PART I. Compressible flow

1. Finite differences for linear scalar problems

We would like to numerically model the behavior of compressible flow. As a first pass, we will start out with the one-dimensional scalar advection equation, i.e., we would like to solve for $u(x, t)$ which can be described by the following partial differential equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \quad (1)$$

where a is a positive constant. The initial conditions are given as:

$$u(x, 0) = \psi(x) \quad (2)$$

The solution to this problem is known to be:

$$u(x, t) = \psi(x - at) \quad (3)$$

It is useful to start with a problem for which the solution is known in order to test the accuracy of our numerical methods. In the following section we discuss finite-difference methods for solving this PDE.

We will assume that the mesh spacing, Δx, and the time step, Δt, are constant. This means that the continuous value of the solution at some location and time, $u(x, t)$, can be expressed as $u(j\Delta x, n\Delta t)$. We will represent the analogous discretized solution at the mesh cell j and time level n as u_j^n. We represent this graphically in x-t space as:

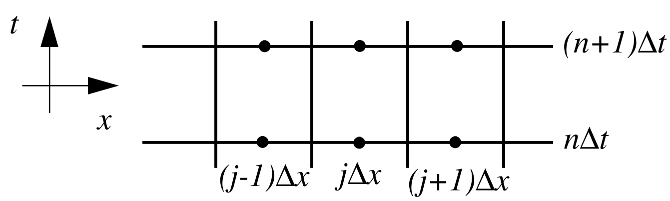


FIGURE 1. Discretization in space and time

The solution u_j^n is defined at the nodes which are placed at the cell centers in physical space.

We discretize the partial differential equation by replacing the derivatives in equation (1) with one-sided finite difference approximations:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + O(\Delta t) + a\frac{u_j^n - u_{j-1}^n}{\Delta x} + O(\Delta x) = 0$$

This numerical scheme uses one-sided forward differencing in time and upwind differencing in space. Dropping the error terms and rewriting this equation, we obtain the discrete evolution equation for u_j^n:

$$u_j^{n+1} = u_j^n + \frac{a\Delta t}{\Delta x}(u_{j-1}^n - u_j^n) \tag{4}$$

The term $\frac{a\Delta t}{\Delta x}$ is the CFL (Courant, Friedrichs, and Lewy) number, sometimes denoted by σ. It plays a pivotal role in determining the stability of a scheme such as this one. More generally, we will consider finite difference schemes of the form

$$u_j^{n+1} = Lu_j^n = \sum_{|s| \leq S} c_s u_{j+s}^n$$

where c_s is independent of u and j and s is an integer set by the specific spatial discretization used. We often write this as a discrete evolution operator acting on sequences, i.e.

$$u^n = \{u^n{}_j\} : u^{n+1} = Lu^n$$

Note that in this case, the operator L is linear, i.e., $L(\alpha v + w) = \alpha L(v) + L(w)$, where α is a real or complex number, and v, w are grid vectors. Also note that the evolution operator is explicit, i.e., u_j^{n+1} is a only function of the values at the current time step, u_{j+s}^n, where $|s| \leq S$.

1.1 Consistency, stability, and convergence

So far, we have adequately expressed the discretized form of the governing equation to define the linear operator L. We would like to evaluate how well this operator stacks up against the exact solution. In addition, we would also like to know whether or not a specific numerical scheme will converge. To address these issues, we will carefully define a benchmark value against which to compare the computed solution. Let us denote $u_e = u_e(x, t)$ as the exact solution to the original differential equation (1):

$$\frac{\partial u_e}{\partial t} + a\frac{\partial u_e}{\partial x} = 0 \tag{5}$$

We can evaluate this exact solution at the discrete points in space and time where the numerical solution is defined. We define $u_{e,j}(t) = u_e(j\Delta x, t)$ and $u_{e,j}^n = u_e(j\Delta x, n\Delta t)$.

We also denote the sequences $u_e(t) = \{u_{e,j}(t)\}_j$ and $u^n{}_e = \{u^n{}_{e,j}\}_j$.

The local truncation error (*LTE*) of a scheme measures how much the discrete evolution differs from the exact solution after one time step and is given by:

$$LTE = u_e^{n+1} - Lu_e^n \tag{6}$$

Graphically, in u_e space:

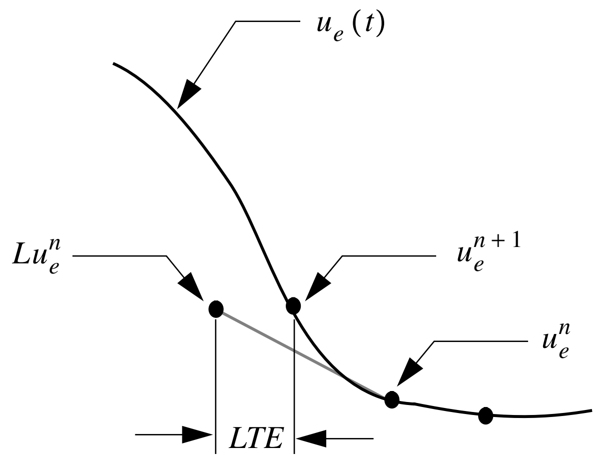


FIGURE 2. Graphical depiction of local truncation error

We say that a numerical scheme is consistent if the $LTE \to 0$ as $\Delta x, \Delta t \to 0$:

$$LTE = u_e^{n+1} - Lu_e^n = O\left(\Delta t \sum_{p,\, q \geq 0,\, p+q=l} \Delta t^p \Delta x^q\right) \tag{7}$$

where $l \geq 1$. The order of the *LTE* is given by *l+1*, while the global error of the numerical scheme is of order *l*.

What is the local truncation error for upwind differencing? We showed earlier in equation (4) that the evolution equation is:

$$(Lu_e^n)_j = u_{e,j}^n + \sigma (u_{e,j-1}^n - u_{e,j}^n) \tag{8}$$

where σ is the CFL number defined by:

$$\sigma = \frac{a\Delta t}{\Delta x} \tag{9}$$

The *LTE* can therefore be written:

$$\begin{aligned} LTE &= u_{e,j}^{n+1} - [u_{e,j}^n + \sigma (u_{e,j-1}^n - u_{e,j}^n)] \\ &= \frac{\partial u_e}{\partial t}\Delta t + O(\Delta t^2) + \sigma \frac{\partial u_e}{\partial x}\Delta x + \sigma O(\Delta x^2) \\ &= \left(\frac{\partial u_e}{\partial t} + a\frac{\partial u_e}{\partial x}\right)\Delta t + O(\Delta x \Delta t + \Delta t^2), \end{aligned}$$

where all derivatives are evaluated at $(j\Delta x, n\Delta t)$. From the PDE, we know that the first term on the right-hand side above is zero. We can thus express the *LTE* as:

$$LTE = \Delta t O(\Delta x + \Delta t) \tag{10}$$

The coefficients multiplying the bounds on the local truncation error are all proportional to second derivatives of the solution (and therefore, of the initial data). They are uniformly bounded as long as those derivatives are also bounded. Looking back at the requirement for consistency above in equation (7), we can say that upwind differencing is consistent to first order globally. Notice that we used in an essential way that u_e^n is the exact solution to the PDE evaluated at discrete points in space and time.

Suppose that we had two discrete solutions at a given time, u_e^n and u^n. Let's apply the operator L to both solutions and look at how far they move apart in one time step. Consider:

$$\begin{aligned} \left\| u_e^{n+1} - u^{n+1} \right\| &= \left\| u_e^{n+1} - Lu_e^n + Lu_e^n - Lu^n \right\| \\ &\leq \left\| u_e^{n+1} - Lu_e^n \right\| + \left\| Lu_e^n - Lu^n \right\| . \end{aligned} \tag{11}$$

The first term on the right-hand side of equation (11) above is by now familiar to us as the local truncation error, a measure of the consistency error. The second term measures how far apart two discrete solutions are after one time step. A method is stable if $\left\| Lu_e^n - Lu^n \right\|$ can be bounded appropriately by $\left\| u_e^n - u^n \right\|$, multiplied by a constant which is independent of both u^n and u_e^n.

This leads to the Lax equivalence theorem, which states that stability and consistency for a given numerical scheme imply its convergence. Analysis of the stability for a given linear difference scheme is simplified by the fact that stability implies boundedness (and vice versa) for any L. So we would like to show that:

$$\begin{aligned} \| Lu - Lv \| &= \| L(u - v) \| \\ &= \| L\varepsilon \| \\ \| Lu - Lv \| &\leq (1 + K\Delta t) \| \varepsilon \| \end{aligned} \tag{12}$$

The following questions arise:

- What is the proper choice for the norm $\| \quad \|$?
- What is the constant K in the stability bound?

For the first question, we can take a hint from the analogous bounds in the exact solution, i.e., that the solution translates with shape unchanged for all time. The L_∞ norm takes as its value the maximum magnitude over the domain, i.e.:

$$\|\delta\|_\infty = max_x|\delta(x)| \tag{13}$$

Looking at the form of the exact solution to the advection equation, we deduce that the L_∞ norm does not change in time. However, the analogous restriction on the discrete formulation is inconvenient for a numerical scheme. Ultimately, the goal is to create a numerical scheme which is at least second-order accurate. We shall show later (Godunov's Theorem) that a linear scheme which is consistent and max-norm stable is only first order accurate. Therefore, we will instead use the L_2 norm, defined by:

$$\|\delta\|_2 = \left[\int|\delta|^2 dx\right]^{\frac{1}{2}} \tag{14}$$

Again, it is clear that this norm does not change in time for the exact solution. We take as the discrete norm

$$\|\delta\|_2 = \left[\sum_j \delta_j^2 \Delta x\right]^{\frac{1}{2}} \tag{15}$$

This also answers the question of what K ought to be in equation (12). Since the exact solution doesn't grow with time, we would like to ensure that no growth occurs in the numerical solution, which can be assured if we choose $K = 0$.

Now let's prove the Lax equivalence theorem, which is so important that we will give its statement a special paragraph all of its own:

Lax Equivalence Theorem: $Stability + Consistency \rightarrow Convergence$

Proof:

$$\begin{aligned}\left\|u^n - u_e^n\right\| &\le \left\|Lu^{n-1} - Lu_e^{n-1}\right\| + \left\|Lu_e^{n-1} - u_e^n\right\| \\ &\le \left\|u^{n-1} - u_e^{n-1}\right\| + \Delta t O\Big(\sum_{p+q=l} (\Delta x)^p (\Delta t)^q\Big) \\ &\le \left\|u^0 - u_e^0\right\| + n\Delta t O\Big(\sum_{p+q=l} (\Delta x)^p (\Delta t)^q\Big)\end{aligned}$$

As $\Delta t \rightarrow 0$, $n\Delta t \rightarrow T$, and the desired result is proven. Recall that l is the global order of the method, while $l+1$ is the order of the LTE.

1.2 Von Neumann stability analysis (discrete Fourier transform)

We will now restrict the problem further: instead of considering solutions on the entire real line, we will consider them on the interval *[0, D]* with periodic boundary conditions and given initial conditions:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \qquad (16)$$

$$u(0, t) = u(D, t) \qquad (17)$$

$$u(x, 0) = \psi(x) \qquad (18)$$

Equivalently, we can consider periodic initial data on the real line; from that point of view, we still have an exact solution $u_e(x, t) = u(x - at, 0) = \psi(x - at)$.

The discrete solution is given by $u^n = \{u_j^n\}$ where the range of *j* is $j = 0, \ldots, M-1$ and the mesh spacing $\Delta x = 1/M$. For the purposes of computing boundary conditions, u^n is extended periodically, so that $u_{j+Mp}^n = u_j^n$ for all integers *p*. We can define the linear finite-difference schemes as the sum of shifts:

$$Lu^n = \sum c_s u_{j+s} \qquad (19)$$

or

$$L = \sum c_s S^s \qquad (20)$$

This shift notation represents location relative to cell *j* and is defined by

$$(Su)_j = u_{j+1}, j = 0, \ldots, M-2 \qquad (21)$$

$$(S^{-1}u)_j = u_{j-1}, j = 1, \ldots, M-1 \qquad (22)$$

At the endpoints of the periodic domain:

$$(Su)_{M-1} = u_0 = u_M \qquad (23)$$

$$(S^{-1}u)_0 = u_{M-1} = u_{-1} \qquad (24)$$

The following complex exponentials form the orthonormal basis for the set of all square-integrable functions on the interval *[0, D]*:

$$W^{(k)}(x) = e^{\frac{2\pi ikx}{D}}, k = 0, \pm 1, \pm 2, \ldots \qquad (25)$$

In addition, these exponentials formally diagonalize the $\frac{\partial}{\partial x}$ operator:

$$\frac{\partial}{\partial x}\left(e^{\frac{2\pi ikx}{D}}\right) = \left(\frac{2\pi ik}{D}\right) e^{\frac{2\pi ikx}{D}} \qquad (26)$$

Can we find an analogous basis for the discrete solutions $\{u_j^n\}_{j=0}^{M-1}$ that diagonalizes *L*?

We define the following inner product on complex *M*-dimensional vectors:

$$\langle u, v \rangle = \frac{1}{M}\sum_{j=0}^{M-1} u_j \bar{v}_j \tag{27}$$

where the overbar denotes the complex conjugate. Assuming that M is even, we define the vectors $w^{(k)}$, $k = -\frac{M}{2}+1, \ldots, \frac{M}{2}$ as the values of the continuous exponentials evaluated at discrete points:

$$w_j^{(k)} = W^{(k)}(j\Delta x) = e^{\frac{2\pi ikj\Delta x}{D}} \tag{28}$$

There are two useful results which follow from this definition:

(1) The vectors $w^{(k)}$ form a complete orthonormal basis with respect to the inner product:

$$\langle w^{(k)}, w^{(k')} \rangle = \begin{cases} 1 \ldots if \ldots k = k' \\ 0 \ldots otherwise \end{cases} \tag{29}$$

Proof:

Compute:

$$\langle w^{(k)}, w^{(k')} \rangle = \frac{1}{M}\sum_{j=0}^{M-1} e^{\frac{2\pi i(k-k')j\Delta x}{D}}$$

$$= \frac{1}{M}\sum_{j=0}^{M-1} z^j$$

where $z = e^{\frac{2\pi i(k-k')\Delta x}{D}}$. Continuing:

$$\langle w^{(k)}, w^{(k')} \rangle = \frac{1}{M}\frac{z^M-1}{z-1}$$

$$= \begin{cases} 1 \ldots if \ldots z = 1 \\ 0 \ldots if \ldots z \neq 1 \end{cases}$$

Saying that $z = 1$ is equivalent to saying that $\frac{(k-k')\Delta x}{D} = p$, where p is an integer. For k and k' in the range of $-\frac{M}{2}+1, \ldots, \frac{M}{2}$, this is possible only if $k = k'$.

(2) The second feature is that the application of the shift operator becomes ridiculously easy:

$$Sw^{(k)} = e^{i\beta}w^{(k)} \tag{30}$$

and

$$S^{-1}w^{(k)} = e^{-i\beta}w^{(k)} \tag{31}$$

where

$$\beta = \frac{2\pi k \Delta x}{D} \tag{32}$$

Proof:

Compute:

$$\begin{aligned}
(Sw^{(k)})_j &= w^{(k)}_{j+1} \\
&= e^{\frac{2\pi i k (j+1)\Delta x}{D}} \\
&= e^{\frac{2\pi i k \Delta x}{D}} e^{\frac{2\pi i k j \Delta x}{D}} \\
&= e^{i\beta} w^{(k)}_j, \; j = 0, \dots, M-2
\end{aligned}$$

This still holds at the edge of the domain

$$\begin{aligned}
(Sw^{(k)})_{M-1} &= w^{(k)}_0 = 1 \\
&= e^{\frac{2\pi i k \Delta x}{D}} e^{\frac{2\pi i k (M-1)\Delta x}{D}} \\
&= e^{i\beta} w^{(k)}_{M-1},
\end{aligned}$$

The procedure for S^{-1} is similar.

Given the basis vectors $w^{(k)}$, we can arrive at a particularly simple criterion for stability. Define the norm to be that given by the inner product: $\|u\|^2 = \langle u, u\rangle$. We also note that the basis $w^{(k)}$ diagonalizes L. In addition, L acts on the $w^{(k)}$'s by simple complex multiplication:

$$S^2 w^{(k)} = S(Sw^{(k)}) = S(e^{i\beta} w^{(k)}) = e^{2i\beta} w^{(k)} \tag{33}$$

This procedure can be repeated so that we can write:

$$Lw^{(k)} = \left(\sum c_s S^s\right) w^{(k)} = \sum c_s e^{i\beta s} w^{(k)} \tag{34}$$

We define the coefficient $\lambda(\beta) = \sum c_s e^{i\beta s}$ to be the *symbol* of L.

Let u be some vector; u can then be expanded in terms of the $w^{(k)}$'s:

$$u = \sum_{k=-\frac{M}{2}+1}^{\frac{M}{2}} b_k w^{(k)} \tag{35}$$

where $b_k = \langle u, w^{(k)}\rangle$ and $\|u\|^2 = \sum |b_k|^2$. Then

$$\begin{aligned}
Lu &= L \sum b_k w^{(k)} \\
&= \sum b_k \lambda(\beta) w^{(k)}
\end{aligned}$$

and

$$\|Lu\|^2 = \sum |b_k|^2 |\lambda(\beta)|^2$$

If $|\lambda(\beta)| \le 1$, then $\|Lu\| \le \|u\|$. Note that the range of β is contained in $-\pi < \beta \le \pi$. To check stability for all $M \to \infty$, we need only look at a single function $\lambda(\beta)$.

Various options for determining stability of a particular numerical discretization are:

(1) Analytically verify a bound of the form $|\lambda(\beta)| \le 1$ for $-\pi < \beta \le \pi$. This is difficult to do in all but the simplest cases.

(2) Verify the bound computationally, i.e., sample $\lambda(\beta)$ and see whether it is less than or equal to 1; or possibly graph $|\lambda(\beta)| - 1$.

(3) Compute a Taylor series around $\beta = 0$ of:

$$|\lambda(\beta)|^2 = 1 + C_q \beta^q + O(\beta^{q+1}) \tag{36}$$

If

- $C_q > 0 \Rightarrow$ the scheme is *unstable* and the solution will grow without limit;
- $C_q < 0 \Rightarrow$ the scheme *may be stable* (*is* stable for long wavelengths)

In practice, a combination of (2) and (3) above give the most insight on "complicated" schemes. When we see instability for $\beta \sim 0$, this means that $k\Delta x \ll 1$ and long-wavelength signals are unstable. This means big trouble. When $\beta \sim \pi$ and we see wavelengths on the order of $2\Delta x$, we may be able to damp these scales.

1.3 Simple examples of finite difference schemes

Let's get into the specifics of stability calculations for a number of common finite-difference schemes.

Upwind Differencing

For this case, the discretized PDE is given as follows:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + O(\Delta t) + a\frac{u_j^n - u_{j-1}^n}{\Delta x} + O(\Delta x) = 0 \tag{37}$$

or rewriting as an explicit evolution equation:

$$u_j^{n+1} = u_j^n + \sigma(u_{j-1}^n - u_j^n) \tag{38}$$

where σ is the CFL number, defined as:

$$\sigma = \frac{a\Delta t}{\Delta x} \tag{39}$$

We can write the discretized evolution equation in shift notation:

$$Lu^n = u^n + \sigma (S^{-1} - I) u^n \tag{40}$$

where I is the identity. The symbol for this scheme is deduced from the form above in equation (40):

$$\lambda(\beta) = 1 + \sigma (e^{-i\beta} - 1)$$

or

$$\lambda(\beta) = (1 - \sigma) + \sigma e^{-i\beta} \tag{41}$$

The constraint on the CFL number arises from the requirement that $|\lambda(\beta)| \le 1$ for all β. For upwinding:

$$\begin{aligned}
\lambda(\beta)\bar{\lambda}(\beta) &= ((1-\sigma) + \sigma e^{-i\beta})((1-\sigma) + \sigma e^{i\beta}) \\
&= (1-\sigma)^2 + 2\sigma(1-\sigma)(e^{i\beta} + e^{-i\beta}) + \sigma^2 \\
&= 1 - 2\sigma + 2\sigma^2 + 2\sigma(1-\sigma)\cos\beta \\
&= 1 - 2\sigma + 2\sigma^2 - 2\sigma(1-\sigma)\left(\left(\cos\frac{\beta}{2}\right)^2 + 1\right) \\
&= 1 - 4\sigma(1-\sigma)\left(\sin\frac{\beta}{2}\right)^2
\end{aligned}$$

From this result, we can see that if $0 < \sigma < 1$, then $|\lambda(\beta)| \le 1$ and the scheme is stable. (For this range of σ, the symbol $|\lambda(\beta)|$ can also be seen to be a linear interpolation between 1 and $e^{-i\beta}$ by a geometric argument.)

Downwind Differencing

For this scheme, we will keep the same forward differencing in time, as in upwind differencing above, but replace the spatial derivative with downwind differencing:

$$\left.\frac{\partial u}{\partial x}\right|_{j\Delta x,\, n\Delta t} = \frac{u^n_{j+1} - u^n_j}{\Delta x} + O(\Delta x) \tag{42}$$

The evolution equation is:

$$u^{n+1}_j = u^n_j + \sigma (u^n_j - u^n_{j+1}) \tag{43}$$

We can see from equation (43) above that the linear operator L should be, in terms of shift notation:

$$L = I + \sigma (I - S) \tag{44}$$

The symbol is:

$$\lambda(\beta) = 1 + \sigma(1 - e^{i\beta}) \tag{45}$$

or

$$\lambda(\beta) = (1 + \sigma) - \sigma e^{i\beta} \tag{46}$$

Although the truncation error indicates that this scheme has the same order accuracy as forward differencing, this scheme has the undesirable property that is *unstable* for all $\sigma > 0$!

<u>Centered Differencing</u>

This time, we'll replace the partial derivative with respect to *x* by the following centered difference approximation:

$$\frac{\partial u}{\partial x} = \frac{u^n_{j+1} - u^n_{j-1}}{2\Delta x} + O(\Delta x^2) \tag{47}$$

resulting in the following evolution equation:

$$u^{n+1}_j = u^n_j + \frac{\sigma}{2}(u^n_{j-1} - u^n_{j+1}) \tag{48}$$

so that the symbol for this scheme is:

and

$$\lambda(\beta) = 1 + \frac{\sigma}{2}(e^{-i\beta} + e^{i\beta}) = 1 - i\sigma\sin\beta$$

$$|\lambda(\beta)| = [1 + \sigma^2(\sin\beta)^2]^{1/2} \geq 1 \tag{49}$$

Therefore, this scheme is also unstable.

<u>Two-Step Lax-Wendroff Scheme</u>

This predictor/corrector method adds an extra step in solving for half-step values in time (the "predictor" step) and a subsequent evolution to the next time level (the "corrector" step) based on those half-step values. The scheme has the dual advantages of a global spatial error of $O(\Delta x^2)$ and is stable for reasonable values of σ. We can think of this graphically as solving for the discrete values of u at the points in space/time shown by the circles in figure +1 below:

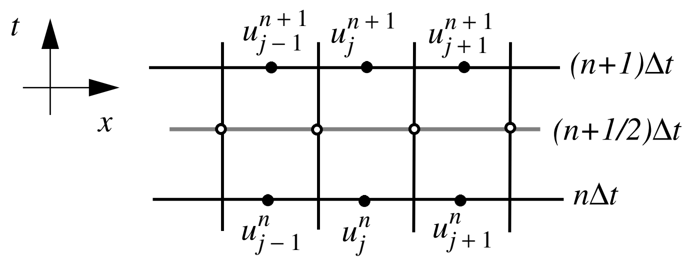


FIGURE 3. Discretization for the Lax-Wendroff predictor/corrector method

The solid circles represent quantities at the full-time step. We know the values at $n\Delta t$ before the start of the time step; we would like to solve for the values at $(n+1)\Delta t$. In order to calculate this, we will have to find values for the points in open circles. How are we going to do this? In the predictor step, we will solve for the half-step values:

$$\tilde{u}^{n+1/2}_{j+1/2} = \frac{1}{2}(u^n_j + u^n_{j+1}) + \frac{\sigma}{2}(u^n_j - u^n_{j+1}) + O(\Delta x^2) \tag{50}$$

In the subsequent corrector step:

$$u^{n+1}_j = u^n_j + \sigma(\tilde{u}^{n+1/2}_{j+1/2} - \tilde{u}^{n+1/2}_{j-1/2}) \tag{51}$$

We would like to show that this scheme is globally second-order accurate, i.e., if u_e is the exact solution, then we must show that the local truncation error is:

$$u^{n+1}_e - L(u_e) = O(\Delta x^3) \tag{52}$$

We will do this in two parts. First we need to show that:

$$u^{n+1/2}_{e,j+1/2} - \tilde{u}^{n+1/2}_{e,j+1/2} = C\left(\left(j+\frac{1}{2}\right)\Delta x\right)\Delta x^2 + O(\Delta x^3) \tag{53}$$

where $u^{n+1/2}_{e,j+1/2} = u_e\left(\left(j+\frac{1}{2}\right)\Delta x,\ \left(n+\frac{1}{2}\right)\Delta t\right)$ and C denotes a smooth function of the quantities in parentheses. Using the expression for the half-step values from equation (50) above:

$$u^{n+1/2}_{e,j+1/2} - \tilde{u}^{n+1/2}_{e,j+1/2} = u^{n+1/2}_{e,j+1/2} - \left(\frac{1}{2}(u^n_{e,j} + u^n_{e,j+1}) + \frac{\sigma}{2}(u^n_{e,j} - u^n_{e,j+1})\right)$$

$$= u^{n+1/2}_{e,j+1/2} - \left[u^n_{e,j+1/2} + \frac{1}{2}\left(\frac{\Delta x}{2}\right)^2 \left.\frac{\partial^2 u_e}{\partial x^2}\right|_{n\Delta t,\ (j+\frac{1}{2})\Delta x} + O(\Delta x^3) \right]$$

$$+\frac{\sigma}{2}\left[\Delta x \left.\frac{\partial u_e}{\partial x}\right|_{n\Delta t,\ (j+\frac{1}{2})\,\Delta x} + O(\Delta x^3)\right]$$

From this point, we will assume implicitly that all derivatives are evaluated at the point $(n\Delta t,\ (j+1/2)\,\Delta x)$. Continuing:

$$u^{n+1/2}_{e,j+1/2} - \tilde{u}^{n+1/2}_{e,j+1/2} = \frac{\Delta t}{2}\frac{\partial u_e}{\partial t} + \frac{\Delta t^2}{8}\frac{\partial^2 u_e}{\partial t^2} + O(\Delta t^3) - \frac{\Delta x^2}{8}\frac{\partial^2 u_e}{\partial x^2} + O(\Delta x^3)$$

$$+\frac{a\Delta t}{2}\frac{\partial u_e}{\partial x} + O(\Delta x^3)$$

$$= \frac{\Delta t}{2}\left[\frac{\partial u_e}{\partial t} + a\frac{\partial u_e}{\partial x}\right] + \frac{\Delta t^2}{8}\frac{\partial^2 u_e}{\partial t^2} + O(\Delta t^3) - \frac{\Delta x^2}{8}\frac{\partial^2 u_e}{\partial x^2} + O(\Delta x^3)$$

$$= \left[\frac{1}{8}\left(\frac{\sigma}{a}\right)^2\frac{\partial^2 u_e}{\partial t^2} - \frac{1}{8}\frac{\partial^2 u_e}{\partial x^2}\right]\Delta x^2 + O(\Delta x^3, \Delta t^3)$$

which means that if u_e is smooth, we have arrived at the desired result:

$$u_{e,j+1/2}{}^{n+1/2} - \tilde{u}^{n+1/2}_{e,j+1/2} = C\left(\left(j+\frac{1}{2}\right)\Delta x^2 + O(\Delta x^3)\right)$$

With this result in hand, we can now show that the local truncation error is of $O(\Delta x^3)$:

$$u^{n+1}_{e,j} - L(u^n_e)_j = u^{n+1}_{e,j} - u^n_{e,j} + \sigma\left(\tilde{u}^{n+1/2}_{e,j+1/2} - \tilde{u}^{n+1/2}_{e,j-1/2}\right)$$

$$= u^{n+1}_{e,j} - u^n_{e,j} + \sigma\left(u^{n+1/2}_{e,j+1/2} - u^{n+1/2}_{e,j-1/2}\right)$$

$$+\sigma\Delta x^2\left(C\left(\left(j+\frac{1}{2}\right)\Delta x\right) - C\left(\left(j-\frac{1}{2}\right)\Delta x\right)\right) + O(\Delta x^3)$$

$$= \Delta t\left.\frac{\partial u_e}{\partial t}\right|_{(n+\frac{1}{2})\,\Delta t,\ j\Delta x} + O(\Delta t^3) + \sigma\Delta x\left.\frac{\partial u_e}{\partial x}\right|_{(n+\frac{1}{2})\,\Delta t,\ j\Delta x} + O(\Delta x^3)$$

$$= \Delta t\left[\frac{\partial u_e}{\partial t} + a\frac{\partial u_e}{\partial x}\right] + O(\Delta x^3, \Delta t^3)$$

Note that the term in brackets is zero. The delicate cancellations of the above procedure depended in a critical way on the following assumptions: (1) that the function u is smooth;

and (2) that the grid spacing is uniform. The first assumption begs the question of what happens at shocks and other discontinuities. The second brings up the question of irregular grids. These topics will be discussed later in sections xxx.xx.x and xxx.xx.x.

Now that we have examined the accuracy of this method, let's take a look at the stability. We can write the equation for the predictor step in shift notation as:

$$\tilde{u}^{n+1/2} = \{\frac{1}{2}[I+S] + \frac{\sigma}{2}[I-S]\}u^n \tag{54}$$

Putting this into the expression for the corrector step:

$$u^{n+1} = \{I + \sigma(S^{-1}-I)\left[\frac{1}{2}(I+S) + \frac{\sigma}{2}(I-S)\right]\}u^n \tag{55}$$

This means that we can write the symbol as:

$$\lambda(\beta) = 1 + \sigma(e^{-i\beta}-1)\left[\frac{1}{2}(1+e^{i\beta}) + \frac{\sigma}{2}(1-e^{i\beta})\right] \tag{56}$$

or in simpler form

$$\lambda(\beta) = \sigma^2\cos\beta + (1-\sigma^2) - i\sigma\sin\beta \tag{57}$$

Recall that $\beta = 2\pi k\Delta x/D$. Is this stable? We can check this by performing a Taylor expansion:

$$|\lambda(\beta)|^2 = 1 + (\sigma^4-\sigma^2)(\cos\beta - 1)^2 \tag{58}$$

$$= 1 + \frac{\sigma^4-\sigma^2}{4}\beta^4 + O(\beta^6) \tag{59}$$

Thus we get lucky, because:

$$1 - |\lambda(\beta)|^2 = (\sigma^4-\sigma^2)(\cos\beta - 1)^2 \tag{60}$$

which will be true for $\sigma \le 1$. Although the Lax-Wendroff scheme is unstable for $\sigma > 1$, it is stable in the long-wavelength limit for $\sigma \le 1$.

1.4 Upwind schemes and geometric interpretation of finite differences.

Suppose we wanted to find the function *u(x,t)* which satisfies the scalar hyperbolic equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{61}$$

and appropriate boundary conditions for some constant $a \ge 0$. In an upwinding scheme, we think of the direction of information transfer as flowing from the upstream to the downstream locations. Thus, to evaluate something at point $j\Delta x$, we only need data from upstream locations. Using single-sided differencing (forward in time, backward in space), the finite-difference representation of equation (61) above is:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j-1}^n - u_j^n}{\Delta x} = 0 \tag{62}$$

Or, rewriting, we might express the discrete evolution of u as:

$$u_j^{n+1} = u_j^n + \sigma\,(u_{j-1}^n - u_j^n) \tag{63}$$

where the CFL number, $\sigma = a\Delta t/\Delta x$, is a positive number. We can interpret u_j^n as the average value of the function $u(x,t)$ over the jth cell at time $n\Delta t$:

$$u_j^n \approx \frac{1}{\Delta x}\int_{(j-\frac{1}{2})\Delta x}^{(j+\frac{1}{2})\Delta x} u\,(x, n\Delta t)\,dx \tag{64}$$

where $u_j^n\Delta x$ is the total amount of the conserved quantity in cell j, e.g., mass, momentum, energy, etc. Graphically:

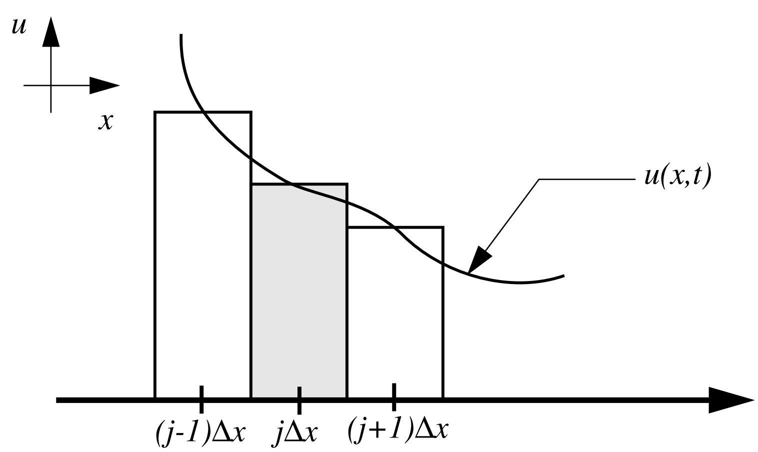


FIGURE 4. Geometric interpretation of $u_j^n\Delta x$.

The area of the shaded rectangle represents an approximation to the area under the function $u(x,t)$, i.e., $u_j^n\Delta x$. We can treat each cell as a control volume and, assuming no generation within the cell, derive a conservation equation in the form:

$$\Delta x u_j^{n+1} = \Delta x u_j^n + F_{in} - F_{out} \tag{65}$$

where F represents a flux through the cell edge. In order to compute $u_j^n \Delta x$, we would like to find a suitable interpolating function $u_I(x)$ at every time step which is easy to calculate but does not violate the conservation constraint. We therefore require that $u_I(x)$ satisfy:

$$u_j^n \Delta x = \int_{(j-\frac{1}{2})\Delta x}^{(j+\frac{1}{2})\Delta x} u_I(x)\, dx \tag{66}$$

The quantity $u_j^n \Delta x$ thus represents a *constraint* on the choice of $u_I(x)$. A trivial example of such an interpolating function might be simply a piecewise constant interpolation, i.e., $u_I(x) = u_j^n$ for $(j-\frac{1}{2})\Delta x < x < (j+\frac{1}{2})\Delta x$.

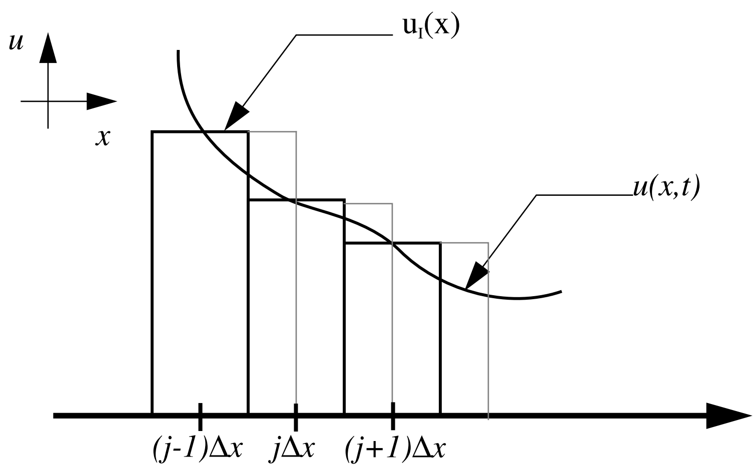


FIGURE 5. Piecewise constant interpolation function $u_I(x)$.

The broad outline for the classical finite-difference scheme can then be broken up into the following steps:

(1) Given u^n, construct $u_I(x)$

(2) Compute the exact solution to the differential equation, using the interpolated function at time $n\Delta t$: $u_I^{n,n+1}(x, (n+1)\Delta t) = u_I(x - a\Delta t)$

(3) Calculate u_j^{n+1} to be the average of $u_I^{n,n+1}(x, (n+1)\Delta t)$ over the finite difference cell:

$$u_j^{n+1} = \frac{1}{\Delta x}\int u_I^{n,n+1}(x, (n+1)\Delta t)\, dx$$

$$= \frac{1}{\Delta x} \int_{(j-\frac{1}{2})\Delta x}^{(j+\frac{1}{2})\Delta x} u_I(x - a\Delta t)\, dx$$

or simply

$$u_j^{n+1} = \frac{1}{\Delta x} \int_{(j-\frac{1}{2})\Delta x - a\Delta t}^{(j+\frac{1}{2})\Delta x - a\Delta t} u_I(x)\, dx \tag{67}$$

A geometric representation of this process would be:

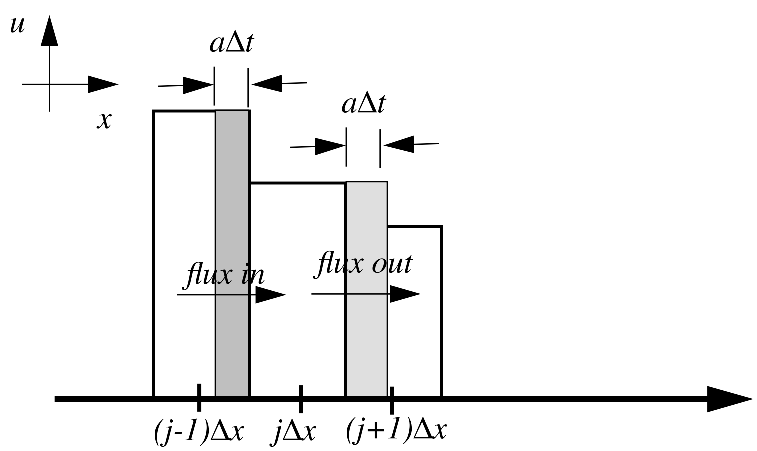


FIGURE 6. Geometric representation of fluxes moving through cell *j*

The advantage of this formulation is that the second and third steps outlined above are *exact*. The error is therefore governed entirely by the interpolation procedure. This observation would suggest that the next logical step would be to improve this step of the numerical method. Fromm's scheme (i.e., second-order upwinding) utilizes a piecewise linear interpolation function for $u_I(x)$, i.e.

$$u_I(x) = u_j^n + \frac{(x - j\Delta x)}{\Delta x} \Delta u_j \tag{68}$$

where Δu_j is the undivided difference, or the finite-difference approximation to:

$$\Delta u_j \cong \left.\frac{\partial u}{\partial x}\right|_{j\Delta x} \Delta x$$

For example, this numerical scheme uses central differencing:

$$\Delta u_j = \frac{u^n_{j+1} - u^n_{j-1}}{2} \tag{69}$$

Note that this method is not necessarily continuous at the cell edges:

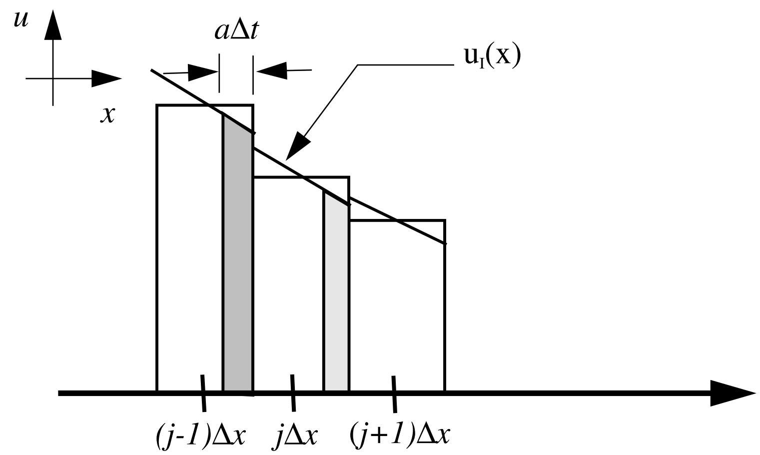


FIGURE 7. Conservation considerations for piecewise linear interpolation

We know that the integral of $u_I(x)$ over each cell is constrained by $u^n_j \Delta x$ so that we can express the conservation form of the finite-difference representation as

$$\Delta x u^{n+1}_j = \Delta x u^n_j + F_{in} - F_{out} \tag{70}$$

where the flux in, F_{in}, is shown in the figure above by the cross-hatched region and is given by:

$$F_{in} = a\Delta t u_I\left[\left(j - \frac{1}{2}\right) \Delta x - \frac{a\Delta t}{2} \right]$$

The interpolation function is evaluated at the location specified in the brackets. An equivalent form is:

$$F_{in} = a\Delta t\left[u^n_{j-1} + \frac{1}{2}\left(1 - \sigma\right) \Delta u_{j-1} \right] \tag{71}$$

Let us denote the term in the brackets above by $u_{j-1/2}$, since it is actually the first-order accurate approximation to the time-centered value of the solution at the cell edge, in the sense of the predictor-corrector formulation of Lax-Wendroff. The flux out is given by an analogous expression (substitute the subscript *j* for *j-1* in the above expression).

Another way to arrive at the evolution equation is by interpolating u_j^n from j to $j+1/2$ at time level $n\Delta t$, and then interpolating from that spatial location to the next time increment $(n+1)\Delta t$:

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} \cong u\left(\left(j+\frac{1}{2}\right)\Delta x,\ \left(n+\frac{1}{2}\right)\Delta t\right)$$

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_j^n + \frac{\Delta x}{2}\frac{\partial u}{\partial x}\bigg|_{j\Delta x} + \frac{\Delta t}{2}\frac{\partial u}{\partial t}\bigg|_{n\Delta t} + \ldots$$

Next, using the PDE to substitute for the temporal derivative

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_j^n + \left[\frac{\Delta x}{2} - \frac{a\Delta t}{2}\right]\frac{\partial u}{\partial x}\bigg|_{j\Delta x} + \ldots$$

or rewriting

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_j^n + \frac{1}{2}\left(1-\sigma\right)\Delta u_j \tag{72}$$

which is the same as before. The local truncation error is $O(\Delta x^3)$, which can be readily proved if enough terms of the Taylor expansion are retained.

Fromm's scheme keeps track of whether the wave speed is positive or negative, and alters the direction of information transfer accordingly.

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_j^n - \frac{1}{2}\left(1+\sigma\right)\Delta u_j \tag{73}$$

This scheme is stable for $\sigma \le 1$, since:

$$\left|\lambda\left(\beta\right)\right|^2 = 1 + \left(\sigma-1\right)\sigma\left(\sigma^2-\sigma+1\right)\frac{\beta^4}{4} + \ldots \tag{74}$$

where $\beta = \left(2\pi k\Delta x\right)/D$. In addition, the local truncation error is $O(\Delta x^3)$.

1.5 Fourier analysis and accuracy

We have derived and analyzed three stable schemes: upwind differencing, the Lax-Wendroff method, and Fromm's method. From our truncation error analysis for these schemes, we know that the global error for the upwind scheme is $O\left(\Delta x\right)$, and that for Lax-Wendroff and Fromm it is $O\left(\Delta x\right)^2$. We can check to see whether these estimates are reliable indicators of the actual performance of these methods. In the following table, we give the values of the error, defined by:

$$\left\| u_e^n - u^n \right\| = \sum_{(-M/2)+1}^{M/2} \left| u_{e,j}^n - u_j^n \right|^2 \Delta x = \varepsilon(\Delta x, n\Delta t) \tag{75}$$

for these three schemes, for different values of M. We also show plots of the solution cor-

Table 1: L2 error in various schemes for smooth Gaussian pulse

Method	M= 32	M=64	M=128
Upwind	4.11e-02	2.42e-02	1.32e-02
Lax-Wendroff	3.66e-02	1.62e-02	3.92e-03
Fromm	1.62e-02	5.95e-03	1.11e-03

responding to the times that those errors were computed. For these schemes, we find that $\frac{\varepsilon(\Delta x, T)}{\varepsilon(2\Delta x, T)} = 2^{-l}$, where l is the global order of accuracy. Thus the errors go to zero at a rate consistent with the assumption that the higher order terms in the truncation error estimate (74) are negligible, relative to the leading order terms. Comparing the actual values of ε for each of the schemes, we also find that for a fixed mesh spacing, the first order upwind method has much larger errors than either of the second order schemes. However, we also notice that the Lax Wendroff and Fromm schemes have considerably different errors from each other for a given mesh spacing; in particular, the Fromm scheme appears to be more accurate. In this section, we will develop more refined analytical tools that will enable us to distinguish between the two second-order schemes and will provide some more general design principles for finite difference methods.

Let $f_j = \sum_{-M/2+1}^{M/2} b_k w_j^k$ be the discrete Fourier transform of the M-vector f. The function

$$f_S(x) = \sum_{-M/2+1}^{M/2} b_k W^k(x) \tag{76}$$

defined on the interval $[0, D]$ is an interpolation function for the values f_j:

$$f_S(j\Delta x) = \sum_{-M/2+1}^{M/2} b_k W^k(j\Delta x) = f_j \tag{77}$$

Furthermore, if $f_j = f(j\Delta x)$ for $j = 0\ldots M-1$, for some q-times continuously differentiable function $f(x)$, then

$$\|f(x) - f_S(x)\| = O\left(\frac{1}{M^{q+1}}\right) \tag{78}$$

uniformly in x.

We can apply this to the discretization of (1). Let $\varphi_j = \varphi(j\Delta x)$, where $\varphi(x)$ is infinitely differentiable data for (1), and let

$$\varphi_j = \sum_{-M/2+1}^{M/2} b_k w_j^k$$

be the discrete Fourier transform of the M-vector φ. Then $\varphi_S(x)$ interpolates the discrete data, and (76) for any fixed q as $M \to \infty$.We can define a spectrally accurate (approximate) solution on M points to be the exact solution to (1), with initial data $\varphi_S(x)$:

$$u_S(x, t) = \varphi_S(x - at). \tag{79}$$

By equation (76) :

$$u_S(x, t) - u(x, t) = \varphi_S(x - at) - \varphi(x - at) = O(\Delta x^{q+1}).$$

Thus, for sufficiently large q, the spectrally accurate solution differs from the exact solution by an amount that is much smaller that the difference between either of them and the solutions obtained from the finite difference methods described above. In particular, $u_S(j\Delta x, n\Delta t) - u_j^n$ is an excellent approximation to $u(j\Delta x, n\Delta t) - u_j^n$.

Second, we can also apply Fourier interpolation to the numerical solution u_j^n. We can define, for each n, the function

$$u_I^n(x) = \sum_{-M/2+1}^{M/2} b_k \lambda(\beta)^n W^k(x) \tag{80}$$

Since the discrete Fourier transform diagonalizes the difference operator L, and by the interpolation property (75), we have $u_I^n(j\Delta x) = u_j^n$. Thus we can view the finite difference method as defining a discrete time dynamics on the set of all sums of finite Fourier series of the form (75).

Finally, we can write $u_s(x, n\Delta t)$ in such a way so as to easily compare it to $u_I^n(x)$:

$$u_S(x, n\Delta t) = \sum_{-M/2+1}^{M/2} b_k e^{i\sigma\beta n} W^k(x) \tag{81}$$

where $\beta = 2\pi k\Delta x/D$ and $\sigma = a\Delta t/\Delta x$. From this expression for u_S, it is clear that the spectrally accurate discrete time dynamics on the Fourier modes is given by multiplying each amplitude by $e^{i\sigma\beta}$, and that the numerical solution gives an error exactly to the extent that the symbol $\lambda(\beta)$ differs from $e^{i\sigma\beta}$.

To illustrate the utility of these ideas, we can use the symbol to compute the local truncation error of a finite difference method. For initial data in the form of a finite Fourier sum (75), the LTE is given by

$$u_e^{n+1} - Lu_e^n = \sum_{-M/2+1}^{M/2} b_k \left(e^{-i\sigma\beta} - \lambda(\beta)\right) e^{-i\sigma\beta n} w^k. \tag{82}$$

In particular, for a single Fourier mode of wavenumber *k*, with *k* fixed, the statement that the LTE is $O((\Delta x)^{l+1})$ is the same as saying that $e^{-i\sigma\beta} - \lambda(\beta) = O(\beta^{l+1})$. In fact, we can infer the form of the truncation error from the symbol. If

$$e^{-i\sigma\beta} - \lambda(\beta) = C(i\beta)^{l+1} + O(\beta^{l+2})$$

then the LTE can be written in the form

$$u_e^{n+1} - Lu_e^n = \Delta t\left(\frac{Ca}{\sigma}(\Delta x)^l \left.\frac{\partial^{l+1} u_e}{\partial x^{l+1}}\right|_{(j\Delta x,\, n\Delta t)} + O((\Delta x)^{l+1})\right) \tag{83}$$

This may look like a somewhat singular expression, due to the division by σ, but it is not since $C = O(\sigma)$ if the scheme is consistent.

<u>*1.5.1 Amplitude error*</u>

To determine the difference between the amplitudes of the numerical and exact solutions, we must determine the degree to which the symbol $\lambda(\beta)$ approximates $e^{-i\sigma\beta}$. The amplitude error, *AE*, is:

$$AE = 1 - |\lambda(\beta)|$$

since $|e^{-i\sigma\beta}| = 1$. Note that for a single Fourier mode $\left|e^{-i\sigma\beta} b_k\right| = |b_k|$ and that $|\lambda(\beta) b_k| \le |b_k|$, so that the amplitudes of Fourier modes tend to decrease as a function of time. In general, the symbol can be expressed as a function of β:

$$|\lambda(\beta)| = 1 + C_q\beta^q + \dots \tag{84}$$

As we have previously discussed, we want C_q to be negative in order to yield a stable numerical scheme. Taylor expansions for several common discretizations yield the results shown in table 2:

Table 2: Amplitude error for various schemes

Scheme	$\lambda(\beta)$
Upwind	$1 + \frac{\sigma^2 - \sigma}{2}\beta^2 + \ldots$
Lax-Wendroff	$1 + \frac{\sigma^4 - \sigma^2}{8}\beta^4 + \ldots$
Fromm	$1 + \frac{(\sigma - 1)\,\sigma\,(\sigma^2 - \sigma + 1)}{8}\beta^4 + \ldots$

Since β is proportional to the mesh spacing Δx for fixed k, the amplitude error for both Fromm and Lax-Wendroff is quite small, of $O(\Delta x^4)$. For reasonable CFL numbers (e.g., for compressible flow, one might set the optimal σ at 0.9), upwind differencing does indeed give a negative C_q, and one can predict that the Fourier modes will be damped by $O(\Delta x^2)$.

1.5.2 Phase errors

The phase error is a measure of how accurately the finite difference scheme is translating each Fourier mode in space. The exact solution at time level $n+1$ looks like a translation of the Fourier mode by $\sigma\beta/k = a\Delta t$:

$$u_s(x, n\Delta t) = \sum_{-M/2+1}^{M/2} b_k W^k(x - an\Delta t) \tag{85}$$

We wish to compute the corresponding quantity for u_I^n. Define η by:

$$\frac{\lambda(\beta)}{|\lambda(\beta)|} = e^{-i\eta(\sigma, \beta)} \tag{86}$$

The range of η is from $-\pi$ to π. Multiplication by $e^{-i\eta(\sigma, \beta)}$ corresponds to shifting the argument of W^k by $\eta(\sigma, \beta)/(2\pi k/D)$, so that

$$u_I^n(x) = \sum_{-M/2+1}^{M/2} b_k |\lambda(\beta)|^n W^k(x - n\eta(\sigma, \beta)/(2\pi k/D)) \tag{87}$$

$$= \sum_{-M/2+1}^{M/2} b_k |\lambda(\beta)|^n e^{-i\eta(\sigma, \beta)n} W^k(x)$$

While the exact solution would yield a multiplier of $e^{-i\sigma\beta}$ in equation (xx) above, the numerical solution gives a multiplier of $e^{-i\eta(\sigma,\beta)}$. However, this definition of phase error does not permit us to treat the time step as a parameter, comparing phase error for different values of σ. A definition that allows us to make such comparisons is to define the error in the phase after a length of time required for the solution to move a fixed distance. A natural choice of such a distance is Δx, the length of a finite difference cell. In that case, we can define the phase error using the following expression:

$$\left(e^{i\sigma\beta}\frac{\lambda(\beta)}{|\lambda(\beta)|}\right)^{1/\sigma} = e^{-i\delta} \tag{88}$$

In this case, $\delta/(2\pi k/D)$ is the error in the displacement of the k-th Fourier component after the solution has evolved for a time $\Delta x/a$.

We can use Taylor expansions of the left-hand side of (87) to evaluate δ in the limit of $\beta \to 0$. To leading order,

$$\left(e^{i\sigma\beta}\frac{\lambda(\beta)}{|\lambda(\beta)|}\right)^{1/\sigma} = 1 - iC_{2q+1}\beta^{2q+1} + O(\beta^{2q+3}) = 1 - i\delta + O(\delta^2), \tag{89}$$

where $2q = l$ if l is even, and $2q = l+1$ if l is odd. Given sufficient effort (or a symbolic manipulation package), one can compute $C_{2q+1}\beta^{2q+1}$, the leading order term in the expansion for each of the schemes, to be the following:

Table 3: Phase error for various schemes

Scheme	Phase error (89)
Upwind	$\left(\frac{-(2\sigma^2 - 3\sigma + 1)}{6}\right)\beta^3 + O(\beta^5)$
Lax-Wendroff	$\frac{(\sigma^2 - 1)}{6}\beta^3 + O(\beta^5)$
Fromm	$(2\sigma^2 - 3\sigma + 1)\frac{\beta^3}{12} + O(\beta^5)$

It is particularly instructive to plot the polynomials in σ that multiply the leading order terms in the phase error (figure xx) for the case of Fromm's scheme and the Lax-Wendroff scheme. Overall, we see that the phase error for Fromm's scheme is smaller than that for the Lax-Wendroff scheme. In addition, we see that, for $0.5 \le \sigma \le 1$, the phase error for

Fromm's scheme is much smaller that for Lax-Wendroff, as well as for that for either scheme for $0 \le \sigma \le 0.5$. In particular, for $\sigma = 0.75$, the value in that range for which the phase error for Fromm has the largest magnitude, the coefficients differ by a factor of 8. Since the leading order term in the phase error is cubic in Δx, this implies that, for this value of the CFL number, it is necessary to reduce Δx by a factor of two in the Lax-Wendroff scheme to obtain the same accuracy as that obtained in Fromm's scheme. This conclusion is supported by the convergence study described at the beginning of the section. As a practical matter, it is usually not possible to run an algorithm at a specified CFL number, since wave speeds vary in space over the grid. However, it is typical of a large class of practical applications that the CFL condition is constrained by the behavior of the solution in the region of greatest interest, and that one runs using a CFL number that is as large as possible. This analysis provides an explanation of what has been observed in much more complicated nonlinear problems in several space dimensions [Woodward and Colella, 1984]: that the improvement in going from Lax-Wendroff to Fromm is worth about a factor of two mesh refinement.

The classical definition of phase error is given by $\lim_{\sigma \to 0} \delta = \delta_c(\beta)$. In terms of the definition (87), it is given by

$$\delta_c = \frac{1}{-i} \frac{d}{d\sigma} \left(e^{i\sigma\beta} \frac{\lambda(\beta)}{|\lambda(\beta)|} \right) \Bigg|_{\sigma = 0} = \frac{d}{d\sigma} (Im(\lambda)) \Bigg|_{\sigma = 0} \tag{90}$$

where $Im(\lambda) = \frac{\bar{\lambda} - \lambda}{2i}$. It is easy to see from the table above that the classical phase error for Lax-Wendroff is, to leading order in β, only twice as large as that for Fromm. Thus it gives only a partial, and sometimes misleading picture of how schemes will behave. The advantage is that it is precisely this part of the phase error that is most easily improved. It is not difficult to show that δ_c depends only on the terms in the finite difference scheme that are first order in σ

$$(Lu^n)_j = u_j^n + \sigma (u_{j-1/2}^n - u_{j+1/2}^n) + O(\sigma^2) .$$

In particular, one can reduce the classical phase error by improving the accuracy with which $\frac{u_{j-1/2}^n - u_{j+1/2}^n}{\Delta x}$ approximates $\left.\frac{\partial u}{\partial x}\right|_{(j\Delta x,\, n\Delta t)}$.

We can use Richardson extrapolation to construct higher order approximations to spatial derivatives. Given a smooth function *f(x)*, say that we want to compute $f'(0)$, given values on a grid with grid spacing Δx. We can use Taylor expansions to obtain the following error estimates for centered difference approximations to $f'(0)$

$$\frac{f_{j+1}-f_{j-1}}{2\Delta x} = f'(0) + \frac{(\Delta x)^2}{6} f'''(0) + O(\Delta x)^4$$

$$\frac{f_{j+2}-f_{j-2}}{4\Delta x} = f'(0) + \frac{(2\Delta x)^2}{6} f'''(0) + O(\Delta x)^4$$

We can take linear combinations of these two expressions to obtain a fourth-order accurate expression for $f'(0)$:

$$f'(0) = \frac{4}{3}\frac{f_{j+1}-f_{j-1}}{2\Delta x} - \frac{1}{3}\frac{f_{j+2}-f_{j-2}}{4\Delta x} + O(\Delta x)^4 \tag{91}$$

We can write this expression in flux form, as follows:

$$f'(0) = \frac{f_{j+1/2}-f_{j-1/2}}{\Delta x}$$

where

$$f_{j+1/2} = \frac{7}{12}(f_j + f_{j+1}) - \frac{1}{12}(f_{j-1} + f_{j+2})$$

This process can be continued to any order of accuracy, and flux forms derived; for details, see, e.g., [Zalesak, 1984].

We can now use this to define a Lax-Wendroff scheme with fourth-order spatial accuracy (LW4) and lower classical phase error.It is given by:

$$u_j^{n+1} = u_j^n + \sigma\left(u_{j-\frac{1}{2}}^{n+\frac{1}{2}} - u_{j+\frac{1}{2}}^{n+\frac{1}{2}}\right) \tag{92}$$

where

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{7}{12}(u_j + u_{j+1}) - \frac{1}{12}(u_{j-2} - u_{j-1}) + \frac{\sigma}{2}(u_j^n - u_{j+1}^n) \tag{93}$$

As compared to second-order Lax-Wendroff (LW2), the first two terms on the right-hand side of equation (93) have replaced $(u_j + u_{j+1})/2$. The global phase error in the long-wavelength limit is given by $\delta = \frac{\sigma^2}{6}\beta^3 + O(\beta^5)$. Thus the use of higher order differencing eliminates the leading order term in σ of the phase error, as claimed at the outset. In general, the use of $2q$-th order centered differences in (xxx) leads to a classical phase error $\delta_c = O(\beta^{2q+1})$.

A Taylor expansion of the amplitude error indicates that the scheme as stated is unstable in the long-wavelength limit: $|\lambda(\beta)| = 1 + C\beta^4$ where $C > 0$. A little artificial viscosity can be added to damp out the linear instability. Even-order derivatives are dissipative, so we might try adding a term like $\Delta t \Delta x \frac{\partial^2 u}{\partial x^2}$; however, this would wreak havoc with our error. Instead, try adding something like $\Delta t \Delta x^3 \frac{\partial^4 u}{\partial x^4}$. Consider the effect of fourth-order derivatives:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + C\frac{\partial^4 u}{\partial x^4} = 0 \tag{94}$$

The last term on the left-hand side of equation (94) above acts to reduce the amplitudes of the Fourier modes. The numerical method is then given as follows:

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{7}{12}(u_j + u_{j+1}) - \frac{1}{12}(u_{j-2} - u_{j-1}) + \frac{\sigma}{2}(u_j^n - u_{j+1}^n) + \\ + \left(\frac{\sigma^3}{4} + \frac{1}{16}\right)(u_{j+2}^n - u_{j-1}^n - 3(u_{j+1}^n - u_j^n)) \tag{95}$$

In the long-wavelength limit, $C < 0$ for $|\sigma| \le 1$. This does not mean that we can neglect the short wavelengths however; if one computes $|\lambda(\beta)|$ for the full range $0 \le \beta \le \pi$, one finds that this scheme is stable only for $|\sigma| \le 0.6$.

The following table summarizes the phase and amplitude errors for a number of numerical schemes, as well as ranges of CFL number for which stability is anticipated.

Table 4: Comparison of accuracy, error and stability of various schemes

Scheme	Accuracy	Phase Error	Amplitude Error	Stability
Upwind	$O(\Delta x)$	$O(\beta^3)$	$O(\beta^2)$	$0 \le \sigma \le 1$
LW2	$O(\Delta x^2)$	$O(\beta^3)$	$O(\beta^4)$	$0 \le \sigma \le 1$
Fromm	$O(\Delta x^2)$	$O(\beta^3)$	$O(\beta^4)$	$0 \le \sigma \le 1$
LW4	$O(\Delta x^2)$	$O(\beta^5)$ $(\sigma \rightarrow 0)$	$O(\beta^4)$	$0 \le \sigma \le 0.6$
		$O(\beta^3)$ (finite σ)		

With the exception of upwinding, the phase error swamps the amplitude error for the above schemes. So, from the standpoint of accuracy, it is the phase error which is of the most concern.

1.6 The modified equation.

Discrete systems are not objects that we have a great deal of intuition about, while we do have a substantial understanding of how solutions to differential equations behave. Thus, in this section, we look to develop an intuition about the behavior of errors in difference schemes by deriving a differential equation - called the modified equation - that models the behavior of the error. We start by expressing the local truncation error in the following form

$$u_{e,j}^{n+1} - (Lu_e^n)_j = \Delta t \left(C (\Delta x)^l \frac{\partial^{l+1} u_e}{\partial x^{l+1}} \bigg|_{(j\Delta x,\, n\Delta t)} + O((\Delta x)^{l+1}) \right) \tag{96}$$

Where the coefficient can be determined from the leading order term in the Taylor expansion of the symbol, for example. From this expression, one observes that locally in time, at least, the effect of replacing the exact solution operator by the result from the difference scheme is to add to the result from the exact solution a quantity of the form Δt times the spatial derivatives in the brackets. For that reason, we define the modified solution $u_M$as the solution to the following differential equation:

$$\frac{\partial u_M}{\partial t} + a\frac{\partial u_M}{\partial x} - C\Delta x^l \frac{\partial^{l+1} u_M}{\partial x^{l+1}} = 0 \tag{97}$$

for which the expression for Lu_e implied by (96) is a consistent discretization in time.- Thus,u_M will better describe what the difference scheme is doing than u_e. If we define $u_{M,j}^n$ to be the solution to this differential equation evaluated at grid points, analogously to $u_{e,j}^n$, then the difference between the modified solution at time level *n+1* and the evolved modified solution is:

$$u_M^{n+1} - Lu_M^n = \Delta t O(\Delta x^{l+1}) \tag{98}$$

For the discrete solution

$$u^{n+1} - Lu^n = 0 \tag{99}$$

For the exact solution

$$u_E^{n+1} - Lu_E^n = \Delta t O(\Delta x^l) \tag{100}$$

thus we find that the discrete dynamics are approximated to one higher order of accuracy by the solution to the modified equation that by the solution to the original differential equation, as illustrated in the figure below:

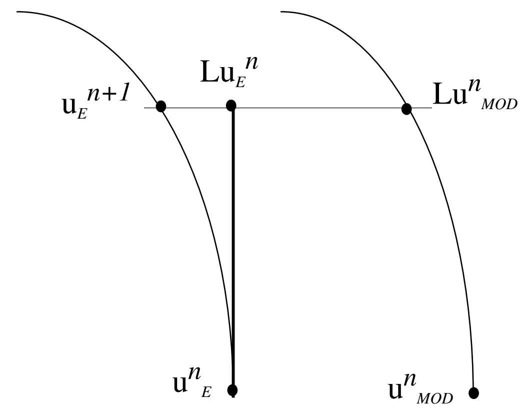


FIGURE 8. Qualitative evolution of the modified and numerical solutions.

Example: Upwind differencing The modified equation for upwind differencing is given as

$$\frac{\partial u_M}{\partial t} + a\frac{\partial u_M}{\partial x} = \frac{a\Delta x}{2}(1-\sigma)\frac{\partial^2 u_M}{\partial x^2}. \tag{101}$$

The effect of the numerical error in upwind differencing is, to leading order, to add a diffusion term to the PDE with a diffusion coefficient of magnitude $\frac{(1+\sigma)a\Delta x}{2}$. Second-order schemes are found to be dispersive in an analogous fashion.

1.7 Discontinuities

We would like to know whether we can design finite difference methods for the case where φ is a discontinuous function, i.e., a function that is continuous, except at a finite number of points, $x_1 \ldots x_p$ such that the following limits exist

$$\lim_{x \to x_p,\, x < x_p} \varphi(x) = \varphi_L(x_p) \neq \varphi_R(x_p) = \lim_{x \to x_p,\, x > x_p} \varphi(x)$$

For such initial data, the exact solution formula (xx) is still well-defined, even though the derivatives appearing in the differential equation are not. Consequently, we can take that formula as the definition of what we mean by a solution to the differential equation. We can construct the discretized initial data φ_j by evaluating φ at grid points, and taking $\varphi_j = \varphi_L(j\Delta x)$ at points for which the data is discontinuous. Using that discrete data, we can define both the Fourier interpolant $\varphi_S(x)$, the spectrally accurate approximate solution $u_S(x,t)$, and the numerical solution u_j^n. We consider first the convergence properties of the spectrally accurate solution. In that case, it is well-known (Dym and McKean) that Fourier interpolation / approximation of discontinuities does not converge uniformly, i.e., that $max_x|\varphi(x) - \varphi_S(x)|$ does not approach zero as $M \to \infty$. Instead, one obtains an overshoot / undershoot in the neighborhood of the discontinuity, with magnitude of approximately ±14\% . This failure of Fourier approximations to converge, and the nature of the lack of convergence, was first observed by J.W. Gibbs, and is known as Gibb's phenomenon; an example is shown in figure (xx).Thus, as the solution evolves in time, $u_S(x, t) = \varphi_S(x - at)$, and we will sample the Gibbs oscillations at the grid points, depending on the location of the grid points relative to the oscillations. In spite of this problem, the Gibbs' oscillations occupy a decreasingly narrow interval in the neighborhood of the discontinuity as $M \to \infty$, so that the solution converges in L^1 or L^2.

Overall, finite difference methods have the same convergence properties as spectral methods in the presence of discontinuities. The numerical solutions converge in L^1 or L^2, but not uniformly. However, the qualitative behavior of these methods varies considerably, depending on the method. Upwind differencing, as we have discussed above, does not introduce any new maxima or minima that were not already present in the solution.

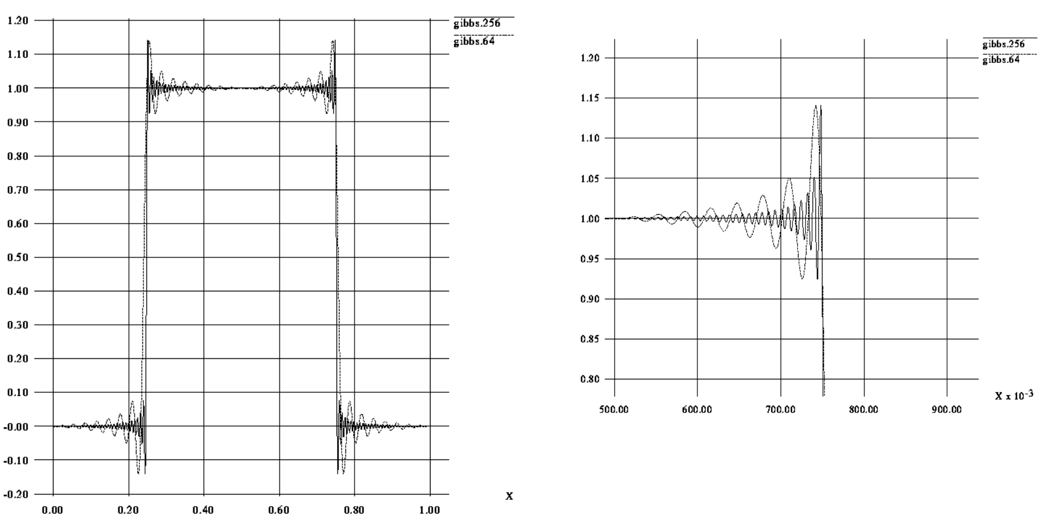


FIGURE 9. Gibbs' phenomenon, 64 and 256 -interpolation points. We plot the Fourier interpolant of step function at ten times the number of interpolation points.

This is considerably different from the behavior of the spectrally accurate solution, which produces new extrema. Both Lax-Wendroff and Fromm also produce oscillations near discontinuities. One can think of these oscillations as coming from two sources. First, to the extent that the finite difference solution $u_I^n(x)$ approximates $u_S^n(x)$, we expect to see Gibbs oscillations in the solution (these do not appear in the upwind differencing results because the short-wavelength components that lead to these oscillations are sufficiently strongly damped). In addition, the amplitudes of the short-wavelength Fourier components are sufficiently large ($b_k \sim k^{-1}$ for k large) that errors in the locations of these Fourier modes, i.e., phase errors, can lead to oscillations that are larger that the Gibbs oscillations.

In figures (xxx) - (xxx) we show several examples of the behavior of difference schemes in the presence of discontinuities. In all cases, we compare the results with the exact solution, as well as to the continuous Fourier interpolant, as in figure 9. Shown in figure (xx). The upwind differencing represents the discontinuity as a smooth monotone transition; however, there are many mesh points in the transition. In fact, one can analytically estimate the width of the transition for this scheme; at time t, it is given by $O((t\Delta x)^{1/2})$, i.e. a transition region which is, for fixed t, $O(M^{1/2})$ mesh points wide. This is consistent with the model of the upwind difference method that comes from the modified equation analysis that models the error in upwind differencing acting like an added diffusion term in the differential equation, with diffusion coefficient proportional to Δx. For both the Fromm and Lax-Wendroff schemes, we see oscillations whose amplitudes and spatial extent exceed those due to Gibbs' phenomenon in the spectrally accurate calculation. In addition, the qualitative character of the oscillation is consistent with the phase error properties of the schemes discussed previously. For both schemes, the phase error is a lagging phase error, causing short-wavelength to propagate a speeds less than the exact propagation speed. This leads to enhanced oscillatory behavior in back of the exact location of the shock; in addition, there is smoothing of the solution in front of the discontinuity due to the displacement of the short wavelengths that had initially represented the sharp transition.

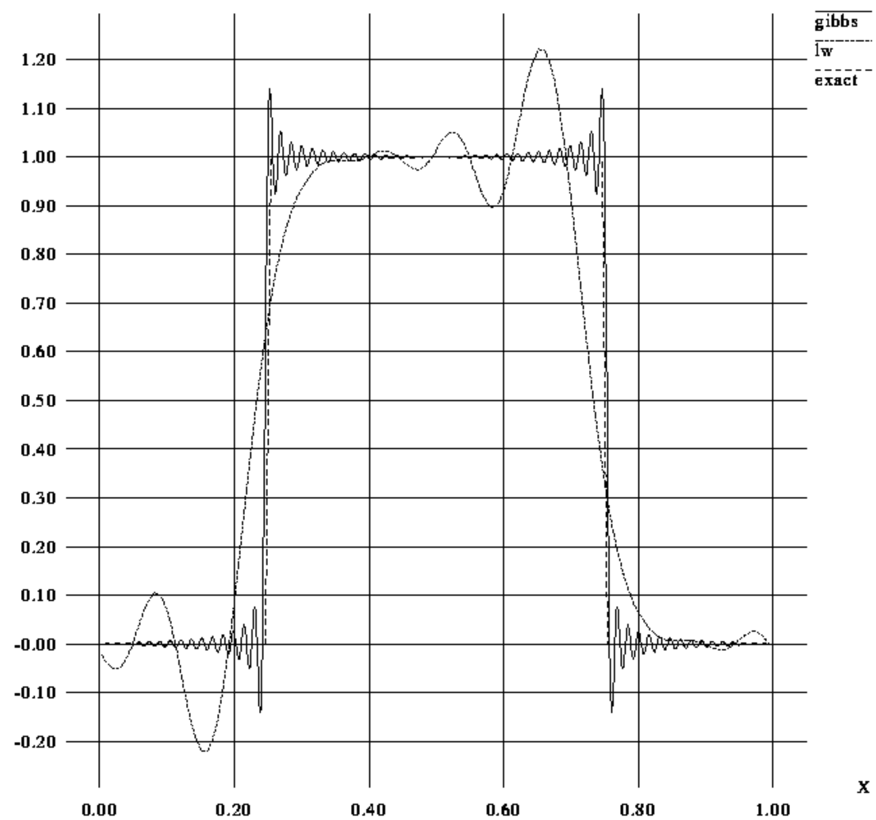


FIGURE 10. Lax Wendroff, the Fourier interpolant, and the exact solution, for 128 grid points. The Fourier interpolant is plotted as in figure 9.

A somewhat different, and more spatially localized explanation of what oscillations occur can be obtained by looking in detail at what the various finite difference schemes produce

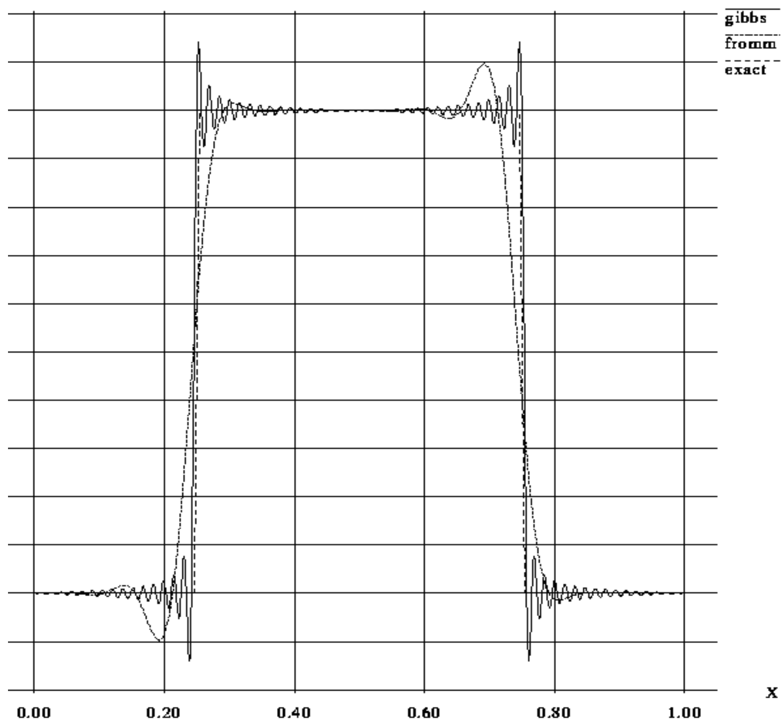


FIGURE 11. Fromm, Fourier interpolant, and exact solution, 128 grid points. Fourier interpolant plotted as in figure 9.

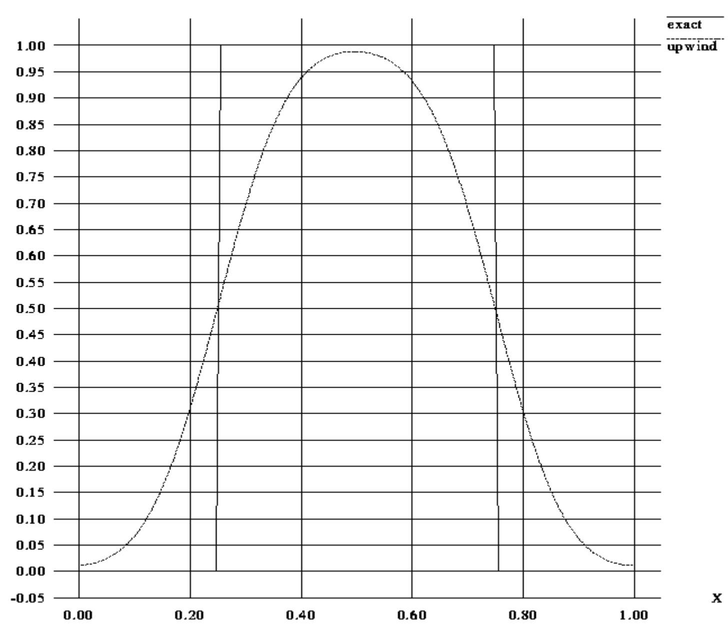
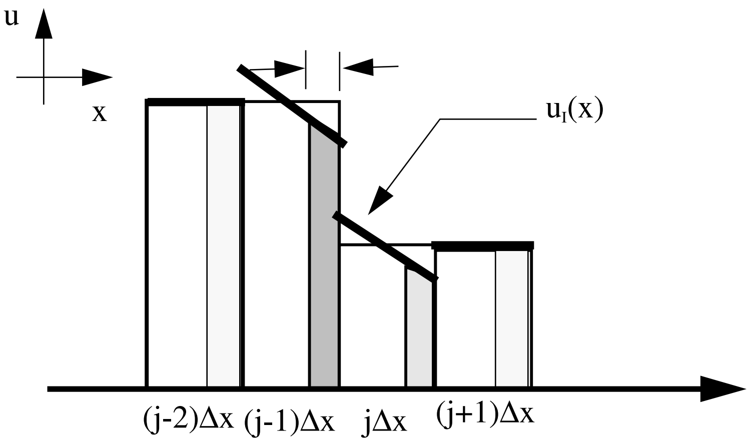


FIGURE 12. Upwind differencing and exact solution, 128 grid points.



(102)

FIGURE 13. Over/undershoot of numerical solution near a discontinuity using Fromm's scheme

for discontinuous initial data at the end of one time step. In figure 13, we show the case of Fromm's scheme. The shaded regions represent the total mass crossing the various cell edges. In the $j-1$ cell, the amount of mass leaving the cell is smaller than the amount of mass coming from the right. since the average of u is already at the maximum value to the left of the discontinuity, this leads to an overshoot. Similarly, in the $j+1$ cell, the amount of mass exiting from the left is greater than the amount of mass entering from the right. Since the initial value of the average is equal to the minimum value to the right of the discontinuity, this leads to an undershoot. in that cell.

In figure 14, we show the corresponding result for the Lax-Wendroff scheme. In this case, we observe only the an overshoot in the $j-1$ cell due to the lower value of the mass moving out across the right edge.

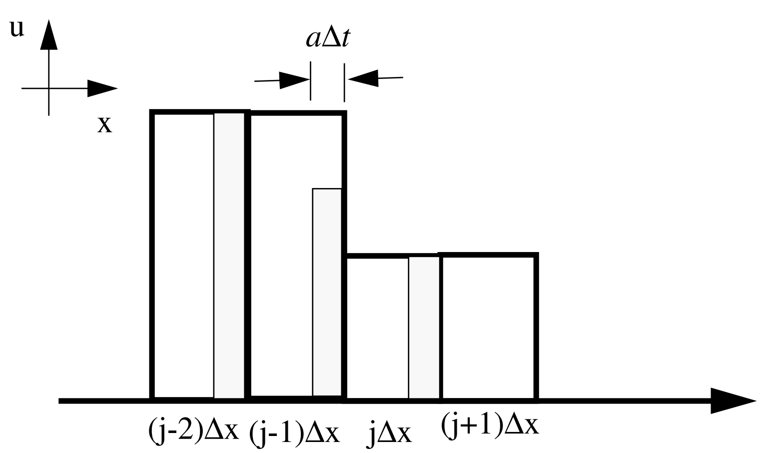


FIGURE 14. Overshoot of numerical solution near a discontinuity using the Lax-Wendroff scheme

On the other hand, upwinding avoids this problem altogether.

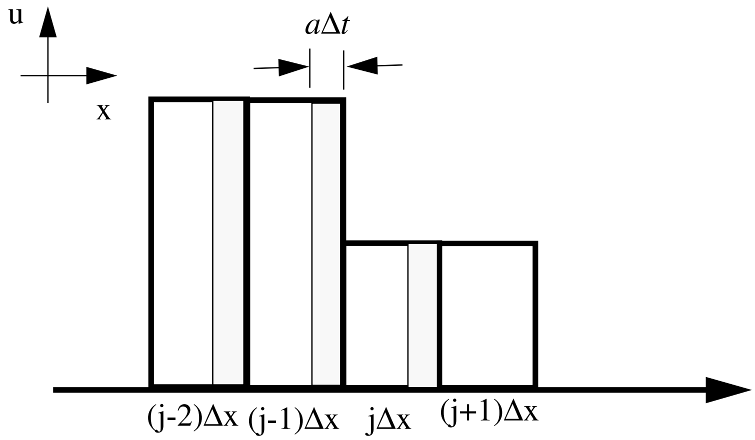


FIGURE 15. Behavior of numerical solution near a discontinuity using upwinding.

1.7.1 Max-norm boundedness and Godunov's Theorem

On physical grounds, it is often important that the solution to an advection equation of the form (1) not exceed the range of values defined by the initial data. If the solution represents the density of a fluid in a compressible flow calculation, the density must remain positive; otherwise, various quantities derived from the density may not be well-defined. Similarly, if the solution represents a concentration or mass fraction, it is important that the values remain between 0 and 1.

There are three ways of specifying this requirement. We say that a scheme is positivity preserving, if the $u_j^n \geq 0$ implies that $u_j^{n+1} \geq 0$. Second, we say that a scheme is monotonicity - preserving if $u_j^n \geq u_{j+1}^n$ for all j implies that $u_j^{n+1} \geq u_{j+1}^{n+1}$. finally, we say that a scheme is max-norm bounded if

$$\|Lu\|_\infty \leq \|u\|_\infty \tag{103}$$

where the max norm is defined by:

$$\|u\|_\infty = max_j(u_j) \tag{104}$$

For linear schemes, all three of these conditions are equivalent.

It is clear from the examples given above, that the Lax-Wendroff and Fromm schemes are not max-norm bounded, while upwind differencing is. On the other hand, the observed empirical accuracy of upwind differencing is much lower than that of the other two schemes, a reflection of the fact that the other two schemes are second-order accurate, while upwind differencing is first order accurate. Could it be that we have been insufficiently clever in constructing our methods, and that if we looked further, we could come up with a second-order accurate linear scheme which is also max-norm bounded? The answer, unfortunately, is no, as indicated by the following theorem.

Godunov's theorem: Let

$$(Lu)_j = \sum_s c_s u_{j+s} \tag{105}$$

be a linear scheme that is consistent and max - norm bounded and assume at least two of the coefficients c_s are nonzero. Then the scheme is first-order accurate.

.**Proof:**

We shall prove this in two steps. First, we show that L being stable in the max norm is equivalent to saying that the coefficients of the scheme are nonnegative; i.e. $c_s \geq 0$ for all s.

Assume that this is *not* true, i.e., that equation (105) holds for some c_s which is negative. For example, if:

$$u_s = \{ \begin{matrix} 1 \ldots (c_s > 0) \\ -1 \ldots (c_s < 0) \end{matrix} \tag{106}$$

Then $u_j = 0$ if $j \in \{s| (c_s = 0)\}$. Then, $\|u\|_\infty = 1$ and

$$(Lu)_0 = \sum c_s u_s = \sum |c_s|.$$

Since $\sum c_s = 1$ and $\exists s$ such that $c_s < 0$, this implies that:

$$\sum |c_s| > 1$$

and

$$\|u\|_\infty < \|Lu\|_\infty$$

which is a contradiction.

The second part of the proof is a simple calculation using the symbol. By (82)

$$|\lambda(\beta)|^2 = 1 + C_q \beta^q + O(\beta^{q+2}) \tag{107}$$

If the scheme is to be better than first order, then $q \geq 4$. For first order

$$|\lambda(\beta)|^2 = 1 + C_2 \beta^2 + O(\beta^4) \tag{108}$$

with $C_2 < 0$ for stability. This can be written as

$$|\lambda(\beta)|^2 = \sum_s c_s e^{i\beta s} \sum_{s'} c_{s'} e^{-i\beta s'}$$

Performing a Taylor expansion on the exponential, we obtain

$$|\lambda(\beta)|^2 = \sum_{s,\, s'} (c_s c_{s'}) \left(1 - i\beta s - \frac{\beta^2}{2} s^2\right) \left(1 + i\beta s' - \frac{\beta^2}{2} s'^2\right) + O(\beta^4)$$

$$1 - \sum_{s,\, s'} \frac{(s - s')^2}{2} c_s c_{s'} \beta^2 + O(\beta^4)\ .$$

If at least two of the c_s are nonzero, then the coefficient multiplying β^2 is nonzero, and the result is proven.

By consistency, in the form of equation (82), we have that $\sum_s c_s s = -\sigma$; consequently, any scheme that is consistent must have at least two nonzero c_s for σ not equal to an integer. Thus for almost all choices of σ, a linear finite difference scheme that is max-norm bounded will be first-order accurate. In addition, it will behave qualitatively like upwind differencing, in the sense that it will tend to smooth out discontinuities. This is easily seen, for example, by performing a modified equation analysis following that of equation (82).

So, how can we develop schemes that are higher-order accurate but still satisfy a maximum principle? The answer is to allow the schemes to be nonlinear, even for the linear problem (1) i.e., to be of the form

$$u^{n+1} = \pounds u^n \tag{109}$$

where £ is a nonlinear operator. However, we pay a price in doing this, in that we lose our linear analysis machinery - equivalence of stability and boundedness, and the use of Fourier analysis. Our strategy will be to allow us to use the results from linear analysis "almost everywhere", in the sense that, for regions where the solution is smooth, our nonlinear schemes will reduce to linear ones, and whose properties we understand. However, to talk about convergence for all parts of the solution, we need to develop some machinery.

1.8 Weak solutions, conservative finite differences and the Lax-Wendroff theorem.

Given initial data $U(x,0)$ for a general nonlinear system of equations:

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0 \tag{110}$$

where $U = U(x, t) \in \Re^n$ and $F = F(U(x, t)) \in \Re^n$, assume that a smooth solution exists over the domain. The engineering definition of a weak solution is that the solution must be integrable. All of the methods (so far) satisfy the discrete analog of the weak solution. By the divergence theorem,

$$\int_{x_L}^{x_R} U(x, t^N)\, dx = \int_{x_L}^{x_R} U(x, t^0)\, dx + \int_{t^0}^{t^N} F(U(x_L, t))\, dt + \int_{t^0}^{t^N} F(U(x_R, t))\, dt \qquad (111)$$

For any x_L, x_R, t^O, t^N. We can take this to be the definition of a solution, even in the case where the derivatives in (109) are not defined. Incidentally, this is just an expression in mathematical terms, of the familiar control-volume formulation of continuum mechanics. We can think of the integral of U over an interval in space as being the total amount of conserved quantity at some time - a total “mass”, in some generalized sense. Then (120) can be interpreted as the usual flux-balance relation over a control volume: mass at the new time is equal to the mass at the old time plus the net fluxes on the left and right edges.

A finite-difference method is in *conservation form* if it can be written as:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}\left(F_{j-\frac{1}{2}} - F_{j+\frac{1}{2}}\right)$$

where F represents the numerical fluxes at the cell boundaries. The weak consistency condition says that if we express the flux as

$$F_{j+\frac{1}{2}} = \Psi(U_{j-s}^n, \ldots, U_{j+s}^n)$$

and if we stuff a constant value U_o into Ψ, i.e., $\Psi(U_o, \ldots, U_o) = F(U_o)$. The discrete analogue of equation (111) above is

$$\sum_{j=j_L}^{j_R} U_j^{n+1} \Delta x = \sum_{j=j_L}^{j_R} U_j^n \Delta x + \Delta t \sum_{j=j_L}^{j_R} \left(F_{j-\frac{1}{2}} - F_{j+\frac{1}{2}}\right) \qquad (112)$$

At any jth cell, the flux coming from the *j-1*th cell is equal to the flux leaving the *j-1*th cell. Consider a domain which is three cells wide as below. The rate of change of the mass of

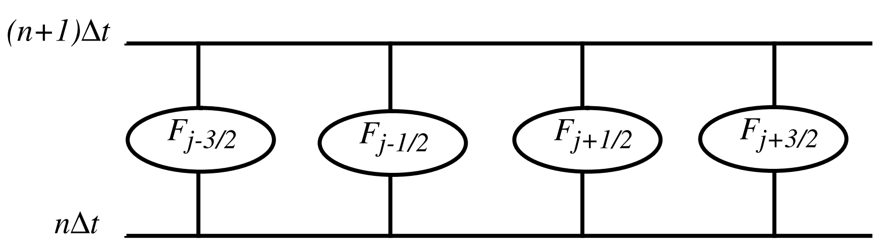


FIGURE 16. Flux through three cells

these three cells is given by

$$\left(F_{j-\frac{3}{2}} - F_{j-\frac{1}{2}}\right) + \left(F_{j-\frac{1}{2}} - F_{j+\frac{1}{2}}\right) + \left(F_{j+\frac{1}{2}} - F_{j+\frac{3}{2}}\right) = F_{j-\frac{3}{2}} - F_{j+\frac{3}{2}}$$

Notice that all the terms on the right-hand side cancel except for the first and last terms. This implies that the only fluxes which count are those at the edges of the domain. More generally, we have the following mass balance relation:

$$\Delta x \sum_{j=j_L}^{j_R} U_j^{N_n} = \Delta x \sum_{j=j_L}^{j_R} U_j^{N_o} + \Delta t \sum_{N_O}^{N_N-1} F_{j_L-1/2} - \Delta t \sum_{N_O}^{N_N-1} F_{j_R+1/} \qquad (113)$$

This is a consistent discretization of the integral relation (110) defining weak solution. In particular, this enables us to compute discontinuous solutions. In figure 17 below, we show the exact solution to a discontinuous initial value problem plotted with the solution obtained by a conservative finite difference scheme. The fact that the scheme is conservative implies that the integral under the numerical solution curve is identical to that for the exact solution.

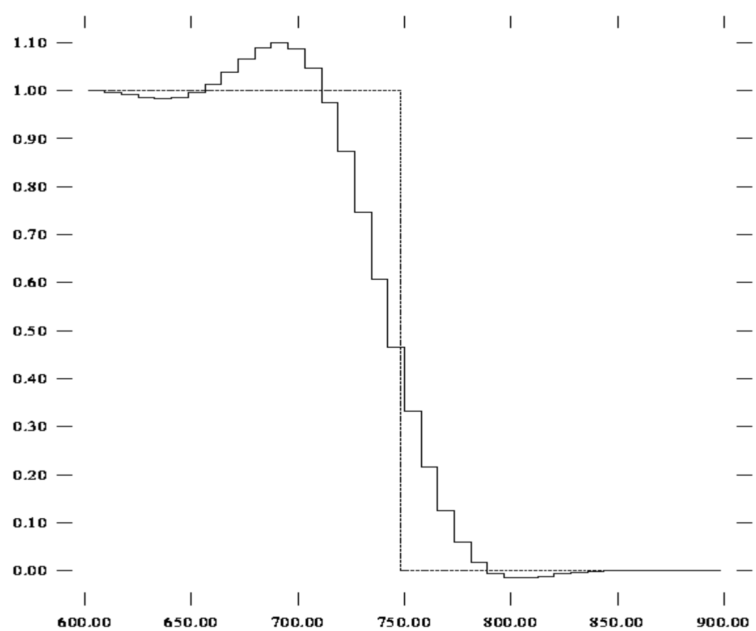


FIGURE 17. Conservative finite difference calculation of a discontinuity versus the exact solution.

The Lax-Wendroff theorem.

Assume that the finite-difference scheme under consideration is consistent and in conservation form. If U_j^n converges to some function, say $U_\infty(x,t)$, in some norm such as L^1 or L^2, then U_∞ is a weak solution to

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0 \tag{114}$$

Note that this is *not* equivalent to saying that if the scheme is consistent and in conservation form, then it will converge; Lax-Wendroff doesn't guarantee existence of a weak solution.

1.9 Limiters

Consider discontinuous solutions to the hyperbolic equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{115}$$

Reasonable design criteria for a numerical scheme which solves this equation would need to include consistency and conservation. The discretized equation could be written in the form:

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{\Delta x}\left(F_{j-\frac{1}{2}} - F_{j+\frac{1}{2}}\right) \tag{116}$$

where $F_{j+\frac{1}{2}} = au_{j+\frac{1}{2}}$. The consistency condition indicates that:

$$\Phi(u_{j-s}, \ldots, u_{j+s}) = F_{j+\frac{1}{2}}$$

$$\Phi(u_0, \ldots, u_0) = au_0$$

We will attempt to obtained the properties of second- order accuracy for smooth solutions and positivity preservation by hybridizing first order and second order schemes. Given a second order scheme with scheme with flux F^H and a first-order scheme with flux F^L, we can take a linear combination of the two fluxes to obtain a hybrid scheme, as follows.

$$\begin{aligned} F_{j+\frac{1}{2}} &= \alpha_{j+1/2} F^H_{j+\frac{1}{2}} + (1-\alpha_{j+1/2}) F^L_{j+\frac{1}{2}} \\ &= F^L_{j+\frac{1}{2}} + \alpha_{j+1/2}\left(F^H_{j+\frac{1}{2}} - F^L_{j+\frac{1}{2}}\right) \end{aligned}$$

Here, $0 \le \alpha \le 1$. The high-order flux comes from our arsenal of linear difference schemes with order greater than two, while the low-order flux satisfies the maximum principle. At the discrete level, this looks like having an exact and a diffusive flux. The goal of linear hybridization is to keep the solution from straying outside some maximum/minimum bounds. Godunov's theorem shows that α=$\alpha(u)$ for the hybridized scheme to be second-order accurate in regions where the solution is smooth; in such regions, we would expect $\alpha = 1$

1.9.1 Flux-corrected transport

Flux-corrected transport was an early version of this approach. The initial work was done by Boris and Book (early 1970's); see Zalesak (1979) **JCP 35**:xxx-xxx for a good discussion of this method. The basic idea is to choose α so that no overshoot or undershoot can occur. Since we don't want *any* high-frequency amplification (a nontrivial requirement), we require F^H to be Fourier stable. Since notation is free, let's denote the anti-diffusive flux as follows.

$$A_{j+\frac{1}{2}} = F^H_{j+\frac{1}{2}} - F^L_{j-\frac{1}{2}} \tag{117}$$

The quantity $A_{j+1/2}$, to leading order in Δx, looks like a flux for a backwards diffusion term, in the sense of a modified equation analysis; hence its name. We also define the transported and diffused velocity by:

$$u^{TD} = u^n_j + \frac{\Delta t}{\Delta x}\left(F^L_{j-\frac{1}{2}} - F^L_{j+\frac{1}{2}}\right) \tag{118}$$

In other words, in order to obtain u^{TD}, simply evolve u for one time step using the lower-order scheme. Finally, we define the corrected flux A^c as follows.

$$A^c_{j+\frac{1}{2}} = \alpha_{j+1/2}\left(F^h_{j+\frac{1}{2}} - F^l_{j-\frac{1}{2}}\right) \tag{119}$$

With all of these new definitions in place, the discrete evolution equation for u is given by

$$u^{n+1}_j = u^{TD} + \frac{\Delta t}{\Delta x}\left(A^c_{j-\frac{1}{2}} - A^c_{j+\frac{1}{2}}\right) \tag{120}$$

The strategy for computing the hybridization parameter $\alpha_{j+1/2}$ or, equivalently, $A^c_{j+1/2}$, is based on the following ideas. First, we want to compute the corrected flux so that it as nearly equal to $A_{j+1/2}$ as possible, subject to the constraint that the outcome of applying (119) will not exceed the range of values defined by nearby values of u^{TD}. Thus, if the low-order scheme satisfies a maximum principle, then the hybrid scheme will do so, as well. In addition, if we choose a sufficiently large window over which to define the range of minimum and maximum values ofu^{TD}, the scheme should reduce to the linear high-order scheme in regions where the solution is sufficiently smooth to be resolved on the grid. Finally, one would like a algorithm that looks at each edge independently from the others. The problem with this last criterion is that it is possible for two fluxes, acting together, to cause a value to exceed the max and min limits, although each flux, acting separately, might remain in bounds. The solution to this problem obtained by Boris and Book is to use the intuition that $A_{j+1/2}$ is an antidiffusive flux, so that it has the same sign as the local gradient.

We illustrate this idea in the following figure. If the antidiffusive fluxes are all in the direction of the local solution gradient, as defined by u^{TD}, the two antidiffusive fluxes will have the same sign, and for that reason partially cancel.

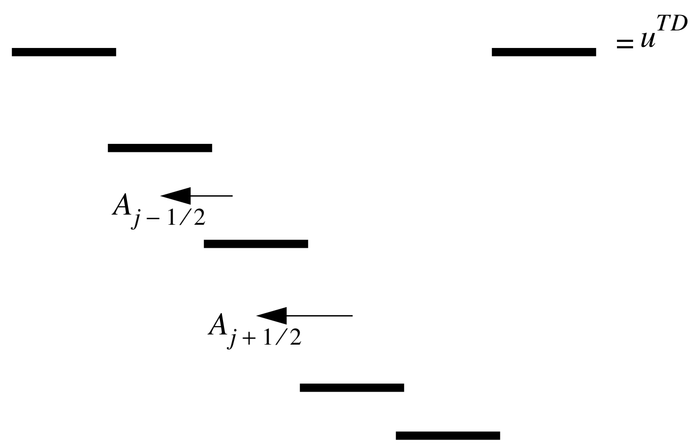


FIGURE 18. Partial cancellation of antidiffusive fluxes at adjacent cell edges

On the other hand, if any antidiffusive flux fails to satisfy this condition, the corresponding corrected flux will be set to zero.

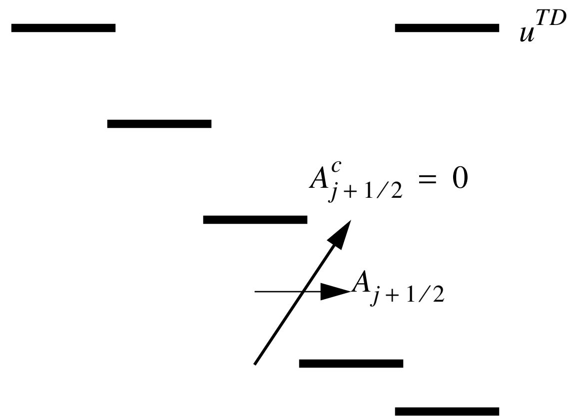


FIGURE 19. Cancellation of a corrected flux that is not antidiffusive.

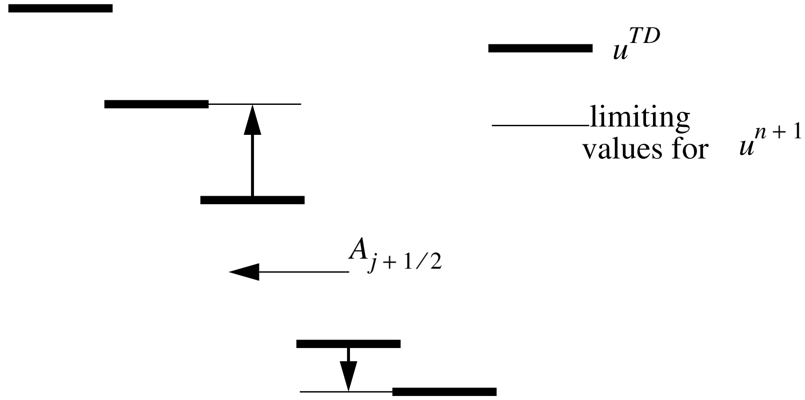


FIGURE 20. Limiting of the corrected antidiffusive flux.

The combination of these two conditions allows us to limit each corrected flux separately, since the extent to which it will cause the solution to exceed the range of values defined by u^{TD} will be bounded by the extent to which that will occur with the fluxes at the other edges set to zero. Specifically, we will take the corrected flux to be as nearly equal to $A_{j+1/2}$, subject to the constraint that its use in (xx) with the fluxes set to zero at the other edges does not allow u_j^{n+1} to go outside the range defined by u_{j-1}^{TD}, u_j^{TD}, nor u_{j+1}^{n+1} to go outside the range defined by u_{j+1}^{TD}, u_{j+2}^{TD} (figure xx).Algebraically, this constraint is given as follows.:

$$\left|A^c_{j+\frac{1}{2}}\right| = min\left(\left|\frac{\Delta x}{\Delta t}\left(u^{TD}_{j-1} - u^{TD}_{j}\right)\right|, \left|\frac{\Delta x}{\Delta t}\left(u^{TD}_{j+2} - u^{TD}_{j+1}\right)\right|, \left|A_{j+1/2}\right|\right) \tag{121}$$

Combining these two conditions, and accounting for all the possible combinations of signs, the corrected flux can be written as follows.

$$A^c_{j+\frac{1}{2}} = S_{j+\frac{1}{2}} \bullet max\left(0, \left|A_{j+\frac{1}{2}}\right|, S_{j+\frac{1}{2}}\frac{\Delta x}{\Delta t}\left(u^{TD}_{j+2} - u^{TD}_{j+1}\right), S_{j+\frac{1}{2}}\frac{\Delta x}{\Delta t}\left(u^{TD}_{j} - u^{TD}_{j-1}\right)\right) \tag{122}$$

where

$$S_{j+\frac{1}{2}} = sign\left(A_{j+\frac{1}{2}}\right) \tag{123}$$

1.9.2 Geometric limiters

The basic idea of geometric limiters is to first apply a limiter to an interpolation function $u_I(x)$, and then use an upwind method to calculate the fluxes and advance the solution. One example is van Leer's scheme, which combines Fromm's method with a geometric limiter. Van Leer's scheme adjusts the slopes, i.e., the Δu_j's, so that the interpolation function $u_I((j+1/2)\Delta x)$ stays in the range defined by u_j^n and $u_{j\pm1}^n$ (Figure 21).

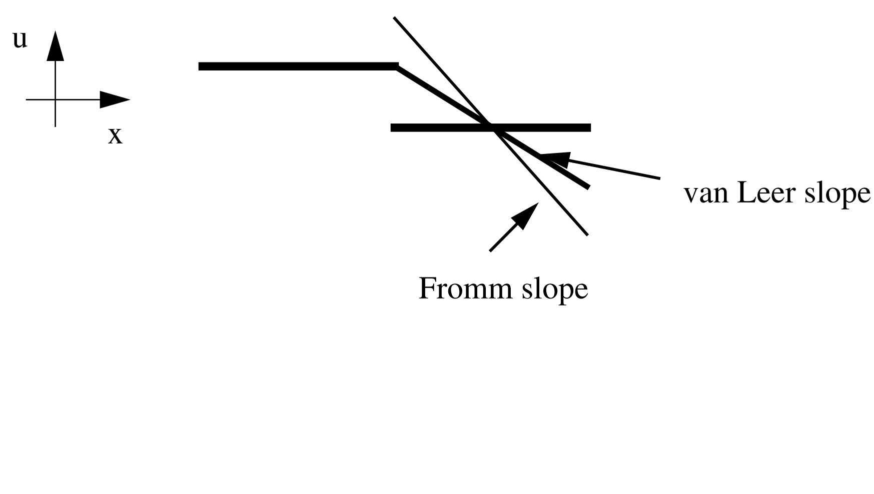


FIGURE 21. Example of slope limiting.

. We can write the flux as:

$$F_{j+\frac{1}{2}} = au_{j+\frac{1}{2}} \tag{124}$$

where

$$u_{j+\frac{1}{2}} = u_j^n + (1-\sigma)\,\Delta u_j^{VL} \tag{125}$$

In the above equation, σ is the CFL number and the van Leer slopes are:

$$\Delta^{VL} u_j = \begin{cases} S_j \bullet min\left(2|u_{j+1} - u_j|, 2|u_j - u_{j-1}|, \frac{1}{2}|u_{j+1} - u_{j-1}|\right) \ldots if \ldots \varphi > 0 \\ 0 \ldots otherwise \end{cases} \tag{126}$$

where S_j now denotes the sign of the gradient

$$S_j = sign\,(u_{j+1}^n - u_{j-1}^n) \tag{127}$$

and the quantity φ is given by

$$\varphi = (u_{j+1}^n - u_j^n) \bullet (u_j^n - u_{j-1}^n) \tag{128}$$

This scheme "naturally" detects a discontinuity and modifies its behavior accordingly. The implications of this are that this method retains the high-order accuracy of Fromm's scheme in smooth regions of the solution, but where discontinuities are detected, the discretized evolution equation drops to first-order accuracy. (This is a reasonable thing to do since you probably don't have enough derivatives in this region, anyways.)

We note finally that the van Leer scheme is actually a linear hybridization of Fromm's scheme and upwinding with a special choice of the hybridization parameter α. Consider

$$F_{j+\frac{1}{2}} = au_{j+\frac{1}{2}}$$

where the cell edge velocity is defined for van Leer's scheme as

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}(1-\sigma)\,\Delta^{VL} u_j \tag{129}$$

The van Leer slope $\Delta^{VL} u_j$ was defined previously in equation (xxx). For Fromm's scheme, the analogous quantity is

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}(1-\sigma)\,\Delta^c u_j \tag{130}$$

where the slope is defined as simply

$$\Delta^c u_j = \frac{1}{2}(u_{j+1} - u_{j-1}) \tag{131}$$

To conform to the notation adopted at the beginning of this section on limiters, if we consider $F^H_{j+\frac{1}{2}} \leftrightarrow F^{Fromm}$ and $F^L_{j+\frac{1}{2}} \leftrightarrow au^n_j$, then we can write the van Leer flux as:

$$\begin{aligned} F^{VL} &= F^L + \alpha(F^{Fromm} - F^L) \\ &= au^n_j + a\frac{1}{2}(1-\sigma)\Delta^{VL}u_j \\ &= au^n_j + a\frac{1}{2}(1-\sigma)\Delta^C u_j \left[\frac{\Delta^{VL}u_j}{\Delta^C u_j}\right] \end{aligned}$$

If we choose α to be the term in brackets above, then we can write F^{VL} as:

$$F^{VL} = au^n_j + \alpha(F^{Fromm} - F^L) \tag{132}$$

1.9.3 Design criteria for schemes with limiters

Thus far, we have identified a general set of techniques for constructing schemes with limiters. The ingredients are:

- A linear low-order scheme that satisfies a maximum principle;
- A linear high-order scheme that is linearly stable;
- A hybridization switch between the two, designed so that it the resulting hybrid scheme introduces no new local extrema, and reduces to the high-order scheme for smooth solutions. In order for this last condition to be satisfied, the hybridization coefficient $\alpha_{j+1/2}$ must depend on the solution values.

The way we have accomplished the third criterion is to construct the hybridization switch so that it blends in, by some measure, the minimum amount of the low-order scheme necessary to prevent new extrema from forming. In the FCT case, the measure was that the solution must remain within bounds set by the transported and diffused solution. For the van Leer scheme, the interpolation profile is constrained so that, independent of the CFL number σ, the overall scheme would general no new extrema. On of the consequences of this kind of construction is that, after a few time steps, the discrete profile representing a discontinuity approaches something very close to a discrete travelling wave solution, with the number of mesh points in the transition across the discontinuity increasing only very slowly as a function of time. This is in contrast with the first-order linear schemes, for which the number of mesh points in the transition representing a discontinuity increases like $O(n^{1/2})$, where n is the number of time steps.

Next, we would like to make the argument that the accuracy with which hybrid schemes of the form discussed above represent discontinuities is intimately tied to the phase error properties of the scheme. In section 1.7, we saw that the effect of the leading-order dispersive errors of second order schemes was to cause a lagging or leading oscillatory wave train (depending on the sign of the phase error) due to the errors in the transport of the large wavenumber Fourier modes. In addition, the extraction of the high-wavenumber modes from one side of the discontinuity caused the side opposite that of the oscillations to be smoothed out. Overall, the lower the phase error of the scheme, the narrower this oscillatory / smooth structure representing the discontinuity was.

When we combine these second-order schemes with limiters, we find that the interaction of the phase error properties of the high-order scheme with the limiter dominate the accuracy with which discontinuities are represented on the grid. Roughly speaking, in the region in which the linear high-order scheme would otherwise be oscillating, the solution is chopped off to be flat, due to the influence of the limiter. However, that high-wavenumber component of the solution is still being extracted from the side opposite the oscillatory region, leading to a smoothed out profile. Thus to a high degree of accuracy, the effect of the limiter is to replace the oscillatory region near a discontinuity with a flat profile, while leaving the smoothed-out portion of the solution from the second-order linear scheme unchanged. In Figure 22 we show comparisons between the Lax-Wendroff and FCT / Lax-Wendroff schemes, and between the Fromm and van Leer schemes. The solution profiles support this scenario for the interaction of the phase error of the high-order scheme with the treatment of discontinuities. Finally, if we compare the FCT/ Lax-Wendroff and van Leer schemes for this run, we find that the number of mesh points in the van Leer representation of the discontinuities is about half that in the FCT / Lax Wendroff representation of the discontinuity. Thus, even in the case of hybrid schemes and discontinuous solutions, the lower phase error of the Fromm scheme is worth about a mesh refinement over the Lax-Wendroff scheme.

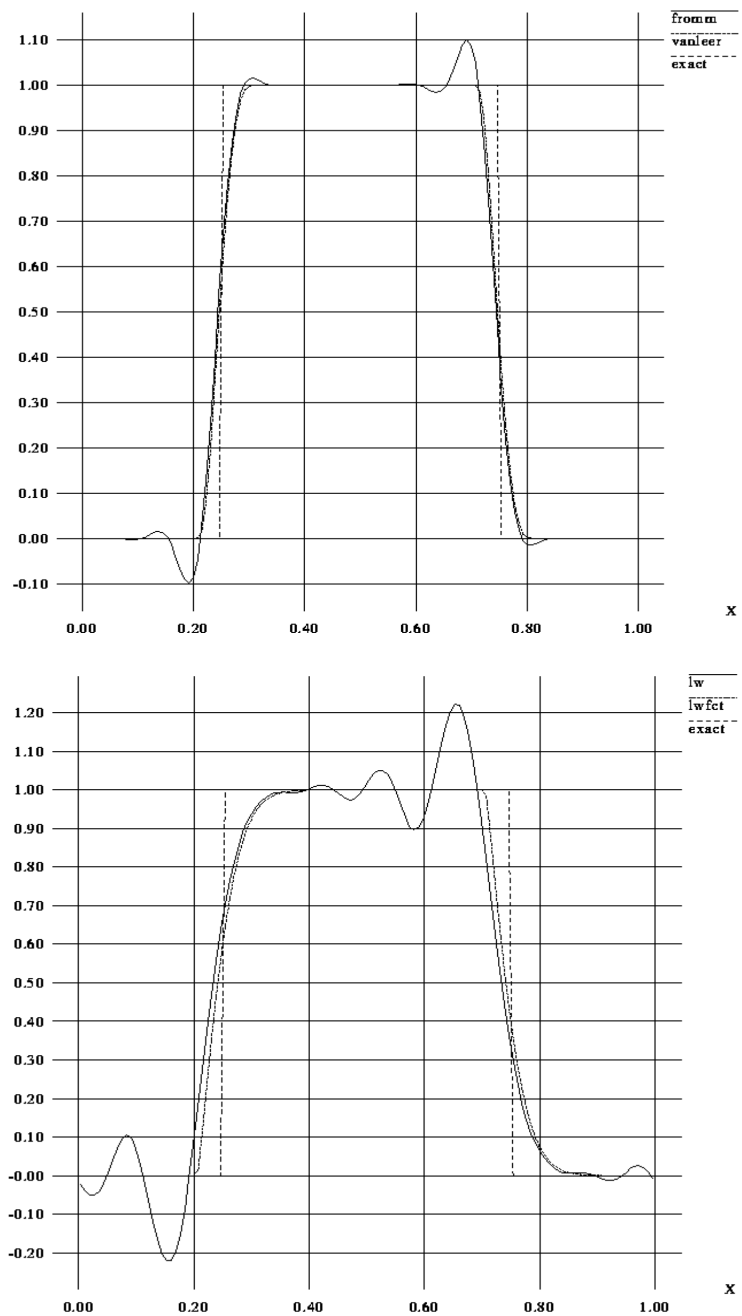


FIGURE 22. Fromm / van Leer comparison (top), Lax-Wendroff / FCT comparison (bottom). 128 mesh points, CFL = 0.75, time = 5.

1.9.4 Total variation diminishing (TVD) schemes.

The total variation norm of u is

$$TV(u) = \sum_j |u_j - u_{j+1}| \tag{133}$$

If this the spatial derivative of u is integrable, this summation should be approximately equivalent to

$$TV(u) \approx \int \left|\frac{\partial u}{\partial x}\right| dx$$

The significance of the total variation norm is that it is the appropriate norm to bound if one wants to guarantee convergence of a function on a grid to something in L^1. Roughly speaking, if $u_j^{(q)}$ is a sequence of functions defined on a grid with grid spacing $\Delta x = D/q$, and $TV(u^{(q)})$ is bounded uniformly in q, then there is a function v such that a subsequence of the $u^{(q)}$'s converges to v, locally in L^1:

$$\sum_{j,\ |j\Delta x| \leq K} \left|u_j^{(q)} - v(j\Delta x)\right| \Delta x \to 0 \tag{134}$$

for any finite K. (Actually, there is a slight additional technical requirement: the interval over which the bulk of the variation takes place does not go out to infinity).

A scheme is said to be TVD if

$$TV(Lu) \leq TV(u) \tag{135}$$

which implies

$$TV(u^n) \leq TV(u^0) \tag{136}$$

The significance of the condition that a scheme is TVD is that the scheme converges to some function, locally in L^1 to some function v. If, in addition, it is consistent and in conservation form, the Lax-Wendroff theorem implies that v is a weak solution to the differential equation. Thus we have the analogue of the Lax equivalence theorem for nonlinear schemes in conservation form: TVD + conservation form / consistency $\rightarrow$ convergence.

which is what we want. If you can write a scheme in conservation form for a linear scalar:

$$F_{j+\frac{1}{2}} = au_j + C^-_{j+\frac{1}{2}} \cdot (u_{j+1} - u_j) \tag{137}$$

and

$$F_{j-\frac{1}{2}} = au_j + C^+_{j-\frac{1}{2}} \cdot (u_j - u_{j-1}) \tag{138}$$

where $C^{-,+}_{j+\frac{1}{2}} \geq 0$, then

$$\frac{\Delta t}{\Delta x}\left(C^{+}_{j+\frac{1}{2}} + C^{-}_{j-\frac{1}{2}}\right) \le 1 \tag{139}$$

One can write a generic TVD scheme as:

$$u_j^{n+1} = u_j^n - C^{-}_{j+\frac{1}{2}} \cdot (u_{j+1} - u_j) + C^{+}_{j-\frac{1}{2}} \cdot (u_j - u_{j-1}) \tag{140}$$

Two obscure comments: all the limiters we've described previously look like the above on the second and third terms. Can we consider the flux-splitting schemes as TVD? For the van Leer scheme,

$$F_{j+\frac{1}{2}} = a\,(u_j^n + \frac{1}{2}(1-\sigma)\,\Delta^{VL}u_j) \tag{141}$$

The analogous expression for TVD is

$$F_{j+\frac{1}{2}} = au_j^n + C^{-}_{j+\frac{1}{2}}(u_{j+1} - u_j) \tag{142}$$

and similarly

$$F_{j-\frac{1}{2}} = au_j^n + C^{+}_{j-\frac{1}{2}} \cdot (u_j - u_{j-1}) \tag{143}$$

Recall that the van Leer limiter looks for the jump between cells *j+1* and *j-1*, not between *j+1* and *j*. Comparing the expressions for $F_{j+1/2}$, equations (-2) and (-3) above, we conclude that the correction to the flux is

$$\frac{a}{2}(1-\sigma)\,\Delta^{VL}u_j = \begin{cases} \dfrac{\frac{a}{2}(1-\sigma)\;(u_{j+1} - u_{j-1})\left|\Delta^{VL}u_j\right|}{\left|u_{j+1} - u_{j-1}\right|} \\ 0 \end{cases} \tag{144}$$

If the top quantity on the right can be shown to be a constant, e.g., η then we're done. Note that

$$\eta = \frac{\left|\Delta^{VL}u_j\right|}{\left|u_{j+1} - u_{j-1}\right|} \le 2$$

From this, we can conclude that

$$C^{-}_{j+\frac{1}{2}} = \left(\begin{matrix} 2\eta\,(1-\sigma) \\ 0 \end{matrix}\right. \tag{145}$$

Secondly, a scheme is positivity-preserving if

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{\Delta x} A_j (u_{j+1}^n - u_j^n) - \frac{\Delta t}{\Delta x} B_j (u_j^n - u_{j-1}^n) \tag{146}$$

or

$$u_j^{n+1} = u_j^n \left(1 - \left(\frac{\Delta t}{\Delta x} A_j + \frac{\Delta t}{\Delta x} B_j\right)\right) + \frac{\Delta t}{\Delta x} A_j u_{j-1}^n + \frac{\Delta t}{\Delta x} B_j u_{j-1}^n \tag{147}$$