: Note that

$$\begin{aligned} Q_{i+1}^{n+1} - Q_i^{n+1} = \left(1 - C_i^n - D_i^n\right)\left(Q_{i+1}^n - Q_i^n\right) + D_{i+1}^n\left(Q_{i+2}^n - Q_{i+1}^n\right) \\ + C_{i-1}^n\left(Q_i^n - Q_{i-1}^n\right). \end{aligned}$$

Sum $|Q_{i+1}^{n+1} - Q_i^{n+1}|$ over i and use the nonnegativity of each coefficient, as in the stability proof of Section 8.3.4.

8.6. Use the method of Section 8.3.4 to show that the method (4.64) is stable in the 1-norm for $\Delta x \le \bar{u}\,\Delta t \le 2\Delta x$.

8.7. View (8.41) as a numerical method for the equation (8.44). Compute the local truncation error, and verify that it is $\mathcal{O}(\Delta t^2)$.

8.8. Derive the modified equation (8.45) for the Lax–Wendroff method.

8.9. Determine the modified equation for the centered method (4.19), and show that the diffusion coefficient is always negative and this equation is hence ill posed. Recall that the method (4.19) is unstable for all fixed $\Delta t/\Delta x$.