11 Computing Discontinuous Solutions

For conservation laws we are naturally interested in the difficulties caused by discontinuities in the solution. In the linear theory presented so far we have assumed smooth solutions, and used this in our discussion of the truncation error and convergence proof. We now consider what happens when we apply a numerical method to a linear problem with discontinuous initial data, e.g., the Riemann problem for the scalar advection equation

\[ u_t + au_x = 0, \quad -\infty < x < \infty, \quad t \geq 0, \]

\[ u_0(x) = \begin{cases} 
1 & x < 0 \\
0 & x > 0.
\end{cases} \quad (11.1) \]

Of course the exact solution is simply \( u_0(x - at) \), but we expect our numerical method to have difficulty near the discontinuity. Note for example that a finite difference approximation to \( u_x \) applied across the discontinuity in the true solution will blow up as \( h \to 0 \), e.g.,

\[ \frac{u(at + h, t) - u(at - h, t)}{2h} \to \frac{0 - 1}{2h} \to -\infty \quad \text{as} \quad h \to 0. \]

The local truncation error does not vanish as \( h \to 0 \) and the proof of convergence presented in the previous chapter breaks down. This proof can be rescued by considering approximations \( u^\varepsilon(x) \) to \( u_0(x) \) that are smooth and approach \( u_0 \) as \( \varepsilon \to 0 \). Although convergence can again be proved for a stable and consistent method, the rate of convergence may be considerably less than what is expected from the "order" of the method as defined by its behavior on smooth solutions. Moreover, the numerical solution may look very unsatisfactory on any particular finite grid.

Figures 11.1 and 11.2 show numerical solutions to the Riemann problem (11.1) computed with some of the standard methods discussed in the previous chapter. In all cases \( a = 1, \ k/h = 0.5 \) and the results are plotted at time \( t = 0.5 \) along with the exact solution. In Figure 11.1, \( h = 0.01 \) while a finer grid with \( h = 0.0025 \) is used in Figure 11.2. Notice
that the first order methods (Lax-Friedrichs and upwind) give very smeared solutions while the second order methods (Lax-Wendroff and Beam-Warming) give oscillations. This behavior is typical.

If we compute the $l$-norm error in these computed solutions we do not see the expected rates of convergence. Instead the "first order" methods converge with an error that is $O(k^{1/2})$ while the "second order" methods have an error that is $O(k^{2/3})$ at best. These convergence rates can be proved for very general initial data by a careful analysis using smooth approximations. An indication of why this should be so will be given later in this chapter.
Figure 11.2. Numerical and exact solution to (11.1) with $h = 0.0025$ and the following methods: (a) Lax-Friedrichs, (b) Upwind, (c) Lax-Wendroff, (d) Beam-Warming.
11.1 Modified equations

A useful technique for studying the behavior of solutions to difference equations is to model the difference equation by a differential equation. Of course the difference equation was originally derived by approximating a PDE, and so we can view the original PDE as a model for the difference equation, but there are other differential equations that are better models. In other words, there are other PDEs that the numerical method solves more accurately than the original PDE.

At first glance it may seem strange to approximate the difference equation by a PDE. The difference equation was introduced in the first place because it is easier to solve than the PDE. This is true if we want to generate numerical approximations, but on the other hand it is often easier to predict the qualitative behavior of a PDE than of a system of difference equations. At the moment it is the qualitative behavior of the numerical methods that we wish to understand. Good descriptions of the theory and use of modified equations can be found in Hedstrom[37] or Warming-Hyett[96].

The derivation of the modified equation is closely related to the calculation of the local truncation error for a given method. Consider Lax-Friedrichs, for example, where we calculated the local truncation error $L_4(x, t)$ from (10.29). Since $u(x, t)$ is taken to be the exact solution to $u_t + Au_x = 0$, we found that $L_4(x, t) = O(k)$. However, if we instead take $u(x, t)$ to be the solution of the PDE

$$u_t + Au_x + \frac{1}{2} \left( k u_{tt} - \frac{h^2}{k} u_{xx} \right) = 0,$$

then the truncation error would be $O(k^2)$. We conclude that the Lax-Friedrichs method produces a second order accurate approximation to the solution of (11.2). This equation is called a modified equation for the Lax-Friedrichs method. (The term “model equation” is also sometimes used.)

If we express the $u_{tt}$ term in (11.2) in terms of $x$-derivatives we obtain an equation that is easier to analyze. Note that we cannot use (10.11) directly since $u$ no longer satisfies the original PDE, but a similar manipulation using (11.2) shows that

$$u_{tt} = -Au_{xx} - \frac{1}{2} \left( k u_{ttt} - \frac{h^2}{k} u_{xtt} \right)$$

$$= -A[-Au_{xx} + O(k)] + O(k)$$

$$= A^2 u_{xx} + O(k).$$

Using $u_{tt} = A^2 u_{xx}$ in (11.2) will give errors that are $O(k^2)$, the same order as other terms that have already been ignored, and consequently the Lax-Friedrichs method is also second order accurate on the modified equation

$$u_t + Au_x = \frac{h^2}{2k} \left( I - \frac{k^2}{h^2} A^2 \right) u_{xx}$$

(11.3)
where \( I \) is the \( m \times m \) identity matrix. For higher order methods this elimination of \( t \)-derivatives in terms of \( x \)-derivatives can also be done, but must be done carefully and is complicated by the need to include higher order terms. Warming and Hyett\cite{1974Warming} present a general procedure.

### 11.1.1 First order methods and diffusion

The modified equation (11.3) is an advection-diffusion equation of the form

\[ u_t + Au_x = Du_{xx} \quad (11.4) \]

with a diffusion matrix \( D \) given by

\[ D = \frac{h^2}{2k} \left( I - \left( \frac{k}{h} A \right)^2 \right). \quad (11.5) \]

We expect solutions of this equation to become smeared out as time evolves, explaining at least the qualitative behavior of the Lax-Friedrichs method seen in the above figures. In fact this equation is even a good quantitative model for how the solution behaves. If we plot the exact solution to (11.3) (with appropriate \( k \) and \( h \)) along with the Lax-Friedrichs numerical solutions shown in Figures 11.1 and 11.2, they are virtually indistinguishable to plotting accuracy.

Note that the modified equation (11.3) varies with \( k \) and \( h \). The diffusive term is \( O(k) \) as \( k \to 0 \) and vanishes in the limit as the grid is refined. The numerical solutions generated by the Lax-Friedrichs method on a sequence of grids are thus good approximations to a sequence of "vanishing viscosity" solutions \( u^e \) that might be used to define the physically relevant weak solution to the conservation law. In the linear case there is only one weak solution, but for nonlinear problems it turns out that the Lax-Friedrichs method satisfies a discrete entropy condition and converges more generally to the vanishing viscosity weak solution as the grid is refined.

The modified equation for the upwind method can be derived similarly, and is found to be

\[ u_t + Au_x = \frac{1}{2} h A \left( I - \frac{k}{h} A \right) u_{xx}. \quad (11.6) \]

This is again an advection-diffusion equation and so we expect similar behavior. Moreover, for the values of \( k/h \) and \( A \) used in the computations presented above, the diffusion coefficient is \( h/4 \) for upwind and \( 3h/4 \) for Lax-Friedrichs, so we expect upwind to be less diffusive than Lax-Friedrichs, as confirmed in the numerical results.

**Relation to stability.** Notice that the equation (11.3) is mathematically well posed only if the diffusion coefficient matrix \( D \) is positive semi-definite. Otherwise it behaves like the backward heat equation which is notoriously ill posed. This requires that the
11.1 Modified equations

eigenvalues of $D$ be nonnegative. The matrix $R$ of eigenvectors of $A$ also diagonalizes $D$
and we see that the eigenvalues of $D$ are

$$\frac{k^2}{2k} \left( 1 - \left( \frac{k\lambda_j}{h} \right)^2 \right).$$

These are all nonnegative if and only if the stability condition (10.48) is satisfied. We
see that the modified equation also gives some indication of the stability properties of the
method.

Similarly, the diffusion matrix for the upwind method has nonnegative eigenvalues if
and only if the stability condition $0 \leq \lambda_k k/h \leq 1$ is satisfied.

**Exercise 11.1.** Compute the modified equation for the method

$$U_j^{n+1} = U_j^n - \frac{k}{2h} A(U_{j+1}^n - U_{j-1}^n).$$

*Explain why this method might be expected to be unstable for all $k/h$ (as in fact it is).*

11.1.2 Second order methods and dispersion

The Lax-Wendroff method, which is second order accurate on $u_t + Au_x = 0$, gives a third
order accurate approximation to the solution of the modified equation

$$u_t + Au_x = \frac{k^2}{6} A \left( \frac{k^2}{h^2} A^2 - I \right) u_{xxx}. \tag{11.7}$$

The Beam-Warming method has a similar modified equation,

$$u_t + Au_x = \frac{h^2}{6} A \left( 2I - \frac{3k}{h} A + \frac{k^2}{h^2} A^2 \right) u_{xxx}. \tag{11.8}$$

In the scalar case, both of these modified equations have the form

$$u_t + 3Au_x = \mu u_{xxx} \tag{11.9}$$

which is a dispersive equation. The theory of dispersive waves is covered in detail in
Whitham[97], for example. The key observation is that if we look at a Fourier series
solution to this equation, taking $u(x,t)$ of the form

$$u(x,t) = \int_{-\infty}^{\infty} \hat{u}(\xi,t) e^{i\xi x} d\xi, \tag{11.10}$$

then the Fourier components with different wave number $\xi$ propagate at different speeds,
*i.e.*, they disperse as time evolves. By linearity it is sufficient to consider each wave number
in isolation, so suppose we look for solutions to (11.9) of the form

$$e^{i(\xi x - \xi t)}. \tag{11.11}$$
Putting this into (11.9) and canceling common terms gives

\[ c(\xi) = a \xi + \mu \xi^3. \]  

(11.12)

This expression is called the dispersion relation for (11.9). The speed at which this oscillating wave propagates is clearly \( c(\xi)/\xi \), which is called the phase velocity \( c_p(\xi) \) for wave number \( \xi \). This is the speed at which wave peaks travel. From (11.12) we find that

\[ c_p(\xi) = c(\xi)/\xi = a + \mu \xi^2. \]  

(11.13)

Note that this varies with \( \xi \) and is close to the propagation speed \( a \) of the original advection equation only for \( \xi \) sufficiently small.

It turns out that for general data composed of many wavenumbers, a more important velocity is the so-called group velocity, defined by

\[ c_g(\xi) = c'(\xi) = a + 3 \mu \xi^2. \]  

(11.14)

This varies even more substantially with \( \xi \) than \( c_p(\xi) \). The importance of the group velocity is discussed in Whitham[97]. See also Brillouin[4] or Lighthill[52]. The utility of this concept in the study of numerical methods has been stressed by Trefethen, in particular in relation to the stability of boundary conditions. A nice summary of some of this theory may be found in Trefethen[86].

A step function, such as the initial data we use in (11.1), has a broad Fourier spectrum: \( \hat{u}(\xi, 0) \) decays only like \( 1/\xi \) as \( |\xi| \to 0 \). (By contrast, a \( C^\infty \) function has an exponentially decaying Fourier transform.) As time evolves these highly oscillatory components disperse, leading to an oscillatory solution as has already been observed in the numerical solution obtained using Lax-Wendroff or Beam-Warming. It can be shown that at time \( t \), wavenumber \( \xi \) will be predominantly visible near \( x = c_g(\xi)t \). So, according to (11.14), the most oscillatory components are found farthest from the correct location \( x = at \). This can also be seen in Figures 11.1 and 11.2.

For the scalar advection equation, the modified equation (11.7) for Lax-Wendroff is of the form (11.9) with

\[ \mu = \frac{1}{6} h^2 a (\nu^2 - 1) \]  

(11.15)

where \( \nu = ak/h \) is the Courant number. Since \( a > 0 \) and \( |\nu| < 1 \) for stability, we have \( \mu < 0 \) and hence \( c_g(\xi) < a \) for all \( \xi \) according to (11.14). All wave numbers travel too slowly, leading to an oscillatory wave train lagging behind the discontinuity in the true solution, as seen in the figures.

For Beam-Warming, on the other hand,

\[ \mu = \frac{1}{6} h^2 a (2 - 3 \nu + \nu^2) \]  

(11.16)

which is easily seen to be positive for \( 0 \leq \nu \leq 1 \). Consequently \( c_g(\xi) > a \) for all \( \xi \) and the oscillations are ahead of the discontinuity.
11.2 Accuracy

We now return to the question of the accuracy of a numerical method when applied to a problem with discontinuities. As an example we will consider the Lax-Friedrichs method applied to the problem (11.1). Since the numerical solution agrees so well with the true solution of the modified equation, we can use the difference between the true solution of the modified equation and the true solution to the advection equation as an estimate of the error in the numerical approximation. This is not a rigorous error estimate, and is only for the particular initial data of (11.1) anyway, but it does give an accurate indication of what can be expected in general. The connection to the modified equation helps explain why we typically lose accuracy near discontinuities.

The true solution to (11.4) with data \( u_0(x) \) from (11.1) is simply

\[
u(x, t) = 1 - \text{erf} \left( \frac{x - at}{\sqrt{4Dt}} \right) \tag{11.17}\]

where the “error function” \( \text{erf} \) is defined by

\[
erf(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-x^2} \, dx \tag{11.18}\]

From this and the fact that the solution to the pure advection problem is simply \( u(x, t) = u_0(x - at) \), we can easily compute that

\[
\|u(\cdot, t) - u^\omega(\cdot, t)\| = 2 \int_{-\infty}^{0} \text{erf} \left( \frac{x}{\sqrt{4Dt}} \right) \, dx
= 2\sqrt{4Dt} \int_{-\infty}^{0} \text{erf}(x) \, dx
= C_1 \sqrt{Dt}
\]

for some constant \( C_1 \) independent of \( D \) and \( t \). For Lax-Friedrichs, \( D \) is given by (11.5) and so we find that

\[
\|u(\cdot, t) - U_h(\cdot, t)\| \approx C_2 \sqrt{ht} \tag{11.19}\]

as \( h \to 0 \), for \( k/h \) fixed. This indicates that the 1-norm of the error decays only like \( h^{1/2} \) even though the method is formally “first order accurate”.

Similar results for nonlinear scalar conservation laws can be found in Kuznetsov[43], Lucier[54], or Sanders[70], for example.
12 Conservative Methods for Nonlinear Problems

When we attempt to solve nonlinear conservation laws numerically we run into additional difficulties not seen in the linear equation. Moreover, the nonlinearity makes everything harder to analyze. In spite of this, a great deal of progress has been made in recent years.

For smooth solutions to nonlinear problems, the numerical method can often be linearized and results from the theory of linear finite difference methods applied to obtain convergence results for nonlinear problems. A very general theorem of this form is due to Strang[80] (see also §5.6 of [63]). We will not pursue this topic here since we are primarily interested in discontinuous solutions, for which very reasonable looking finite difference methods can easily give disastrous results that are obviously (or sometimes, not so obviously) incorrect.

We have already seen some of the difficulties caused by discontinuous solutions even in the linear case. For nonlinear problems there are additional difficulties that can arise:

- The method might be "nonlinearly unstable", i.e., unstable on the nonlinear problem even though linearized versions appear to be stable. Often oscillations will trigger nonlinear instabilities.

- The method might converge to a function that is not a weak solution of our original equation (or that is the wrong weak solution, i.e., does not satisfy the entropy condition).

The fact that we might converge to the wrong weak solution is not so surprising — if there is more than one weak solution why should we necessarily converge to the right one? (We must make sure that the finite difference approximations satisfy some discrete form of the entropy condition, as we will do later.)

The fact that we might converge to a function that is not a weak solution at all is more puzzling, but goes back to the fact that it is possible to derive a variety of conservation laws that are equivalent for smooth solutions but have different weak solutions. For
example, the PDEs
\[ u_t + \left( \frac{1}{2} u^2 \right)_x = 0 \quad (12.1) \]
and
\[ (u^2)_t + \left( \frac{2}{3} u^3 \right)_x = 0 \quad (12.2) \]
have exactly the same smooth solutions, but the Rankine-Hugoniot condition gives different shock speeds, and hence different weak solutions. (Recall Example 3.3.)

Consider a finite difference method that is consistent with one of these equations, say (12.1), using the same definition of consistency as for linear problems (expand in Taylor series, etc.). Then the method is also consistent with (12.2) since the Taylor series expansion (which assumes smoothness) gives the same result in either case. So the method is consistent with both (12.1) and (12.2) and while we might then expect the method to converge to a function that is a weak solution of both, that is impossible when the two weak solutions differ.

**Example 12.1.** If we write Burgers' equation (12.1) in the quasilinear form
\[ u_t + uu_x = 0 \quad (12.3) \]
then a natural finite difference method, obtained by a minor modification of the upwind method for \( u_t + au_x = 0 \) (and assuming \( U^n_j = 0 \) for all \( j, n \)) is
\[ U^{n+1}_j = U^n_j - \frac{k}{h} U^n_j (U^n_j - U^n_{j-1}). \quad (12.4) \]

The method (12.4) is adequate for smooth solutions but will not, in general, converge to a discontinuous weak solution of Burgers' equation (12.1) as the grid is refined. Consider, for example, the data from (11.1), which in discrete form gives
\[ U^n_j = \begin{cases} 1 & j < 0 \\ 0 & j \geq 0. \end{cases} \quad (12.5) \]
Then it is easy to verify from (12.4) that \( U^n_j = U^n_j \) for all \( j \). This happens in every successive step as well and so \( U^n_j = U^n_j \) for all \( j \) and \( n \), regardless of the step sizes \( k \) and \( h \). As the grid is refined, the numerical solution thus converges very nicely to the function \( u(x, t) = u_0(x) \). This is not a weak solution of (12.1) (or of (12.2) either).

In this example the solution is obviously wrong, but similar behavior is seen with other initial data that may give reasonable looking results that are incorrect. Figure 12.1 shows the true and computed solutions at time \( t = 1 \) with Riemann data \( u_0 = 1.2 \) and \( u_r = 0.4 \).
We get a nice looking solution propagating at entirely the wrong speed.

**Exercise 12.1.** Show that (12.4) is consistent with both (12.1) and (12.2).
12.1 Conservative methods

Luckily, there turns out to be a very simple and natural requirement we can impose on our numerical methods which will guarantee that we do not converge to non-solutions. This is the requirement that the method be in conservation form, which means it has the form

\[ U_{j}^{n+1} = U_{j}^{n} + \frac{k}{h} [F(U_{j-\frac{1}{2}}^{n}, U_{j+\frac{1}{2}}^{n}, \ldots, U_{j+q}^{n}) - F(U_{j-p-\frac{1}{2}}^{n}, U_{j-p}^{n}, \ldots, U_{j+q-p-\frac{1}{2}}^{n})] \]  \hspace{1cm} (12.6)

for some function \( F \) of \( p + q + 1 \) arguments. \( F \) is called the numerical flux function. In the simplest case, \( p = 0 \) and \( q = 1 \) so that \( F \) is a function of only two variables and (12.6) becomes

\[ U_{j}^{n+1} = U_{j}^{n} - \frac{k}{h} [F(U_{j}^{n}, U_{j+1}^{n}) - F(U_{j-1}^{n}, U_{j}^{n})]. \]  \hspace{1cm} (12.7)

This form is very natural if we view \( U_{j}^{n} \) as an approximation to the cell average \( \bar{u}_{j}^{n} \) defined by (10.3). We know that the weak solution \( u(x, t) \) satisfies the integral form of the conservation law,

\[ \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_{n+1}) \, dx = \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_{n}) \, dx \]

\[ - \left[ \int_{t_{n}}^{t_{n+1}} f(u(x_{j+1/2}, t)) \, dt - \int_{t_{n}}^{t_{n+1}} f(u(x_{j-1/2}, t)) \, dt \right] \]

Dividing by \( h \) and using the cell averages defined in (10.3), this gives

\[ \bar{u}_{j}^{n+1} = \bar{u}_{j}^{n} - \frac{1}{h} \left[ \int_{t_{n}}^{t_{n+1}} f(u(x_{j+1/2}, t)) \, dt - \int_{t_{n}}^{t_{n+1}} f(u(x_{j-1/2}, t)) \, dt \right] \]  \hspace{1cm} (12.9)
12.1 Conservative methods

Comparing this to (12.7), we see that the numerical flux function \( F(U_j, U_{j+1}) \) plays the role of an average flux through \( x_{j+1/2} \) over the time interval \([t_n, t_{n+1}]\):

\[
F(U_j, U_{j+1}) \sim \frac{1}{h} \int_{t_n}^{t_{n+1}} f(u(x_{j+1/2}, t)) \, dt. \tag{12.10}
\]

One way to derive numerical methods in conservation form is to use standard finite difference discretizations but to start with the conservative form of the PDE rather than the quasilinear form. For example, if we generalize the upwind method to Burgers' equation using the form (12.1) rather than (12.3), we obtain

\[
U_j^{n+1} = U_j^n - \frac{k}{h} \left[ \frac{1}{2} (U_j^n)^2 - \frac{1}{2} (U_{j-1}^n)^2 \right]. \tag{12.11}
\]

This is of the form (12.7) with

\[
F(v, w) = \frac{1}{2} v^2. \tag{12.12}
\]

Here we again assume that \( U_j^n \geq 0 \) for all \( j, n \), so that the "upwind" direction is always to the left. More generally, for a nonlinear system \( u_t + f(u)_x = 0 \) for which the Jacobian matrix \( f'(U_j^n) \) has nonnegative eigenvalues for all \( U_j^n \), the upwind method is of the form (12.7) with

\[
F(v, w) = f(v). \tag{12.13}
\]

Of course in general the Jacobian will have eigenvalues of mixed sign and it will not be possible to use a completely one-sided method. Also note that for the nonlinear problem, unlike the linear advection equation, the "upwind" direction depends on the data \( U_j^n \) and may vary from point to point. Even in the scalar case we need to introduce some way of switching the directional bias based on the data. We will consider various generalizations of (12.13) to handle this situation later. For now we simply note that if \( f'(U_j^n) \) has only nonpositive eigenvalues for all \( U_j^n \), then the upwind method always uses the point to the right, and the flux becomes

\[
F(v, w) = f(w). \tag{12.14}
\]

Lax-Friedrichs. The generalization of the Lax-Friedrichs method to nonlinear systems takes the form

\[
U_j^{n+1} = \frac{1}{2} (U_{j-1}^n + U_{j+1}^n) - \frac{k}{2h} \left( f(U_{j+1}^n) - f(U_{j-1}^n) \right). \tag{12.15}
\]

This method can be written in the conservation form (12.7) by taking

\[
F(U_j, U_{j+1}) = \frac{k}{2h} (U_j - U_{j+1}) + \frac{1}{2} (f(U_j) + f(U_{j+1})). \tag{12.16}
\]
12.2 Consistency

The method (12.7) is consistent with the original conservation law if the numerical flux function \( F \) reduces to the true flux \( f \) for the case of constant flow. If \( u(x, t) \equiv \bar{u} \), say, then by the correspondence (12.16) we expect

\[
F(\bar{u}, \bar{u}) = f(\bar{u}) \quad \forall \bar{u} \in \mathbb{R}.
\]

Some smoothness is also required, so that as the two arguments of \( F \) approach some common value \( \hat{u} \), the value of \( F \) approaches \( f(\hat{u}) \) smoothly. For consistency it suffices to have \( F \) a Lipschitz continuous function of each variable. We say that \( F \) is Lipschitz at \( \bar{u} \) if there is a constant \( K \geq 0 \) (which may depend on \( \bar{u} \)) such that

\[
|F(v, w) - f(\bar{u})| \leq K \max(|v - \bar{u}|, |w - \bar{u}|)
\]

for all \( v, w \) with \( |v - \bar{u}| \) and \( |w - \bar{u}| \) sufficiently small. We say that \( F \) is a Lipschitz continuous function if it is Lipschitz at every point.

**EXAMPLE 12.2.** The upwind flux (12.13) is consistent since it clearly satisfies (12.17) and is Lipschitz continuous provided \( f \) is Lipschitz. Since we are always assuming \( f \) is smooth, this will be the case (Note that any differentiable function is Lipschitz).

**EXERCISE 12.2.** Verify that the Lax-Friedrichs flux (12.15) is consistent (including Lipschitz continuity).

More generally, if the flux \( F \) depends on more than two arguments, as in (12.6), the method is consistent if \( F(\bar{u}, \bar{u}, \ldots, \bar{u}) = f(\bar{u}) \) and the Lipschitz condition requires the existence of a constant \( K \) such that

\[
|F(U_{j-p}, \ldots, U_{j+p}) - f(\bar{u})| \leq K \max_{-p \leq i \leq p} |U_{j+i} - \bar{u}|
\]

for all \( U_{j+i} \) sufficiently close to \( \bar{u} \).

**Notation.** It is often convenient to write (12.6) in the form

\[
U^{n+1}_j = U^n_j - \frac{k}{h} \left[ F(U^n; j) - F(U^n; j - 1) \right]
\]

where \( F(U^n; j) \) is the flux function which is allowed to depend on any (finite) number of elements of the vector \( U^n \), "centered" about the \( j \)th point. For example, the upwind flux (12.13) would simply take the form

\[
F(U^n; j) = f(U^n_j).
\]

In this notation, we can think of \( F(U^n; j) \) as approximating the average flux

\[
F(U^n; j) \approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} f(u(x_{j+1/2}, t)) \, dt.
\]
12.2 Consistency

The notation $F(U^n; j)$ is consistent with the notation $U_{j}^{n+1} = \mathcal{H}_{k}(U^n; j)$ introduced in Chapter 10, and for a method in conservation form we have

$$\mathcal{H}_{k}(U^n; j) = U^n_{j} - \frac{k}{h}[F(U^n; j) - F(U^n; j - 1)].$$  \hspace{1cm} (12.23)

**Lax-Wendroff and MacCormack methods.** Recall that the Lax-Wendroff method for a constant coefficient linear hyperbolic system $u_t + Au_x = 0$ has the form

$$U_{j}^{n+1} = U_{j}^{n} - \frac{k}{2h}A(U_{j+1}^{n} - U_{j-1}^{n}) + \frac{k^2}{2h^2}A^2(U_{j+1}^{n} - 2U_{j}^{n} + U_{j-1}^{n}).$$  \hspace{1cm} (12.24)

There are various ways that this can be extended to give a second order method for nonlinear conservation laws. If we let $A(u) = f'(u)$ be the Jacobian matrix, then a conservative form of Lax-Wendroff is

$$U_{j}^{n+1} = U_{j}^{n} - \frac{k}{2h}(f(U_{j+1}^{n}) - f(U_{j-1}^{n})) + \frac{k^2}{2h^2} \left[ A_{j+1/2}(f(U_{j+1}^{n}) - f(U_{j}^{n})) - A_{j-1/2}(f(U_{j}^{n}) - f(U_{j-1}^{n})) \right],$$  \hspace{1cm} (12.25)

where $A_{j\pm 1/2}$ is the Jacobian matrix evaluated at $\frac{1}{2}(U_{j}^{n} + U_{j\pm 1}^{n})$. The difficulty with this form is that it requires evaluating the Jacobian matrix, and is more expensive to use than other forms that only use the function $f(u)$.

One way to avoid using $A$ is to use a two-step procedure. This was first proposed by Richtmyer, and the Richtmyer two-step Lax-Wendroff method is

$$U_{j+1/2}^{n+1/2} = \frac{1}{2}(U_{j}^{n} + U_{j+1}^{n}) - \frac{k}{2h}[f(U_{j+1}^{n}) - f(U_{j}^{n})]$$  \hspace{1cm} (12.26)

$$U_{j+1}^{n+1} = U_{j}^{n} - \frac{k}{h}[f(U_{j+1}^{n+1/2}) - f(U_{j-1}^{n+1/2})].$$

Another method of this same type was proposed by MacCormack[55]. MacCormack's method uses first forward differencing and then backward differencing to achieve second order accuracy:

$$U_{j}^{n} = U_{j}^{n} - \frac{k}{h}[f(U_{j+1}^{n+1}) - f(U_{j}^{n})]$$  \hspace{1cm} (12.27)

$$U_{j}^{n+1} = \frac{1}{2}(U_{j}^{n} + U_{j}^{n}) - \frac{k}{2h}[f(U_{j}^{n}) - f(U_{j-1}^{n})].$$

Alternatively, we could use backward differencing in the first step and then forward differencing in the second step.

**Exercise 12.3.** Each of the methods (12.25), (12.26) and (12.27) reduces to (12.24) in the constant coefficient linear case and is second order accurate on smooth solutions (to nonlinear problems) and conservative. Verify these statements for at least one of these methods and write it in conservation form, determining the numerical flux function.
12.3 Discrete conservation

The basic principle underlying a conservation law is that the total quantity of a conserved variable in any region changes only due to flux through the boundaries. This gave the integral form of the conservation law,

$$\int_a^b u(x, t_2) \, dx = \int_a^b u(x, t_1) \, dx - \left( \int_{t_1}^{t_2} f(u(b, t)) \, dt - \int_{t_1}^{t_2} f(u(a, t)) \, dt \right),$$  \hspace{1cm} (12.28)

which holds for any \(a, b, t_1\) and \(t_2\). Notice in particular that if \(u\) is identically constant outside some finite interval over the time interval \(t_1 \leq t \leq t_2\), say \(u \equiv u_{-\infty}\) for \(x \leq a\) and \(u \equiv u_{+\infty}\) for \(x \geq b\), then we obtain

$$\int_a^b u(x, t_2) \, dx = \int_a^b u(x, t_1) \, dx - (t_2 - t_1)(f(u_{+\infty}) - f(u_{-\infty})).$$  \hspace{1cm} (12.29)

(Note that by finite propagation speed this will be the case if the initial data is constant outside some finite interval.) If \(u_{+\infty} = u_{-\infty}\) (e.g., if the data has compact support in which case \(u_{\pm \infty} = 0\)), then the flux terms drop out altogether and the integral \(\int_a^b u(x, t) \, dx\) is constant in time over any time interval for which the solution remains constant at \(a\) and \(b\).

We have already used (12.28) to motivate the conservation form (12.6) by applying this with \(a = x_{j-1/2}, \ b = x_{j+1/2}\). But we can easily show that a numerical solution generated by a conservative method will also have a more global form of conservation, analogous to (12.28) for arbitrary \(a, b\). In the discrete case, if we let \(J < K\) be arbitrary cell indices and sum (12.20) over \(j\), we find that

$$h \sum_{j=J}^{K} U_{j}^{n+1} = h \sum_{j=J}^{K} U_{j}^{n} - k \sum_{j=J}^{K} [F(U_{j}; j) - F(U_{j}; j-1)].$$  \hspace{1cm} (12.30)

The last sum here telescopes, and all fluxes drop out except for the fluxes at the extreme cell boundaries \(x_{J-1/2}\) and \(x_{K+1/2}\), so that (12.30) reduces to a discrete form of (12.28) over the time interval \([t_n, t_{n+1}]\):

$$h \sum_{j=J}^{K} U_{j}^{n+1} = h \sum_{j=J}^{K} U_{j}^{n} - k[F(U_{J}; K) - F(U_{J}; J-1)].$$  \hspace{1cm} (12.31)

In particular, if \(u_0\) is constant outside some finite interval then so is \(U^n\), since explicit numerical methods have finite domain of dependence, and so for \(J\) and \(K\) sufficiently far out, we can use consistency of the flux function \(F\) to obtain

$$h \sum_{j=J}^{K} U_{j}^{n+1} = h \sum_{j=J}^{K} U_{j}^{n} - k[f(u_{+\infty}) - f(u_{-\infty})].$$  \hspace{1cm} (12.32)
12.4 The Lax-Wendroff Theorem

Applying this recursively gives, for \( N > n \),

\[
\begin{align*}
    h \sum_{j=0}^{K} U_j^{N} &= h \sum_{j=0}^{K} U_j^{n} - (t_N - t_n) [f(u_{+\infty}) - f(u_{-\infty})],
\end{align*}
\]

(12.33)

a discrete form of (12.29).

It follows that if

\[
    h \sum_{j=0}^{K} U_j^{0} = \int_{x_{j-1/2}}^{x_{j+1/2}} u_0(x) \, dx,
\]

(12.34)

which will hold for example if we take \( U_j^{0} = \bar{u}_j^{0} \), the cell averages, then we also have

\[
    h \sum_{j=0}^{K} U_j^{n} = \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_n) \, dx,
\]

(12.35)

for all \( n \) small enough that the solution remains constant in the neighborhood of \( x_{j-1/2} \) and \( x_{j+1/2} \). Using the notation \( U_k(x, t) \) for the piecewise constant function defined by \( U_j^{n} \), we have

\[
    \int_{x_{j-1/2}}^{x_{j+1/2}} U_k(x, t_n) \, dx = \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_n) \, dx,
\]

(12.36)

and so we say that the discrete method is conservative.

This discrete conservation means that any shocks we compute must, in a sense, be in the "correct" location. Consider, for example, the test problem for Burgers' equation with data (12.5). We have seen in Example 12.1 that a nonconservative method can give a solution with the shock propagating at the wrong speed. This could not happen with a conservative method, since the integral of \( U_k(x, t) \) is obviously increasing at the wrong rate. The solution computed with a conservative method might have the shock smeared out, but since the integral (12.36) is correct, it must at least be smeared about the correct location.

Figure 12.2 shows the same test case as seen in Figure 12.1, but now using the conservative upwind method (12.11). Note that the total integral of \( U_k - u \) appears to be zero, as expected.

12.4 The Lax-Wendroff Theorem

The above discussion suggests that we can hope to correctly approximate discontinuous weak solutions to the conservation law by using a conservative method.

Lax and Wendroff[46] proved that this is true, at least in the sense that if we converge to some function \( u(x, t) \) as the grid is refined, through some sequence \( k_i, h_i \to 0 \), then this function will in fact be a weak solution of the conservation law. The theorem does not guarantee that we do converge. For that we need some form of stability, and even then
Figure 12.2. True and computed solutions to Burgers' equation using the conservative upwind method.

if there is more than one weak solution it might be that one sequence of approximations will converge to one weak solution, while another sequence converges to a different weak solution (and therefore a third sequence, obtained for example by merging together the first two sequences, will not converge at all!). See Exercise 12.4 below for an example of this.

Nonetheless, this is a very powerful and important theorem, for it says that we can have confidence in solutions we compute. In practice we typically do not consider a whole sequence of approximations. Instead we compute a single approximation on one fixed grid. If this solution looks reasonable and has well-resolved discontinuities (an indication that the method is stable and our grid is sufficiently fine), then we can believe that it is in fact a good approximation to some weak solution.

**Theorem 12.1 (Lax and Wendroff[46]).** Consider a sequence of grids indexed by \( l = 1, 2, \ldots \), with mesh parameters \( k_l, h_l \to 0 \) as \( l \to \infty \). Let \( U_l(x,t) \) denote the numerical approximation computed with a consistent and conservative method on the \( l \)th grid. Suppose that \( U_l \) converges to a function \( u \) as \( l \to \infty \), in the sense made precise below. Then \( u(x,t) \) is a weak solution of the conservation law.

We will assume that we have convergence of \( U_l \) to \( u \) in the following sense:

1. Over every bounded set \( \Omega = [a, b] \times [0, T] \) in \( x \)-\( t \) space,
\[
\int_0^T \int_a^b |U_l(x,t) - u(x,t)| \, dx \, dt \to 0 \quad \text{as} \quad l \to \infty. \tag{12.37}
\]
12.4 The Lax-Wendroff Theorem

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

$$
\|U_l - u\|_{1,\Omega} \to 0 \quad \text{as } l \to \infty.
$$

(12.38)

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

$$
TV(U_l(\cdot, t)) < R \quad \text{for all } 0 \leq t \leq T, \ l = 1, 2, \ldots.
$$

(12.39)

Here $TV$ denotes the total variation function,

$$
TV(v) = \sup \sum_{j=1}^{N} |v(\xi_j) - v(\xi_{j-1})|
$$

(12.40)

where the supremum is taken over all subdivisions of the real line $-\infty = \xi_0 < \xi_1 < \cdots < \xi_N = \infty$. Note that for the total variation to be finite $v$ must approach constant values $v_{\pm \infty}$ as $x \to \pm \infty$.

Another possible definition is

$$
TV(v) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{-\infty}^{\infty} |v(x) - v(x - \epsilon)| \, dx.
$$

(12.41)

If $v$ is differentiable then this reduces to

$$
TV(v) = \int_{-\infty}^{\infty} |v'(x)| \, dx.
$$

(12.42)

We can use (12.42) also for nondifferentiable functions (distributions) if we interpret $v'(x)$ as the distribution derivative (which includes delta functions at points where $v$ is discontinuous).

Lax and Wendroff assumed a slightly different form of convergence, namely that $U_l$ converges to $u$ almost everywhere (i.e., except on a set of measure zero) in a uniformly bounded manner. Using the fact that each $U_l$ is a piecewise constant function, it can be shown that this requirement is essentially equivalent to (12.38) and (12.39) above. The advantage of assuming (12.38) and (12.39) is twofold: (a) it is these properties that are really needed in the proof, and (b) for certain important classes of methods (e.g., the "total variation diminishing" methods), it is this form of convergence that we can most directly prove.

**Proof.** We will show that the limit function $u(x,t)$ satisfies the weak form (3.22), i.e., for all $\phi \in C_0^\infty$,

$$
\int_0^\infty \int_{-\infty}^{x(t)} \left[ \phi_t u + \phi_x f(u) \right] \, dx \, dt = -\int_{-\infty}^{x(t)} \phi(x,0) u(x,0) \, dx.
$$

(12.43)
Let $\phi$ be a $C^1_0$-test function and multiply the numerical method (12.20) by $\phi(x_j,t_n)$, yielding

$$\phi(x_j,t_n)U_j^{n+1} = \phi(x_j,t_n)U_j^n - \frac{k}{h} \phi(x_j,t_n)[F(U^n;j) - F(U^n;j-1)].$$  \hspace{1cm} (12.44)

This is true for all values of $j$ and $n$ on each grid $l$. (Dropping the subscript $l$ on $k$ and $h$ and letting $U^n_j$ represent the pointwise values of $U_l$ in order to make the formulas more readable.)

If we now sum (12.44) over all $j$ and $n \geq 0$, we obtain

$$\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \phi(x_j,t_n)(U_j^{n+1} - U_j^n)$$

$$= \frac{k}{h} \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \phi(x_j,t_n)[F(U^n;j) - F(U^n;j-1)].$$  \hspace{1cm} (12.45)

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

$$\sum_{j=1}^{m} a_j(b_j - b_{j-1}) = (a_1b_1 - a_1b_0) + (a_2b_2 - a_2b_1) + \cdots + (a_mb_m - a_mb_{m-1})$$

$$= -a_1b_0 + (a_1b_1 - a_2b_1) + (a_2b_2 - a_3b_2) +$$

$$\cdots + (a_{m-1}b_{m-1} - a_mb_{m-1}) + a_mb_m$$

$$= a_mb_m - a_1b_0 - \sum_{j=1}^{m-1} (a_{j+1} - a_j)b_j.$$  \hspace{1cm} (12.46)

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

We will use this on each term in (12.45) (on the $n$-sum in the first term and on the $j$-sum in the second term). By our assumption that $\phi$ has compact support, $\phi(x_j,t_n) = 0$ for $|j|$ or $n$ sufficiently large, and hence the boundary terms at $j = \pm \infty, n = \infty$ all drop out. The only boundary term that remains is at $n = 0$, $t_0 = 0$. This gives

$$- \sum_{j=-\infty}^{\infty} \phi(x_j,t_0)U_j^0 - \sum_{n=1}^{\infty} (\phi(x_j,t_n) - \phi(x_j,t_{n-1}))U_j^n$$

$$- \frac{k}{h} \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} (\phi(x_{j+1},t_n) - \phi(x_j,t_n))F(U^n;j) = 0.$$  \hspace{1cm} (12.47)

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

$$hk \sum_{n=1}^{\infty} \sum_{j=-\infty}^{\infty} \left( \frac{\phi(x_j,t_n) - \phi(x_j,t_{n-1})}{k} \right) U_j^n.$$  \hspace{1cm} (12.47)
12.5 The entropy condition

\[ + \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \left( \frac{\phi(x_{j+1}, t_n) - \phi(x_j, t_n)}{h} \right) F'(U^n; j) \right] = -h \sum_{j=-\infty}^{\infty} \phi(x_j, 0) U^n_j. \]

This transformation using summation by parts is completely analogous to the derivation of (3.22) from (3.21).

Now let \( l \to \infty \), so that \( k_1, h_l \to 0 \). At this point our simplified notation becomes difficult; the term \( U^l_0 \) in (12.47) should be replaced by \( U_l(x_j, t_n) \), for example, to explicitly show the dependence on \( l \). It is reasonably straightforward, using the 1-norm convergence of \( U_l \) to \( u \) and the smoothness of \( \phi \), to show that the term on the top line of (12.47) converges to \( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x, t) u(x, t) \, dx \, dt \) as \( l \to \infty \). If we take initial data \( U^0_0 = \tilde{u}_0 \), for example, then the right-hand side converges to \( -\int_{-\infty}^{\infty} \phi(x, 0) u(x, 0) \, dx \) as well.

The remaining term in (12.47), involving \( F(U^n; j) \), is more subtle and requires the additional assumptions on \( F \) and \( U \) that we have imposed. The value of the numerical flux function \( F \) appearing here, which depends on some \( p + q + 1 \) values of \( U_l \), can be written more properly as

\[ F(U_l(x_j - ph, t_n), \ldots, U_l(x_j + qh, t_n)). \quad (12.48) \]

Since \( F \) is consistent with \( f \), we have that

\[ |F(U_l(x_j - ph, t_n), \ldots, U_l(x_j + qh, t_n)) - f(U_l(x_j, t_n))| \leq K \max_{-\infty < i < \infty} |U_l(x_j + ih, t_n) - U_l(x_j, t_n)| \]

where \( K \) is the Lipschitz constant. Moreover, since \( U_l(\cdot, t) \) has bounded total variation, uniformly in \( l \), it must be that

\[ \max_{-\infty < i < \infty} |U_l(x + ih, t) - U_l(x, t)| \to 0 \quad \text{as} \quad l \to \infty \]

for almost all values of \( x \). Consequently, we have that the numerical flux function (12.48) can be approximated by \( f(U_l(x_j, t_n)) \) with errors that vanish uniformly almost everywhere. This is the critical step in the proof, and together with some additional standard estimates gives the convergence of (12.47) to the weak form (12.43). Since this is true for any test function \( \phi \in C^1_0 \), we have proved that \( u \) is in fact a weak solution.

12.5 The entropy condition

This theorem does not guarantee that weak solutions obtained in this manner satisfy the entropy condition, and there are many examples of conservative numerical methods that converge to weak solutions violating the entropy condition.

**Example 12.3.** Consider Burgers' equation with data

\[ u_0(x) = \begin{cases} -1 & x < 0 \\ +1 & x > 0 \end{cases} \quad (12.49) \]
which we might naturally discretize by setting
\[ U_j^0 = \begin{cases} 
-1 & j \leq 0 \\
+1 & j > 0
\end{cases} \] (12.50)

The entropy satisfying weak solution consists of a rarefaction wave, but the stationary discontinuity \( u(x,t) = u_0(x) \) (for all \( x \) and \( t \)) is also a weak solution. The Rankine-Hugoniot condition with \( s = 0 \) is satisfied since \( f(-1) = f(1) \) for Burgers' equation. There are very natural conservative methods that converge to this latter solution rather than to the physically correct rarefaction wave.

One example is a natural generalization of the upwind methods (12.13) and (12.14) given by
\[ F(v,w) = \begin{cases} 
f(v) & \text{if} \ (f(v) - f(w))/(v - w) \geq 0 \\
f(w) & \text{if} \ (f(v) - f(w))/(v - w) < 0
\end{cases} \] (12.51)

This attempts to use the appropriate "upwind" direction even for problems where this direction changes from point to point. In many cases this works adequately. However, for the problem we consider here we will obtain \( U_j^{n+1} = U_j^n \) for all \( j \) and \( n \) since \( f(-1) = f(1) \) and so all the flux differences cancel out. Consequently \( U_j(x,t) = U_j(x,0) \) for all \( t \), and we converge to \( u(x,t) = u_0(x) \) as the grid is refined.

Note the sensitivity of this numerical solution to our choice of initial data. If we instead take a different discretization of (12.49), say
\[ U_j^0 = \begin{cases} 
-1 & j < 0 \\
0 & j = 0 \\
+1 & j > 0
\end{cases} \] (12.52)

then it turns out that the upwind method (12.51) gives the proper rarefaction wave solution.

The physically correct solution to this problem is often called a transonic rarefaction because of the fact that the wave speed passes through zero within the rarefaction wave. (In the Euler equations, this only happens when an eigenvalue \( v \pm c \) passes through zero, where \( v \) is the fluid velocity and \( c \) is the sound speed. This means that the flow is subsonic to one side and supersonic to the other.) It is in precisely this situation where entropy violating shocks are most frequently computed.

For some numerical methods, it is possible to show that this can never happen, and that any weak solution obtained by refining the grid must in fact satisfy the entropy condition. Of course this supposes that we have a suitable entropy condition for the system to begin with, and the most convenient form is typically the entropy inequality. Recall that this requires a scalar entropy function \( \eta(u) \) and entropy flux \( \psi(u) \) for which
\[ \frac{\partial}{\partial t} \eta(u(x,t)) + \frac{\partial}{\partial x} \psi(u(x,t)) \leq 0 \] (12.53)
in the weak sense (version IV of the entropy condition). This is equivalent to the statement that

$$\int_0^\infty \int_\infty^\infty \phi_t \eta(u) + \phi_x \psi(u) \, dx \, dt \leq -\int_{-\infty}^\infty \phi(x, 0) \eta(u(x, 0)) \, dx$$  \hspace{1cm} (12.54)

for all $\phi \in C^1_0$ with $\phi(x, t) \geq 0$ for all $x, t$.

In order to show that the weak solution $u(x, t)$ obtained as the limit of $U_l(x, t)$ satisfies this inequality, it suffices to show that a discrete entropy inequality holds, of the form

$$\eta(U_j^{n+1}) \leq \eta(U_j^n) - \frac{k}{h} [\Psi(U^n; j) - \Psi(U^n; j - 1)].$$  \hspace{1cm} (12.55)

Here $\Psi$ is some numerical entropy flux function that must be consistent with $\psi$ in the same manner that we require $F$ to be consistent with $f$. If we can show that (12.55) holds for a suitable $\Psi$, then mimicking the proof of the Lax-Wendroff Theorem (i.e., multiplying (12.55) by $\phi(x, t)$, summing over $j$ and $n$, and using summation by parts), we can show that the limiting weak solution $u(x, t)$ obtained as the grid is refined satisfies the entropy inequality (12.54).

In the next chapter we will study a version of the upwind method for which a discrete entropy inequality of this form can be easily proved.

**Exercise 12.4.** Consider the upwind method with flux (12.51). Take $k/h = 1/2$ and apply it to Burgers' equation with initial data

$$u_0(x) = \begin{cases} 
-1 & x < 1 \\
+1 & x > 1 
\end{cases}$$  \hspace{1cm} (12.56)

discretized using $U_j^0 = u_j^0$ (cell averages). Based on the behavior of this method as described above, justify the following statements:

1. The sequence $U_l(x, t)$ for $k_l = 1/2l$ converges to the correct rarefaction wave solution as $l \to \infty$.

2. The sequence $U_l(x, t)$ for $k_l = 1/(2l + 1)$ converges to an entropy violating shock as $l \to \infty$.

3. The sequence $U_l(x, t)$ for $k_l = 1/l$ does not converge as $l \to \infty$. 