15

Finite Volume Methods for Nonlinear Systems

15.1 Godunov's Method

Godunov's method has already been introduced in the context of linear systems in Chapter 4 and for scalar nonlinear problems in Chapter 12. The method is easily generalized to nonlinear systems if we can solve the nonlinear Riemann problem at each cell interface, and this gives the natural generalization of the first-order upwind method to general systems of conservation laws.

Recall that Q_i^n represents an approximation to the cell average of $q(x, t_n)$ over cell $\mathcal{C}_i$,

$$Q_i^n \approx \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x, t_n)\, dx,$$

and the idea is to use the piecewise constant function defined by these cell values as initial data $\tilde{q}^n(x, t_n)$ for the conservation law. Solving over time Δt with this data gives a function $\tilde{q}^n(x, t_{n+1})$, which is then averaged over each cell to obtain

$$Q_i^{n+1} = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{q}^n(x, t_{n+1})\, dx. \tag{15.1}$$

If the time step Δt is sufficiently small, then the exact solution $\tilde{q}^n(x, t)$ can be determined by piecing together the solutions to the Riemann problem arising from each cell interface, as indicated in Figure 15.1(a).

Recall from Section 4.11 that we do not need to perform the integration in (15.1) explicitly, which might be difficult, since $\tilde{q}^n(x, t_{n+1})$ may be very complicated as a function of x. Instead, we can use the fact that $\tilde{q}^n(x_{i-1/2}, t)$ is constant in time along each cell interface, so that the integral (4.5) can be computed exactly. Hence the cell average is updated by the formula

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+1/2}^n - F_{i-1/2}^n\right), \tag{15.2}$$

with

$$F_{i-1/2}^n = \mathcal{F}\left(Q_{i-1}^n, Q_i^n\right) = f\left(q^{\Downarrow}\left(Q_{i-1}^n, Q_i^n\right)\right). \tag{15.3}$$

As usual, $q^{\Downarrow}(q_l, q_r)$ denotes the solution to the Riemann problem between states q_l and q_r, evaluated along $x/t = 0$.

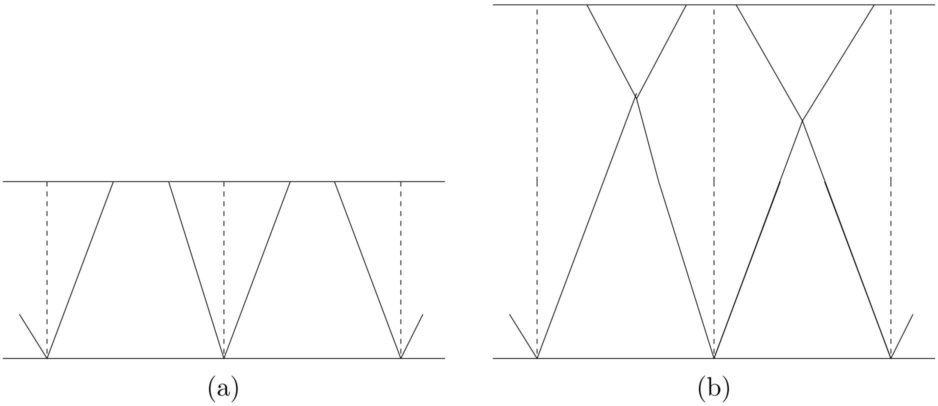


Fig. 15.1. Solving the Riemann problems at each interface for Godunov's method. (a) With Courant number less than $1/2$ there is no interaction of waves. (b) With Courant number less than 1 the interacting waves do not reach the cell interfaces, so the fluxes are still constant in time.

In Figure 15.1(a) the time step is taken to be small enough that there is no interaction of waves from neighboring Riemann problems. This would be necessary if we wanted to construct the solution at $\tilde{q}^n(x, t_{n+1})$ in order to explicitly calculate the cell averages (15.1). However, in order to use the flux formula (15.3) it is only necessary that the edge value $\tilde{q}^n(x_{i-1/2}, t)$ remain constant in time over the entire time step, which allows a time step roughly twice as large, as indicated in Figure 15.1(b). If $s_{\max}$ represents the largest wave speed that is encountered, then on a uniform grid with the cell interfaces distance Δx apart, we must require

$$\frac{s_{\max}\,\Delta t}{\Delta x} \le 1 \tag{15.4}$$

in order to insure that the formula (15.3) is valid. Note that this is precisely the CFL condition required for stability of this three-point method, as discussed in Section 4.4. In general $s_{\max}\Delta t/\Delta x$ is called the *Courant number*. Figure 15.1(a) shows a case for Courant number less than 1/2; Figure 15.1(b), for Courant number close to 1. Note that for a linear system of equations, $s_{\max} = \max_p |\lambda^p|$, and this agrees with the previous definition of the Courant number in Chapter 4.

To implement Godunov's method we do not generally need to determine the full structure of the Riemann solution at each interface, only the value $Q^{\Downarrow}_{i-1/2} = q^{\Downarrow}(Q^n_{i-1}, Q^n_i)$ at the cell interface. Normally we only need to determine the intermediate states where the relevant Hugoniot loci and/or integral curves intersect, and $Q^{\Downarrow}_{i-1/2}$ will equal one of these states. In particular we usually do not need to determine the structure of the solution within rarefaction waves at all. The only exception to this is if one of the rarefactions is transonic, so that the value of $Q^{\Downarrow}_{i-1/2}$ falls within the rarefaction fan rather than being one of the intermediate states. Even in this case we only need to evaluate one value from the fan, the value corresponding to $\xi = x/t = 0$ in the theory of Section 13.8.5, since this is the value that propagates with speed 0 and gives $Q^{\Downarrow}_{i-1/2}$.

Godunov's method can again be implemented in the wave propagation form

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(\mathcal{A}^+\Delta Q_{i-1/2} + \mathcal{A}^-\Delta Q_{i+1/2}\right) \tag{15.5}$$

if we define the fluctuations by

$$\begin{aligned} \mathcal{A}^-\Delta Q_{i-1/2} &= f\left(Q_{i-1/2}^{\vee|}\right) - f(Q_{i-1}), \\ \mathcal{A}^+\Delta Q_{i-1/2} &= f(Q_i) - f\left(Q_{i-1/2}^{\vee|}\right). \end{aligned} \tag{15.6}$$

The fact that so little information from the Riemann solution is used in Godunov's method suggests that one may be able to approximate the Riemann solution and still obtain reasonable solutions. This is often done in practice, and some approaches are discussed in Section 15.3. In Section 15.4 we will see how to extend Godunov's method to high-resolution methods.

15.2 Convergence of Godunov's Method

The Lax–Wendroff theorem of Section 12.10 applies to conservative methods for nonlinear systems of conservation laws as well as to scalar equations. Hence if a sequence of numerical approximations converges in the appropriate sense to a function $q(x, t)$ as the grid is refined, then the limit function $q(x, t)$ must be a weak solution of the conservation law. This is a powerful and useful result, since it gives us confidence that if we compute a reasonable-looking solution on a fine grid, then it is probably close to some weak solution. In particular, we expect that shocks will satisfy the right jump conditions and be propagating at the correct speeds. This would probably not be true if we used a nonconservative method – a reasonable-looking solution might be completely wrong, as we have observed in Section 12.9.

Of course we also need some additional entropy conditions on the numerical method to conclude that the discontinuities seen are physically correct. As we have seen in Section 12.2, it is quite possible that a conservative method will converge to entropy-violating weak solutions if we don't pay attention to this point. However, if we have an entropy function for the system, as described in Section 11.14, and if the Riemann solutions we use in Godunov's method all satisfy the entropy condition (11.51), then the limiting solution produced by Godunov's method will also satisfy the entropy condition, as discussed in Section 12.11. In particular, for the Euler equations of gas dynamics the physical entropy provides an entropy function. Hence any limiting weak solution obtained via Godunov's method will be physically correct.

The limitation of the Lax–Wendroff theorem is that it doesn't guarantee that convergence will occur; it only states that *if* a sequence converges, then the limit is a weak solution. Showing that convergence occurs requires some form of *stability*. We have seen in Section 12.12 that TV-stability is one appropriate form for nonlinear problems. For scalar equations Godunov's method is total variation diminishing (TVD), and we could also develop high-resolution methods that can be shown to be TVD, and hence stable and convergent.

Unfortunately, for nonlinear systems of equations this notion cannot easily be extended. In general the true solution is not TVD in any reasonable sense, and so we cannot expect the numerical solution to be. This is explored in more detail in Section 15.8. Even for classical problems such as the shallow water equations or Euler equations, there is no proof that Godunov's method converges in general. In spite of the lack of rigorous results, this method and high-resolution variants are generally successful in practice and extensively used.

15.3 Approximate Riemann Solvers

To apply Godunov's method on a system of equations we need only determine $q^{\Downarrow}(q_l, q_r)$, the state along $x/t = 0$ based on the Riemann data q_l and q_r. We do not need the entire structure of the Riemann problem. However, to compute $q^{\Downarrow}$ we must typically determine something about the full wave structure and wave speeds in order to determine where $q^{\Downarrow}$ lies in state space. Typically $q^{\Downarrow}$ is one of the intermediate states in the Riemann solution obtained in the process of connecting q_l to q_r by a sequence of shocks or rarefactions, and hence is one of the intersections of Hugoniot loci and/or integral curves. In the special case of a transonic rarefaction, the value of $q^{\Downarrow}$ will lie along the integral curve somewhere between these intersections, and additional work will be required to find it.

The process of solving the Riemann problem is thus often quite expensive, even though in the end we use very little information from this solution in defining the flux. We will see later that in order to extend the high-resolution methods of Chapter 6 to nonlinear systems of conservation laws, we must use more information, since all of the waves and wave speeds are used to define second-order corrections. Even so, it is often true that it is not necessary to compute the exact solution to the Riemann problem in order to obtain good results.

A wide variety of *approximate Riemann solvers* have been proposed that can be applied much more cheaply than the exact Riemann solver and yet give results that in many cases are equally good when used in the Godunov or high-resolution methods. In this section we will consider a few possibilities. For other surveys of approximate Riemann solvers, see for example [156], [245], [450].

Note that speeding up the Riemann solver can have a major impact on the efficiency of Godunov-type methods, since we must solve a Riemann problem at every cell interface in each time step. This will be of particular concern in more than one space dimension, where the amount of work grows rapidly. On a modest 100 × 100 grid in two dimensions, for example, one must solve roughly 20,000 Riemann problems in every time step to implement the simplest two-dimensional generalization of Godunov's method. Methods based on solving Riemann problems are notoriously expensive relative to other methods. The expense may pay off for problems with discontinuous solutions, if it allows good results to be obtained with far fewer grid points than other methods would require, but it is crucial that the Riemann solutions be computed or approximated as efficiently as possible.

For given data Q_{i-1} and Q_i, an approximate Riemann solution might define a function $\hat{Q}_{i-1/2}(x/t)$ that approximates the true similarity solution to the Riemann problem with data Q_{i-1} and Q_i. This function will typically consist of some set of M_w waves $\mathcal{W}^p_{i-1/2}$ propagating at some speeds $s^p_{i-1/2}$, with

$$Q_i - Q_{i-1} = \sum_{p=1}^{M_w} \mathcal{W}^p_{i-1/2}. \tag{15.7}$$

These waves and speeds will also be needed in defining high-resolution methods based on the approximate Riemann solver in Section 15.4.

To generalize Godunov's method using this function, we might take one of two different approaches:

1. Define the numerical flux by

$$F_{i-1/2} = f\big(\hat{Q}^{\vee}_{i-1/2}\big),$$

where

$$\hat{Q}^{\vee}_{i-1/2} = \hat{Q}_{i-1/2}(0) = Q_{i-1} + \sum_{p:s^p_{i-1/2}<0} \mathcal{W}^p_{i-1/2} \tag{15.8}$$

is the value along the cell interface. Then we proceed as in Section 15.1 and set

$$\begin{aligned} \mathcal{A}^-\Delta Q_{i-1/2} &= f\big(\hat{Q}^{\vee}_{i-1/2}\big) - f(Q_{i-1}), \\ \mathcal{A}^+\Delta Q_{i-1/2} &= f(Q_i) - f\big(\hat{Q}^{\vee}_{i-1/2}\big), \end{aligned} \tag{15.9}$$

in order to use the updating formula (15.5). Note that this amounts to evaluating the true flux function at an approximation to $Q^{\vee}_{i-1/2}$.

2. Use the waves and speeds from the approximate Riemann solution to define

$$\begin{aligned} \mathcal{A}^-\Delta Q_{i-1/2} &= \sum_{p=1}^{M_w} \big(s^p_{i-1/2}\big)^- \mathcal{W}^p_{i-1/2}, \\ \mathcal{A}^+\Delta Q_{i-1/2} &= \sum_{p=1}^{M_w} \big(s^p_{i-1/2}\big)^+ \mathcal{W}^p_{i-1/2}, \end{aligned} \tag{15.10}$$

and again use the updating formula (15.5). Note that this amounts to implementing the REA Algorithm 4.1 with the approximate Riemann solution in place of the true Riemann solutions, averaging these solutions over the grid cells to obtain Q^{n+1}.

If the all-shock Riemann solution is used (e.g., Section 13.7.1), then these two approaches yield the same result. This follows from the fact that the Rankine–Hugoniot condition is then satisfied across each wave $\mathcal{W}^p_{i-1/2}$. In general this will not be true if an approximate Riemann solution is used. In fact, the second approach may not even be conservative unless special care is taken in defining the approximate solution. (The first approach is always conservative, since it is based on an interface flux.)

15.3.1 Linearized Riemann Solvers

One very natural approach to defining an approximate Riemann solution is to replace the nonlinear problem $q_t + f(q)_x = 0$ by some linearized problem defined locally at each cell interface,

$$\hat{q}_t + \hat{A}_{i-1/2}\hat{q}_x = 0. \tag{15.11}$$

The matrix $\hat{A}_{i-1/2}$ is chosen to be some approximation to $f'(q)$ valid in a neighborhood of the data Q_{i-1} and Q_i. The matrix $\hat{A}_{i-1/2}$ should satisfy the following conditions:

$$\hat{A}_{i-1/2} \ \text{ is diagonalizable with real eigenvalues,} \tag{15.12}$$

so that (15.11) is hyperbolic, and

$$\hat{A}_{i-1/2} \to f'(\bar{q}) \qquad \text{as } Q_{i-1},\ Q_i \to \bar{q}, \tag{15.13}$$

so that the method is consistent with the original conservation law. The approximate Riemann solution then consists of m waves proportional to the eigenvectors $\hat{r}^p_{i-1/2}$ of $\hat{A}_{i-1/2}$, propagating with speeds $s^p_{i-1/2} = \hat{\lambda}^p_{i-1/2}$ given by the eigenvalues. Since this is a linear problem, the Riemann problem can generally be solved more easily than the original nonlinear problem, and often there are simple closed-form expressions for the eigenvectors and hence for the solution, which is obtained by solving the linear system

$$Q_i - Q_{i-1} = \sum_{p=1}^{m} \alpha^p_{i-1/2} \hat{r}^p_{i-1/2} \tag{15.14}$$

for the coefficients $\alpha^p_{i-1/2}$ and then setting $\mathcal{W}^p_{i-1/2} = \alpha^p_{i-1/2}\hat{r}^p_{i-1/2}$.

We might take, for example,

$$\hat{A}_{i-1/2} = f'\big(\hat{Q}_{i-1/2}\big), \tag{15.15}$$

where $\hat{Q}_{i-1/2}$ is some average of Q_{i-1} and Q_i. In particular, the Roe linearization described in the next section has this form for the Euler or shallow water equations, with a very special average. This special averaging leads to some additional nice properties, but for problems where a Roe linearization is not available it is often possible to simply use $\hat{Q}_{i-1/2} = (Q_{i-1} + Q_i)/2$. Note that for any choice of $\hat{A}_{i-1/2}$ satisfying (15.12) and (15.13), we can obtain a consistent and conservative method if we use the formulas (15.8) and (15.9). The formulas (15.10) will not give a conservative method unless $\hat{A}_{i-1/2}$ satisfies an additional condition described in the next section, since (15.10) leads to a conservative method only if the condition

$$f(Q_i) - f(Q_{i-1}) = \sum_{p=1}^{M_w} s^p_{i-1/2}\mathcal{W}^p_{i-1/2} \tag{15.16}$$

is satisfied. This may not hold in general. (See Section 15.5 for an alternative approach to obtaining conservative fluctuations by directly splitting the flux difference.)

Another obvious linearization is to take

$$\hat{A}_{i-1/2} = \frac{1}{2}\big[f'(Q_{i-1}) + f'(Q_i)\big], \tag{15.17}$$

or some other average of the Jacobian matrix between the two states. But note that in general this matrix could fail to satisfy condition (15.12) even if $f'(Q_{i-1})$ and $f'(Q_i)$ have real eigenvalues.

Using a linearized problem can be easily motivated and justified at most cell interfaces. The solution to a conservation law typically consists of at most a few isolated shock waves or contact discontinuities separated by regions where the solution is smooth. In these regions, the variation in Q from one grid cell to the next has $\|Q_i - Q_{i-1}\| = \mathcal{O}(\Delta x)$ and the Jacobian matrix is nearly constant, $f'(Q_{i-1}) \approx f'(Q_i)$. Zooming in on the region of state space near

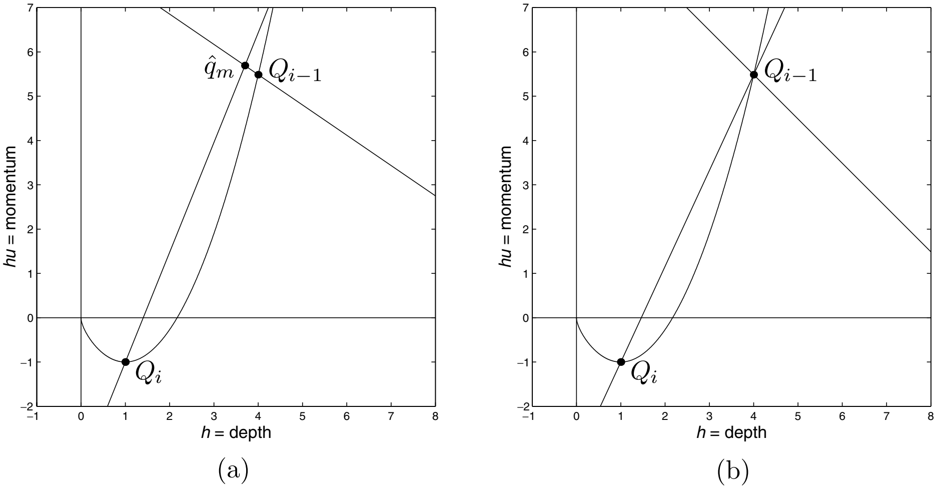


Fig. 15.2. States Q_{i-1} and Q_i are connected by a 2-shock in the shallow water equations, and lie on the same curved Hugoniot locus. The straight lines are in the directions of the eigenvectors of an averaged Jacobian matrix of the form (15.15). (a) Using the average $\hat{Q}_{i-1/2} = (Q_{i-1} + Q_i)/2$. (b) Using the Roe average.

these points, we would find that the Hugoniot loci and integral curves needed to find the exact Riemann solution are nearly straight lines pointing in the directions of the eigenvectors of these matrices. Defining $\hat{A}_{i-1/2}$ as any reasonable approximation to these matrices will yield essentially the same eigenvectors and a Riemann solution that agrees very well with the true solution. This is made more precise in Section 15.6, where we consider the truncation error of methods based on approximate Riemann solutions.

It is only near shocks that we expect Q_{i-1} and Q_i to be far apart in state space, in which case it is harder to justify the use of a Riemann solver that is linearized about one particular point such as $(Q_{i-1}+Q_i)/2$. The true nonlinear structure in state space will look very different from the eigenstructure of any one Jacobian matrix. For example, Figure 15.2(a) shows two states Q_{i-1} and Q_i that should be connected by a single 2-shock in the shallow water equations, since Q_i lies on the 2-Hugoniot locus of Q_{i-1}. If instead the linearized Riemann problem is solved using (15.15) with $\hat{Q}_{i-1/2} = (Q_{i-1} + Q_i)/2$, the state $\hat{q}_m$ indicated in the figure is obtained, with a spurious 1-wave.

15.3.2 Roe Linearization

It is important to notice that even near a shock wave the Riemann problems arising at cell interfaces will typically have a large jump in at most one wave family, say $\mathcal{W}^p$, with $\|\mathcal{W}^j\| = \mathcal{O}(\Delta x)$ for all other waves $j \neq p$. Most of the time a shock in one family is propagating through smooth flow in the other families. It is only at isolated instants in time when two shock waves collide that we expect to observe Riemann problems whose solutions contain more than one strong wave.

For this reason, the situation illustrated in Figure 15.2 is the most important to consider, along with the case where all waves are weak. This suggests the following property that we would like a linearized matrix $\hat{A}_{i-1/2}$ to possess:

If Q_{i-1} and Q_i are connected by a single wave $\mathcal{W}^p = Q_i - Q_{i-1}$ in the true Riemann solution, then $\mathcal{W}^p$ should also be an eigenvector of $\hat{A}_{i-1/2}$.

If this holds, then the "approximate" Riemann solution will also consist of this single wave and will agree with the exact solution. This condition is easy to rewrite in a more useful form using the Rankine–Hugoniot condition (11.21). If Q_{i-1} and Q_i are connected by a single wave (shock or contact discontinuity), then

$$f(Q_i) - f(Q_{i-1}) = s(Q_i - Q_{i-1}),$$

where s is the wave speed. If this is also to be a solution to the linearized Riemann problem, then we must have

$$\hat{A}_{i-1/2}(Q_i - Q_{i-1}) = s(Q_i - Q_{i-1}).$$

Combining these, we obtain the condition

$$\hat{A}_{i-1/2}(Q_i - Q_{i-1}) = f(Q_i) - f(Q_{i-1}). \tag{15.18}$$

In fact this is a useful condition to impose on $\hat{A}_{i-1/2}$ in general, for any Q_{i-1} and Q_i. It guarantees that (15.10) yields a conservative method, and in fact agrees with what is obtained by (15.9). This can be confirmed by recalling that (15.10) will be conservative provided that (6.57) is satisfied,

$$\mathcal{A}^-\Delta Q_{i-1/2} + \mathcal{A}^+\Delta Q_{i-1/2} = f(Q_i) - f(Q_{i-1}), \tag{15.19}$$

which is satisfied for the approximate Riemann solver if and only if the condition (15.18) holds.

Another nice feature of (15.18) is that it states that the matrix $\hat{A}$, which approximates the Jacobian matrix $\partial f/\partial q$, should at least have the correct behavior in the one direction where we know the change in f that results from a change in q.

The problem now is to obtain an approximate Jacobian that will satisfy (15.18) along with (15.12) and (15.13). One way to obtain a matrix satisfying (15.18) is by integrating the Jacobian matrix over a suitable path in state space between Q_{i-1} and Q_i. Consider the straight-line path parameterized by

$$q(\xi) = Q_{i-1} + (Q_i - Q_{i-1})\xi \tag{15.20}$$

for $0 \le \xi \le 1$. Then $f(Q_i) - f(Q_{i-1})$ can be written as the line integral

$$\begin{aligned} f(Q_i) - f(Q_{i-1}) &= \int_0^1 \frac{df(q(\xi))}{d\xi}\,d\xi \\ &= \int_0^1 \frac{df(q(\xi))}{dq}\,q'(\xi)\,d\xi \\ &= \left[\int_0^1 f'(q(\xi))\,d\xi\right](Q_i - Q_{i-1}), \end{aligned} \tag{15.21}$$

since $q'(\xi) = Q_i - Q_{i-1}$ is constant and can be pulled out of the integral. This shows that we can define $\hat{A}_{i-1/2}$ as the average

$$\hat{A}_{i-1/2} = \int_0^1 f'(q(\xi))\,d\xi. \tag{15.22}$$

This average always satisfies (15.18) and (15.13), but in general there is no guarantee that (15.12) will be satisfied, even if the original problem is hyperbolic at each point q. An additional problem with attempting to use (15.22) is that it is generally not possible to evaluate this integral in closed form for most nonlinear problems of interest. So it cannot be used as the basis for a practical algorithm that is more efficient than using the true Riemann solver.

Roe [375] made a significant breakthrough by discovering a way to surmount this difficulty for the Euler equations (for a polytropic ideal gas) by a more clever choice of integration path. Moreover, the resulting $\hat{A}_{i-1/2}$ is of the form (15.15) and hence satisfies (15.12). His approach can also be applied to other interesting systems, and will be demonstrated for the shallow water equations in the next section. See Section 15.3.4 for the Euler equations.

Roe introduced a *parameter vector* $z(q)$, a change of variables that leads to integrals that are easy to evaluate. We assume this mapping is invertible so that we also know $q(z)$. Using this mapping, we can also view f as a function of z, and will write $f(z)$ as shorthand for $f(q(z))$.

Rather than integrating on the path (15.20), we will integrate along the path

$$z(\xi) = Z_{i-1} + (Z_i - Z_{i-1})\xi, \tag{15.23}$$

where $Z_j = z(Q_j)$ for $j = i-1, i$. Then $z'(\xi) = Z_i - Z_{i-1}$ is independent of ξ, and so

$$\begin{aligned} f(Q_i) - f(Q_{i-1}) &= \int_0^1 \frac{df(z(\xi))}{d\xi}\,d\xi \\ &= \int_0^1 \frac{df(z(\xi))}{dz}\,z'(\xi)\,d\xi \\ &= \left[\int_0^1 \frac{df(z(\xi))}{dz}\,d\xi\right](Z_i - Z_{i-1}). \end{aligned} \tag{15.24}$$

We hope that this integral will be easy to evaluate. But even if it is, this does not yet give us what we need, since the right-hand side involves $Z_i - Z_{i-1}$ rather than $Q_i - Q_{i-1}$. However, we can relate these using another path integral,

$$\begin{aligned} Q_i - Q_{i-1} &= \int_0^1 \frac{dq(z(\xi))}{d\xi}\,d\xi \\ &= \int_0^1 \frac{dq(z(\xi))}{dz}\,z'(\xi)\,d\xi \\ &= \left[\int_0^1 \frac{dq(z(\xi))}{dz}\,d\xi\right](Z_i - Z_{i-1}). \end{aligned} \tag{15.25}$$

The goal is now to find a parameter vector $z(q)$ for which *both* the integral in (15.24) and the integral in (15.25) are easy to evaluate. Then we will have

$$\begin{aligned} f(Q_i) - f(Q_{i-1}) &= \hat{C}_{i-1/2}(Z_i - Z_{i-1}), \\ Q_i - Q_{i-1} &= \hat{B}_{i-1/2}(Z_i - Z_{i-1}), \end{aligned} \tag{15.26}$$

where $\hat{C}_{i-1/2}$ and $\hat{B}_{i-1/2}$ are these integrals. From these we can obtain the desired relation (15.18) by using

$$\hat{A}_{i-1/2} = \hat{C}_{i-1/2}\hat{B}^{-1}_{i-1/2}. \tag{15.27}$$

Harten and Lax (see [187]) showed that an integration procedure of this form can always be used to define a matrix $\hat{A}$ satisfying (15.12) provided that the system has a convex entropy function $\eta(q)$ as described in Section 11.14. The choice $z(q) = \eta'(q)$ then works, where $\eta'(q)$ is the gradient of η with respect to q. It is shown in [187] that the resulting matrix $\hat{A}_{\text{HLL}}$ is then similar to a symmetric matrix and hence has real eigenvalues.

To make the integrals easy to evaluate, however, we generally wish to choose z in such a way that both $\partial q/\partial z$ and $\partial f/\partial z$ have components that are polynomials in the components of z. Then they will be polynomials in ξ along the path (15.23) and hence easy to integrate.

15.3.3 Roe Solver for the Shallow Water Equations

As an example, we derive the Roe matrix for the shallow water equations. In [281] the isothermal equations, which have similar structure, are used as an example.

For the shallow water equations (see Chapter 13) we have

$$q = \begin{bmatrix} h \\ hu \end{bmatrix} = \begin{bmatrix} q^1 \\ q^2 \end{bmatrix}, \qquad f(q) = \begin{bmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{bmatrix} = \begin{bmatrix} q^2 \\ (q^2)^2/q^1 + \frac{1}{2}g(q^1)^2 \end{bmatrix}$$

and

$$f'(q) = \begin{bmatrix} 0 & 1 \\ -(q^2/q^1)^2 + gq^1 & 2q^2/q^1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -u^2 + gh & 2u \end{bmatrix}.$$

As a parameter vector we choose

$$z = h^{-1/2}q, \qquad \text{so that} \quad \begin{bmatrix} z^1 \\ z^2 \end{bmatrix} = \begin{bmatrix} \sqrt{h} \\ \sqrt{h}u \end{bmatrix}. \tag{15.28}$$

This is analogous to the parameter vector introduced by Roe for the Euler equations (see Section 15.3.4), in which case $z = \rho^{-1/2}q$. Note that the matrix $f'(q)$ involves the quotient q^2/q^1, and hence integrating along the path (15.20) would require integrating rational functions of ξ. The beauty of this choice of variables z is that the matrices we must integrate in (15.24) and (15.25) involve only polynomials in ξ. We find that

$$q(z) = \begin{bmatrix} (z^1)^2 \\ z^1z^2 \end{bmatrix} \implies \frac{\partial q}{\partial z} = \begin{bmatrix} 2z^1 & 0 \\ z^2 & z^1 \end{bmatrix} \tag{15.29}$$

and

$$f(z)=\begin{bmatrix} z^1z^2 \\ (z^2)^2+\frac{1}{2}g(z^1)^4 \end{bmatrix} \implies \frac{\partial f}{\partial z}=\begin{bmatrix} z^2 & z^1 \\ 2g(z^1)^3 & 2z^2 \end{bmatrix}. \tag{15.30}$$

We now set

$$z^p = Z^p_{i-1} + \left(Z^p_i - Z^p_{i-1}\right)\xi \quad \text{for } p=1,\,2$$

and integrate each element of these matrices from $\xi = 0$ to $\xi = 1$. All elements are linear in ξ except the (2,1) element of $\partial f/\partial z$, which is cubic.

Integrating the linear terms $z^p(\xi)$ yields

$$\int_0^1 z^p(\xi)\,d\xi = \frac{1}{2}\left(Z^p_{i-1}+Z^p_i\right) \equiv \bar{Z}^p,$$

simply the average between the endpoints. For the cubic term we obtain

$$\begin{aligned}\int_0^1 (z^1(\xi))^3\,d\xi &= \frac{1}{4}\left(\frac{\left(Z^1_i\right)^4-\left(Z^1_{i-1}\right)^4}{Z^1_i - Z^1_{i-1}}\right)\\ &= \frac{1}{2}\left(Z^1_{i-1}+Z^1_i\right)\cdot\frac{1}{2}\left[\left(Z^1_{i-1}\right)^2+\left(Z^1_i\right)^2\right]\\ &= \bar{Z}^1\bar{h},\end{aligned} \tag{15.31}$$

where

$$\bar{h} = \frac{1}{2}(h_{i-1}+h_i). \tag{15.32}$$

Hence we obtain

$$\hat{B}_{i-1/2}=\begin{bmatrix} 2\bar{Z}^1 & 0 \\ \bar{Z}^2 & \bar{Z}^1 \end{bmatrix}, \qquad \hat{C}_{i-1/2}=\begin{bmatrix} \bar{Z}^2 & \bar{Z}^1 \\ 2g\bar{Z}^1\bar{h} & 2\bar{Z}^2 \end{bmatrix} \tag{15.33}$$

and so

$$\hat{A}_{i-1/2}=\hat{C}_{i-1/2}\hat{B}^{-1}_{i-1/2}=\begin{bmatrix} 0 & 1 \\ -(\bar{Z}^2/\bar{Z}^1)^2+g\bar{h} & 2\bar{Z}^2/\bar{Z}^1 \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ -\hat{u}^2+g\bar{h} & 2\hat{u} \end{bmatrix}. \tag{15.34}$$

Here $\bar{h}$ is the arithmetic average of h_{i-1} and h_i given in (15.32), but $\hat{u}$ is a different sort of average of the velocities, the *Roe average*:

$$\hat{u}=\frac{\bar{Z}^2}{\bar{Z}^1}=\frac{\sqrt{h_{i-1}}u_{i-1}+\sqrt{h_i}u_i}{\sqrt{h_{i-1}}+\sqrt{h_i}}. \tag{15.35}$$

Note that the matrix $\hat{A}_{i-1/2}$ in (15.34) is simply the Jacobian matrix $f'(\hat{q})$ evaluated at the special state $\hat{q}=(\bar{h},\,\bar{h}\hat{u})$. In particular, if $Q_{i-1}=Q_i$ then $\hat{A}_{i-1/2}$ reduces to $f'(Q_i)$.

The eigenvalues and eigenvectors of $\hat{A}_{i-1/2}$ are known from (13.9) and (13.10):

$$\hat{\lambda}^1=\hat{u}-\hat{c}, \qquad \hat{\lambda}^2=\hat{u}+\hat{c}, \tag{15.36}$$

and

$$\hat{r}^1 = \begin{bmatrix} 1 \\ \hat{u} - \hat{c} \end{bmatrix}, \qquad \hat{r}^2 = \begin{bmatrix} 1 \\ \hat{u} + \hat{c} \end{bmatrix}, \tag{15.37}$$

where $\hat{c} = \sqrt{g\bar{h}}$. To use the approximate Riemann solver we decompose $Q_i - Q_{i-1}$ as in (15.14),

$$Q_i - Q_{i-1} = \alpha^1_{i-1/2}\hat{r}^1 + \alpha^2_{i-1/2}\hat{r}^2 \equiv \mathcal{W}^1_{i-1/2} + \mathcal{W}^2_{i-1/2}. \tag{15.38}$$

The coefficients $\alpha^p_{i-1/2}$ are computed by solving this linear system, which can be done explicitly by inverting the matrix $\hat{R}$ of right eigenvectors to obtain

$$\hat{L} = \hat{R}^{-1} = \frac{1}{2\hat{c}} \begin{bmatrix} \hat{u} + \hat{c} & -1 \\ -(\hat{u} - \hat{c}) & 1 \end{bmatrix}.$$

Multiplying this by the vector $\delta \equiv Q_i - Q_{i-1}$ gives the vector of α-coefficients, and hence

$$\begin{aligned} \alpha^1_{i-1/2} &= \frac{(\hat{u} + \hat{c})\delta^1 - \delta^2}{2\hat{c}}, \\ \alpha^2_{i-1/2} &= \frac{-(\hat{u} - \hat{c})\delta^1 + \delta^2}{2\hat{c}}, \end{aligned} \tag{15.39}$$

The fluctuations (15.10) are then used in Godunov's method, with the speeds s given by the eigenvalues λ of (15.36).

Alternatively, we could compute the numerical flux $F_{i-1/2}$ by

$$F_{i-1/2} = f(Q_{i-1}) + \hat{A}^-_{i-1/2}(Q_i - Q_{i-1})$$

or by

$$F_{i-1/2} = f(Q_i) - \hat{A}^+_{i-1/2}(Q_i - Q_{i-1}).$$

Averaging these two expressions gives a third version, which is symmetric in Q_{i-1} and Q_i,

$$F_{i-1/2} = \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\left|\hat{A}_{i-1/2}\right|(Q_i - Q_{i-1}). \tag{15.40}$$

This form is often called *Roe's method* (see Section 4.14) and has the form of the unstable centered flux plus a viscous correction term.

15.3.4 Roe Solver for the Euler Equations

For the Euler equations with the equation of state (14.23), Roe [375] proposed the parameter vector $z = \rho^{-1/2}q$, leading to the averages

$$\hat{u} = \frac{\sqrt{\rho_{i-1}}\, u_{i-1} + \sqrt{\rho_i}\, u_i}{\sqrt{\rho_{i-1}} + \sqrt{\rho_i}} \tag{15.41}$$

for the velocity,

$$\hat{H} = \frac{\sqrt{\rho_{i-1}}\, H_{i-1} + \sqrt{\rho_i}\, H_i}{\sqrt{\rho_{i-1}} + \sqrt{\rho_i}} = \frac{(E_{i-1} + p_{i-1})/\sqrt{\rho_{i-1}} + (E_i + p_i)/\sqrt{\rho_i}}{\sqrt{\rho_{i-1}} + \sqrt{\rho_i}} \tag{15.42}$$

for the total specific enthalpy, and

$$\hat{c} = \sqrt{(\gamma - 1)\left(\hat{H} - \frac{1}{2}\hat{u}^2\right)} \tag{15.43}$$

for the sound speed. The eigenvalues and eigenvectors of the Roe matrix are then obtained by evaluating (14.45) and (14.46) at this averaged state. The coefficients $\alpha^p_{i-1/2}$ in the wave decomposition

$$\delta \equiv Q_i - Q_{i-1} = \alpha^1 \hat{r}^1 + \alpha^2 \hat{r}^2 + \alpha^3 \hat{r}^3$$

can be obtained by inverting the matrix of right eigenvectors, which leads to the following formulas:

$$\begin{aligned}
\alpha^2 &= (\gamma - 1)\frac{(\hat{H} - \hat{u}^2)\delta^1 + \hat{u}\delta^2 - \delta^3}{\hat{c}^2}, \\
\alpha^3 &= \frac{\delta^2 + (\hat{c} - \hat{u})\delta^1 - \hat{c}\alpha^2}{2\hat{c}}, \\
\alpha^1 &= \delta^1 - \alpha^2 - \alpha^3.
\end{aligned} \tag{15.44}$$

For other equations of state and more complicated gas dynamics problems it may also be possible to derive Roe solvers; see for example [91], [149], [172], [208], [451].

15.3.5 Sonic Entropy Fixes

One disadvantage of using a linearized Riemann solver is that the resulting approximate Riemann solution consists only of discontinuities, with no rarefaction waves. This can lead to a violation of the entropy condition, as has been observed previously for scalar conservation laws in Section 12.3.

In fact, it is worth noting that in the scalar case the Roe condition (15.18) can be satisfied by choosing the scalar $\hat{A}_{i-1/2}$ as

$$\hat{A}_{i-1/2} = \frac{f(Q_i) - f(Q_{i-1})}{Q_i - Q_{i-1}}, \tag{15.45}$$

which is simply the shock speed resulting from the scalar Rankine–Hugoniot condition. Hence using the Roe linearization in the scalar case and solving the resulting advection equation with velocity (15.45) is equivalent to always using the shock-wave solution to the scalar problem, as discussed in Section 12.2. In this scalar case (15.40) reduces to the flux for Murman's method, (12.12) with $a_{i-1/2}$ given by (12.14).

Recall that in the scalar case, the use of an entropy-violating Riemann solution leads to difficulties only in the case of a transonic rarefaction wave, in which $f'(q_l) < 0 < f'(q_r)$.

This is also typically true when we use Roe's approximate Riemann solution for a system of conservation laws. It is only for sonic rarefactions, those for which $\lambda^p < 0$ to the left of the wave while $\lambda^p > 0$ to the right of the wave, that entropy violation is a problem. In the case of a sonic rarefaction wave, it is necessary to modify the approximate Riemann solver in order to obtain entropy-satisfying solutions.

For the shallow water equations, a system of two equations, there is a single intermediate state $\hat{Q}_m$ in the approximate Riemann solution between Q_{i-1} and Q_i. We can compute the characteristic speeds in each state as

$$\begin{aligned} \lambda^1_{i-1} &= u_{i-1} - \sqrt{gh_{i-1}}, \qquad & \lambda^1_m &= u_m - \sqrt{g\hat{h}_m}, \\ \lambda^2_m &= \hat{u}_m + \sqrt{g\hat{h}_m}, \qquad & \lambda^2_i &= u_i + \sqrt{gh_i}. \end{aligned} \tag{15.46}$$

If $\lambda^1_{i-1} < 0 < \lambda^1_m$, then we should suspect that the 1-wave is actually a transonic rarefaction and make some adjustment to the flux, i.e., to $\mathcal{A}^-\Delta Q_{i-1/2}$ and $\mathcal{A}^+\Delta Q_{i-1/2}$, in this case. Similarly, if $\lambda^2_m < 0 < \lambda^2_i$, then we should fix the flux to incorporate a 2-rarefaction. Note that at most one of these situations can hold, since $\lambda^1_m < \lambda^2_m$.

For sufficiently simple systems, such as the shallow water equations, it may be easy to evaluate the true intermediate state $Q^{\Downarrow}_{i-1/2}$ that lies along the interface in the Riemann solution at $x_{i-1/2}$, once we suspect it lies on a transonic rarefaction wave in a particular family. If we suspect that there should be a transonic 1-rarefaction, for example, then we can simply evaluate the state at $\xi = x/t = 0$ in the 1-rarefaction wave connected to Q_{i-1}. This is easily done using the formulas in Section 13.8.5. Evaluating (13.52) at $\xi = 0$ and then using (13.33) gives

$$\begin{aligned} h^{\Downarrow}_{i-1/2} &= \left(u_{i-1} + 2\sqrt{gh_{i-1}}\right)^2/9g, \\ u^{\Downarrow}_{i-1/2} &= u_{i-1} - 2\left(\sqrt{gh_{i-1}} - \sqrt{gh^{\Downarrow}_{i-1/2}}\right). \end{aligned} \tag{15.47}$$

We can now evaluate the flux at this point and set $F_{i-1/2} = f(Q^{\Downarrow}_{i-1/2})$. Finally, the formulas (15.9) can be used to define $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$.

Using the structure of the exact rarefaction wave is not always possible or desirable for more general systems of equations where we hope to avoid determining the exact Riemann solution. A variety of different approaches have been developed as approximate entropy fixes that work well in practice. Several of these are described below. See also [115], [349], [351], [355], [378], [382], [432], or [434] for some other discussions.

The Harten–Hyman Entropy Fix

A more general procedure was taken by Harten and Hyman [184] and modified slightly in [281]. This approach is used in many of the standard CLAWPACK solvers.

Suppose there appears to be a transonic rarefaction in the k-wave, i.e., $\lambda^k_l < 0 < \lambda^k_r$, where $\lambda^k_{l,r}$ represents the kth eigenvalue of the matrix $f'(q)$ computed in the states $q^k_{l,r}$ just to the left and right of the k-wave in the approximate Riemann solution, i.e.,

$$q_l^k = Q_{i-1} + \sum_{p=1}^{k-1} \mathcal{W}^p, \qquad q_r^k = q_l^k + \mathcal{W}^k. \tag{15.48}$$

(We suppress the subscripts $i - 1/2$ here and below for clarity, since we need to add subscripts l and r.) Then we replace the single wave $\mathcal{W}^k$ propagating at speed $\hat{\lambda}^k$ by a pair of waves $\mathcal{W}_l^k = \beta \mathcal{W}^k$ and $\mathcal{W}_r^k = (1-\beta)\mathcal{W}^k$ propagating at speeds λ_l^k and λ_r^k. To maintain conservation we require that

$$\lambda_l^k \mathcal{W}_l^k + \lambda_r^k \mathcal{W}_r^k = \hat{\lambda}^k \mathcal{W}^k$$

and hence

$$\beta = \frac{\lambda_r^k - \hat{\lambda}^k}{\lambda_r^k - \lambda_l^k}. \tag{15.49}$$

In practice it is simpler to leave the wave $\mathcal{W}^k$ alone (and continue to use this single wave in the high-resolution correction terms; see Section 15.4) and instead modify the values $(\hat{\lambda}^k)^\pm$ used in defining $\mathcal{A}^\pm \Delta Q_{i-1/2}$ via (15.10). The formula (15.10) can still be used (with $\hat{s}^k = \hat{\lambda}^k$) if, instead of the positive and negative parts of $\hat{\lambda}^k$, we use the values

$$\begin{aligned} (\hat{\lambda}^k)^- &\equiv \beta \lambda_l^k, \\ (\hat{\lambda}^k)^+ &\equiv (1-\beta)\lambda_r^k \end{aligned} \tag{15.50}$$

in the kth field. These still sum to $\hat{\lambda}^k$ but are both nonzero in the transonic case. We continue to use the standard definitions (4.40) or (4.63) of $(\lambda^k)^\pm$ in any field where λ_l^k and λ_r^k have the same sign.

In the scalar case this entropy fix can be interpreted as using a piecewise linear approximation to the flux function $f(q)$ in the neighborhood of the sonic point. This approximation lies below the true flux function in the convex case, a fact that is used in [281] to show that Roe's method with this entropy fix is an E-scheme (see Section 12.7) and hence converges to the entropy-satisfying weak solution. See also [115] for some related results.

Numerical Viscosity

From (15.40), the flux for Roe's method is

$$\begin{aligned} F_{i-1/2} &= \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\left|\hat{A}_{i-1/2}\right|(Q_i - Q_{i-1}) \\ &= \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\sum_p \left|\hat{\lambda}_{i-1/2}^p\right| \mathcal{W}_{i-1/2}^p. \end{aligned} \tag{15.51}$$

One way to view the need for an entropy fix is to recognize that the viscous term in this flux is too small in the case of a transonic rarefaction. With sufficient viscosity we should not observe entropy-violating shocks. Note that in the transonic case, where the characteristic speeds span 0, we might expect the eigenvalue $\hat{\lambda}_{i-1/2}^p$ to be close to zero. The corresponding term of the sum in (15.51) will then be close to zero, corresponding to no viscosity in this field. In fact a transonic entropy-violating shock typically has speed zero, as can be observed in Figure 12.2(a). The asymmetric portion of the rarefaction fan is moving with nonzero average speed and has positive viscosity due to the averaging process. It is only the symmetric part with speed 0 that remains as a stationary jump.

When we implement the method using $\mathcal{A}^{\pm}\Delta Q$ from (15.10), we find that the numerical flux is actually given by

$$F_{i-1/2} = \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\sum_p \left[\left(\hat{\lambda}^p_{i-1/2}\right)^+ - \left(\hat{\lambda}^p_{i-1/2}\right)^-\right]\mathcal{W}^p_{i-1/2}. \tag{15.52}$$

With the usual definition (4.40) of $\lambda^{\pm}$, this agrees with (15.51). However, if we apply the entropy fix defined in the last section and redefine these values as in (15.50), then we have effectively increased the numerical viscosity specifically in the kth field when this field contains a transonic rarefaction.

Harten's Entropy Fix

Harten [179] proposed an entropy fix based on increasing the viscosity by modifying the absolute-value function in (15.51), never allowing any eigenvalue to be too close to zero. In this simple approach one replaces each value $|\hat{\lambda}^p_{i-1/2}|$ in (15.51) by a value $\phi_\delta(\hat{\lambda}^p_{i-1/2})$, where $\phi_\delta(\lambda)$ is a smoothed version of the absolute-value function that is always positive, staying above some value $\delta/2$:

$$\phi_\delta(\lambda) = \begin{cases} |\lambda| & \text{if } |\lambda| \ge \delta, \\ (\lambda^2 + \delta^2)/(2\delta) & \text{if } |\lambda| < \delta. \end{cases} \tag{15.53}$$

A disadvantage of this approach is that the parameter δ must typically be tuned to the problem.

To implement this in the context of fluctuations $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$, we can translate this modification of the absolute value into modifications of $(\hat{\lambda}^p_{i-1/2})^+$ and $(\hat{\lambda}^p_{i-1/2})^-$. Again we can continue to use the form (15.10) if we redefine

$$\begin{aligned} (\lambda)^- &\equiv \frac{1}{2}[\lambda - \phi_\delta(\lambda)], \\ (\lambda)^+ &\equiv \frac{1}{2}[\lambda + \phi_\delta(\lambda)]. \end{aligned} \tag{15.54}$$

Note that this agrees with the usual definition (4.63) of $\lambda^{\pm}$ if $\phi_\delta(\lambda) = |\lambda|$.

The LLF Entropy Fix

Another approach to introducing more numerical viscosity is to use the approximate Riemann solver in conjunction with an extension of the local Lax–Friedrichs (LLF) method to systems of equations. The formula (12.12) generalizes naturally to systems of equations as

$$F_{i-1/2} = \frac{1}{2}[f(Q_{i-1}) + f(Q_i)] - \frac{1}{2}\sum_p a^p_{i-1/2}\mathcal{W}^p_{i-1/2}, \tag{15.55}$$

where

$$a_{i-1/2}^p = \max\left(|\lambda_{i-1}^p|, |\lambda_i^p|\right). \tag{15.56}$$

Here λ_{i-1}^p and λ_i^p are eigenvalues of the Jacobians $f'(Q_{i-1})$ and $f'(Q_i)$ respectively, while $\mathcal{W}_{i-1/2}^p$ is the wave resulting from the Roe solver. We can implement this using

$$\begin{aligned} \mathcal{A}^-\Delta Q_{i-1/2} &= \frac{1}{2}\sum_p \left(\hat{\lambda}_{i-1/2}^p - a_{i-1/2}^p\right)\mathcal{W}_{i-1/2}^p, \\ \mathcal{A}^+\Delta Q_{i-1/2} &= \frac{1}{2}\sum_p \left(\hat{\lambda}_{i-1/2}^p + a_{i-1/2}^p\right)\mathcal{W}_{i-1/2}^p. \end{aligned} \tag{15.57}$$

Again this can be viewed as a redefinition of $\lambda^\pm$ similar to (15.54). A disadvantage of this approach is that it generally adds numerical viscosity to all fields, whether or not there is a transonic rarefaction. However, wherever the solution is smooth we have $\hat{\lambda}_{i-1/2}^p \approx \lambda_{i-1}^p \approx \lambda_i^p$ and so (15.57) essentially reduces to the standard definition of $\mathcal{A}^\pm\Delta Q_{i-1/2}$.

15.3.6 Failure of Linearized Solvers

In some situations linearized Riemann solvers such as those based on the Roe average can fail completely, giving a nonphysical solution such as negative depth in the shallow water equation or negative pressures or density in the Euler equations. This can happen in particular for data that is near the "vacuum state" or in situations where there is a strong expansion.

Figure 15.3 shows an example for the shallow water equation in the case of a Riemann problem yielding two symmetric rarefaction waves, as studied in Section 13.8.6. In Figure 15.3(a) the data is $h_l = h_r = 1$ and $u_r = u_l = 0.8$ (with $g = 1$). The straight lines

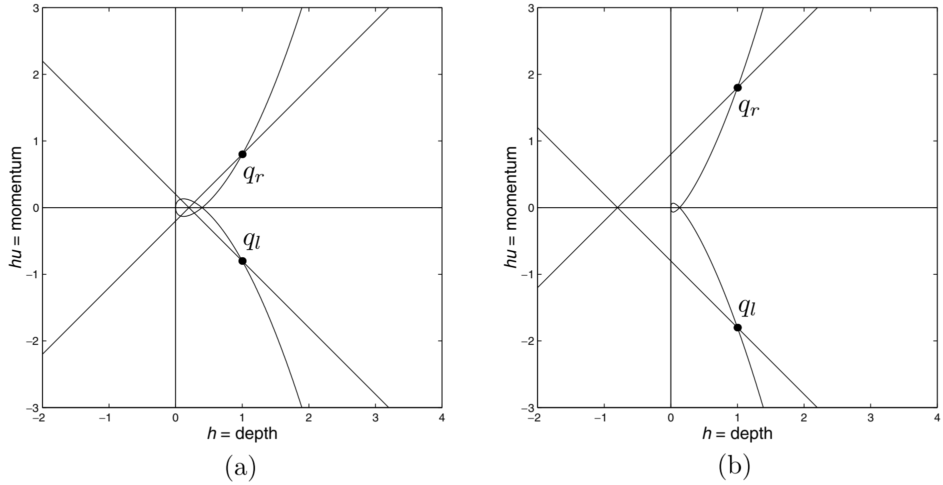


Fig. 15.3. The straight lines are in the directions of the eigenvectors of Roe-averaged Jacobian matrices $\hat{A}$ in the example of Section 15.3.6. (a) For $u_r = -u_l = 0.8$. (b) For $u_r = -u_l = 1.8$, in which case the lines intersect at an unphysical point with $\hat{h}_m < 0$.

show the eigendirections of the Roe matrix $\hat{A}$, which intersect at a point with $u_m = 0$ (as in the true solution) but with h_m about half the correct value. The solution is still positive, however, and reasonable computational results can be obtained for this data: as the solution smoothes out in later time steps the Roe average gives increasingly better estimates of the true solution, and convergence is obtained.

In Figure 15.3(b) the outflow velocity is increased to $u_r = -u_l = 1.8$. In this case the true Riemann solution still has a positive value of h_m, but in the approximate Riemann solution the curves intersect at a negative value of $\hat{h}_m$. The code will typically then fail when the sound speed is computed as $\sqrt{gh}$.

Other averages give similar results. For the Euler equations it has been shown by Einfeldt, Munz, Roe, & Sjogreen [122] that for certain Riemann problems there is no linearization that will preserve positivity, and other approaches to approximating the Riemann solution must be used. They call a method *positively conservative* for the Euler equations if the density and internal energy always remain positive for any physical data. They show that Godunov's method with the exact Riemann solver is positively conservative, and also show this for methods based on the HLLE approximate Riemann solver described in the next section.

15.3.7 The HLL and HLLE Solvers

A simple approximate Riemann solver can be based on estimating the smallest and largest wave speeds arising in the Riemann solution and then taking $\hat{Q}(x/t)$ to consist of only two waves propagating at these speeds $s^1_{i-1/2}$ and $s^2_{i-1/2}$. There will then be a single new state $\hat{Q}_{i-1/2}$ in between, and as waves we use

$$\mathcal{W}^1_{i-1/2} = \hat{Q}_{i-1/2} - Q_{i-1} \quad \text{and} \quad \mathcal{W}^2_{i-1/2} = Q_i - \hat{Q}_{i-1/2}.$$

We can determine the state $\hat{Q}_{i-1/2}$ by requiring that the approximate solution be conservative, which requires

$$s^1_{i-1/2}\big(\hat{Q}_{i-1/2} - Q_{i-1}\big) + s^2_{i-1/2}\big(Q_i - \hat{Q}_{i-1/2}\big) = f(Q_i) - f(Q_{i-1}) \tag{15.58}$$

and so

$$\hat{Q}_{i-1/2} = \frac{f(Q_i) - f(Q_{i-1}) - s^2_{i-1/2}Q_i + s^1_{i-1/2}Q_{i-1}}{s^1_{i-1/2} - s^2_{i-1/2}}. \tag{15.59}$$

Approximate Riemann solvers of this type were studied by Harten, Lax, and van Lear [187] and further developed by Einfeldt [121], who suggested a choice of s^1 and s^2 in the context of gas dynamics that can be generalized to

$$\begin{aligned} s^1_{i-1/2} &= \min_p\big(\min\big(\lambda^p_i, \hat{\lambda}^p_{i-1/2}\big)\big),\\ s^2_{i-1/2} &= \max_p\big(\max\big(\lambda^p_{i+1}, \hat{\lambda}^p_{i-1/2}\big)\big). \end{aligned} \tag{15.60}$$

Here λ^p_j is the pth eigenvalue of the Jacobian $f'(Q_j)$, and $\hat{\lambda}^p_{i-1/2}$ is the pth eigenvalue of the Roe average (for problems where this average is easily defined). In the original HLL method of [187], the values $s^1_{i-1/2}$ and $s^2_{i-1/2}$ are chosen as some lower and upper bounds on all the characteristic speeds that might arise in the true Riemann solution. The choice (15.60) might not satisfy this, but in practice gives sharper results for shock waves since the shock speed is smaller than the characteristic speed behind the shock and in this case (15.60) reduces to the Roe approximation of the shock speed. In particular, the HLLE method shares the nice property of the Roe solver that for data connected by a single shock wave, the approximate solution agrees with the true solution. In the case where the slowest or fastest wave is a rarefaction wave, the formula (15.60) will use the corresponding characteristic speed, which is faster than the Roe average speed in this case. In general it is not necessary to use an "entropy fix" when using this solver. It is also shown in [122] that this method is positively conservative, as discussed in the previous section, and hence may be advantageous for problems where low densities are expected.

A disadvantage of this solver is that the full Riemann solution structure is modeled by only two waves based on approximate speeds of the fastest and slowest waves in the system. For a system of more than two equations this may lead to a loss of resolution for waves traveling at intermediate speeds. For the Euler equations, for example, this approximation is based only on the two acoustic waves while the contact discontinuity is ignored. The resulting numerical solutions show relatively poor resolution of the contact discontinuity as a result.

A modified HLLE method (denoted by HLLEM) is proposed in [121] that attempts to capture a contact discontinuity more accurately by introducing a piecewise linear function as the approximate solution, where the constant intermediate state (15.59) is replaced by a linear function with the same total integral for conservation. This function is based on information about the contact discontinuity. The HLLEC method described in [450] is another approach that introduces a third wave into the approximation.

The introduction of a third wave and hence two intermediate states is also discussed by Harten, Lax, and van Lear [187]. They suggest choosing the speed of this third wave as

$$V = \frac{[\eta'(Q_i) - \eta'(Q_{i-1})] \cdot [f(Q_i) - f(Q_{i-1})]}{[\eta'(Q_i) - \eta'(Q_{i-1})] \cdot [Q_i - Q_{i-1}]} \tag{15.61}$$

for problems with a convex entropy function $\eta(q)$. It can then be shown that V lies between the smallest and largest eigenvalues of the matrix $\hat{A}_{\text{HLL}}$ discussed at the end of Section 15.3.2. Moreover, if Q_{i-1} and Q_i are connected by a single wave, i.e., if $f(Q_i) - f(Q_{i-1}) = s(Q_i - Q_{i-1})$ for some scalar s, then $V = s$, and so this solver, like the Roe solver, will reproduce the exact Riemann solution in this case. Linde [300] has recently developed this approach further.

15.4 High-Resolution Methods for Nonlinear Systems

Godunov's method (or one of the variants based on approximate Riemann solvers) can be extended to high-resolution methods for nonlinear systems using essentially the same approach as was introduced in Section 6.13 for linear systems. The formulas have already been introduced in Section 6.15. The method takes the form

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\big(\mathcal{A}^- \Delta Q_{i+1/2} + \mathcal{A}^+ \Delta Q_{i-1/2}\big) - \frac{\Delta t}{\Delta x}\big(\tilde{F}_{i+1/2} - \tilde{F}_{i-1/2}\big), \tag{15.62}$$

where $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ are the fluctuations corresponding to Godunov's method or one of its variants. The flux $\tilde{F}_{i-1/2}$ is the high-resolution correction given by

$$\tilde{F}_{i-1/2} = \frac{1}{2}\sum_{p=1}^{M_w} \left|s_{i-1/2}^p\right| \left(1 - \frac{\Delta t}{\Delta x}\left|s_{i-1/2}^p\right|\right) \widetilde{\mathcal{W}}_{i-1/2}^p. \tag{15.63}$$

Recall that $\mathcal{W}_{i-1/2}^p$ is the pth wave arising in the solution to the Riemann problem at $x_{i-1/2}$ and $\widetilde{\mathcal{W}}_{i-1/2}^p$ is a limited version of this wave. In the constant-coefficient linear case this limited wave is computed by comparing the magnitude of $\mathcal{W}_{i-1/2}^p$ to $\mathcal{W}_{I-1/2}^p$, the corresponding wave from the neighboring Riemann problem in the upwind direction ($I = i \pm 1$ depending on the sign of the wave speed s^p as in (6.61)).

If a linearized Riemann solver such as the Roe solver is used, then $M_w = m$, and we will generally assume this case below. However, high-resolution corrections of this form can also be applied to other Riemann solvers, for example the HLLE method for which $M_w = 2$. See Example 15.1 for a demonstration of the improvement this makes over the Godunov method with this solver.

Several difficulties arise with nonlinear systems that are not seen with a linear system. In the linear case each wave $\mathcal{W}^p$ is a sharp discontinuity traveling at a single speed $s^p = \lambda^p$. For a nonlinear system, shock waves and contact discontinuities have this form, but rarefaction waves do not. We can still define a wave strength $\mathcal{W}^p$ as the total jump across the wave,

$$\mathcal{W}^p = q_r^p - q_l^p, \tag{15.64}$$

where q_l^p and q_r^p are the states just to the left and right of the wave. However, there is not a single wave speed s^p to use in the formula (15.63). Instead, the characteristic speed λ^p varies continuously through the rarefaction wave. In practice an average speed, e.g.,

$$s^p = \frac{1}{2}\left(\lambda_l^p + \lambda_r^p\right),$$

can generally be used successfully. Recall that the form of the correction terms in (15.62) guarantees that the method will be conservative regardless of the manner in which s^p is chosen, so this is not an issue.

Another way to deal with rarefaction waves is to simply use a discontinuous approximation instead, such as an entropy-violating shock or, more commonly, the approximate solution obtained with a linearized Riemann solver (e.g., the Roe solver) as described in Section 15.3.2. In particular, if a linearized solver is being used to determine the fluctuations $\mathcal{A}^{\pm}\Delta Q$, these same waves $\mathcal{W}^p$ and the corresponding eigenvalues $s^p = \hat{\lambda}^p$ can be used directly in (15.63). If an entropy fix is applied to modify $\mathcal{A}^{\pm}\Delta Q$, the original waves can generally still be used for the high-resolution correction terms with good results. Even if the exact Riemann solver is used for the first-order fluctuations, as may be necessary for some difficult problems where the linearized solver does not suffice, it may still be possible to use a linearized solver to obtain the waves and speeds needed for the high-resolution corrections.

Another issue that arises for nonlinear systems is that the waves $\mathcal{W}^p_{i-1/2}$ and $\mathcal{W}^p_{I-1/2}$ are generally not collinear vectors in state space, and so applying a limiter based on comparing the magnitude of these vectors is not as simple as for a constant-coefficient linear system (where the eigenvectors r^p of $A = f'(q)$ are constant). This difficulty has already been discussed in the context of variable-coefficient linear systems in Section 9.13. Similar approaches can be taken for nonlinear systems. The default approach in CLAWPACK is to project the wave $\mathcal{W}^p_{I-1/2}$ from the neighboring Riemann problem onto $\mathcal{W}^p_{i-1/2}$ in order to obtain a vector that can be directly compared to $\mathcal{W}^p_{i-1/2}$ as described in Section 9.13.

Example 15.1. As an example we solve one standard test problem using several different methods. The Euler equations are solved with initial data $\overset{\circ}{\rho}(x) \equiv 1$, $\overset{\circ}{u}(x) \equiv 0$, and pressure

$$\overset{\circ}{p}(x) = \begin{cases} 1000 & \text{if } 0 \le x \le 0.1, \\ 0.01 & \text{if } 0.1 \le x \le 0.9, \\ 100 & \text{if } 0.9 \le x \le 1.0. \end{cases} \tag{15.65}$$

This problem was first used as a test problem by Woodward & Colella [487] and is often referred to as the *Woodward–Colella blast-wave problem*. The two discontinuities in the initial data each have the form of a shock-tube problem and yield strong shock waves and contact discontinuities going inwards and rarefaction waves going outwards. The boundaries $x = 0$ and $x = 1$ are both reflecting walls and the reflected rarefaction waves interact with the other waves.

Figure 15.4 illustrates the structure of the solution in the x–t plane, showing contour lines of both density and pressure. The two shock waves collide at about time $t = 0.27$ and generate an additional contact discontinuity. The right-going shock then collides with the contact discontinuity arising from the Riemann problem at $x = 0.9$, deflecting it to the right. Solutions are often compared at time $t = 0.038$, when the solution consists of contact discontinuities near $x = 0.6$, $x = 0.76$, and $x = 0.8$ and shock waves near $x = 0.65$ and $x = 0.87$. This is a challenging test problem because of the strength of the shocks involved and the interaction of the different waves.

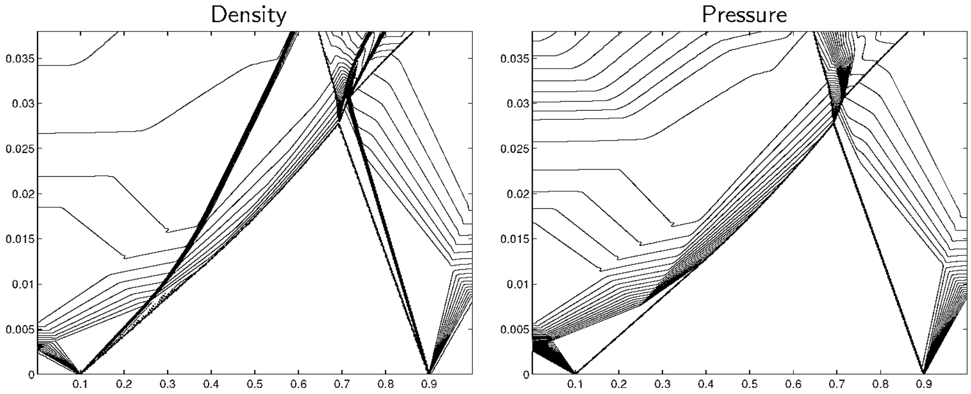


Fig. 15.4. Solution to the Woodward–Colella blast-wave problem in the x–t plane.

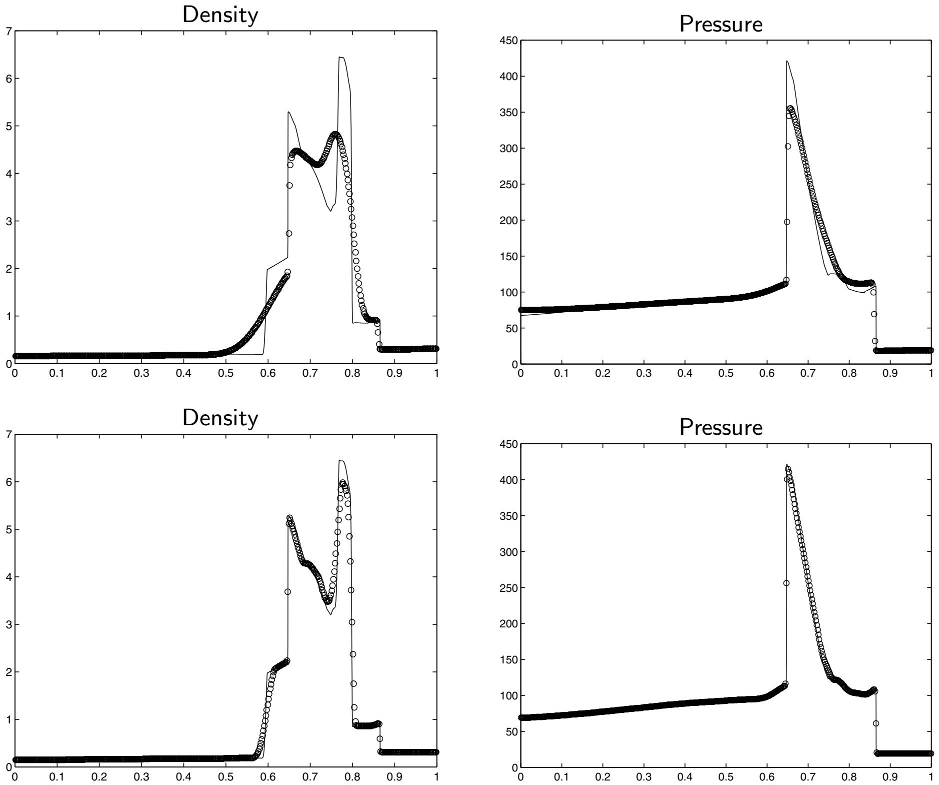


Fig. 15.5. Solution to the Woodward–Colella blast-wave problem at time $t = 0.38$ computed with the Roe solver. Top: Godunov method. Bottom: High-resolution method. [claw/book/chap15/wcblast]

Figure 15.5 shows results obtained on a grid with 500 cells using the Roe solver and either the first-order Godunov method (top) or the high-resolution method with the MC limiter (bottom). The solid line shows results obtained with the same method on a grid with 4000 cells. Note that with the high-resolution method the shocks are captured very sharply and are in the correct locations. The contact discontinuities are considerably more smeared out, however (even in the computation on the finer grid). This is typically seen in computations with the Euler equations. The nonlinearity that causes a shock wave to form also tends to keep it sharp numerically. A contact discontinuity is a linearly degenerate wave for which the characteristics are parallel to the wave on each side. This wave simply continues to smear further in each time step with no nonlinear sharpening effect. Notice that the pressure is continuous across the contact discontinuities and is well captured in spite of the errors in the density.

Figure 15.6 shows results obtained on a grid with 500 cells using the simpler HLLE solver and either the first-order Godunov method (top) or the high-resolution method with the MC limiter (bottom). Recall that this solver only uses two waves with speeds that approximate the acoustic speeds and hence does not attempt to model the contact discontinuity at all. In spite of this the solution has the correct structure, although with considerably more smearing of the contact discontinuities and less accuracy overall than the Roe solver provides.

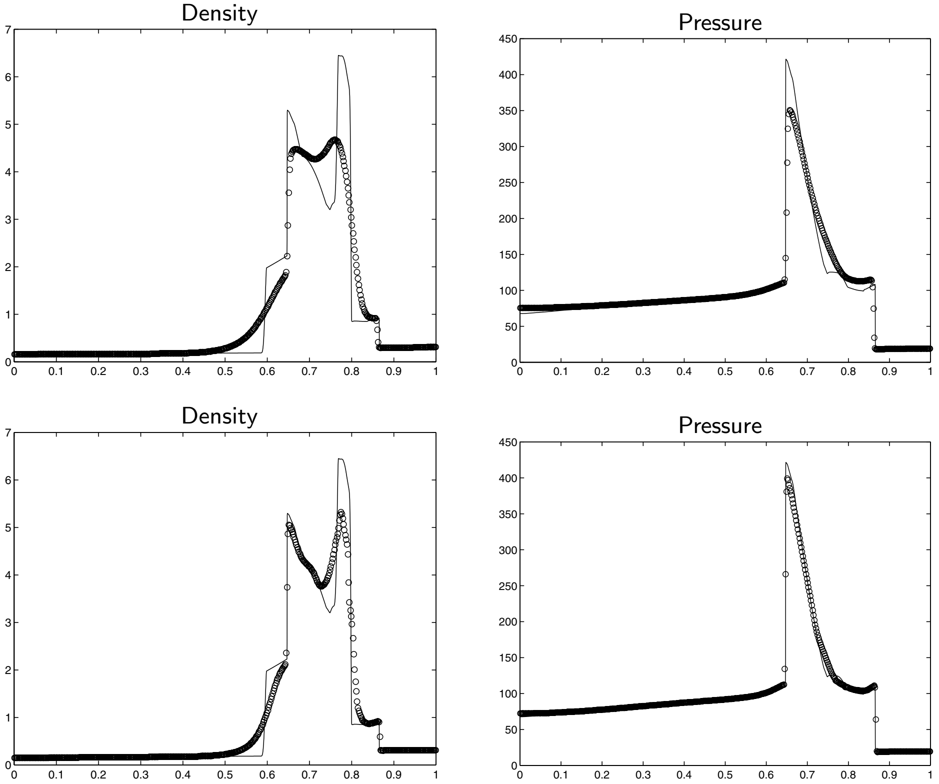


Fig. 15.6. Solution to the Woodward–Colella blast-wave problem at time $t = 0.38$ computed with the HLLE solver. Top: Godunov method. Bottom: High-resolution method. [claw/book/chap15/wcblast]

15.5 An Alternative Wave-Propagation Implementation of Approximate Riemann Solvers

The high-resolution wave-propagation method (15.62) is based on the assumption that $Q_i - Q_{i-1}$ has been split into waves as in (15.7) and the fluctuations $\mathcal{A}^{\pm}\Delta Q_{i-1/2}$ defined using either (15.9) or (15.10). An alternative approach is to first split the jump in f into "waves"

$$f(Q_i) - f(Q_{i-1}) = \sum_{p=1}^{M_w} \mathcal{Z}_{i-1/2}^p \tag{15.66}$$

moving at speed $s_{i-1/2}^p$, and then define the fluctuations and correction terms directly from the $\mathcal{Z}^p$. This viewpoint is useful in applying some approximate Riemann solvers, and will be used in showing the second-order accuracy of wave-propagation methods in the next section. It also appears to be quite useful in the context of spatially-varying flux functions (see Section 16.4), as explored in [18]. See also [288] for a more general approach where the jumps in both Q and $f(Q)$ are simultaneously split into waves.

If a linearized Riemann solver is used, then the vector $f(Q_i) - f(Q_{i-1})$ can be decomposed as a linear combination of the eigenvectors $\hat{r}_{i-1/2}^p$ of the linearized matrix $\hat{A}_{i-1/2}$.

Instead of solving the system (15.14), we solve

$$f(Q_i) - f(Q_{i-1}) = \sum_{p=1}^{m} \beta_{i-1/2}^p \hat{r}_{i-1/2}^p \tag{15.67}$$

for the coefficients $\beta_{i-1/2}^p$ and then define

$$\mathcal{Z}_{i-1/2}^p = \beta_{i-1/2}^p \hat{r}_{i-1/2}^p. \tag{15.68}$$

If the wave speeds are all nonzero, then we can recover waves $\mathcal{W}_{i-1/2}^p$ by setting

$$\mathcal{W}_{i-1/2}^p = \frac{1}{s_{i-1/2}^p} \mathcal{Z}_{i-1/2}^p, \tag{15.69}$$

and view this as an alternative way to obtain an approximate Riemann solution in the standard form needed for wave propagation. An advantage of this approach is that using the condition (15.66) to define the $\mathcal{Z}_{i-1/2}^p$ guarantees that the method will be conservative when the fluctuations (15.10) are used. This is true for any linearization, for example the simple arithmetic average $\hat{A}_{i-1/2} = f'(\frac{1}{2}(Q_{i-1} + Q_i))$, whereas (15.16) may not be satisfied if the wave splitting is based on (15.14) unless $\hat{A}_{i-1/2}$ is chosen to be a special average such as the Roe average. When the Roe average is used, for which (15.18) is satisfied, the two approaches give exactly the same splitting, since (15.67) then yields

$$\hat{A}_{i-1/2}(Q_i - Q_{i-1}) = \sum_{p=1}^{m} \beta_{i-1/2}^p \hat{r}_{i-1/2}^p$$

and applying $\hat{A}_{i-1/2}$ to (15.14) shows that $\beta_{i-1/2}^p = s_{i-1/2}^p \alpha_{i-1/2}^p$.

The wave-propagation methods can be written directly in terms of the waves $\mathcal{Z}_{i-1/2}^p$ in a manner that avoids needing to form the $\mathcal{W}_{i-1/2}^p$ at all, which is more satisfying in cases where a wave speed is near zero and (15.69) might break down. The fluctuations can be rewritten as

$$\begin{aligned} \mathcal{A}^- \Delta Q_{i-1/2} &= \sum_{p:s_{i-1/2}^p < 0} \mathcal{Z}_{i-1/2}^p, \\ \mathcal{A}^+ \Delta Q_{i-1/2} &= \sum_{p:s_{i-1/2}^p > 0} \mathcal{Z}_{i-1/2}^p. \end{aligned} \tag{15.70}$$

The second-order correction terms (15.63) can also be rewritten in terms of the $\mathcal{Z}_{i-1/2}^p$ by combining one factor of $s_{i-1/2}^p$ with $\mathcal{W}_{i-1/2}^p$, at least in the case where no limiter is used so that $\widetilde{\mathcal{W}}_{i-1/2}^p = \mathcal{W}_{i-1/2}^p$ in (15.63). We obtain

$$\tilde{F}_{i-1/2} = \frac{1}{2} \sum_{p=1}^{M_w} \operatorname{sgn}\left(s_{i-1/2}^p\right) \left(1 - \frac{\Delta t}{\Delta x} \left|s_{i-1/2}^p\right|\right) \mathcal{Z}_{i-1/2}^p. \tag{15.71}$$

In Section 15.6 we will show that the resulting method is second-order accurate with a reasonable consistency condition on the Riemann solver.

To obtain high-resolution results, limiters can now be applied to the vectors $\mathcal{Z}^p_{i-1/2}$ rather than to the $\mathcal{W}^p_{i-1/2}$, using any standard limiting techniques. This appears to work as well in practice as the standard approach and allows a broader range of linearizations to be used. If the Roe linearization is used, then the two approaches give identical methods in the unlimited case, though they will be slightly different if the limiters are applied to the $\mathcal{Z}^p_{i-1/2}$ rather than to the $\mathcal{W}^p_{i-1/2}$.

15.6 Second-Order Accuracy

For smooth solutions we would like to confirm second-order accuracy of the method (15.62), at least if the limiters are suppressed and $\widetilde{\mathcal{W}}^p_{i-1/2}$ is replaced by $\mathcal{W}^p_{i-1/2}$ in (15.63), or the unlimited waves $\mathcal{Z}^p_{i-1/2}$ are used in the formulation of Section 15.5. Having built this method up from Godunov's method (based on Riemann solutions) and correction terms in each characteristic field (based on scalar theory), it is not obvious that second-order accuracy will be obtained for nonlinear systems, especially when approximate Riemann solvers are used. To confirm that it is, one must compute the local truncation error or, equivalently, compare the numerical updating formula with the Taylor series expansion. For the conservation law $q_t + f(q)_x = 0$, we have

$$\begin{aligned} q_t &= -f(q)_x, \\ q_{tt} &= -(f'(q)q_t)_x = \left[f'(q)f(q)_x\right]_x, \end{aligned} \tag{15.72}$$

and so

$$q(x_i, t_{n+1}) = q(x_i, t_n) - \Delta t f(q)_x + \frac{1}{2}\Delta t^2\left[f'(q)f(q)_x\right]_x + \mathcal{O}(\Delta t^3), \tag{15.73}$$

where all terms on the right are evaluated at (x_i, t_n).

To obtain an expression that matches this to the desired order from the numerical method, we will need to make an assumption on the accuracy of the Riemann solver. For arbitrary data Q_{i-1} and Q_i we assume that the method uses a flux-difference splitting of the form

$$f(Q_i) - f(Q_{i-1}) = \sum_{p=1}^{m} \mathcal{Z}^p_{i-1/2} \tag{15.74}$$

where the vectors $\mathcal{Z}^p_{i-1/2}$ are the eigenvectors of some matrix $\hat{A}_{i-1/2} = \hat{A}(Q_{i-1}, Q_i)$ corresponding to eigenvalues $s^p_{i-1/2}$. Either the $\mathcal{Z}^p_{i-1/2}$ are computed directly as described in Section 15.5, or else we define

$$\mathcal{Z}^p_{i-1/2} = s^p_{i-1/2}\mathcal{W}^p_{i-1/2} \tag{15.75}$$

in terms of the waves $\mathcal{W}^p_{i-1/2}$ computed from the decomposition (15.14).

To obtain second-order accuracy, we must make a mild assumption on the consistency of the matrix-valued function $\hat{A}(q_l, q_r)$ with the Jacobian $f'(q)$. If $q(x)$ is a smooth function of x, then we require that

$$\hat{A}(q(x),\, q(x + \Delta x)) = f'(q(x + \Delta x/2)) + E(x, \Delta x), \tag{15.76}$$

where the error $E(x, \Delta x)$ satisfies

$$E(x, \Delta x) = \mathcal{O}(\Delta x) \quad \text{as } \Delta x \to 0 \tag{15.77}$$

and

$$\frac{E(x + \Delta x, \Delta x) - E(x, \Delta x)}{\Delta x} = \mathcal{O}(\Delta x) \quad \text{as } \Delta x \to 0. \tag{15.78}$$

In particular, if

$$\hat{A}(q(x),\ q(x + \Delta x)) = f'(q(x + \Delta x/2)) + \mathcal{O}(\Delta x^2), \tag{15.79}$$

then both of these conditions will be satisfied. Hence we can choose

$$\hat{A}(Q_{i-1}, Q_i) = f'\big(\hat{Q}_{i-1/2}\big) \tag{15.80}$$

with $\hat{Q}_{i-1/2} = \frac{1}{2}(Q_{i-1} + Q_i)$ or with the Roe average and obtain a second-order method. The form of the conditions (15.77) and (15.78) allows more flexibility, however. The matrix $\hat{A}$ need only be a first-order accurate approximation to f' at the midpoint provided that the error is smoothly varying. This allows, for example, taking $\hat{Q}_{i-1/2} = Q_{i-1}$ or Q_i in (15.80), provided the same choice is made at all grid points.

To verify the second-order accuracy of a method satisfying this consistency condition, we write out the updating formula (15.62) for Q_i^{n+1} using the fluctuations (15.70) and the corrections (15.71). This gives

$$\begin{aligned}
Q_i^{n+1} &= Q_i^n - \frac{\Delta t}{\Delta x}\left[\sum_{p:s_{i-1/2}^p>0} \mathcal{Z}_{i-1/2}^p + \sum_{p:s_{i-1/2}^p<0} \mathcal{Z}_{i+1/2}^p\right] \\
&\quad - \frac{\Delta t}{2\,\Delta x}\left[\sum_{p=1}^m \operatorname{sgn}\big(s_{i+1/2}^p\big)\left(1 - \frac{\Delta t}{\Delta x}\big|s_{i+1/2}^p\big|\right)\mathcal{Z}_{i+1/2}^p \right. \\
&\qquad \left. - \sum_{p=1}^m \operatorname{sgn}\big(s_{i-1/2}^p\big)\left(1 - \frac{\Delta t}{\Delta x}\big|s_{i-1/2}^p\big|\right)\mathcal{Z}_{i-1/2}^p\right] \\
&= Q_i^n - \frac{\Delta t}{2\,\Delta x}\left[\sum_{p=1}^m \mathcal{Z}_{i-1/2}^p + \sum_{p=1}^m \mathcal{Z}_{i+1/2}^p\right] \\
&\quad + \frac{\Delta t^2}{2\,\Delta x^2}\left[\sum_{p=1}^m s_{i+1/2}^p \mathcal{Z}_{i+1/2}^p - \sum_{p=1}^m s_{i-1/2}^p \mathcal{Z}_{i-1/2}^p\right] \\
&= Q_i^n - \frac{\Delta t}{2\,\Delta x}\left[\sum_{p=1}^m \mathcal{Z}_{i-1/2}^p + \sum_{p=1}^m \mathcal{Z}_{i+1/2}^p\right] \\
&\quad + \frac{\Delta t^2}{2\,\Delta x^2}\left[\hat{A}_{i+1/2}\sum_{p=1}^m \mathcal{Z}_{i+1/2}^p - \hat{A}_{i-1/2}\sum_{p=1}^m \mathcal{Z}_{i-1/2}^p\right].
\end{aligned} \tag{15.81}$$

To obtain the last line we have used the fact that each $\mathcal{Z}^p$ is an eigenvector of the corresponding $\hat{A}$ with eigenvalue s^p. We can now use the assumption (15.66) to rewrite this as

$$\begin{aligned} Q_i^{n+1} &= Q_i^n - \frac{\Delta t}{2\,\Delta x}[f(Q_{i+1}) - f(Q_{i-1})] \\ &\quad - \frac{\Delta t^2}{2\,\Delta x^2}\left\{\hat{A}_{i+1/2}[f(Q_{i+1}) - f(Q_i)] - \hat{A}_{i-1/2}[f(Q_i) - f(Q_{i-1})]\right\}. \end{aligned} \tag{15.82}$$

This agrees with the Taylor series expansion (15.73) to sufficient accuracy that a standard computation of the truncation error now shows that the method is second-order accurate, provided that $\hat{A}$ is a consistent approximation to $f'(q)$ as described above. This follows because the conditions (15.77) and (15.78) guarantee that

$$\begin{aligned} &\hat{A}(q(x),\, q(x+\Delta x))\left(\frac{f(q(x+\Delta x)) - f(q(x))}{\Delta x}\right) \\ &\quad = f'(q(x+\Delta x/2))\, f(q(x+\Delta x/2))_x + E_2(x, \Delta x) \end{aligned} \tag{15.83}$$

with $E_2(x, \Delta x)$ satisfying the same conditions as $E(x, \Delta x)$. This in turn is sufficient to show that the final term in (15.82) agrees with the $\mathcal{O}(\Delta t^2)$ term in (15.73) to $\mathcal{O}(\Delta t^2 \Delta x)$, as required for second-order accuracy. Note that in place of the assumptions (15.77) and (15.78) on $E(x, \Delta x)$, it would be sufficient to simply assume that (15.83) holds with $E_2(x, \Delta x)$ satisfying these conditions. This is looser in the sense that only the product of $\hat{A}$ with one particular vector is required to be well behaved, not the entire matrix. This proof carries over to spatially-varying flux functions as well, as presented in [18].

15.6.1 Two-Step Lax–Wendroff Methods

It is worth noting that there are other ways to achieve second-order accuracy that do not require approximating the Jacobian matrix or its eigenstructure, in spite of the fact that the Taylor series expansion (15.73) appears to require this for the second-order terms. The need for the Jacobian can be avoided by taking a two-step approach. One example is the *Richtmyer method* of Section 4.7. When (4.23) is inserted in (4.22) and a Taylor series expansion performed, the required Jacobian terms appear, but these are not explicitly computed in the implementation where only the flux $f(q)$ is evaluated.

Another popular variant is *MacCormack's method*, originally introduced in [318]:

$$\begin{aligned} Q_i^* &= Q_i^n - \frac{\Delta t}{\Delta x}\big[f(Q_{i+1}^n) - f(Q_i^n)\big], \\ Q_i^{**} &= Q_i^* - \frac{\Delta t}{\Delta x}\big[f(Q_i^*) - f(Q_{i-1}^*)\big], \\ Q_i^{n+1} &= \frac{1}{2}\big(Q_i^n + Q_i^{**}\big). \end{aligned} \tag{15.84}$$

Note that one-sided differencing is used twice, first to one side and then to the other. The order in which the two directions are used can also be switched, or one can alternate between the two orderings in successive time steps, yielding a more symmetric method. Again, Taylor series expansion shows that the method is second-order accurate, while the explicit use of Jacobian matrices or Riemann solvers is avoided.

The problem with both the Richtmyer and MacCormack methods is that they typically produce spurious oscillations unless artificial viscosity is explicitly added, as is done in most practical calculations with these methods. But adding artificial viscosity often results in the addition of too much diffusion over most of the domain. The advantage of the high-resolution methods based on Riemann solvers is that we can tune this viscosity much more carefully to the behavior of the solution. By applying limiter functions to each characteristic field separately, we are in essence applying the optimal amount of artificial viscosity at each cell interface, and only to the fields where it is needed. This often results in much better solutions, though at some expense relative to the simpler Richtmyer or MacCormack methods.

15.7 Flux-Vector Splitting

Our focus has been on flux-difference splitting methods for conservation laws, where the flux difference $f(Q_i) - f(Q_{i-1})$ is split into fluctuations $\mathcal{A}^-\Delta Q_{i-1/2}$ (which modifies Q_{i-1}) and $\mathcal{A}^+\Delta Q_{i-1/2}$ (which modifies Q_i). This splitting is typically determined by solving a Riemann problem between the states Q_{i-1} and Q_i. There is another related approach, already introduced in Section 4.13, where instead each flux vector $f(Q_i)$ is split into a left-going part $f_i^{(-)}$ and a right-going part $f_i^{(+)}$, so we have

$$f(Q_i) = f_i^{(-)} + f_i^{(+)} . \tag{15.85}$$

We can then define the interface flux

$$F_{i-1/2} = f_{i-1}^{(+)} + f_i^{(-)} \tag{15.86}$$

based on the portion of each cell-centered flux approaching the interface. A method of this form is called a *flux-vector splitting* method, since it is the flux vector $f(Q_i)$ that is split instead of the flux difference.

As noted in Section 4.13, for constant-coefficient linear systems these two approaches are identical, but for nonlinear problems they typically differ. From a flux-vector splitting method it is possible to define fluctuations $\mathcal{A}^\pm\Delta Q_{i-1/2}$ as described in Section 4.13, so that these methods can also be expressed in the form (15.5) and implemented in CLAWPACK if desired.

15.7.1 The Steger–Warming flux

Steger and Warming [424] introduced the idea of flux-vector splitting for the Euler equations and used a special property of this system of equations, that it is *homogeneous of degree 1* (at least for certain equations of state such as that of an ideal polytropic gas). This means that $f(\alpha q) = \alpha f(q)$ for any scalar α. From this it follows, by Euler's identity for homogeneous functions, that

$$f(q) = f'(q)q \tag{15.87}$$

for any state q, which is not true for nonlinear functions that do not have this homogeneity. Hence if A_i is the Jacobian matrix $f'(Q_i)$ (or some approximation to it), then a natural flux-vector splitting is given by

$$\begin{aligned} f_i^{(-)} &= A_i^- Q_i = \sum_{p=1}^{m} \left(\lambda_i^p\right)^- \omega_i^p r_i^p, \\ f_i^{(+)} &= A_i^+ Q_i = \sum_{p=1}^{m} \left(\lambda_i^p\right)^+ \omega_i^p r_i^p, \end{aligned} \tag{15.88}$$

where the notation (4.45) is used and

$$Q_i = \sum_{p=1}^{m} \omega_i^p r_i^p \tag{15.89}$$

is the eigendecomposition of Q_i. By (15.86) we thus have

$$F_{i-1/2} = A_{i-1}^+ Q_{i-1} + A_i^- Q_i \ . \tag{15.90}$$

This is a natural generalization of (4.56) to nonlinear systems that are homogeneous of degree 1. For systems that don't have this property, we can still define a flux-vector splitting using the eigenvectors r_i^p of A_i. Instead of decomposing Q_i into the eigenvectors and then using (15.88), we can directly decompose $f(Q_i)$ into these eigenvectors,

$$f(Q_i) = \sum_{p=1}^{m} \phi_i^p r_i^p \tag{15.91}$$

and then define the flux splitting by

$$f_i^{(-)} = \sum_{p=1}^{m} \phi_i^{p(-)} r_i^p, \qquad f_i^{(+)} = \sum_{p=1}^{m} \phi_i^{p(+)} r_i^p, \tag{15.92}$$

where

$$\begin{aligned} \phi_i^{p(-)} &= \begin{cases} \phi_i^p & \text{if } \lambda_i^p < 0, \\ 0 & \text{if } \lambda_i^p \ge 0, \end{cases} \\ \phi_i^{p(+)} &= \begin{cases} 0 & \text{if } \lambda_i^p < 0, \\ \phi_i^p & \text{if } \lambda_i^p \ge 0. \end{cases} \end{aligned} \tag{15.93}$$

If the system is homogeneous of degree 1, then $\phi_i^p = \lambda_i^p r_i^p$ and (15.92) reduces to the previous expression (15.88).

The above splitting for the Euler equations is called *Steger–Warming flux-vector splitting* in the aerodynamics community. An equivalent method, known as the *beam scheme*, was introduced earlier in astrophysics [392] from a different viewpoint: each state Q_i is decomposed into distinct beams of particles traveling at the different wave speeds.

For transonic flow problems in aerodynamics, the flux-vector splitting given above suffers from the fact that the splitting does not behave smoothly as the Mach number passes through

1 (where the characteristic speed $u - c$ or $u + c$ changes sign). This can cause convergence problems when solving for a steady state. A smoother flux-vector splitting was introduced by van Leer [469]. Many variants and improvements have since been introduced, such as the AUSM method of Liou and coworkers [301], [302], [479], the Marquina flux [112], [287], [323], and kinetic flux-vector splittings [39], [63], [107], [320], [488], [489].

Use of the flux function (15.90) would only give a first-order accurate method. To obtain better accuracy one might use this flux function in an ENO-based semi discrete method and then apply Runge–Kutta time stepping (see Section 10.4) to achieve higher-order accuracy.

The flux-vector splitting can also be used in conjunction with the high-resolution method developed in Section 15.4 (and CLAWPACK) by using $F_{i-1/2}$ to define fluctuations as in (4.58) and then also defining waves $\mathcal{W}^p_{i-1/2}$ and speeds $s^p_{i-1/2}$ for use in the second-order correction terms of (15.5). From (15.89) we have

$$Q_i - Q_{i-1} = \sum_{p=1}^{m} \left(\omega^p_i r^p_i - \omega^p_{i-1} r^p_{i-1}\right),$$

which suggests defining the pth wave $\mathcal{W}^p_{i-1/2}$ as

$$\mathcal{W}^p_{i-1/2} = \omega^p_i r^p_i - \omega^p_{i-1} r^p_{i-1}.$$

The corresponding speed $s^p_{i-1/2}$ might then be defined as

$$s^p_{i-1/2} = \frac{1}{2}\left(\lambda^p_{i-1} + \lambda^p_i\right).$$

Alternatively, the formulation of Section 15.5 can be used with this same wave speed and the waves

$$\mathcal{Z}^p_{i-1/2} = \phi^p_i r^p_i - \phi^p_{i-1} r^p_{i-1},$$

in which case the fluctuations can be defined using (15.70).

15.8 Total Variation for Systems of Equations

As noted in Section 15.2, there is no proof that even the first-order Godunov method converges on general systems of nonlinear conservation laws. This is because in general there is no analogue of the TVD property for scalar problems that allows us to prove compactness and hence stability. In fact, there is not even a proof of existence of the "true solution" for general nonlinear systems of conservation laws, unless the initial data is severely restricted, even for problems where physically we know a solution exists. In a sense, this results from our inability to prove convergence of numerical methods, since one standard way of proving existence theorems is to construct a sequence of approximations (i.e., define some algorithm that could also be used numerically) and then prove that this sequence converges to a solution of the equation. Recall, for example, that this was the context in which Courant, Friedrichs, and Lewy first developed the CFL condition [93].

In this section we will briefly explore some issues related to variation and oscillations in nonlinear systems. We start by looking at some of the difficulties inherent in trying to obtain total variation bounds. For a system of m equations, we might try to define the total variation by

$$\mathrm{TV}(q) = \sup \sum_{j=1}^{N} \|q(\xi_j) - q(\xi_{j-1})\|, \tag{15.94}$$

where the supremum is taken over all subdivisions of the real line $-\infty = \xi_0 < \xi_1 < \cdots < \xi_N = \infty$, generalizing (6.19) by replacing the absolute value by some vector norm in $\mathbb{R}^m$. We will be particularly concerned with piecewise constant grid functions, in which case (15.94) reduces to

$$\mathrm{TV}(Q) = \sum_{i=-\infty}^{\infty} \|Q_i - Q_{i-1}\|. \tag{15.95}$$

With this definition of TV, and similar replacement of absolute value by the vector norm elsewhere in the notions of convergence used in the Lax–Wendroff theorem, this theorem continues to hold. We might also hope that numerical methods will produce solutions that have bounded total variation in this sense, in which case we could prove stability and convergence.

In general, however, we cannot hope to develop methods that are TVD with this definition of TV, because the true solution is itself not TVD. In fact, the total variation can increase by an arbitrarily large amount over an arbitrarily short time if we choose suitable data, and so we cannot even hope to obtain a bound of the form $\mathrm{TV}(Q^{n+1}) \le (1 + \alpha\,\Delta t)\,\mathrm{TV}(Q^n)$. To see this, consider the simple example in Section 13.4, the Riemann problem for the shallow water equations in which $h_l = h_r$ and $u_l = -u_r > 0$ (two equal streams of water smashing into one another). The initial data has no variation in h, and the variation in hu is $2h_l u_l$. For any time $t > 0$ the variation in hu is still $2h_l u_l$, but the depth near $x = 0$ increases to $h_m > h_l$, and so h has total variation $2(h_m - h_l) > 0$. By choosing u_l large enough we can make this increase in variation arbitrarily large, because h_m increases with u_l.

For certain systems of equations it is possible to prove stability by measuring the total variation in terms of *wave strengths* instead of using standard vector norms in $\mathbb{R}^m$. A simple example is a constant-coefficient linear system, as considered in the next section.

15.8.1 Total Variation Estimates for Linear Systems

Consider a linear system $q_t + Aq_x = 0$. Note that linear systems can exhibit exactly the same growth in TV as seen in the shallow water example above. Consider the Riemann problem for the the acoustics equations with $p_l = p_r$ and $u_l = -u_r > 0$, for example, which behaves just as described above. In spite of this, for a constant-coefficient linear system we can easily prove convergence of standard methods by diagonalizing the system, decoupling it into independent scalar advection equations that have the TVD property. For example, Godunov's method for a linear system can be written as

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left[A^+\left(Q_i^n - Q_{i-1}^n\right) + A^-\left(Q_{i+1}^n - Q_i^n\right)\right].$$

Multiplying by R^{-1} (the matrix of left eigenvectors) and defining $W_i^n = R^{-1}Q_i^n$, we obtain

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left[\Lambda^+\left(W_i^n - W_{i-1}^n\right) + \Lambda^-\left(W_{i+1}^n - W_i^n\right)\right],$$

where $\Lambda^\pm = R^{-1}A^\pm R$. This is an uncoupled set of m first-order upwind algorithms for the characteristic variables, each of which is TVD and convergent. It follows that $Q_i^n = RW_i^n$ is also convergent.

This suggests that we define the total variation of Q by using a vector norm such as

$$\|Q\|_W \equiv \|R^{-1}Q\|_1, \tag{15.96}$$

where $\|\cdot\|_1$ is the standard 1-norm in $\mathbb{R}^m$. Since R^{-1} is nonsingular, this defines a vector norm. Then we can define the corresponding total variation by

$$\begin{aligned}\mathrm{TV}_W(Q) &= \sum_{i=-\infty}^{\infty} \|Q_i - Q_{i-1}\|_W\\ &= \sum_{i=-\infty}^{\infty} \|R^{-1}(Q_i - Q_{i-1})\|_1\\ &= \sum_{i=-\infty}^{\infty} \|W_i - W_{i-1}\|_1\\ &= \sum_{i=-\infty}^{\infty}\sum_{p=1}^{m} \left|W_i^p - W_{i-1}^p\right|\\ &= \sum_{p=1}^{m} \mathrm{TV}(W^p),\end{aligned} \tag{15.97}$$

where $\mathrm{TV}(W^p)$ is the scalar total variation of the pth characteristic component. Since the scalar upwind method is TVD, we have

$$\mathrm{TV}((W^p)^{n+1}) \le \mathrm{TV}((W^p)^n),$$

and hence, with the definition (15.97) of total variation, we can show that

$$\mathrm{TV}_W(Q^{n+1}) \le \mathrm{TV}_W(Q^n). \tag{15.98}$$

For the exact solution to a linear system, this same approach shows that $\mathrm{TV}_W(q(\cdot, t))$ remains constant in time, since each scalar advection equation for the characteristic variable $w^p(x, t)$ maintains constant variation.

From this result we can also show that other forms of the total variation, such as (15.94), remain uniformly bounded even if they are not diminishing. We have

$$\begin{aligned}
\mathrm{TV}_1(Q^n) &= \sum_i \|Q_i - Q_{i-1}\|_1 \\
&= \sum_i \left\| R\left(W_i^n - W_{i-1}^n\right)\right\|_1 \\
&\le \|R\|_1 \sum_i \left\| W_i^n - W_{i-1}^n \right\|_1 \\
&= \|R\|_1 \,\mathrm{TV}_W(Q^n) \\
&\le \|R\|_1 \,\mathrm{TV}_W(Q^0) \\
&= \|R\|_1 \sum_i \left\| R^{-1}\left(Q_i^0 - Q_{i-1}^0\right)\right\|_1 \\
&\le \|R\|_1 \, \|R^{-1}\|_1 \,\mathrm{TV}_1(Q^0).
\end{aligned} \tag{15.99}$$

Hence TV_1 grows by at most a factor of $\|R\|_1 \|R^{-1}\|_1$, the condition number of the eigenvector matrix, over any arbitrary time period. It is natural for this condition number to appear in the bound, since it measures how nearly linearly dependent the eigenvectors of A are. Recalling the construction of the Riemann solution from Chapter 3, we know that eigenvectors that are nearly linearly dependent can give rise to large variation in the Riemann solutions. (See also the example in Section 16.3.1.)

Note that if we write the Riemann solution as a sum of waves,

$$Q_i - Q_{i-1} = \sum_p \mathcal{W}_{i-1/2}^p = \sum_p \alpha_{i-1/2}^p r^p,$$

as introduced in Section 3.8, then we can also express $\mathrm{TV}_W(Q)$ as

$$\mathrm{TV}_W(Q) = \sum_i \sum_p \left|\alpha_{i-1/2}^p\right|. \tag{15.100}$$

Recall that in defining the basis of eigenvectors r^p (and hence the matrix R) we could choose any convenient normalization. For our present purposes it is most convenient to assume that r^p is chosen to have $\|r^p\|_1 = 1$. Then

$$\left|\alpha_{i-1/2}^p\right| = \left\|\alpha_{i-1/2}^p r^p\right\|_1 = \left\|\mathcal{W}_{i-1/2}^p\right\|_1,$$

and we simply have

$$\mathrm{TV}_W(Q) = \sum_i \sum_p \left\|\mathcal{W}_{i-1/2}^p\right\|_1. \tag{15.101}$$

Thus another interpretation of TV_W is that it is the sum of the wave strengths over all waves arising in all Riemann problems at cell interfaces. This is illustrated in Figure 15.7. For a linear system all waves simply pass through one another (linear superposition) with no change in their strength as time advances. This is not true for nonlinear systems.

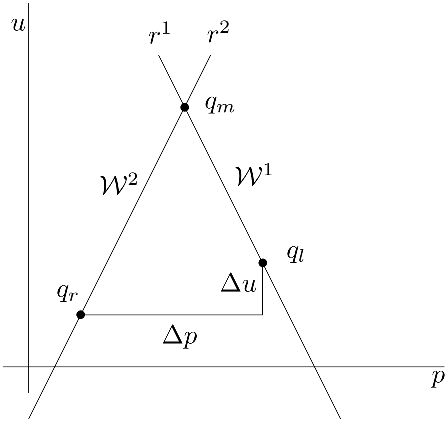


Fig. 15.7. Illustration of $\text{TV}_W(Q)$ for Riemann-problem data in the acoustics equations. $\|q_r - q_l\|_1 = |\Delta p| + |\Delta u|$, whereas $\|q_r - q_l\|_W = \|\mathcal{W}^1\|_1 + \|\mathcal{W}^2\|_1$.

15.8.2 Wave-Strength Estimates for Nonlinear Systems

For a linear system of equations we have just seen that we can prove the solution is total variation bounded by measuring the variation of a piecewise constant solution in terms of the metric

$$d_W(q_l, q_r) \equiv \sum_p \|\mathcal{W}^p\|_1, \tag{15.102}$$

where the $\mathcal{W}^p$ represent waves arising in the Riemann solution between states q_l and q_r. Then

$$\text{TV}_W(Q) = \sum_i d_W(Q_{i-1}, Q_i). \tag{15.103}$$

We can make the same definition of total variation in the case of a nonlinear hyperbolic system. For a linear system this metric can be rewritten in terms of a norm $\|\cdot\|_W$, since the value of $d_W(q_l, q_r)$ depends only on the value $q_r - q_l$. This is what was used in the previous section. For a nonlinear problem this is no longer true. The wave strengths in the Riemann solution between two states q_l' and q_r' might be quite different from those in the Riemann solution between q_l and q_r even if $q_r' - q_l' = q_r - q_l$.

If we attempt to use TV_W as defined in (15.103) to study the stability of Godunov's method on nonlinear problems we run into two difficulties:

1. In general the total variation TV_W may not be diminishing even in the true solution. Consider, for example, data that consists of two approaching shocks that collide and produce two outgoing shocks. For a nonlinear problem it might happen that the outgoing shocks are *stronger* than the incoming shocks (with larger $\|\mathcal{W}^p\|_1$).
2. The averaging process in Godunov's method introduces new states into the approximate solution that may not appear in the true solution. Because the structure of the Hugoniot loci and integral curves can vary rapidly in state space, this may introduce additional unphysical variation when Riemann problems are solved between these new states.

For certain special systems of equations these difficulties can be overcome and stability proved. For example, for certain systems of two equations the integral curves and Hugoniot loci are identical. In this case it can be shown that that these curves are in fact straight lines in state space. Such systems have been studied by Temple [448], [447], who obtained total variation bounds for the true solution and used these to prove existence of solutions. It is also possible to obtain TV estimates when Godunov's method is applied to such a system, and hence convergence can be proved [291]. Similar results have also been obtained for more general nonlinear systems with straight-line fields [46]. For general nonlinear systems, however, such results are not currently available.

15.8.3 Glimm's Random-Choice Method

For systems such as the Euler equations it is not possible to prove convergence of Godunov's method. In fact, it is not even possible to prove that weak solutions exist for all time if we allow arbitrary initial data. However, if the initial data is constrained to have sufficiently small total variation, then there is a famous proof of existence due to Glimm [152]. This proof is based on a constructive method that can also be implemented numerically, and is often called the *random-choice method*. Several variants of this method are described in various sources e.g., [43], [78], [98], [307], [316], [420], [450], [499]. Here we only summarize the major features:

- A finite volume grid with piecewise-constant data is introduced as in Godunov's method, and Riemann problems are solved at each cell interface to define a function $\tilde{q}^n(x, t)$ as in Section 4.10. However, instead of averaging the resulting solutions over grid cells, the value of $\tilde{q}^n(x, t_{n+1})$ at a random point x in each cell is chosen. This avoids difficulty 2 mentioned in the previous subsection.
- A functional similar to (15.103) is used to measure the solution, but an additional quadratic term is introduced that measures the potential for future interaction between each pair of waves that are approaching one another. This is necessary so as to take into account the potential increase in variation that can arise when two waves interact. The quadratic functional is the crux of Glimm's proof, since he was able to show that a suitable choice results in a functional that is nonincreasing in time (for data with sufficiently small variation).

Computationally, the random-choice method has the advantage that shocks remain sharp, since the solution is sampled rather than averaged. Of course, the method cannot be exactly conservative, for this same reason, but Glimm showed convergence to a weak solution with probability 1. Another disadvantage computationally is that smooth flow is typically not very smoothly represented. There also appear to be problems extending the method to more than one space dimension [78], and for these reasons the method is not widely used for practical problems, though it may be very useful for some special ones.

Since Glimm's paper, other approaches have been introduced to prove existence and in some cases uniqueness results for certain systems of conservation laws (under suitable restrictions on the initial data), using techniques such as compensated compactness [109], [403], [446], front tracking [44], [97], [371], and semigroup methods [35],[45], [46], [48], [49], [205], [312]. For an overview of many recent theoretical results, see Dafermos [98].

15.8.4 Oscillation in Shock Computations

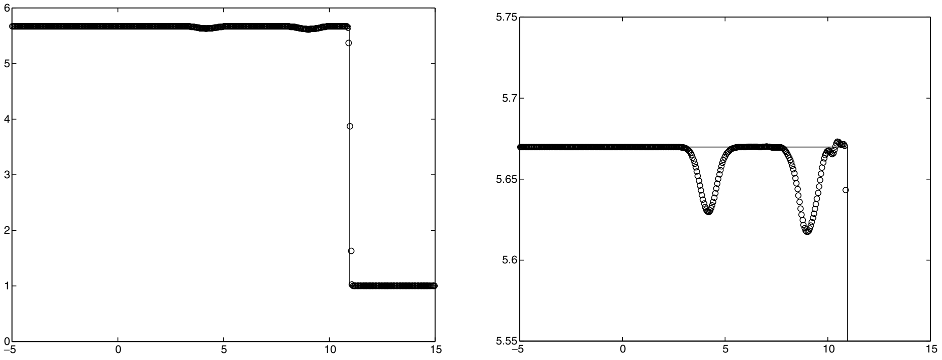
Consider a simple Riemann problem in which the data lies on a Hugoniot curve, so that the solution consists of a single shock wave. In this case the exact solution is monotonically varying and has constant total variation. In this case, at least, we might hope that a sensible numerical method based on scalar TVD theory would behave well and not introduce spurious oscillations. In practice high-resolution methods of the type we have developed do work very well on problems of this sort, most of the time. However, even the simplest first-order Godunov method can produce small spurious oscillations, which can be noticeable in some computations. Here we briefly consider two common situations where this can arise.

Start-up Errors

Figure 15.8 shows the numerical solution to the Riemann problem for the Euler equations with the data

$$\begin{aligned} &\rho_l = 5.6698, \quad u_l = 9.0299, \quad p_l = 100 \qquad \text{for } x < 0,\\ &\rho_r = 1.0, \qquad\;\; u_r = 0.0, \qquad\;\; p_r = 1.0 \qquad \text{for } x > 0. \end{aligned} \tag{15.104}$$

This data has been chosen to satisfy the Rankine–Hugoniot jump relations (with $\gamma = 1.4$), so that the solution is a single 3-shock propagating with speed $s = 10.9636$. Figure 15.8 shows the computed density at time $t = 1$ on a grid with 400 cells, using the minmod limiter. We see in Figure 15.8(a) that the solution is resolved quite sharply, with only three points in the shock. The shock is in the correct location and the solution is roughly constant away from it. However, some small oscillations are visible. These are seen much more clearly in Figure 15.8(b), which shows the same solution on a different scale. Most noticeable are two dips in the density. These are waves that arose from the initial discontinuity at $x = 0$. The leftmost dip is an acoustic wave moving at speed $u_l - c_l \approx 4.06$, and the other dip is an entropy wave moving at the fluid velocity $u_l \approx 9.03$. This sort of *start-up error* is frequently seen when an exact discontinuity is used as initial data.



(a) (b)

Fig. 15.8. Start-up error with Godunov's method on a 3-shock in the Euler equations. (a) Density at $t = 1$. (b) Magnification of the same results. `[claw/book/chap15/startup]`

How does this oscillation arise? The initial data is

$$Q_i^0 = \begin{cases} q_l & \text{if } i < I, \\ q_r & \text{if } i \geq I \end{cases}$$

for some I. In the first time step, the Riemann problem at $x_{I-1/2}$ gives rise to a single wave that is computed exactly (even if the Roe solver is used instead of the exact solver computationally). However, this wave travels a distance less than Δx in this time step, and the averaging process of Godunov's method produces a new state Q_I^1 in one grid cell that is a convex combination of q_l and q_r,

$$Q_I^1 = \frac{s\,\Delta t}{\Delta x} q_l + \left(1 - \frac{s\,\Delta t}{\Delta x}\right) q_r.$$

This state lies on the straight line connecting q_l and q_r in phase space. In the next time step, there will be two Riemann problems at $x_{I-1/2}$ and at $x_{I+1/2}$ that have nontrivial solutions. If the Hugoniot locus joining q_l to q_r happens to be a straight line (as in Temple-class systems, for example; see Section 15.8.2), then each these two Riemann problems will result in a single shock wave, since Q_I^1 will lie on the same Hugoniot locus, and the two shocks together make up the original shock. As in a scalar problem, the numerical solution will be smeared over more grid cells as time evolves, but we will still be approximating the single shock well, and no oscillations will appear.

However, for most nonlinear systems the Hugoniot curve is not a straight line, and so the state Q_I^1 does not lie on the same Hugoniot curve as $Q_{I-1}^1 = q_l$ or $Q_{I+1}^1 = q_r$. Solving these Riemann problems then results in the generation of waves in *all* families, and not just a 3-wave as expected from the initial data. It is these other waves that lead to the oscillations observed in Figure 15.8. See [14] for more analysis and some plots showing how these oscillations evolve in state space.

Slow-Moving Shocks

The start-up error discussed above tends to be damped by the numerical viscosity in Godunov's method, and so is often not very visible. In situations where the numerical viscosity is small, however, the oscillations can become significant.

In particular, oscillations are frequently seen if the shock is moving very slowly, in the sense that the shock speed is very small relative to the fastest characteristic speeds in the problem. Then it takes several time steps to cross a single grid cell even when the Courant number is close to one. Recall from (15.51) that the numerical viscosity of Roe's method vanishes as the wave speed goes to zero, and the same is true for Godunov's method based on the exact Riemann solver in the case where the solution is a single shock with speed close to zero. This can be an advantage for scalar problems: nearly stationary shocks are captured very sharply. But for nonlinear systems this lack of viscosity can lead to increased oscillations.

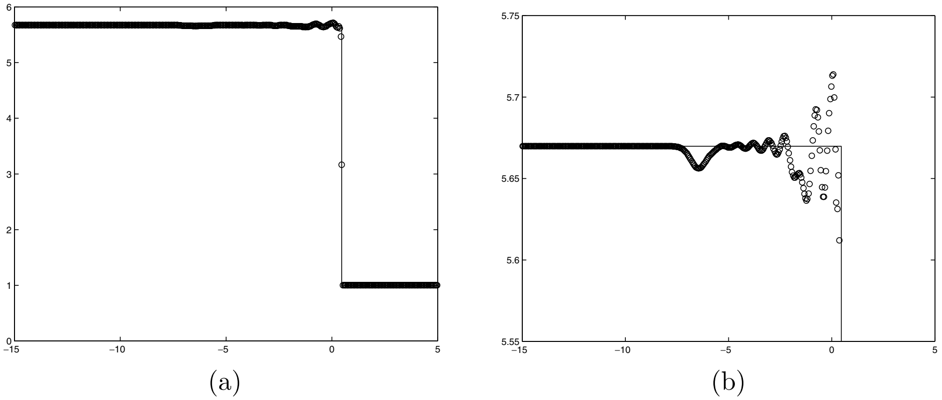


Fig. 15.9. Oscillations arising in a slow shock computed with Godunov's method. (a) Density at $t = 1$. (b) Magnification of the same results. `[claw/book/chap15/slowshock]`

As an example, suppose we take the same data (15.104) as in the previous example, but shift the velocity of u_l and u_r by an amount close to the previous shock speed s:

$$\begin{aligned} &\rho_l = 5.6698, \quad u_l = -1.4701, \quad p_l = 100 \quad &&\text{for } x < 0,\\ &\rho_r = 1.0, \quad u_r = -10.5, \quad p_r = 1.0 \quad &&\text{for } x > 0. \end{aligned} \tag{15.105}$$

This is simply a shift in the reference frame, and so the Riemann solution is exactly the same as before, but with all velocities shifted by the same amount -10.5. So this again gives a single shock wave, now propagating with velocity $s = 0.4636$. Computational results are shown in Figure 15.9, illustrating the oscillations that appear in this case, again with the first-order Godunov method. Similar oscillations arise in high-resolution methods, even when limiters are used. In this case the shock continues to shed oscillations as it moves along.

For some discussions of oscillations due to slow-moving shocks, and ways to improve the situation by introducing additional dissipation, see [14], [112], [224], [232], [343], [373]. The lack of numerical dissipation in Godunov-type methods can also lead to some other numerical problems, particularly for multidimensional computations in regions where strong shocks are nearly aligned with the grid. This can lead to a *cross-flow instability* or *odd–even decoupling*, and to the appearance of unphysical extrusions from the shock that are often called *carbuncles*. Again the addition of more numerical dissipation may be necessary to improve the results. For some discussions of such numerical problems and their relation to physical instabilities see, for example, [195], [287], [353], [364], [374], [487].

Exercises

15.1. Consider the p-system

$$\begin{aligned} v_t - u_x &= 0,\\ u_t + p(v)_x &= 0. \end{aligned} \tag{15.106}$$

(a) Show that for this system the integral in (15.21) can be evaluated in order to obtain the following Roe linearization:

$$\hat{A}_{i-1/2} = \begin{bmatrix} 0 & -1 \\ \dfrac{p_i - p_{i-1}}{V_i - V_{i-1}} & 0 \end{bmatrix}.$$

(b) In particular, determine the Roe solver for $p(v) = a^2/v$, modeling isothermal flow in Lagrangian coordinates, where a is the constant sound speed.

(c) Implement and test this isothermal Riemann solver in CLAWPACK.

(d) Does this solver require an entropy fix?

15.2. Suppose an HLL approximate Riemann solver of the form discussed in Section 15.3.7 is used, but with $s^1_{i-1/2} = -\Delta x/\Delta t$ and $s^2_{i-1/2} = \Delta x/\Delta t$. These are the largest speeds that can be used with this grid spacing and still respect the CFL condition, so these should be upper bounds on the physical speeds. Show that if this approximate Riemann solver is used in the first-order Godunov method, then the result is the Lax–Friedrichs method (4.20).