# 12

# *Finite Volume Methods for Nonlinear Scalar Conservation Laws*

We now turn to the development of finite volume methods for nonlinear conservation laws. We will build directly on what has already been developed for linear systems in Chapter 4, concentrating in this chapter on scalar equations, although much of what is developed will also apply to nonlinear systems of equations. In Chapter 15 we consider some additional issues arising for systems of equations, particularly the need for efficient *approximate* Riemann solvers in Section 15.3.

Nonlinearity introduces a number of new difficulties not seen for the linear problem. Stability and convergence theory are more difficult than in the linear case, particularly in that we are primarily interested in discontinuous solutions involving shock waves. This theory is taken up in Section 12.10. Moreover, we must ensure that we are converging to the *correct* weak solution of the conservation law, since the weak solution may not be unique. This requires that the numerical method be consistent with a suitable entropy condition; see Section 12.11.

For a nonlinear conservation law $q_t + f(q)_x = 0$ it is very important that the method be in *conservation form*, as described in Section 4.1,

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+1/2}^n - F_{i-1/2}^n\right), \tag{12.1}$$

in order to insure that weak solutions to the conservation law are properly approximated. Recall that this form is derived directly from the integral form of the conservation laws, which is the correct equation to model when the solution is discontinuous. In Section 12.9 an example is given to illustrate that methods based instead on the quasilinear form $q_t + f'(q)q_x = 0$ may be accurate for smooth solutions but may completely fail to approximate a weak solution when the solution contains shock waves.

## 12.1 Godunov's Method

Recall from Section 4.11 that Godunov's method is obtained by solving the Riemann problem between states $Q_{i-1}^n$ and $Q_i^n$ in order to determine the flux $F_{i-1/2}^n$ as

$$F_{i-1/2}^n = f\left(Q_{i-1/2}^{\Downarrow}\right).$$

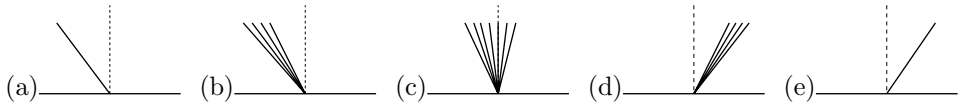


Fig. 12.1. Five possible configurations for the solution to a scalar Riemann problem between states $Q_{i-1}$ and $Q_i$, shown in the $x$–$t$ plane: (a) left-going shock, $Q^{\Downarrow}_{i-1/2} = Q_i$; (b) left-going rarefaction, $Q^{\Downarrow}_{i-1/2} = Q_i$; (c) transonic rarefaction, $Q^{\Downarrow}_{i-1/2} = q_s$; (d) right-going rarefaction, $Q^{\Downarrow}_{i-1/2} = Q_{i-1}$; (e) right-going shock, $Q^{\Downarrow}_{i-1/2} = Q_{i-1}$.

The value $Q^{\Downarrow}_{i-1/2} = q^{\Downarrow}(Q^n_{i-1}, Q^n_i)$ is the value obtained along the ray $x \equiv x_{i-1/2}$ in this Riemann solution. This value is constant for $t > t_n$, since the Riemann solution is a similarity solution. To keep the notation less cluttered, we will often drop the superscript $n$ on $Q$ below.

To begin, we assume the flux function $f(q)$ is *convex* (or concave), i.e., $f''(q)$ does not change sign over the range of $q$ of interest. Then the Riemann solution consists of a single shock or rarefaction wave. (See Section 16.1.3 for the nonconvex case.) For a scalar conservation law with a convex flux function there are five possible forms that the Riemann solution might take, as illustrated in Figure 12.1. In most cases the solution $Q^{\Downarrow}_{i-1/2}$ is either $Q_i$ (if the solution is a shock or rarefaction wave moving entirely to the left, as in Figure 12.1(a) or (b)), or $Q_{i-1}$ (if the solution is a shock or rarefaction wave moving entirely to the right, as in Figure 12.1(d) or (e)).

The only case where $Q^{\Downarrow}_{i-1/2}$ has a different value than $Q_i$ or $Q_{i-1}$ is if the solution consists of a rarefaction wave that spreads partly to the left and partly to the right, as shown in Figure 12.1(c). Suppose for example that $f''(q) > 0$ everywhere, in which case $f'(q)$ is increasing with $q$, so that a rarefaction wave arises if $Q_{i-1} < Q_i$. In this case the situation shown in Figure 12.1(c) occurs only if

$$Q_{i-1} < q_s < Q_i,$$

where $q_s$ is the (unique) value of $q$ for which $f'(q_s) = 0$. This is called the *stagnation point*, since the value $q_s$ propagates with velocity 0. It is also called the *sonic point*, since in gas dynamics the eigenvalues $u \pm c$ can take the value 0 only when the fluid speed $|u|$ is equal to the sound speed $c$. The solution shown in Figure 12.1(c) is called a *transonic rarefaction* since in gas dynamics the fluid is accelerated from a subsonic velocity to a supersonic velocity through such a rarefaction. In a transonic rarefaction the value along $x/t = x_{i-1/2}$ is simply $q_s$.

For the case $f''(q) > 0$ we thus see that the Godunov flux function for a convex scalar conservation law is

$$F^n_{i-1/2} = \begin{cases} f(Q_{i-1}) & \text{if } Q_{i-1} > q_s \text{ and } s > 0, \\ f(Q_i) & \text{if } Q_i < q_s \text{ and } s < 0, \\ f(q_s) & \text{if } Q_{i-1} < q_s < Q_i. \end{cases} \tag{12.2}$$

Here $s = [f(Q_i) - f(Q_{i-1})]/(Q_i - Q_{i-1})$ is the shock speed given by (11.21).

Note in particular that if $f'(q) > 0$ for both $Q_{i-1}$ and $Q_i$ then $F^n_{i-1/2} = f(Q_{i-1})$ and Godunov's method reduces to the *first-order upwind* method

$$Q^{n+1}_i = Q_i - \frac{\Delta t}{\Delta x}[f(Q_i) - f(Q_{i-1})]. \tag{12.3}$$

The natural upwind method is also obtained if $f'(q) < 0$ for both values of $Q$, involving one-sided differences in the other direction. Only in the case where $f'(q)$ changes sign between $Q_{i-1}$ and $Q_i$ is the formula more complicated, as we should expect, since the "upwind direction" is ambiguous in this case and information must flow both ways.

The formula (12.2) can be written more compactly as

$$F^n_{i-1/2} = \begin{cases} \min\limits_{Q_{i-1} \le q \le Q_i} f(q) & \text{if } Q_{i-1} \le Q_i, \\ \max\limits_{Q_i \le q \le Q_{i-1}} f(q) & \text{if } Q_i \le Q_{i-1}, \end{cases} \tag{12.4}$$

since the stagnation point $q_s$ is the global minimum or maximum of $f$ in the convex case. This formula is valid also for the case $f''(q) < 0$ and even for nonconvex fluxes, in which case there may be several stagnation points at each maximum and minimum of $f$ (see Section 16.1).

Note that there is one solution structure not illustrated in Figure 12.1, a stationary shock with speed $s = 0$. In this case the value $Q^{\Downarrow}_{i-1/2}$ is ambiguous, since the Riemann solution is discontinuous along $x = x_{i-1/2}$. However, if $s = 0$ then $f(Q_{i-1}) = f(Q_i)$ by the Rankine–Hugoniot condition, and so $F^n_{i-1/2}$ is still well defined and the formula (12.4) is still valid.

## 12.2 Fluctuations, Waves, and Speeds

Godunov's method can be implemented in our standard form

$$Q^{n+1}_i = Q_i - \frac{\Delta t}{\Delta x}\left(\mathcal{A}^+ \Delta Q_{i-1/2} + \mathcal{A}^- \Delta Q_{i+1/2}\right) \tag{12.5}$$

if we define the fluctuations $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ by

$$\begin{aligned} \mathcal{A}^+ \Delta Q_{i-1/2} &= f(Q_i) - f\left(Q^{\Downarrow}_{i-1/2}\right), \\ \mathcal{A}^- \Delta Q_{i-1/2} &= f\left(Q^{\Downarrow}_{i-1/2}\right) - f(Q_{i-1}). \end{aligned} \tag{12.6}$$

In order to define high-resolution correction terms, we also wish to compute a wave $\mathcal{W}_{i-1/2}$ and speed $s_{i-1/2}$ resulting from this Riemann problem. The natural choice is

$$\begin{aligned} \mathcal{W}_{i-1/2} &= Q_i - Q_{i-1}, \\ s_{i-1/2} &= \begin{cases} [f(Q_i) - f(Q_{i-1})]/(Q_i - Q_{i-1}) & \text{if } Q_{i-1} \ne Q_i, \\ f'(Q_i) & \text{if } Q_{i-1} = Q_i, \end{cases} \end{aligned} \tag{12.7}$$

although the value of $s_{i-1/2}$ is immaterial when $Q_{i-1} = Q_i$. The speed chosen is the Rankine–Hugoniot shock speed (11.21) for this data. If the Riemann solution is a shock wave, this is clearly the right thing to do. If the solution is a rarefaction wave, then the wave is not simply a jump discontinuity propagating at a single speed. However, this is still a suitable definition of the wave and speed to use in defining correction terms that yield second-order accuracy in the smooth case. This can be verified by a truncation-error analysis of the resulting method; see Section 15.6. Note that when a smooth solution is being approximated, we expect $\mathcal{W}_{i-1/2} = \mathcal{O}(\Delta x)$, and there is very little spreading of the rarefaction wave in any case. Moreover, a wave consisting of this jump discontinuity propagating with speed $s_{i-1/2}$

does define a weak solution to the Riemann problem, although it is an expansion shock that does not satisfy the entropy condition. However, provided it is not a transonic rarefaction, the same result will be obtained in Godunov's method whether we use the entropy-satisfying rarefaction wave or an expansion shock as the solution to the Riemann problem. When the wave lies entirely within one cell, the cell average is determined uniquely by conservation and does not depend on the structure of the particular weak solution chosen. We can compute the fluctuations $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ using

$$\begin{aligned} \mathcal{A}^{+}\Delta Q_{i-1/2} &= s^{+}_{i-1/2}\mathcal{W}_{i-1/2}, \\ \mathcal{A}^{-}\Delta Q_{i-1/2} &= s^{-}_{i-1/2}\mathcal{W}_{i-1/2} \end{aligned} \tag{12.8}$$

in place of (12.6), provided that $Q^{\Downarrow}_{i-1/2} = Q_{i-1}$ or $Q_i$. Only in the case of a transonic rarefaction is it necessary to instead use the formulas (12.6) with $Q^{\Downarrow}_{i-1/2} = q_s$.

## 12.3 Transonic Rarefactions and an Entropy Fix

Note that if we were to always use (12.8), even for transonic rarefactions, then we would still be applying Godunov's method using an exact solution to the Riemann problem; the Rankine–Hugoniot conditions are always satisfied for the jump and speed determined in (12.7). The only problem is that the entropy condition would not be satisfied and we would be using the wrong solution. For this reason the modification to $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ required in the transonic case is often called an *entropy fix*.

This approach is used in the Riemann solver `[claw/book/chap12/efix/rp1.f]`. The wave and speed are first calculated, and the fluctuations are set using (12.8). Only in the case of a transonic rarefaction is the entropy fix applied to modify the fluctuations. If $f'(Q_{i-1}) < 0 < f'(Q_i)$ then the fluctuations in (12.8) are replaced by

$$\begin{aligned} \mathcal{A}^{+}\Delta Q_{i-1/2} &= f(Q_i) - f(q_s), \\ \mathcal{A}^{-}\Delta Q_{i-1/2} &= f(q_s) - f(Q_{i-1}). \end{aligned} \tag{12.9}$$

We will see in Section 15.3 that this approach generalizes to nonlinear systems of equations in a natural way. An *approximate Riemann solver* can often be used that gives an approximate solution involving a finite set of waves that are jump discontinuities $\mathcal{W}^{p}_{i-1/2}$ propagating at some speeds $s^{p}_{i-1/2}$. These are used to define fluctuations and also high-resolution correction terms, as indicated in Section 6.15. A check is then performed to see if any of the waves should really be transonic rarefactions, and if so, the fluctuations are modified by performing an entropy fix.

For the scalar equation this approach may seem a rather convoluted way to specify the true Godunov flux, which is quite simple to determine directly. For nonlinear systems, however, it is generally too expensive to determine the rarefaction wave structure exactly and this approach is usually necessary. This is discussed further in Section 15.3, where various entropy fixes are presented.

Dealing with transonic rarefactions properly is an important component in the development of successful methods. This is illustrated in Figure 12.2, which shows several computed solutions to Burgers' equation (11.13) at time $t = 1$ for the same set of data, a Riemann problem with $u_l = -1$ and $u_r = 2$. (Note that for Burgers' equation the sonic point is at

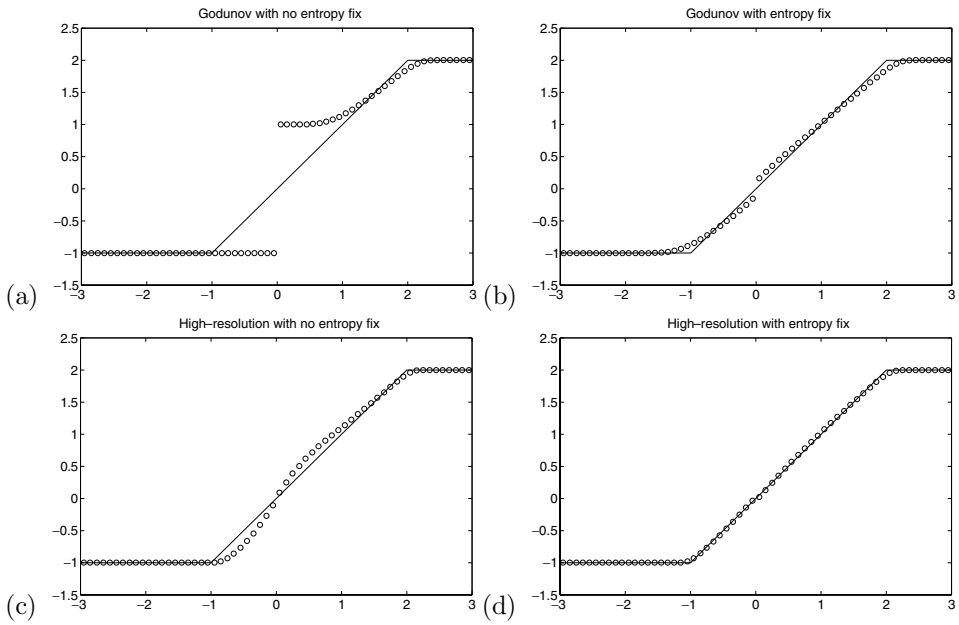


Fig. 12.2. The solid line is the entropy-satisfying solution to Burgers' equation with a transonic rarefaction wave. The circles show computed solutions. (a) Godunov's method using expansion-shock solutions to each Riemann problem. This method converges to a weak solution that does not satisfy the entropy condition. (b) Godunov's method using entropy-satisfying solutions to each Riemann problem. (c) High-resolution corrections added to the expansion-shock Godunov method. (d) High-resolution corrections added to the entropy-satisfying Godunov method. `[claw/book/chap12/efix]`

$u_s = 0$.) The top row shows results obtained with Godunov's method and the bottom row shows results with the high-resolution method, using the MC limiter. The plots on the left were obtained using (12.8) everywhere, with no entropy fix. The plots on the right were obtained using the entropy fix, which means that the fluctuations were redefined at a single grid interface each time step, the one for which $U^n_{i-1} < 0$ and $U^n_i > 0$. This modification at a single grid interface makes a huge difference in the quality of the results. In particular, the result obtained using Godunov's method with the expansion-shock Riemann solution looks entirely wrong. In fact it is a reasonable approximation to a weak solution to the problem, the function

$$u(x,t) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } 0 < x \le t, \\ x/t & \text{if } t \le x \le 2t, \\ 2 & \text{if } x \ge 2t. \end{cases}$$

This solution consists of an entropy-violating stationary shock at $x = 0$ and also a rarefaction wave. If the grid is refined, the computed solution converges nicely to this weak solution. However, this is not the physically relevant vanishing-viscosity solution that we had hoped to compute.

When the correct rarefaction-wave solution to each Riemann problem is used (i.e., the entropy fix (12.9) is employed), Godunov's method gives a result that is much closer to the weak solution we desire. There is still a small expansion shock visible in Figure 12.2(b)

near $x = 0$, but this is of magnitude $\mathcal{O}(\Delta x)$ and vanishes as the grid is refined. This feature (sometimes called an *entropy glitch*) is a result of Godunov's method lacking sufficient numerical viscosity when the wave speed is very close to zero. See [160] for one analysis of this. In Section 12.11.1 we will prove that Godunov's method converges to the correct solution provided that transonic rarefactions are properly handled.

Adding in the high-resolution correction terms (as discussed in Section 12.8) produces better results, even when the entropy fix is not used (Figure 12.2(c)), and convergence to the proper weak solution is obtained. Even in this case, however, better results are seen if the first-order flux for the transonic rarefaction is properly computed, as shown in Figure 12.2(d).

## 12.4 Numerical Viscosity

The weak solution seen in Figure 12.2(a), obtained with Godunov's method using the expansion shock solution to each Riemann problem, contains a portion of the correct rarefaction wave along with a stationary expansion shock located at $x = 0$. Why does it have this particular structure? In the first time step a Riemann problem with $u_l = -1$ and $u_r = 2$ is solved, resulting in an expansion shock with speed $\frac{1}{2}(u_l + u_r) = \frac{1}{2}$. This wave propagates a distance $\frac{1}{2}\Delta t$ and is then averaged onto the grid, resulting in some smearing of the initial discontinuity. The *numerical viscosity* causing this smearing acts similarly to the *physical viscosity* of the viscous Burgers equation (11.14), and tends to produce a rarefaction wave. However, unlike the viscosity of fixed magnitude $\epsilon$ appearing in (11.14), the magnitude of the numerical viscosity depends on the local Courant number $s_{i-1/2}\Delta t/\Delta x$, since it results from the averaging process (where $s_{i-1/2}$ is the Rankine–Hugoniot shock speed for the data $Q_{i-1}$ and $Q_i$). In particular, if $s_{i-1/2} = 0$, then there is no smearing of the discontinuity and no numerical viscosity. For Burgers' equation this happens whenever $Q_{i-1} = -Q_i$, in which case the expansion-shock weak solution is stationary. The solution shown in Figure 12.2(a) has just such a stationary shock.

This suggests that another way to view the entropy fix needed in the transonic case is as the addition of extra numerical viscosity in the neighborhood of such a point. This can be examined further by noting that the fluctuations (12.8) result in the numerical flux function

$$F_{i-1/2} = \frac{1}{2}\big[f(Q_{i-1}) + f(Q_i) - |s_{i-1/2}|(Q_i - Q_{i-1})\big], \tag{12.10}$$

as in the derivation of (4.61). Recall from Section 4.14 that this is the central flux (4.18) with the addition of a viscous flux term. However, when $s_{i-1/2} = 0$ this viscous term disappears. This viewpoint is discussed further in Section 15.3.5 where a variety of entropy fixes for nonlinear systems are discussed.

## 12.5 The Lax–Friedrichs and Local Lax–Friedrichs Methods

The Lax-Friedrichs (LxF) method was introduced in Section 4.6. The flux function (4.21) for this method,

$$F_{i-1/2} = \frac{1}{2}[f(Q_{i-1}) + f(Q_i) - a(Q_i - Q_{i-1})], \tag{12.11}$$

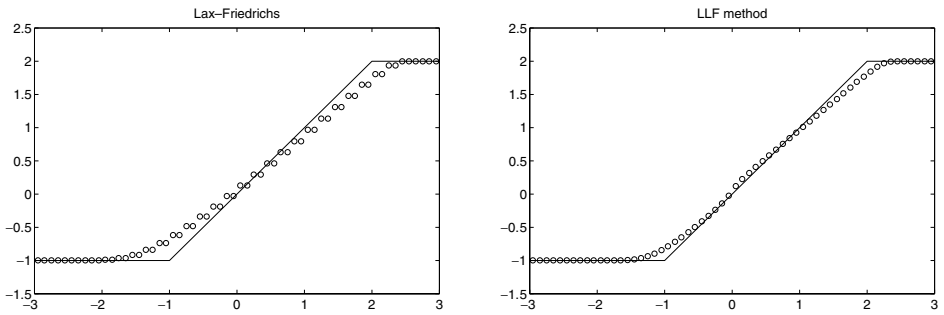


Fig. 12.3. The solid line is the entropy-satisfying solution to Burgers' equation with a transonic rarefaction wave. The symbols show computed solutions. (a) Lax–Friedrichs method. (b) Local Lax–Friedrichs (LLF) method. `[claw/book/chap12/llf]`

has a numerical viscosity $a = \Delta x/\Delta t$ with a fixed magnitude that does not vanish near a sonic point. As a result, this method always converges to the correct vanishing viscosity solution as the grid is refined; see Section 12.7.

If the LxF method is applied to the same transonic rarefaction problem considered in Figure 12.2, we obtain the results shown in Figure 12.3(a). Note that this method is more dissipative than Godunov's method. It also exhibits a curious stair-step pattern in which $Q_{2j} = Q_{2j+1}$ for each value of $j$. This results from the fact that the formula (4.20) for $Q_i^{n+1}$ involves only $Q_{i-1}^n$ and $Q_{i+1}^n$, so there is a decoupling of even and odd grid points. With the piecewise constant initial data used in this example, the even and odd points evolve in exactly the same manner, so each solution value appears twice. (See Section 10.5 for a discussion of the LxF method on a staggered grid in which only half the points appear. This viewpoint allows it to be related more directly to Godunov's method.)

An improvement to the LxF method is obtained by replacing the value $a = \Delta x/\Delta t$ in (12.11) by a locally determined value,

$$F_{i-1/2} = \frac{1}{2}\big[f(Q_{i-1}) + f(Q_i) - a_{i-1/2}(Q_i - Q_{i-1})\big], \tag{12.12}$$

where

$$a_{i-1/2} = \max(|f'(q)|) \quad \text{over all } q \text{ between } Q_{i-1} \text{ and } Q_i. \tag{12.13}$$

For a convex flux function this reduces to

$$a_{i-1/2} = \max(|f'(Q_{i-1})|,\ |f'(Q_i)|).$$

This resulting method is *Rusanov's method* [387], though recently it is often called the *local Lax–Friedrichs* (LLF) *method* because it has the same form as the LxF method but the viscosity coefficient is chosen locally at each Riemann problem. It can be shown that this is sufficient viscosity to make the method converge to the vanishing-viscosity solution; see Section 12.7.

Note that if the CFL condition is satisfied (which is a necessary condition for stability), then $|f'(q)|\,\Delta t/\Delta x \le 1$ for each value of $q$ arising in the whole problem, and so

$$|f'(q)| \le \frac{\Delta x}{\Delta t}.$$

Hence using $a = \Delta x/\Delta t$ in the standard LxF method amounts to taking a uniform viscosity that is sufficient everywhere, at the expense of too much smearing in most cases. Figure 12.3(b) shows the results on the same test problem when the LLF method is used.

Another related method is *Murman's method*, in which (12.12) is used with (12.13) replaced by

$$a_{i-1/2} = \left| \frac{f(Q_i) - f(Q_{i-1})}{Q_i - Q_{i-1}} \right|. \tag{12.14}$$

Unlike the LLF scheme, solutions generated with this method may fail to satisfy the entropy condition because $a_{i-1/2}$ vanishes for the case of a stationary expansion shock. In fact, this is exactly the method (12.5) with $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ defined by (12.8), expressed in a different form (see Exercise 12.1).

Note that all these methods are easily implemented in CLAWPACK by taking

$$\begin{aligned} \mathcal{A}^-\Delta Q_{i-1/2} &= \frac{1}{2}\big[f(Q_i) - f(Q_{i-1}) - a_{i-1/2}(Q_i - Q_{i-1})\big], \\ \mathcal{A}^+\Delta Q_{i-1/2} &= \frac{1}{2}\big[f(Q_i) - f(Q_{i-1}) + a_{i-1/2}(Q_i - Q_{i-1})\big], \end{aligned} \tag{12.15}$$

as is done in `[claw/book/chap12/llf/rp1.f]`.

## 12.6 The Engquist–Osher method

We have seen that the first-order method (12.5) with the fluctuations (12.8) can be interpreted as an implementation of Godunov's method in which we always use the shock-wave solution to each Riemann problem, even when this violates the entropy condition. The *Engquist–Osher method* [124] takes the opposite approach and always assumes the solution is a "rarefaction wave", even when this wave must be triple-valued as in Figure 11.4(b). This can be accomplished by setting

$$\begin{aligned} \mathcal{A}^+\Delta Q_{i-1/2} &= \int_{Q_{i-1}}^{Q_i} (f'(q))^+\,dq, \\ \mathcal{A}^-\Delta Q_{i-1/2} &= \int_{Q_{i-1}}^{Q_i} (f'(q))^-\,dq. \end{aligned} \tag{12.16}$$

Here the $\pm$ superscript on $f'(q)$ means the positive and negative part as in (4.40). These fluctuations result in an interface flux $F_{i-1/2}$ that can be expressed in any of the following

ways:

$$\begin{aligned} F_{i-1/2} &= f(Q_{i-1}) + \int_{Q_{i-1}}^{Q_i} (f'(q))^- \, dq \\ &= f(Q_i) - \int_{Q_{i-1}}^{Q_i} (f'(q))^+ \, dq \\ &= \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\int_{Q_{i-1}}^{Q_i} |f'(q)| \, dq. \end{aligned} \tag{12.17}$$

If $f'(q)$ does not change sign between $Q_{i-1}$ and $Q_i$, then one of the fluctuations in (12.16) will be zero and these formulas reduce to the usual upwind fluxes as in (12.2). In the sonic rarefaction case both fluctuations are nonzero and we obtain the desired value $F_{i-1/2} = f(q_s)$ as in (12.2). It is only in the *transonic shock* case, when $f'(Q_{i-1}) > 0 > f'(Q_i)$, that the Engquist–Osher method gives a value different from (12.2). In this case both fluctuations are again nonzero and we obtain

$$F_{i-1/2} = f(Q_{i-1}) + f(Q_i) - f(q_s) \tag{12.18}$$

rather than simply $f(Q_{i-1})$ or $f(Q_i)$. This is because the triple-valued solution of Figure 11.4(b) spans the interface $x_{i-1/2}$ in this case, so that the integral picks up three different values of $f$. This flux is still consistent with the conservation law, however, and by assuming the rarefaction structure, the entropy condition is always satisfied. This approach can be extended to systems of equations to derive approximate Riemann solvers that satisfy the entropy condition, giving the *Osher scheme* [349], [352].

## 12.7 E-schemes

Osher [349] introduced the notion of an *E-scheme* as one that satisfies the inequality

$$\operatorname{sgn}(Q_i - Q_{i-1})\left[F_{i-1/2} - f(q)\right] \le 0 \tag{12.19}$$

for all $q$ between $Q_{i-1}$ and $Q_i$. In particular, Godunov's method with flux $F^G_{i-1/2}$ defined by (12.4) is clearly an E-scheme. In fact it is the limiting case, in the sense that E-schemes are precisely those for which

$$\begin{aligned} F_{i-1/2} &\le F^G_{i-1/2} \quad \text{if} Q_{i-1} \le Q_i, \\ F_{i-1/2} &\ge F^G_{i-1/2} \quad \text{if} Q_{i-1} \ge Q_i. \end{aligned} \tag{12.20}$$

It can be shown that any E-scheme is TVD if the Courant number is sufficiently small (Exercise 12.3). Osher [349] proves that E-schemes are convergent to the entropy-satisfying weak solution. In addition to Godunov's method, the LxF, LLF, and Engquist–Osher methods are all E-schemes. Osher also shows that E-schemes are at most first-order accurate.

## 12.8 High-Resolution TVD Methods

The methods described so far are only first-order accurate and not very useful in their own right. They are, however, used as building blocks in developing certain high-resolution

methods. Godunov's method, as described in Section 12.2, can be easily extended to a high-resolution method of the type developed in Chapter 6,

$$Q_i^{n+1} = Q_i - \frac{\Delta t}{\Delta x}\left(A^+ \Delta Q_{i-1/2} + A^- \Delta Q_{i+1/2}\right) - \frac{\Delta t}{\Delta x}\left(\tilde{F}_{i+1/2} - \tilde{F}_{i-1/2}\right). \tag{12.21}$$

We set

$$\tilde{F}_{i-1/2} = \frac{1}{2}|s_{i-1/2}|\left(1 - \frac{\Delta t}{\Delta x}|s_{i-1/2}|\right)\widetilde{\mathcal{W}}_{i-1/2}, \tag{12.22}$$

just as we have done for the variable-coefficient advection equation in Section 9.3.1. Again $\widetilde{\mathcal{W}}_{i-1/2}$ is a limited version of the wave (12.7) obtained by comparing $\widetilde{\mathcal{W}}_{i-1/2}$ to $\widetilde{\mathcal{W}}_{i-3/2}$ or $\widetilde{\mathcal{W}}_{i-1/2}$, depending on the sign of $s_{i-1/2}$. Note that the method remains conservative with such a modification, which is crucial in solving nonlinear conservation laws (see Section 12.9).

It is also possible to prove that the resulting method will be TVD provided that one of the TVD limiters presented in Section 6.9 is used. This is a very important result, since it means that these methods can be applied with confidence to nonlinear problems with shock waves where we wish to avoid spurious oscillations. Moreover, this TVD property allows us to prove stability and hence convergence of the methods as the grid is refined, as shown in Section 12.12. (Unfortunately, these claims are valid only for *scalar* conservation laws. Extending these ideas to systems of equations gives methods that are often very successful in practice, but for which much less can be proved in general.)

Here we will prove that the limiter methods are TVD under restricted conditions to illustrate the main ideas. (See [160] for more details.) We assume that data is monotone (say nonincreasing) and that $f'(q)$ does not change sign over the range of the data (say $f'(Q_i^n) > 0$). A similar approach can be used near extreme points of $Q^n$ and sonic points, but more care is required, and the formulas are more complicated. We will also impose the time-step restriction

$$\frac{\Delta t}{\Delta x} \max |f'(q)| < \frac{1}{2}, \tag{12.23}$$

although this can also be relaxed to the usual CFL limit of 1 with some modification of the method.

The main idea is to again use the REA Algorithm 4.1 to interpret the high-resolution method. The first step is to reconstruct the piecewise linear function $\tilde{q}^n(x, t_n)$ from the cell averages $Q_i^n$. This is where the limiters come into play, and this reconstruction does not increase the total variation of the data. The second step is to evolve the conservation law with this data. If we were to solve the conservation law *exactly* and then average onto the grid, then the resulting method would clearly be TVD, because the exact solution operator for a scalar conservation law is TVD (and so is the averaging process). But unlike the methods we developed in Chapter 6 for the advection equation, we are not able to solve the original conservation law exactly in step 2 of Algorithm 4.1 (except in the case of zero slopes, where Godunov's method does this). In principle one could do so also for more general slopes, but the resulting correction formula would be much more complicated than (12.22). However, the formula (12.22) can be interpreted as what results from solving a slightly different conservation law exactly and then averaging onto the grid. Exact solutions of this modified conservation law also have the TVD property, and it follows that the method is TVD.

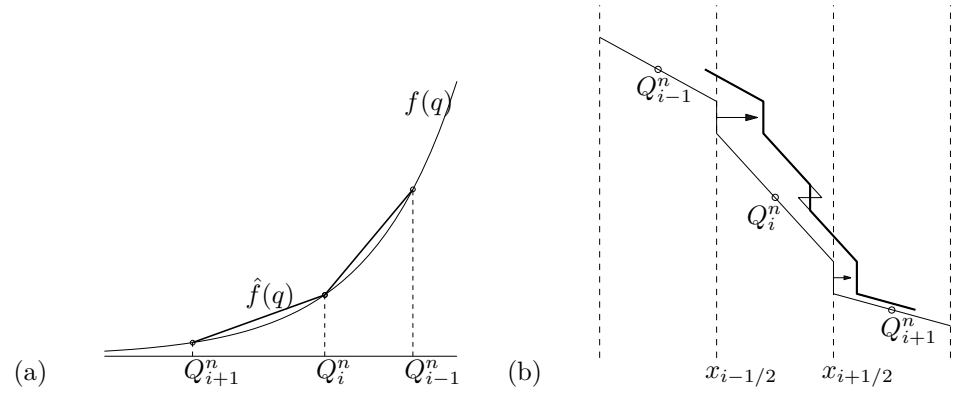


Fig. 12.4. (a) The flux function $f(q)$ is approximated by a piecewise linear function $\hat{f}(q)$. (b) The piecewise linear data $\tilde{q}^n(x, t_n)$ is evolved by solving the conservation law with flux $\hat{f}(q)$. The result is the heavy line. Note that a shock forms near the value $Q^n_i$ where the characteristic velocity $\hat{f}'(q)$ is discontinuous.

The modified conservation law used in the time step from $t_n$ to $t_{n+1}$ is obtained by replacing the flux function $f(q)$ by a piecewise linear function $\hat{f}(q)$ that interpolates the values $(Q^n_i, f(Q^n_i))$, as shown in Figure 12.4(a). Then in step 2 we solve the conservation law with this flux function to evolve $\tilde{q}^n(x, t_n)$, as shown in Figure 12.4(b). This flux is still nonlinear, but the nonlinearity has been concentrated at the points $Q^n_i$. Shocks form immediately at the points $x_i$ (the midpoints of the grid cells), but because of the time-step restriction (12.23), these shocks do not reach the cell boundary during the time step.

Near each interface $x_{i-1/2}$ the data lies between $Q^n_{i-1}$ and $Q^n_i$, and so the flux function is linear with constant slope $s_{i-1/2} = [f(Q^n_i) - f(Q^n_{i-1})]/(Q^n_i - Q^n_{i-1})$, as arises from the piecewise linear interpolation of $f$. Hence the conservation law with flux $\hat{f}$ behaves locally like the scalar advection equation with velocity $s_{i-1/2}$. This is exactly the velocity that appears in the updating formula (12.22), and it can be verified that this method produces the correct cell averages at the end of the time step for this modified conservation law.

In this informal analysis we have assumed the data is monotone near $Q^n_i$. The case where $Q^n_i$ is a local extreme point must be handled differently, since we would not be able to define a single function $\hat{f}$ in the same manner. However, in this case the slope in cell $\mathcal{C}_i$ is zero if a TVD limiter is used, and we can easily show that the total variation can not increase in this case.

Some more details may be found in Goodman & LeVeque [160]. Recently Morton [332] has performed a more extensive analysis of this type of method, including also similar methods with piecewise quadratic reconstructions as well as methods on nonuniform grids and multidimensional versions.

## 12.9 The Importance of Conservation Form

In Section 4.1 we derived the conservative form of a finite volume method based on the integral form of the conservation law. Using a method in this form guarantees that the discrete solution will be conservative in the sense that (4.8) will be satisfied. For weak solutions involving shock waves, this integral form is more fundamental than the differential equation and forms the basis for the mathematical theory of weak solutions, including the derivation

of the Rankine–Hugoniot conditions (see Section 11.8) that govern the form and speed of shock waves. It thus makes sense that a conservative method based on the integral form might be more successful than other methods based on the differential equation. In fact, we will see that the use of conservative finite volume methods is essential in computing weak solutions to conservation laws. Nonconservative methods can fail, as illustrated below. With conservative methods, one has the satisfaction of knowing that if the method converges to some limiting function as the grid is refined, then this function is a weak solution. This is further explained and proved in Section 12.10 in the form of the *Lax–Wendroff theorem*.

In Section 12.11 we will see that similar ideas can be used to show that the limiting function also satisfies the entropy condition, provided the numerical method satisfies a natural discrete version of the entropy condition.

Consider Burgers' equation $u_t + \frac{1}{2}(u^2)_x = 0$, for example. If $u > 0$ everywhere, then the conservative upwind method (Godunov's method) takes the form

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\left(\frac{1}{2}\left(U_i^n\right)^2 - \frac{1}{2}\left(U_{i-1}^n\right)^2\right). \tag{12.24}$$

On the other hand, using the quasilinear form $u_t + uu_x = 0$, we could derive the nonconservative upwind method

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x} U_i^n\left(U_i^n - U_{i-1}^n\right). \tag{12.25}$$

On smooth solutions, both of these methods are first-order accurate, and they give comparable results. When the solution contains a shock wave, the method (12.25) fails to converge to a weak solution of the conservation law. This is illustrated in Figure 12.5. The conservative method (12.24) gives a slightly smeared approximation to the shock, but it is smeared about the correct location. We can easily see that it must be, since the method has the discrete conservation property (4.8). The nonconservative method (12.25), on the other hand, gives the results shown in Figure 12.5(b). These clearly do not satisfy (4.8), and as the grid is

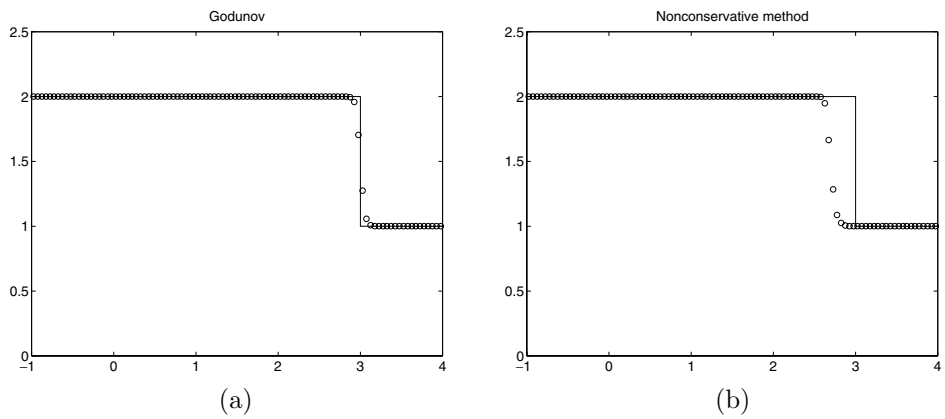


Fig. 12.5. True and computed solutions to a Riemann problem for Burgers' equation with data $u_l = 2$, $u_r = 1$, shown at time $t = 2$: (a) using the conservative method (12.24), (b) using the nonconservative method (12.25). `[claw/book/chap12/nonconservative]`

refined the approximation converges to a discontinuous function that is not a weak solution to the conservation law.

This is not surprising if we recall that it is possible to derive a variety of conservation laws that are equivalent for smooth solutions but have different weak solutions. For example, the equations (11.34) and (11.35) have exactly the same smooth solutions, but the Rankine–Hugoniot condition gives different shock speeds, and hence different weak solutions. Consider a finite difference method that is consistent with one of these equations, say (11.34), using the definition of consistency introduced in Section 8.2 for linear problems (using the local truncation error derived by expanding in Taylor series). Then the method is also consistent with (11.35), since the Taylor series expansion (which assumes smoothness) gives the same result in either case. So the method is consistent with both (11.34) and (11.35), and while we might then expect the method to converge to a function that is a weak solution of both, that is impossible when the two weak solutions differ. Similarly, if we use a nonconservative method based on the quasilinear form, then there is no reason to expect to obtain the correct solution, except in the case of smooth solutions.

Note that the nonconservative method (12.25) can be rewritten as

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\left(\frac{1}{2}\left(U_i^n\right)^2 - \frac{1}{2}\left(U_{i-1}^n\right)^2\right) + \frac{1}{2}\Delta t\,\Delta x\left(\frac{U_i^n - U_{i-1}^n}{\Delta x}\right)^2. \quad (12.26)$$

Except for the final term, this is identical to the conservative method (12.24). The final term approximates the time integral of $\frac{1}{2}\Delta x\,(u_x)^2$. For smooth solutions, where $u_x$ is bounded, the effect of this term can be expected to vanish as $\Delta x \to 0$. For a shock wave, however, it does not. Just as in the derivation of the weak form of the entropy inequality (11.51), this term can give a finite contribution in the limit, leading to a different shock speed.

The final term in (12.26) can also be viewed as a singular source term that is being added to the conservation law, an approximation to a delta function concentrated at the shock. This leads to a change in the shock speed as discussed in Section 17.12. See [203] for further analysis of the behavior of nonconservative methods.

## 12.10 The Lax–Wendroff Theorem

The fact that conservative finite volume methods are based on the integral conservation law suggests that we can hope to correctly approximate discontinuous weak solutions to the conservation law by using such a method. Lax and Wendroff [265] proved that this is true, at least in the sense that *if* the approximation converges to some function $q(x, t)$ as the grid is refined, through some sequence $\Delta t^{(j)}, \Delta x^{(j)} \to 0$, then this function will in fact be a weak solution of the conservation law. The theorem does not guarantee that convergence occurs. For that we need some form of stability, and even then, if there is more than one weak solution, it might be that one sequence of approximations will converge to one weak solution, while another sequence converges to a different weak solution (and therefore a third sequence, obtained for example by merging the first two sequences, will not converge at all!).

Nonetheless, this is a very powerful and important theorem, for it says that we can have confidence in solutions we compute. In practice we typically do not consider a whole

sequence of approximations. Instead we compute a single approximation on one fixed grid. If this solution looks reasonable and has well-resolved discontinuities (an indication that the method is stable and our grid is sufficiently fine), then we can believe that it is in fact a good approximation to *some* weak solution.

Before stating the theorem, we note that it is valid for systems of conservation laws $q_t + f(q)_x = 0$ as well as for scalar equations.

***Theorem 12.1 (Lax and Wendroff [265]).*** *Consider a sequence of grids indexed by* $j = 1, 2, \ldots,$ *with mesh parameters* $\Delta t^{(j)}$, $\Delta x^{(j)} \to 0$ *as* $j \to \infty$. *Let* $Q^{(j)}(x,t)$ *denote the numerical approximation computed with a consistent and conservative method on the* $j$*th grid. Suppose that* $Q^{(j)}$ *converges to a function* $q$ *as* $j \to \infty$, *in the sense made precise below. Then* $q(x,t)$ *is a weak solution of the conservation law.*

The proof of this theorem does not use smoothness of the solution, and so we do not define consistency in terms of Taylor series expansions. Instead we need the form of consistency discussed in Section 4.3.1.

In the statement of this theorem, $Q^{(j)}(x,t)$ denotes a piecewise constant function that takes the value $Q_i^n$ on the space–time mesh cell $(x_{i-1/2}, x_{i+1/2}) \times [t_n, t_{n+1})$. It is indexed by $j$ corresponding to the particular mesh used, with $\Delta x^{(j)}$ and $\Delta t^{(j)}$ both approaching zero as $j \to \infty$. We assume that we have convergence of the function $Q^{(j)}(x,t)$ to $q(x,t)$ in the following sense:

1. Over every bounded set $\Omega = [a,b] \times [0,T]$ in $x$–$t$ space,

$$\int_0^T \int_a^b \left| Q^{(j)}(x,t) - q(x,t) \right| dx\, dt \to 0 \quad \text{as } j \to \infty. \tag{12.27}$$

This is the 1-norm over the set $\Omega$, so we can simply write

$$\left\| Q^{(j)} - q \right\|_{1,\Omega} \to 0 \quad \text{as } j \to \infty. \tag{12.28}$$

2. We also assume that for each $T$ there is an $R > 0$ such that

$$\mathrm{TV}\left(Q^{(j)}(\cdot,t)\right) < R \quad \text{for all } 0 \le t \le T, \qquad j = 1, 2, \ldots, \tag{12.29}$$

where TV denotes the total variation function introduced in Section 6.7.

Lax and Wendroff assumed a slightly different form of convergence, namely that $Q^{(j)}$ converges to $q$ almost everywhere (i.e., except on a set of measure zero) in a uniformly bounded manner. Using the fact that each $Q^{(j)}$ is a piecewise constant function, it can be shown that this requirement is essentially equivalent to (12.28) and (12.29) above. The advantage of assuming (12.28) and (12.29) is twofold: (a) it is these properties that are really needed in the proof, and (b) for certain important classes of methods (e.g., the total variation diminishing methods), it is this form of convergence that we can most directly prove.

*Proof.* We will show that the limit function $q(x,t)$ satisfies the weak form (11.33), i.e., for all $\phi \in C_0^1$,

$$\int_0^\infty \int_{-\infty}^{+\infty} [\phi_t q + \phi_x f(q)]\, dx\, dt = -\int_{-\infty}^{\infty} \phi(x,0) q(x,0)\, dx. \tag{12.30}$$

Let $\phi$ be a $C_0^1$ test function. On the $j$th grid, define the discrete version $\Phi^{(j)}$ by $\Phi_i^{(j)n} = \phi(x_i^{(j)}, t_n^{(j)})$ where $(x_i^{(j)}, t_n^{(j)})$ is a grid point on this grid. Similarly $Q_i^{(j)n}$ denotes the numerical approximation on this grid. To simplify notation, we will drop the superscript $(j)$ below and simply use $\Phi_i^n$ and $Q_i^n$, but remember that $(j)$ is implicitly present, since in the end we must take the limit as $j \to \infty$.

Multiply the conservative numerical method

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\big(F_{i+1/2}^n - F_{i-1/2}^n\big)$$

by $\Phi_i^n$ to obtain

$$\Phi_i^n Q_i^{n+1} = \Phi_i^n Q_i^n - \frac{\Delta t}{\Delta x}\Phi_i^n\big(F_{i+1/2}^n - F_{i-1/2}^n\big). \tag{12.31}$$

This is true for all values of $i$ and $n$ on each grid $j$. If we now sum (12.31) over all $i$ and $n \ge 0$, we obtain

$$\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty} \Phi_i^n\big(Q_i^{n+1} - Q_i^n\big) = -\frac{\Delta t}{\Delta x}\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty} \Phi_i^n\big(F_{i+1/2}^n - F_{i-1/2}^n\big). \tag{12.32}$$

We now use *summation by parts*, which just amounts to recombining the terms in each sum. A simple example is

$$\begin{aligned}\sum_{i=1}^{m} a_i(b_i - b_{i-1}) &= (a_1b_1 - a_1b_0) + (a_2b_2 - a_2b_1) + \cdots + (a_mb_m - a_mb_{m-1})\\ &= -a_1b_0 + (a_1b_1 - a_2b_1) + (a_2b_2 - a_3b_2)\\ &\quad + \cdots + (a_{m-1}b_{m-1} - a_mb_{m-1}) + a_mb_m\\ &= a_mb_m - a_1b_0 - \sum_{i=1}^{m-1}(a_{i+1} - a_i)b_i. \end{aligned}\tag{12.33}$$

Note that the original sum involved the product of $a_i$ with differences of $b$'s, whereas the final sum involves the product of $b_i$ with differences of $a$'s. This is completely analogous to integration by parts, where the derivative is moved from one function to the other. Just as in integration by parts, there are also boundary terms $a_mb_m - a_1b_0$ that arise.

We will use this on both sides of (12.32) (for the $n$-sum on the left and for the $i$-sum on the right). By our assumption that $\phi$ has compact support, $\Phi_i^n = 0$ for $|i|$ or $n$ sufficiently large, and hence the boundary terms at $i = \pm\infty$, $n = \infty$ all drop out. The only boundary term that remains is at $n = 0$, where $t_0 = 0$. This gives

$$-\sum_{i=-\infty}^{\infty} \Phi_i^0 Q_i^0 - \sum_{n=1}^{\infty}\sum_{i=-\infty}^{\infty} \big(\Phi_i^n - \Phi_i^{n-1}\big)Q_i^n = \frac{\Delta t}{\Delta x}\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty} \big(\Phi_{i+1}^n - \Phi_i^n\big)F_{i-1/2}^n.$$

Note that each of these sums is in fact a finite sum, since $\phi$ has compact support. Multiplying by $\Delta x$ and rearranging this equation gives

$$\Delta x\,\Delta t\left[\sum_{n=1}^{\infty}\sum_{i=-\infty}^{\infty}\left(\frac{\Phi_i^n-\Phi_i^{n-1}}{\Delta t}\right)Q_i^n + \sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty}\left(\frac{\Phi_{i+1}^n-\Phi_i^n}{\Delta x}\right)F_{i-1/2}^n\right] = -\Delta x\sum_{i=-\infty}^{\infty}\Phi_i^0 Q_i^0. \quad (12.34)$$

This transformation using summation by parts is completely analogous to the derivation of (11.33) from (11.31).

Now let $j\to\infty$, so that $\Delta t^{(j)}, \Delta x^{(j)}\to 0$ in (12.34). (Recall that all of the symbols in that equation should also be indexed by $(j)$ as we refine the grid.) It is reasonably straightforward, using the 1-norm convergence of $Q^{(j)}$ to $q$ and the smoothness of $\phi$, to show that the term on the top line of (12.34) converges to $\int_0^\infty\int_{-\infty}^\infty \phi_t(x,t)q(x,t)\,dx$ as $j\to\infty$. If we define initial data $Q_i^0$ by taking cell averages of the data $\mathring{q}(x)$, for example, then the right-hand side converges to $-\int_{-\infty}^\infty \phi(x,0)q(x,0)\,dx$ as well.

The remaining term in (12.34), involving $F_{i-1/2}^n$, is more subtle and requires the additional assumptions on $F$ and $Q$ that we have imposed. For a three-point method (such as Godunov's method), we have

$$F_{i-1/2}^n \equiv F_{i-1/2}^{(j)n} = \mathcal{F}\big(Q_{i-1}^{(j)n}, Q_i^{(j)n}\big),$$

and the consistency condition (4.15), with the choice $\bar q = Q_i^{(j)n}$, gives

$$\big|F_{i-1/2}^{(j)n} - f\big(Q_i^{(j)n}\big)\big| \le L\big|Q_i^{(j)n} - Q_{i-1}^{(j)n}\big|, \quad (12.35)$$

where $L$ is the Lipschitz constant for the numerical flux function. Since $Q^{(j)n}$ has bounded total variation, uniformly in $j$, it must be that

$$\big|F_{i-1/2}^{(j)n} - f\big(Q_i^{(j)n}\big)\big| \to 0 \quad \text{as } j\to\infty$$

for almost all values of $i$. Using this and the fact that $Q^{(j)n}$ converges to $q$, it can be shown that

$$\Delta x\,\Delta t\sum_{n=1}^{\infty}\sum_{i=-\infty}^{\infty}\left(\frac{\Phi_{i+1}^n-\Phi_i^n}{\Delta x}\right)F_{i-1/2}^n \to \int_0^\infty\int_{-\infty}^\infty \phi_x(x,t)\,f(q(x,t))\,dx\,dt$$

as $j\to\infty$, which completes the demonstration that (12.34) converges to the weak form (12.30). Since this is true for any test function $\phi\in C_0^1$, we have proved that $q$ is in fact a weak solution. □

For simplicity we assumed the numerical flux $F_{i-1/2}^n$ depends only on the two neighboring values $Q_{i-1}^n$ and $Q_i^n$. However, the proof is easily extended to methods with a wider stencil provided a more general consistency condition holds, stating that the flux function is uniformly Lipschitz-continuous in all values on which it depends.

## 12.11 The Entropy Condition

The Lax–Wendroff theorem does not guarantee that weak solutions obtained using conservative methods satisfy the entropy condition. As we have seen in Section 12.3, some additional care is required to insure that the correct weak solution is obtained.

For some numerical methods, it is possible to show that any weak solution obtained by refining the grid will satisfy the entropy condition. Of course this supposes that we have a suitable entropy condition for the system to begin with, and the most convenient form is typically the entropy inequality introduced in Section 11.14. Recall that this requires a convex scalar entropy function $\eta(q)$ and entropy flux $\psi(q)$ for which

$$\frac{\partial}{\partial t}\eta\left(q(x,t)\right) + \frac{\partial}{\partial x}\psi\left(q(x,t)\right) \le 0 \tag{12.36}$$

in the weak sense, i.e., for which the inequality (11.51) holds for all $\phi \in C_0^1$ with $\phi(x,t) \ge 0$ for all $x, t$:

$$\int_0^\infty \int_{-\infty}^\infty [\phi_t(x,t)\eta(q(x,t)) + \phi_x(x,t)\psi(q(x,t))]\,dx\,dt + \int_{-\infty}^\infty \phi(x,0)\eta(q(x,0))\,dx \ge 0. \tag{12.37}$$

In order to show that the weak solution $q(x,t)$ obtained as the limit of $Q^{(j)}$ satisfies this inequality, it suffices to show that a discrete entropy inequality holds, of the form

$$\eta\big(Q_i^{n+1}\big) \le \eta\big(Q_i^n\big) - \frac{\Delta t}{\Delta x}\big(\Psi_{i+1/2}^n - \Psi_{i-1/2}^n\big). \tag{12.38}$$

Here $\Psi_{i-1/2}^n = \Psi(Q_{i-1}^n, Q_i^n)$, where $\Psi(q_l, q_r)$ is some numerical entropy flux function that must be consistent with $\psi$ in the same manner that we require $F$ to be consistent with $f$. If we can show that (12.38) holds for a suitable $\Psi$, then mimicking the proof of the Lax–Wendroff theorem (i.e., multiplying (12.38) by $\Phi_i^n$, summing over $i$ and $n$, and using summation by parts), we can show that the limiting weak solution $q(x,t)$ obtained as the grid is refined satisfies the entropy inequality (12.37).

### *12.11.1 Entropy Consistency of Godunov's Method*

For Godunov's method we can show that the numerical approximation will always satisfy the entropy condition provided that the Riemann solution used to define the flux at each cell interface satisfies the entropy condition. Recall that we can interpret Godunov's method as an implementation of the REA Algorithm 4.1. The piecewise constant function $\tilde{q}^n(x, t_n)$ is constructed from the data $Q^n$, and the exact solution $\tilde{q}^n(x, t_{n+1})$ to the conservation law is then averaged on the grid to obtain $Q^{n+1}$. What we now require is that the solution $\tilde{q}^n(x,t)$ satisfy the entropy condition. If so, then integrating (12.36) over the rectangle

$(x_{i-1/2}, x_{i+1/2}) \times (t_n, t_{n+1})$ gives

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \eta(\tilde{q}^n(x, t_{n+1}))\, dx \le \int_{x_{i-1/2}}^{x_{i+1/2}} \eta(\tilde{q}^n(x, t_n))\, dx + \int_{t_n}^{t_{n+1}} \psi\big(\tilde{q}^n(x_{i-1/2}, t)\big)\, dt - \int_{t_n}^{t_{n+1}} \psi\big(\tilde{q}^n(x_{i+1/2}, t)\big)\, dt.$$

This is almost what we need. Since $\tilde{q}^n$ is constant along three of the four sides of this rectangle, all integrals on the right-hand side can be evaluated. Doing so, and dividing by $\Delta x$, yields

$$\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \eta(\tilde{q}^n(x, t_{n+1}))\, dx \le \eta\big(Q_i^n\big) - \frac{\Delta t}{\Delta x}\big[\psi\big(Q_{i+1/2}^{\Downarrow}\big) - \psi\big(Q_{i-1/2}^{\Downarrow}\big)\big]. \tag{12.39}$$

Again $Q_{i-1/2}^{\Downarrow}$ represents the value propagating with velocity 0 in the solution of the Riemann problem. If we define the numerical entropy flux by

$$\Psi_{i-1/2}^n = \psi\big(Q_{i-1/2}^{\Downarrow}\big), \tag{12.40}$$

then $\Psi$ is consistent with $\psi$, and the right-hand side of (12.39) agrees with that of (12.38).

The left-hand side of (12.39) is not equal to $\eta(Q_i^{n+1})$, because $\tilde{q}^n$ is not constant in this interval. However, since the entropy function $\eta$ is convex with $\eta''(q) > 0$, we can use *Jensen's inequality*. This states that the value of $\eta$ evaluated at the average value of $\tilde{q}^n$ is less than or equal to the average value of $\eta(\tilde{q}^n)$, i.e.,

$$\eta\left(\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{q}^n(x, t_{n+1})\, dx\right) \le \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \eta(\tilde{q}^n(x, t_{n+1}))\, dx. \tag{12.41}$$

The left-hand side here is simply $\eta(Q_i^{n+1})$, while the right-hand side is bounded by (12.39). Combining (12.39), (12.40), and (12.41) thus gives the desired entropy inequality (12.38).

This shows that weak solutions obtained by Godunov's method satisfy the entropy condition, provided we use entropy-satisfying Riemann solutions at each cell interface. This result is valid not only for scalar conservation laws. It holds more generally for any nonlinear system for which we have an entropy function.

For the special case of a convex scalar conservation law, this simply means that we must use a rarefaction wave when possible rather than an expansion shock in defining the state $Q_{i-1/2}^{\Downarrow}$ used to compute the Godunov flux. However, as we have seen in Section 12.2, this affects the value of $Q_{i-1/2}^{\Downarrow}$ only in the case of a transonic rarefaction. So we conclude that Godunov's method will always produce the vanishing-viscosity solution to a convex scalar conservation law provided that transonic rarefactions are handled properly.

## 12.12 Nonlinear Stability

The Lax–Wendroff theorem presented in Section 12.10 does not say anything about whether the method converges, only that if a sequence of approximations converges, then the limit is a weak solution. To guarantee convergence, we need some form of stability, just as for

linear problems. Unfortunately, the Lax equivalence theorem mentioned in Section 8.3.2 no longer holds, and we cannot use the same approach (which relies heavily on linearity) to prove convergence.

The convergence proof of Section 8.3.1 can be used in the nonlinear case if the numerical method is contractive in some norm. In particular, this is true for the class of *monotone methods*. These are methods with the property that

$$\frac{\partial Q_i^{n+1}}{\partial Q_j^n} \geq 0 \tag{12.42}$$

for all values of $j$. This means that if we increase the value of any $Q_j^n$ at time $t_n$, then the value of $Q_i^{n+1}$ at the next time step cannot decrease as a result. This is suggested by the fact that the true vanishing-viscosity solution of a scalar conservation law has an analogous property: If $\mathring{q}(x)$ and $\mathring{p}(x)$ are two sets of initial data and $\mathring{q}(x) \geq \mathring{p}(x)$ for all $x$, then $q(x,t) \geq p(x,t)$ for all $x$ at later times as well. Unfortunately, this monotone property holds only for certain first-order accurate methods, and so this approach cannot be applied to the high-resolution methods of greatest interest. For more details on monotone methods see, for example, [96], [156], [185], [281].

In this chapter we consider a form of nonlinear stability based on total-variation bounds that allows us to prove convergence results for a wide class of practical TVD or TVB methods. So far, this approach has been completely successful only for scalar problems. For general systems of equations with arbitrary initial data no numerical method has been proved to be stable or convergent in general, although convergence results have been obtained in some special cases (see Section 15.8.2).

### 12.12.1 Convergence Notions

To discuss the convergence of a grid function with discrete values $Q_i^n$ to a function $q(x,t)$, it is convenient to define a piecewise-constant function $Q^{(\Delta t)}(x,t)$ taking the values

$$Q^{(\Delta t)}(x,t) = Q_i^n \quad \text{for } (x,t) \in \left[x_{i-1/2}, x_{i+1/2}\right) \times \left[t_n, t_{n+1}\right). \tag{12.43}$$

We index this function by $\Delta t$ because it depends on the particular grid being used. It should really be indexed by $\Delta x$ as well, but to simplify notation we suppose there is a fixed relation between $\Delta x$ and $\Delta t$ as we refine the grid and talk about convergence as $\Delta t \to 0$. In Section 12.10 a similar sequence of functions was considered and labeled $Q^{(j)}$, corresponding to a a grid with mesh spacing $\Delta t^{(j)}$ and $\Delta x^{(j)}$. The same notation could be used here, but it will be more convenient below to use $\Delta t$ as the index rather than $j$.

One difficulty immediately presents itself when we contemplate the convergence of a numerical method for conservation laws. The global error $Q^{(\Delta t)}(x,t) - q(x,t)$ is not well defined when the weak solution $q$ is not unique. Instead, we measure the global error in our approximation by the distance from $Q^{(\Delta t)}(x,t)$ to the set of *all* weak solutions $\mathcal{W}$,

$$\mathcal{W} = \{q : q(x,t) \text{ is a weak solution to the conservation law}\}. \tag{12.44}$$

To measure this distance we need a norm, for example the 1-norm over some finite time

interval $[0, T]$, denoted by

$$\|v\|_{1,T} = \int_0^T \|v(\cdot, t)\|_1 \, dt$$
$$= \int_0^T \int_{-\infty}^{\infty} |v(x, t)| \, dx \, dt. \tag{12.45}$$

The global error is then defined by

$$\text{dist}\left(Q^{(\Delta t)}, \mathcal{W}\right) = \inf_{w \in \mathcal{W}} \left\| Q^{(\Delta t)} - q \right\|_{1,T}. \tag{12.46}$$

The convergence result we would now like to prove takes the following form:

*If $Q^{(\Delta t)}$ is generated by a numerical method in conservation form, consistent with the conservation law, and if the method is stable in some appropriate sense, then*

$$\text{dist}\left(Q^{(\Delta t)}, \mathcal{W}\right) \to 0 \quad \text{as } \Delta t \to 0.$$

Note that there is no guarantee that $\|Q^{(\Delta t)} - q\|_{1,T} \to 0$ as $\Delta t \to 0$ for any fixed weak solution $q(x, t)$. The computed $Q^{(\Delta t)}$ might be close to one weak solution for one value of the time step $\Delta t$ and close to a completely different weak solution for a slightly smaller value of $\Delta t$. This is of no great concern, since in practice we typically compute only on one particular grid, not a sequence of grids with $\Delta t \to 0$, and what the convergence result tells us is that by taking a fine enough grid, we can be assured of being arbitrarily close to *some* weak solution.

Of course, in situations where there is a unique physically relevant weak solution satisfying some entropy condition, we would ultimately like to prove convergence to this particular weak solution. This can be done if we also know that the method satisfies a discrete form of the entropy condition, such as (12.38). For then we know that any limiting solution obtained by refining the grid must satisfy the entropy condition (see Section 12.11). Since the entropy solution $q(x, t)$ to the conservation law is unique, this can be used to prove that in fact any sequence $Q^{(\Delta t)}$ must converge to this function $q$ as $\Delta t \to 0$.

### *12.12.2 Compactness*

In order to prove a convergence result of the type formulated above for nonlinear problems, we must define an appropriate notion of stability. For nonlinear problems one very useful tool for proving convergence is *compactness*, and so we will take a slight detour to define this concept and indicate its use.

There are several equivalent definitions of a compact set within some normed space. One definition, which describes the most important property of compact sets in relation to our goals of defining stability and proving convergence, is the following.

***Definition 12.1.*** *$\mathcal{K}$ is a compact set in some normed space if any infinite sequence of elements of $\mathcal{K}$, $\{\kappa_1, \kappa_2, \kappa_3, \dots\}$, contains a subsequence that converges to an element of $\mathcal{K}$.*

This means that from the original sequence we can, by selecting certain elements from this sequence, construct a new infinite sequence

$$\{\kappa_{i_1}, \kappa_{i_2}, \kappa_{i_3}, \dots\} \qquad (\text{with } i_1 < i_2 < i_3 < \cdots)$$

that converges to some element $\kappa \in \mathcal{K}$,

$$\|\kappa_{i_j} - \kappa\| \to 0 \qquad \text{as } j \to \infty.$$

The fact that compactness guarantees the existence of convergent subsequences, combined with the Lax–Wendroff theorem 12.1, will give us a convergence proof of the type formulated above.

**Example 12.1.** In the space $\mathbb{R}$ with norm given by the absolute value, any closed interval is a compact set. So, for example, any sequence of real numbers in $[0, 1]$ contains a subsequence that converges to a number between 0 and 1. Of course, there may be several different subsequences one could extract, converging perhaps to different numbers. For example, the sequence

$$\{0, 1, 0, 1, 0, 1, \dots\}$$

contains subsequences converging to 0 and subsequences converging to 1.

**Example 12.2.** In the same space as the previous example, an open interval is *not* compact. For example, the sequence

$$\{1, 10^{-1}, 10^{-2}, 10^{-3}, \dots\}$$

of elements lying in the open interval $(0, 1)$ contains no subsequences convergent to an element of $(0, 1)$. Of course the whole sequence, and hence every subsequence, converges to 0, but this number is not in (0,1).

**Example 12.3.** An unbounded set, e.g., $[0, \infty)$, is *not* compact, since the sequence $\{1, 2, 3, \dots\}$ contains no convergent subsequence.

Generalizing these examples, it turns out that in any finite-dimensional normed linear space, any closed and bounded set is compact. Moreover, these are the only compact sets.

**Example 12.4.** In the $n$-dimensional space $\mathbb{R}^n$ with any vector norm $\|\cdot\|$, the closed ball

$$B_R = \{x \in \mathbb{R}^n : \|x\| \le R\}$$

is a compact set.

### 12.12.3 Function Spaces

Since we are interested in proving the convergence of a sequence of functions $Q^{(\Delta t)}(x, t)$, our definition of stability will require that all the functions lie within some compact set in some normed *function space*. Restricting our attention to the time interval $[0, T]$, the natural function space is the space $L_{1,T}$ consisting of all functions of $x$ and $t$ for which the norm (12.45) is finite,

$$L_{1,T} = \{v : \|v\|_{1,T} < \infty\}.$$

This is an infinite-dimensional space, and so it is not immediately clear what constitutes a compact set in this space. Recall that the *dimension* of a linear space is the number of elements in a basis for the space, and that a *basis* is a linearly independent set of elements with the property that any element can be expressed as a linear combination of the basis elements. Any space with $n$ linearly independent elements has dimension at least $n$.

**Example 12.5.** The space of functions of $x$ alone with finite 1-norm is denoted by $L_1$,

$$L_1 = \{v(x) : \|v\|_1 < \infty\}.$$

This space is clearly infinite-dimensional, since the functions

$$v_j(x) = \begin{cases} 1 & \text{if } j < x < j + 1, \\ 0 & \text{otherwise} \end{cases} \tag{12.47}$$

for $j = 0, 1, 2, \ldots$ are linearly independent, for example.

Unfortunately, in an infinite-dimensional space, a closed and bounded set is not necessarily compact, as the next example shows.

**Example 12.6.** The sequence of functions $\{v_1, v_2, \ldots\}$ with $v_j$ defined by (12.47) all lie in the closed and bounded unit ball

$$B_1 = \{v \in L_1 : \|v\|_1 \le 1\},$$

and yet this sequence has no convergent subsequences.

The difficulty here is that the support of these functions is nonoverlapping and marches off to infinity as $j \to \infty$. We might try to avoid this by considering a set of the form

$$\{v \in L_1 : \|v\|_1 \le R \quad \text{and} \quad \text{Supp}(v) \subset [-M, M]\}$$

for some $R, M > 0$, where $\text{Supp}(v)$ denotes the support of the function $v$, i.e., $\text{Supp}(v) \subset [-M, M]$ means that $v(x) \equiv 0$ for $|x| > M$. However, this set is also not compact, as shown by the sequence of functions $\{v_1, v_2, \ldots\}$ with

$$v_j(x) = \begin{cases} \sin(jx) & \text{if } |x| \le 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

Again this sequence has no convergent subsequences, now because the functions become more and more oscillatory as $j \to \infty$.

### 12.12.4 Total-Variation Stability

In order to obtain a compact set in $L_1$, we will put a bound on the total variation of the functions, a quantity already defined in (6.19) through (6.21). The set

$$\{v \in L_1 : \mathrm{TV}(v) \le R \quad \text{and} \quad \mathrm{Supp}(v) \subset [-M, M]\} \tag{12.48}$$

is a compact set, and any sequence of functions with uniformly bounded total variation and support must contain convergent subsequences. (Note that the 1-norm will also be uniformly bounded as a result, with $\|v\|_1 \le MR$.)

Since our numerical approximations $Q^{(\Delta t)}$ are functions of $x$ and $t$, we need to bound the total variation in both space and time. We define the total variation over $[0, T]$ by

$$\begin{aligned}\mathrm{TV}_T(q) = &\limsup_{\epsilon \to 0} \frac{1}{\epsilon} \int_0^T \int_{-\infty}^{\infty} |q(x+\epsilon, t) - q(x,t)|\, dx\, dt \\ &+ \limsup_{\epsilon \to 0} \frac{1}{\epsilon} \int_0^T \int_{-\infty}^{\infty} |q(x, t+\epsilon) - q(x,t)|\, dx\, dt.\end{aligned} \tag{12.49}$$

It can be shown that the set

$$\mathcal{K} = \{q \in L_{1,T} : \mathrm{TV}_T(q) \le R \text{ and } \mathrm{Supp}(q(\cdot, t)) \subset [-M, M] \ \ \forall t \ \in [0, T]\} \tag{12.50}$$

is a compact set in $L_{1,T}$.

Since our functions $Q^{(\Delta t)}(x, t)$ are always piecewise constant, the definition (12.49) of $\mathrm{TV}_T$ reduces to simply

$$\mathrm{TV}_T\big(Q^{(\Delta t)}\big) = \sum_{n=0}^{T/\Delta t} \sum_{j=-\infty}^{\infty} \big[\Delta t \big| Q_{i+1}^n - Q_i^n\big| + \Delta x \big| Q_i^{n+1} - Q_i^n\big|\big]. \tag{12.51}$$

Note that we can rewrite this in terms of the one-dimensional total variation and 1-norm as

$$\mathrm{TV}_T\big(Q^{(\Delta t)}\big) = \sum_{n=0}^{T/\Delta t} [\Delta t\, \mathrm{TV}(Q^n) + \|Q^{n+1} - Q^n\|_1]. \tag{12.52}$$

***Definition 12.2.*** *We will say that a numerical method is* total-variation-stable*, or simply* TV-stable*, if all the approximations $Q^{(\Delta t)}$ for $\Delta t < \Delta t_0$ lie in some fixed set of the form (12.50) (where $R$ and $M$ may depend on the initial data $\mathring{q}(x)$, the time $T$, and the flux function $f(q)$, but not on $\Delta t$).*

Note that our requirement in (12.50) that $\mathrm{Supp}(q)$ be uniformly bounded over $[0, T]$ is always satisfied for any explicit method if the initial data $\mathring{q}$ has compact support and $\Delta t/\Delta x$ is constant as $\Delta t \to 0$. This follows from the finite speed of propagation for such a method.

The other requirement for TV-stability can be simplified considerably by noting the following theorem. This says that for the special case of functions generated by conservative

numerical methods, it suffices to insure that the *one-dimensional* total variation at each time $t_n$ is uniformly bounded (independent of $n$). Uniform boundedness of $\mathrm{TV}_T$ then follows.

***Theorem 12.2.*** *Consider a conservative method with a Lipschitz-continuous numerical flux $F^n_{i-1/2}$, and suppose that for each initial data $\mathring{q}$ there exist some $\Delta t_0, R > 0$ such that*

$$\mathrm{TV}(Q^n) \le R \qquad \forall n, \Delta t \quad \text{with } \Delta t < \Delta t_0, \quad n\,\Delta t \le T. \tag{12.53}$$

*Then the method is TV-stable.*

To prove this theorem we use the following lemma, which is proved below.

***Lemma 1.*** *If $Q^n$ is generated by a conservative method with a Lipschitz-continuous numerical flux function, then the bound (12.53) implies that there exists $\alpha > 0$ such that*

$$\|Q^{n+1} - Q^n\|_1 \le \alpha\,\Delta t \qquad \forall n, \Delta t \quad \text{with } \Delta t < \Delta t_0, \quad n\,\Delta t \le T. \tag{12.54}$$

*Proof of Theorem 12.2.* Using (12.53) and (12.54) in (12.52) gives

$$\begin{aligned}
\mathrm{TV}_T\big(Q^{(\Delta t)}\big) &= \sum_{n=0}^{T/\Delta t} [\Delta t\ \mathrm{TV}(Q^n) + \|Q^{n+1} - Q^n\|_1] \\
&\le \sum_{n=0}^{T/\Delta t} [\Delta t\ R + \alpha\ \Delta t] \\
&\le \Delta t\,(R+\alpha)T/\Delta t = (R+\alpha)T
\end{aligned}$$

for all $\Delta t < \Delta t_0$, showing that $\mathrm{TV}_T(Q^{(\Delta t)})$ is uniformly bounded as $\Delta t \to 0$. This, together with the finite-speed-of-propagation argument outlined above, shows that all $Q^{(\Delta t)}$ lie in a set of the form (12.50) for all $\Delta t < \Delta t_0$ and the method is TV-stable. □

*Proof of Lemma 1.* Recall that a method in conservation form has

$$Q^{n+1}_i - Q^n_i = \frac{\Delta t}{\Delta x}\big(F^n_{i+1/2} - F^n_{i-1/2}\big)$$

and hence

$$\|Q^{n+1} - Q^n\|_1 = \Delta t \sum_{j=-\infty}^{\infty} \big|F^n_{i+1/2} - F^n_{i-1/2}\big|. \tag{12.55}$$

The flux $F^n_{i-1/2}$ depends on a finite number of values $Q_{i-p}, \ldots, Q_{i+r}$. The bound (12.53) together with the compact support of each $Q^n$ easily gives

$$\big|Q^n_i\big| \le R/2 \qquad \forall i, n \quad \text{with } n\,\Delta t \le T. \tag{12.56}$$

This uniform bound on $Q^n_i$, together with the Lipschitz continuity of the flux function, allows us to derive a bound of the form

$$\big|F^n_{i+1/2} - F^n_{i-1/2}\big| \le K \max_{-p\le j\le r} \big|Q^n_{i+j} - Q^n_{i+j-1}\big|. \tag{12.57}$$

It follows that

$$\left|F^n_{i+1/2} - F^n_{i-1/2}\right| \le K \sum_{j=-p}^{r} \left|Q^n_{i+j} - Q^n_{i+j-1}\right|,$$

and so (12.55) gives

$$\|Q^{n+1} - Q^n\|_1 \le \Delta t\, K \sum_{j=-p}^{r} \sum_{i=-\infty}^{\infty} \left|Q^n_{i+j} - Q^n_{i+j-1}\right|$$

after interchanging sums. But now the latter sum is simply $\mathrm{TV}(Q^n)$ for any value of $j$, and so

$$\begin{aligned}\|Q^{n+1} - Q^n\|_1 &\le \Delta t\, K \sum_{j=-p}^{r} \mathrm{TV}(Q^n) \\ &\le \Delta t\, K(p + r + 1)R,\end{aligned}$$

yielding the bound (12.54). □

We are now set to prove our convergence theorem, which requires TV-stability along with consistency.

***Theorem 12.3.*** *Suppose $Q^{(\Delta t)}$ is generated by a numerical method in conservation form with a Lipschitz continuous numerical flux, consistent with some scalar conservation law. If the method is TV-stable, i.e., if $\mathrm{TV}(Q^n)$ is uniformly bounded for all $n$, $\Delta t$ with $\Delta t < \Delta t_0$, $n\,\Delta t \le T$, then the method is convergent, i.e., $\mathrm{dist}(Q^{(\Delta t)}, \mathcal{W}) \to 0$ as $\Delta t \to 0$.*

*Proof.* To prove this theorem we suppose that the conclusion is false, and obtain a contradiction. If $\mathrm{dist}(Q^{(\Delta t)}, \mathcal{W})$ does not converge to zero, then there must be some $\epsilon > 0$ and some sequence of approximations $\{Q^{(\Delta t_1)}, Q^{(\Delta t_2)}, \dots\}$ such that $\Delta t_j \to 0$ as $j \to \infty$ while

$$\mathrm{dist}\left(Q^{(\Delta t_j)}, \mathcal{W}\right) > \epsilon \quad \text{for all } j. \tag{12.58}$$

Since $Q^{(\Delta t_j)} \in \mathcal{K}$ (the compact set of (12.50)) for all $j$, this sequence must have a convergent subsequence, converging to some function $v \in \mathcal{K}$. Hence far enough out in this subsequence, $Q^{(\Delta t_j)}$ must satisfy

$$\left\|Q^{(\Delta t_j)} - v\right\|_{1,T} < \epsilon \quad \text{for all } j \text{ sufficiently large} \tag{12.59}$$

for the $\epsilon$ defined above. Moreover, since the $Q^{(\Delta t)}$ are generated by a conservative and consistent method, it follows from the Lax–Wendroff theorem (Theorem 12.1) that the limit $v$ must be a weak solution of the conservation law, i.e., $v \in \mathcal{W}$. But then (12.59) contradicts (12.58), and hence a sequence satisfying (12.58) cannot exist, and we conclude that $\mathrm{dist}(Q^{(\Delta t)}, \mathcal{W}) \to 0$ as $\Delta t \to 0$. □

There are other ways to prove convergence of approximate solutions to nonlinear scalar problems that do not require TV-stability. This is particularly useful in more than one dimension, where the total variation is harder to bound; see Section 20.10.2. For nonlinear systems

of equations it is impossible to bound the total variation in most cases, and convergence results are available only for special systems; see Section 15.8.2.

## Exercises

12.1. (a) Show that the method (12.5) with $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ defined by (12.8) corresponds to the flux function (12.12) with $a_{i-1/2}$ given by Murman's formula (12.14).

(b) Show that this has entropy-violating solutions by computing the flux $F_{i-1/2}$ everywhere for Burgers' equation with Riemann data $q_l = -1$ and $q_r = 1$.

12.2. Show that the LLF method (12.12) is an E-scheme.

12.3. Show that any E-scheme is TVD if the Courant number is sufficiently small. Hint: Use Theorem 6.1 with

$$C_{i-1} = \frac{\Delta t}{\Delta x}\left(\frac{f(Q_{i-1}) - F_{i-1/2}}{Q_i - Q_{i-1}}\right) \tag{12.60}$$

and a suitable choice for $D_i$.

12.4. Show that Jensen's inequality (12.41) need not hold if $\eta(q)$ is not convex.