Materials from

LeVeque R. J. 1992. *Numerical Methods for Conservation Laws*, Birkhäuser.

LeVeque R. J. 2002. *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press.

Hirsch C. 1988-1990 *Numerical Computation of Internal and External Flows*, Volumes 1 and 2, Wiley.

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 model. In other words, there are 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 qualitative behavior of the numerical methods we wish to understand. Our model linear equation is

$$u_t + au_x = Du_{xx} + \mu u_{xxx} + \dots \ ,$$

where a, D and μ are constants, and the x and the t subscripts indicate partial derivatives. The second order term on the RHD is the diffusion term and the third order term the dispersive term. Its solution is composed of $e^{i(kx-wt)}$, where $w = ak - iDk^2 + \mu k^3$.

Consider the linear convection equation

$$u_t + au_x = 0 \ , \tag{1.1}$$

where $a > 0$. The unstable, forward in time, central scheme for linear convection equation is

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) = 0. \tag{1.2}$$

The derivation of the modified equation is closely related to the calculation of the local truncation error for a given model. Consider the explicit CDS scheme for the linear convection equation (1.2): If the function $u(x,t)$ is sufficiently smooth we can write the following Taylor series expansions:

$$u_i^{n+1} = u_i^n + \Delta t (u_t)_i^n + \frac{\Delta t^2}{2}(u_{tt})_i^n + \ldots \tag{1.3}$$

$$u_{i+1}^n = u_i^n + \Delta x (u_x)_i^n + \frac{\Delta x^2}{2}(u_{xx})_i^n + \frac{\Delta x^3}{6}(u_{xxx})_i^n + \ldots \tag{1.4}$$

$$u_{i-1}^n = u_i^n - \Delta x (u_x)_i^n + \frac{\Delta x^2}{2}(u_{xx})_i^n - \frac{\Delta x^3}{6}(u_{xxx})_i^n + \ldots \tag{1.5}$$

Substituting these developments in Eqn. (1.2) we obtain

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) - (u_t + a u_x)_i^n = \frac{\Delta t}{2}(u_{tt})_i^n + \frac{\Delta x^2}{6} a (u_{xxx})_i^n + O(\Delta t^2, \Delta x^4) \tag{1.6}$$

It is clearly seen from the above equation that the right-hand side vanishes when Δt and Δx tend to zero, and therefor scheme (1.2) is consistent. As expected, the accuracy of the scheme is first order in time and second order in space. The consistency equation 1.6 can be interpreted in the following ways. If the value u_i^n are the exact solution of the numerical scheme, Eqn (1.6) reduces to

$$(u_t + a u_x)_i^n = -\frac{\Delta t}{2}(u_{tt})_i^n - \frac{\Delta x^2}{6} a (u_{xxx})_i^n + O(\Delta t^2, \Delta x^4) \tag{1.7}$$

showing that the exact solution of the difference equation does not satisfy exactly the differential equation at finite values of Δt and Δx (which is always the case in practical computations.) However, the solution of the difference equation satisfies an equivalent differential equation (also called a modified differential equation), which differs from the original (differential) equation by a truncation error represented by the terms on the right-hand side.

In the present example the truncation error ϵ_T is equal to

$$\epsilon_T = -\frac{\Delta t}{2}(u_{tt})_i^n - \frac{\Delta x^2}{6}(u_{xxx})_i^n + O(\Delta t^2, \Delta x^4) \tag{1.8}$$

and can be written in an equivalent form, up to higher-order correction terms, by applying the equivalent differential equation (1.6) to eliminate the time derivatives; for instant,

$$(u_t)_i^n = -a(u_x)_i^n + O(\Delta t, \Delta x^2) \tag{1.9}$$

and differentiating the above equation,

$$(u_{tt})_i^n = -a(u_{xt})_i^n + O(\Delta t, \Delta x^2) = (a^2 u_{xx})_i^n + O(\Delta t, \Delta x^2) \tag{1.10}$$

Hence the truncation error can be written as

$$\epsilon_T = -\frac{\Delta t}{2}a^2(u_{xx})_i^n - \frac{\Delta x^2}{6}(u_{xxx})_i^n + O(\Delta t^2, \Delta x^2) \tag{1.11}$$

Up to the truncation error the equivalent differential equation becomes

$$u_t + au_x = -\frac{\Delta t}{2}a^2 u_{xx} + O(\Delta t^2, \Delta x^2) \tag{1.12}$$

and this shows why the corresponding scheme is unstable. Indeed, the right-hand side represents a viscosity term, with a negative viscosity coefficient equal to $-\frac{\Delta t}{2}a^2$. A positive viscosity is known to damp oscillations and strong gradients; a negative viscosity, on the other hand, will amplify any disturbance, its behavior is unstable. Therefore the determination of the equivalent differential equation and in particular, the truncation error provides essential information as to the behavior of the numerical solution, and will generally lead to necessary conditions for the stability.

2 First order methods and diffusion

The first order upwind method scheme of an linear convection-diffusion equation uses one-side approximation for u_x:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{a}{\Delta x}(u_i^n - u_{i-1}^n) = 0 \ . \tag{2.13}$$

The modified equation of this scheme can be derived similarly, and is found to be of the form

$$u_t + au_x = Du_{xx} + O(\Delta t^2, \Delta x^2) \tag{2.14}$$

with a diffusion constant given by

$$D = \frac{a\Delta x}{2}\left(1 - \left(\frac{\Delta t}{\Delta x}a\right)\right) \tag{2.15}$$

Note that the modified equation varies with Δt and Δx.

We expect solution of these equation to become smeared out as time evolves, explaining at least the qualitative behavior of the upwind method method seen in the Fig. 1. In fact this equation is even a good quantitative model for how the solution behaves. If we plot the exact solution to (2.14) along with the upwind method numerical solution, they are virtually indistinguishable to plotting accuracy.

Relation to Stability Notice that the equation is mathematically well posed only if the diffusion coefficient is positive. Otherwise it behaves like the backward heat equation which is notoriously ill posed. This requires of D be nonnegative. This is nonnegative if and only if the stability condition is satisfied. We see that the modified equation also gives some indication of the stability properties of the method.

3 Second order method and dispersion

The Lax-Wendroff method for the linear system $u_t + au_x = 0$ is based on the Taylor series expansion (1.3). From the differential equation we have that $u_t = -au_x$, and differentiating this gives

$$u_{tt} = -au_{xt} = a^2u_{xx} \ .$$

Using these expressions for u_t and u_{tt} in (1.3) gives

$$u_i^{n+1} = u_i^n - a\Delta t (u_x)_i^n + \frac{a^2 \Delta t^2}{2} (u_{xx})_i^n + \ldots \tag{3.16}$$

Keeping only the first three terms on the right-hand side and replacing the spatial derivatives by central finite difference approximations gives the Lax-Wendroff method,

---

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{2} \left(u_{i+1}^n - u_{i-1}^n\right) + \frac{a^2 \Delta t^2}{2\Delta x^2} \left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) \ . \tag{3.17}$$

By matching three terms in the Taylor series and using centered approximations, we obtain a second-order accurate method.

In place of the centered formula for u_x and u_{xx}, we might use one-sided formulas:

$$(u_x)_i^n = \frac{1}{\Delta x} \left(3u_i^n - 4u_{i-1}^n + u_{i-2}^n\right)$$

$$(u_{xx})_i^n = \frac{1}{\Delta x^2} \left(u_i^n - 2u_{i-1}^n + u_{i-2}^n\right)$$

Using these in Eqn. (3.16) gives a method that is again second-order accurate,

---

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{2} \left(3u_i^n - 4u_{i-1}^n + u_{i-2}^n\right) + \frac{a^2 \Delta t^2}{2\Delta x^2} \left(u_i^n - 2u_{i-1}^n + u_{i-2}^n\right) \ . \tag{3.18}$$

This is known as the Beam-Warming method.

The Lax-Wendroff method gives a third order accurate approximation to the solution of the modified equation:

$$u_t + a u_x = \mu u_{xxx}. \tag{3.19}$$

---

where

$$\mu = \frac{\Delta x^2}{6} a \left(\frac{\Delta t^2}{\Delta x^2} a^2 - 1 \right) \tag{3.20}$$

This is a dispersive equation. The theory of dispersive wave is covered in detail in Whitham[1] and Module 4M12. So suppose that we look for solutions to (3.19) of the form

$$e^{\mathrm{i}(kx - w(k)t)} \tag{3.21}$$

where k is the wavenumber and w the frequency. Putting this into (3.19) and canceling the common terms gives

$$w = ak + \mu k^3 \tag{3.22}$$

This expression is called the dispersive relation. The speed at which this oscillating wave propagates is clearly

$$c_p(k) = w(k)/k = a + \mu k^2 \tag{3.23}$$

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

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

$$c_g(k) = \frac{dw}{dk} = a + 3\mu k^2 \tag{3.24}$$

This varies even more substantially with k than $c_p(k)$. The importance of the group velocity is discussed in Whitham and in Module 4M12.

A step function, such as the initial data we use, has a broad Fourier spectrum. As time evolves these highly oscillatory components disperse, leading to an oscillatory solution as has been observed in the numerical solution obtained using Lax-Wendroff. The modified equation for Lax-Wendroff is of the form with

[1] Whitham, Linear and nonlinear waves, 1974

$$\mu = \frac{1}{6}\Delta x^2 a(\sigma^2 - 1) \tag{3.25}$$

where $\sigma = a\Delta t/\Delta x$ is the CFL number. Since $a > 0$ and $|\sigma| < 1$ for stability, we have $\mu < 0$, and hence $c_g(k) < a$. All wave number travel too slowly, leading to an oscillatory wave train lagging behind the discontinuity in the true solution, as seen in the figure.

The second-order BeamWarming method has a modified equation similar to that of the Lax-Wendroff method,

$$\mu = \frac{\Delta x^2}{6} a \left(\sigma^2 - 3\sigma + 2\right) \tag{3.26}$$

Note that the 2 roots of $\sigma^2 - 3\sigma + 2 = 0$ are 1 and 2. In this case the group velocity is greater than a for all wave numbers in the case $0 < \sigma < 1$, so that the oscillations move ahead of the main hump. This can be observed in Fig. 1, where $\sigma = 0.8$ was used. If $1 < \sigma < 2$, then the group velocity is less than and the oscillations will fall behind.

4 Conservation Laws

The simplest example of a one-dimensional conservation law is the partial differential equation (PDE):

$$q_t + [f(q)]_x = 0, \tag{4.27}$$

where $q(x,t)$ is a vector of m conserved quantities, and $f(q)$ the flux function. Rewriting this in the quasilinear form

$$q_t + f'(q)q_x = 0, \tag{4.28}$$

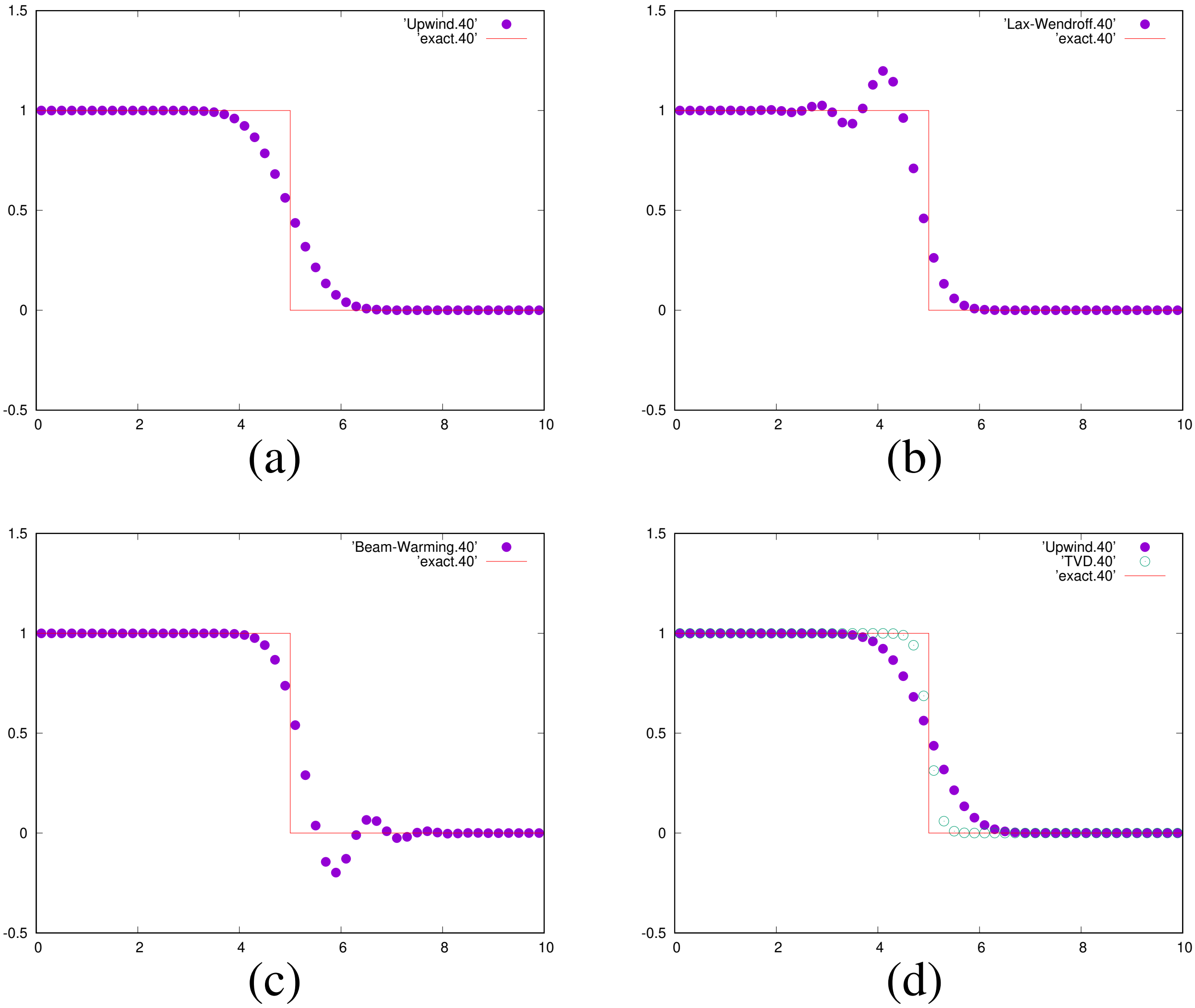


Figure 1: Comparison of four schemes on the linear convection equation for propagating discontinuity: (a) upwind method, (b) Lax-Wendroff method, (c) Beam-Warming method and (d) TVD methods with superbee and minimod limiters.

where the Jacobian matrix $f'(q)$ satisfies certain conditions. For example, the one-dimensional Euler equations are

$$q = \begin{bmatrix} \rho \\ \rho u \\ \rho E \end{bmatrix}, \quad f(q) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (\rho E + p)u \end{bmatrix}. \tag{4.29}$$

To develop high-resolution methods for the Euler equations, one can start from a one-dimensional scalar linear advection equation and extend the method in the following steps:

1. The first order upwind method for one-dimensional scalar equation.
2. Second order methods for one-dimensional scalar equation.
3. High order (TVD) methods for one-dimensional scalar equation.
4. One-dimensional linear hyperbolic system.
5. One-dimensional nonlinear hyperbolic system (Euler equation).
6. Two-dimensional nonlinear hyperbolic system (two-dimensional Euler equation) on Cartesian meshes using *directional splitting operators*.
7. Two-dimensional nonlinear hyperbolic system (two-dimensional Euler equation) on curvilinear meshes.

Module 4A2 is just a course for beginners of CFD. It is a launching platform from which you can progress to more advanced concepts which constitute the essence of modern algorithms in CFD. It is well beyond the scope of this course to present in detail such advanced topics. We will cover only the first 3 items of the above 7 extensions.

5 Finite Volume Methods

In one space dimension, a finite volume method is based on subdividing the spatial domain into intervals (the finite volumes, also called grid cells) and keeping track of an approximation to the integral of q over each of these volumes. In each time step we update these values using approximations to the flux through the endpoints of the intervals.

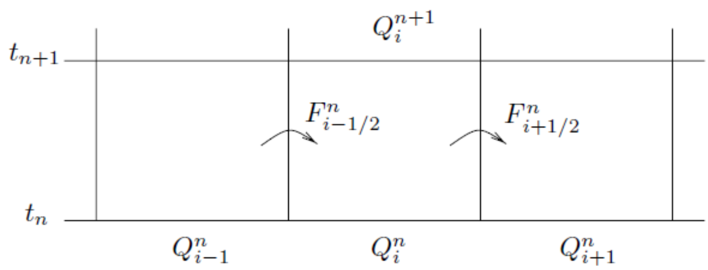


Figure 2: Illustration of a finite volume method for updating the cell average Q_i^n by fluxes at the cell edges. Shown in $x-t$ space. Courtesy of LeVeque R. J. 2002.

Denote the ith grid cell by

$$\mathcal{C}_i = (x_{i-1/2}, x_{i+1/2}) \tag{5.30}$$

as shown in Fig. 2. The value Q_i^n will approximate the average value over the ith interval at time t_n:

$$Q_i^n \approx \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x,t_n)dx \equiv \frac{1}{\Delta x}\int_{\mathcal{C}_i} q(x,t_n)dx, \tag{5.31}$$

where$\Delta x = x_{i+1/2} - x_{i-1/2}$ is the length of the cell. For simplicity we will generally assume a uniform grid, but this is not required.

The integral form of the conservation law gives

$$\frac{d}{dt}\int_{\mathcal{C}_i} q(x,t)dx = f(q(x_{i-1/2},t)) - f(q(x_{i+1/2},t)) \tag{5.32}$$

Integrating (5.32) in time from t_n to t_{n+1} yields

$$\begin{aligned} &\int_{\mathcal{C}_i} q(x,t_{n+1})dx - \int_{\mathcal{C}_i} q(x,t_n)dx \\ &= \int_{t_n}^{t_{n+1}} f(q(x_{i-1/2},t))dt - \int_{t_n}^{t_{n+1}} f(q(x_{i+1/2},t))dt \end{aligned}$$

Rearranging this and dividing by Δx gives

$$\begin{aligned} \frac{1}{\Delta x}\int_{\mathcal{C}_i} q(x,t_{n+1})dx = \frac{1}{\Delta x}\int_{\mathcal{C}_i} q(x,t_n)dx - \frac{1}{\Delta x} \\ -\left[\int_{t_n}^{t_{n+1}} f(q(x_{i+1/2},t))dt - \int_{t_n}^{t_{n+1}} f(q(x_{i-1/2},t))dt\right] \end{aligned} \tag{5.33}$$

This does suggest that we should study numerical methods of the form

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+1/2}^n - F_{i-1/2}^n\right), \tag{5.34}$$

where $F_{i-1/2}^n$ is some approximation to the average flux along $x = x_{i-1/2}$:

$$F_{i-1/2}^n \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(q(x_{i-1/2}, t)dt. \tag{5.35}$$

If we can approximate this average flux based on the values Q^n, then we will have a fully discrete method. See Fig. 2 for a schematic of this process.

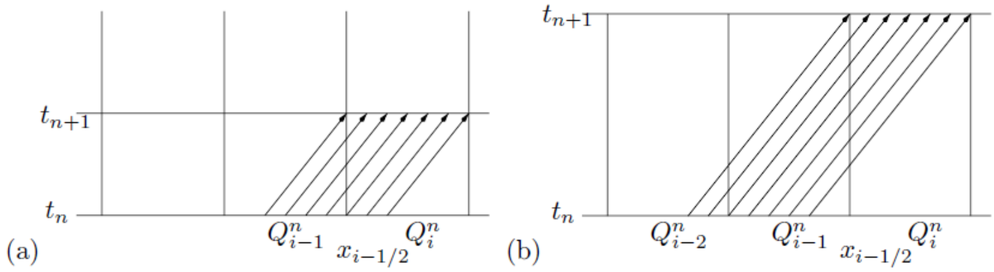


Figure 3: Characteristics for the advection equation, showing the information that flows into cell $\mathcal{C}_i$ during a single time step. (a) For a small enough time step, the flux at $x_{i-1/2}$ depends only on the values in the neighboring cells - only on Q_{i-1}^n in this case where $\bar{u} > 0$. (b) For a larger time step, the flux should depend on values farther away. Courtesy of LeVeque R. J. 2002.

5.1 The Upwind Method for Advection

For the constant-coefficient advection equation $q_t + \bar{u}q_x = 0$ Fig. 3 (a) indicates that the flux through the left edge of the cell is entirely determined by the value Q_{i-1}^n in the cell to the left of this cell. This suggests defining the numerical flux as

$$F_{i-1/2}^n = \bar{u}Q_{i-1}^n \tag{5.36}$$

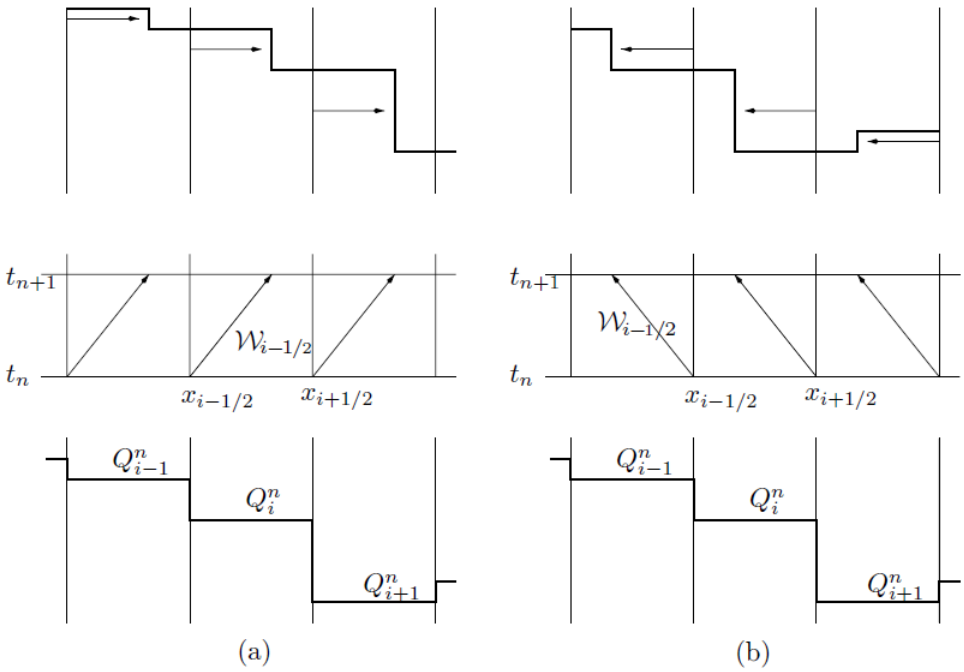


Figure 4: Wave-propagation interpretation of the upwind method for advection. The bottom pair of graphs shows data at time t_n , represented as a piecewise constant function. Over time Δt this function shifts by a distance $\bar{u}\Delta t$ as indicated in the middle pair of graphs. We view the discontinuity that originates at $x_{i-1/2}$ as a wave $\mathcal{W}_{i-1/2}$. The top pair shows the piecewise constant function at the end of the time step after advecting. The new cell averages Q_i^{n+1} in each cell are then computed by averaging this function over each cell. (a) shows a case with $\bar{u} > 0$, while (b) shows $\bar{u} < 0$. Courtesy of LeVeque R. J. 2002.

This leads to the standard first-order upwind method for the advection equation,

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\Delta t}{\Delta x}(Q_i^n - Q_{i-1}^n). \tag{5.37}$$

Note that this can be rewritten as

$$\frac{Q_i^{n+1} - Q_i^n}{\Delta t} + \bar{u}\left(\frac{Q_i^n - Q_{i-1}^n}{\Delta x}\right) = 0, \tag{5.38}$$

We are primarily interested in finite volume methods, and so other interpretations of the upwind method is valuable. Fig. 4 show a geometric viewpoint. We approximate

q as a constant function within each cell at time t_n. This defines a piecewise constant function at time t^n with the value Q_i^n in cell $\mathcal{C}_i$. As time evolves, this piecewise constant function advects to the right with velocity $\bar{u}$, and the jump between states Q_{i-1}^n and Q_i^n shifts a distance $\bar{u}\Delta t$ into cell $\mathcal{C}_i$. At the end of the time step we compute a new cell average Q_i^{n+1} in order to repeat this process. To compute Q_i^{n+1} we must average the piecewise constant function shown in the top of Fig. 4 over the cell. This results in a convex combination of Q_{i-1}^n and Q_i^n (i.e., the weights are both nonnegative and sum to 1):

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

This is simply the upwind method, since a rearrangement gives (5.37).

Above, the upwind method was derived as a special case of the approach referred to as the **REA algorithm**, for *reconstruct-evolve-average*. This is indeed the famous Godunov method named after its inventor. These are one-word summaries of the three steps involved.

Algorithm (REA)

1. Reconstruct a piecewise polynomial function $\tilde{q}^n(x, t_n)$ defined for all x, from the cell averages Q_i^n. In the simplest case this is a piecewise constant function that takes the value Q_i^n in the ith grid cell, i.e.,

$$\tilde{q}^n(x, t_n) = Q_i^n \quad \text{for all} \quad x \in \mathcal{C}_i.$$

2. Evolve the equation exactly (or approximately) with this initial data to obtain $\tilde{q}^n(x, t_{n+1})$a time Δt later.

3. Average this function over each grid cell to obtain new cell averages

$$Q_i^{n+1} = \frac{1}{\Delta x} \int_{\mathcal{C}_i} \tilde{q}^n(x, t_{n+1}) dx.$$

This whole process is then repeated in the next time step.

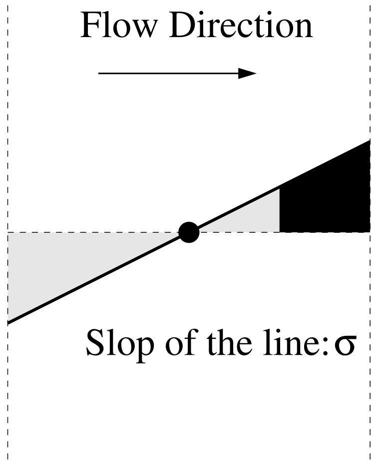


Figure 5: Piecewise linear reconstruction and correction of the flux: dark shaded area flows to the right cell and light shaded area remains in the same cell.

5.2 The REA Algorithm with Piecewise Linear Reconstruction

In the previous example, we derived the upwind method by reconstructing a piecewise constant function $\tilde{q}^n(x, t_n)$ from the cell averages Q_i^n. To achieve better than first-order accuracy, we must use a better reconstruction than a piecewise constant function. From the cell averages Q_i^n we can construct a piecewise linear function of the form

$$\tilde{q}^n(x, t_n) = Q_i^n + \sigma_i^n(x - x_i) \quad \text{for} \quad x_{i-1/2} \le x \le x_{i+1/2}, \tag{5.39}$$

where

$$x_i = \frac{1}{2}(x_{i-1/2} + x_{i+1/2}) = x_{i-1/2} + \frac{1}{2}\Delta x \tag{5.40}$$

is the center of the ith grid cell and σ_i^n is the slope on the ith cell. The linear function (5.39) on the ith cell is defined in such away that its value at the cell center x_i is Q_i^n. More importantly, the average value of $\tilde{q}^n(x, t_n)$ over cell $\mathcal{C}_i$ is Q_i^n (regardless of the slope σ_i^n), so that the reconstructed function has the cell average Q_i^n. This is crucial in developing conservative methods for conservation laws. Note that steps 2 and 3 are conservative in general, and so Algorithm (REA) is conservative provided we use a conservative reconstruction in step 1, as we have in (5.39).

For the scalar advection equation $q_t + \bar{u} q_x = 0$, we can easily solve the equation with this data, and compute the new cell averages as required in step 3 of Algorithm (REA). We have

$$\tilde{q}^n(x, t_{n+1}) = \tilde{q}^n(x - \bar{u}\Delta t, t_n) \ .$$

Until further notice we will assume that $\bar{u} > 0$ and present the formulas for this particular case. The corresponding formulas for $\bar{u} < 0$ should be easy to derive, and we will see a better way to formulate the methods in the general case. Suppose also that $|\bar{u}\Delta t/\Delta x| \leq 1$, as is required by the CFL condition. Then it is straightforward to compute that (see Fig. 5)

$$\begin{aligned} Q_i^{n+1} &= \frac{\bar{u}\Delta t}{\Delta x}\left(Q_{i-1}^n + \frac{1}{2}(\Delta x - \bar{u}\Delta t)\sigma_{i-1}^n\right) + \left(1 - \frac{\bar{u}\Delta t}{\Delta x}\right)\left(Q_i^n - \frac{1}{2}\bar{u}\Delta t \sigma_i^n\right) \\ &= Q_i^n - \frac{\bar{u}\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right) - \frac{1}{2}\frac{\bar{u}\Delta t}{\Delta x}(\Delta x - \bar{u}\Delta t)\left(\sigma_i^n - \sigma_{i-1}^n\right) \ . \end{aligned} \tag{5.41}$$

This is the upwind method with a correction term that depends on the slopes. See Fig. 6.

5.3 Choice of Slopes

Choosing slopes $\sigma_i^n = 0$ gives the upwind method for the advection equation, since the final term in (5.41) drops out. To obtain a second-order accurate method we want to choose nonzero slopes in such a way that σ_i^n approximates the derivative q_x over the ith grid cell. Three obvious possibilities are

$$\text{Centred slope}: \quad \sigma_i^n = \frac{Q_{i+1}^n - Q_{i-1}^n}{2\Delta x} \quad \text{(Fromm)}, \tag{5.42}$$

$$\text{Upwind slope}: \quad \sigma_i^n = \frac{Q_i^n - Q_{i-1}^n}{\Delta x} \quad \text{(Beam – Warming)}, \tag{5.43}$$

$$\text{Downwind slope}: \quad \sigma_i^n = \frac{Q_{i+1}^n - Q_i^n}{\Delta x} \quad \text{(Lax – Wendroff)}. \tag{5.44}$$

The centered slope might seem like the most natural choice to obtain second-order accuracy, but in fact all three choices give the same formal order of accuracy, and it is the other two choices that give methods we have already derived using the Taylor series expansion. Only the downwind slope results in a centered three-point method, and this

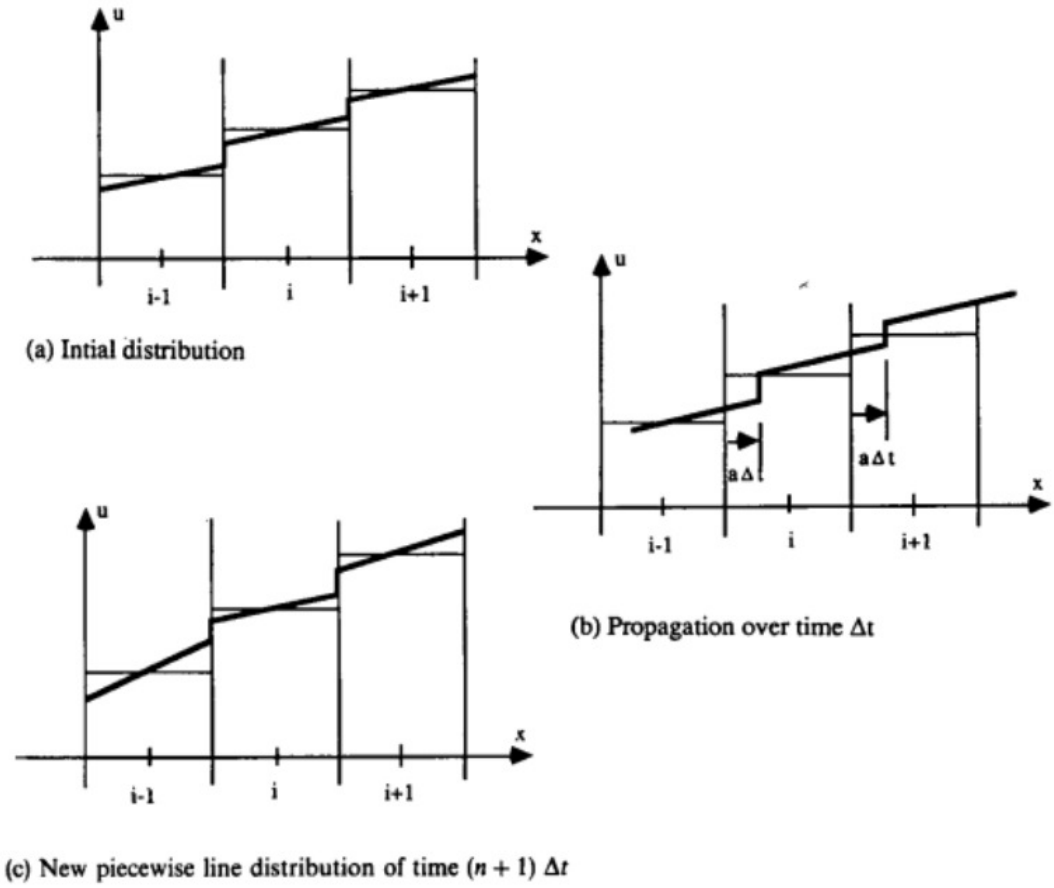


Figure 6: Second-order Godunov-type scheme (REA algorithm) for the linear convection equation. Courtesy of Hirsch C. 1988-1990.

choice gives the Lax-Wendroff method. The upwind slope gives a fully-upwinded 3-point method, which is simply the Beam-Warming method.

The centered slope (5.42) may seem the most symmetric choice at first glance, but because the reconstructed function is then advected in the positive direction, the final updating formula turns out to be a nonsymmetric four-point formula. This method is known as Fromm's method.

To compare the typical behavior of the upwind and Lax-Wendroff methods, Fig. 7 shows numerical solutions to the scalar advection equation $q_t + q_x = 0$, which is solved on the unit interval up to time $t = 1$ with periodic boundary conditions. Hence the solution should agree with the initial data, translated back to the initial location. The data, shown as a solid line in each plot, consists of both a smooth pulse and a square-

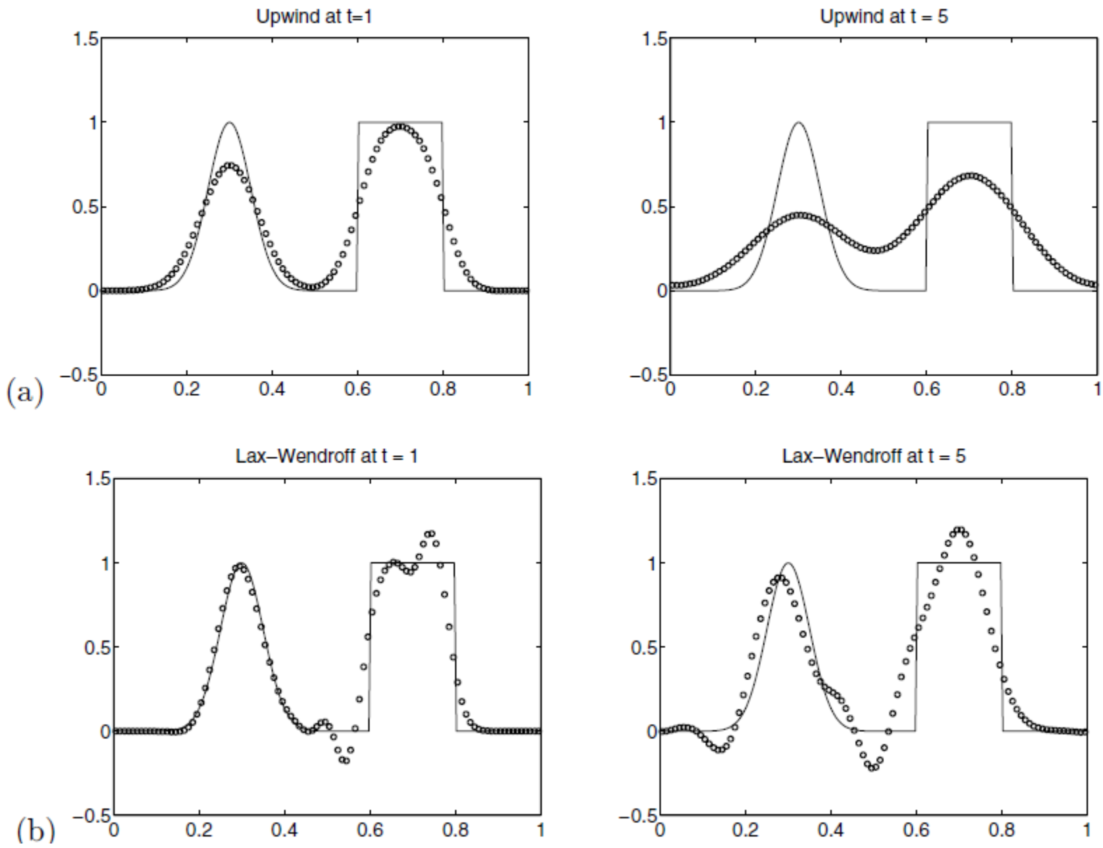


Figure 7: Tests on the advection equation with different linear methods. Results at time $t = 1$ and $t = 5$ are shown, corresponding to 1 and 5 revolutions through the domain in which the equation $q_t + q_x = 0$ is solved with periodic boundary conditions: (a) upwind, (b) Lax-Wendroff. Courtesy of LeVeque R. J. 2002.

wave pulse. Fig. 7(a) shows the results when the upwind method is used. Excessive dissipation of the solution is evident. Fig. 7(b) shows the results when the LaxWendroff method is used instead. The smooth pulse is captured much better, but the square wave gives rise to an oscillatory solution.

5.4 Oscillations

Second-order methods such as the Lax-Wendroff or Beam-Warming (and also Fromms method) give oscillatory approximations to discontinuous solutions. This can be easily understood using the interpretation of Algorithm (REA).

Consider the Lax-Wendroff method, for example, applied to piecewise constant data

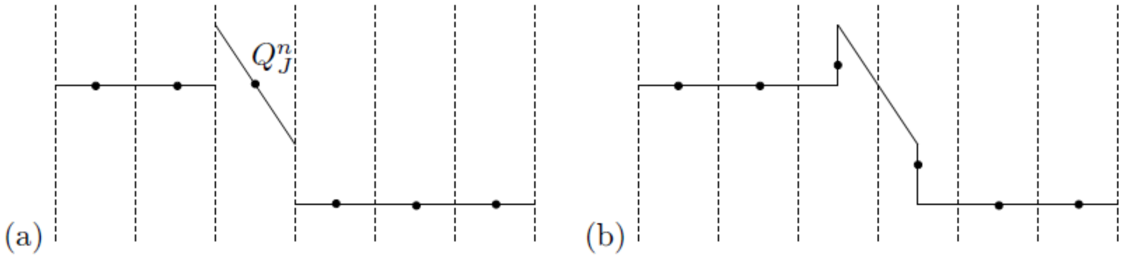


Figure 8: (a)Grid values Q^n and reconstructed $\tilde{q}^n(., t_n)$ using Lax-Wendroff slopes. (b)After advection with $\bar{u}\Delta t = \Delta x/2$. The dots show the new cell averages Q^{n+1}. Note the overshoot. Courtesy of LeVeque R. J. 2002.

with values

$$Q_i^n = \begin{cases} 1 & \text{if} \quad i \leq J, \\ 0 & \text{if} \quad i > J. \end{cases}$$

Choosing slopes in each grid cell based on the Lax-Wendroff prescription (5.44) gives the piecewise linear function shown in Fig. 8(a). The slope σ_i^n is nonzero only for $i = J$.

The function $\tilde{q}^n(x, t_n)$ has an overshoot with a maximum value of 1.5 regardless of Δx. When we advect this profile a distance $\bar{u}\Delta t$, and then compute the average over the Jth cell, we will get a value that is greater than 1 for any Δt with $0 < \bar{u}\Delta t < \Delta x$. The worst case is when $\bar{u}\Delta t = \Delta x/2$, in which case $\tilde{q}^n(x, t_{n+1})$ is shown in Fig 8(b) and $Q_J^{n+1} = 1.125$. In the next time step this overshoot will be accentuated, while in cell $J-1$ we will now have a positive slope, leading to a value Q_{J-1}^{n+1} that is less than 1. This oscillation then grows with time.

The slopes proposed in the previous section were based on the assumption that the solution is smooth. Near a discontinuity there is no reason to believe that introducing this slope will improve the accuracy. On the contrary, if one of our goals is to avoid nonphysical oscillations, then in the above example we must set the slope to zero in the Jth cell. Any $\sigma_J^n < 0$ will lead to $Q_J^{n+1} > 1$, while a positive slope wouldn't make much sense. On the other hand we don't want to set all slopes to zero all the time, or we simply have the first-order upwind method. Where the solution is smooth we want second-order accuracy. Moreover, we will see below that even near a discontinuity, once the solution is somewhat smeared out over more than one cell, introducing nonzero slopes can help keep the solution from smearing out too far, and hence will significantly increase the resolution and keep discontinuities fairly sharp, as long as care is taken to avoid oscillations.

This suggests that we must pay attention to how the solution is behaving near the ith cell in choosing our formula for σ_i^n. (And hence the resulting updating formula will be nonlinear even for the linear advection equation). Where the solution is smooth, we want to choose something like the Lax-Wendroff slope. Near a discontinuity we may want to limit this slope, using a value that is smaller in magnitude in order to avoid oscillations. Methods based on this idea are known as slope-limiter methods. This approach was introduced by van Leer in a series of papers where he developed the approach known as MUSCL (monotonic upstream-centered scheme for conservation laws) for nonlinear conservation laws.

5.5 Total Variation

How much should we limit the slope? Ideally we would like to have a mathematical prescription that will allow us to use the Lax-Wendroff slope whenever possible, for second-order accuracy, while guaranteeing that no nonphysical oscillations will arise. To achieve this we need a way to measure oscillations in the solution. This is provided by the notion of the total variation of a function. For a grid function Q we define

$$\mathrm{TV}(Q) = \sum_{i=-\infty}^{\infty} |Q_i - Q_{i-1}|. \tag{5.45}$$

For an arbitrary function $q(x)$ we can define

$$\mathrm{TV}(q) = \sup \sum_{j=1}^{N} |q(\xi_j) - q(\xi)_{j-1}|, \tag{5.46}$$

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

Definition A two-level method is called total variation diminishing (TVD) if, for any set of data Q^n, the values Q^{n+1} computed by the method satisfy

$$\mathrm{TV}\left(Q^{n+1}\right) \leq \mathrm{TV}\left(Q^n\right). \tag{5.47}$$

If a method is TVD, then in particular data that is initially monotone, say

$$Q_i^n \geq Q_{i+1}^n \quad \text{for all} \quad i,$$

will remain monotone in all future time steps. Hence if we discretize a single propagating discontinuity (as in Fig. 8), the discontinuity may become smeared in future time steps but cannot become oscillatory. This property is especially useful, and we make the following definition.

Definition A method is called monotonicity-preserving if

$$Q_i^n \geq Q_{i+1}^n \quad \text{for all} \quad i,$$

implies that

$$Q_i^{n+1} \geq Q_{i+1}^{n+1} \quad \text{for all} \quad i.$$

Any TVD method is monotonicity-preserving.

5.6 TVD Methods Based on the REA Algorithm

How can we derive a method that is TVD? One easy way follows from the reconstruct-evolve-average approach to deriving methods described by Algorithm (REA). Suppose that we perform the reconstruction in such a way that

$$\mathrm{TV}(\tilde{q}^n(.,t_n)) \leq \mathrm{TV}(Q^n). \tag{5.48}$$

Then the method will be TVD. The reason is that the evolving and averaging steps cannot possibly increase the total variation, and so it is only the reconstruction that we need to worry about.

In the evolve step we clearly have

$$\mathrm{TV}(\tilde{q}^n(.,t_{n+1})) = \mathrm{TV}(\tilde{q}^n(.,t_n)) \tag{5.49}$$

for the advection equation, since $\tilde{q}^n$ simply advects without changing shape. The total variation turns out to be a very useful concept in studying nonlinear problems as well, because a wide class of nonlinear scalar conservation laws also have this property, that the true solution has a non-increasing total variation.

It is a simple exercise to show that the averaging step gives

$$\mathrm{TV}(Q^{n+1}) \leq \mathrm{TV}(\tilde{q}^n(.,t_{n+1})). \tag{5.50}$$

Combining (5.48), (5.49) and (5.50) then shows that the method is TVD.

5.7 Slope-Limiter Methods

Setting $\sigma_i^n = 0$, the first-order upwind method is TVD for the advection equation. The upwind method may smear solutions but cannot introduce oscillations.

One choice of slope that gives second-order accuracy for smooth solutions while still satisfying the TVD property is the minmod slope (Fig. 9 (a)):

$$\sigma_i^n = \mathrm{minmod}\left(\frac{Q_i^n - Q_{i-1}^n}{\Delta x}, \frac{Q_{i+1}^n - Q_i^n}{\Delta x}\right), \tag{5.51}$$

where the minmod function of two arguments is defined by

$$\mathrm{minmod}(a,b) = \begin{cases} a \ \text{ if } \ |a| < |b| \ \text{ and } \ ab > 0, \\ b \ \text{ if } \ |b| < |a| \ \text{ and } \ ab > 0, \\ 0 \qquad\qquad \text{if } \ ab < 0. \end{cases} \tag{5.52}$$

If a and b have the same sign, then this selects the one that is smaller in modulus, else it returns zero.

Rather than defining the slope on the ith cell by always using the downwind difference (which would give the Lax-Wendroff method), or by always using the upwind difference (which would give the Beam-Warming method), the minmod method compares the two slopes and chooses the one that is smaller in magnitude. If the two slopes have different sign, then the value Q_i^n must be a local maximum or minimum, and in this case that we must set $\sigma_i^n = 0$.

Fig. 10(a) shows results using the minmod method for the advection problem considered previously. We see that the minmod method does a fairly good job of maintaining

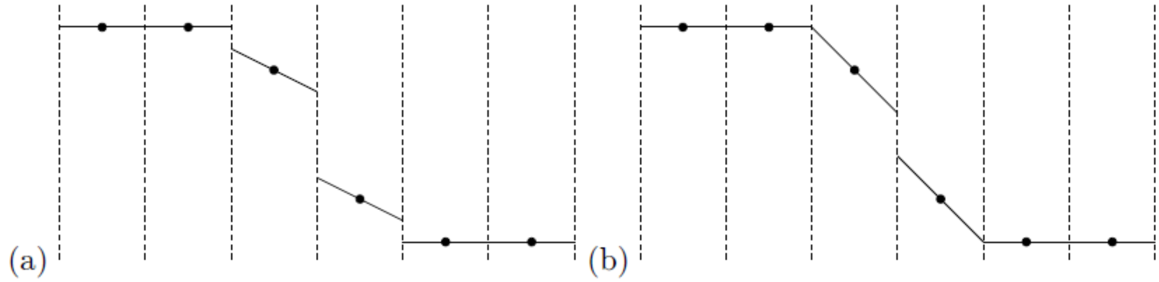


Figure 9: Grid values Q^n and reconstructed $\tilde{q}^n(., t_n)$ using (a) minmod slopes, (b) superbee or MC slopes. Note that these steeper slopes can be used and still have the TVD property. Courtesy of LeVeque R. J. 2002.

good accuracy in the smooth hump and also sharp discontinuities in the square wave, with no oscillations.

Sharper resolution of discontinuities can be achieved with other limiters that do not reduce the slope as severely as minmod near a discontinuity. Fig. 9 (a) shows some sample data representing a discontinuity smeared over two cells, along with the minmod slopes. Fig. 9 (b) shows that we can increase the slopes in these two cells to twice the value of the minmod slopes and still have (5.48) satisfied. This sharper reconstruction will lead to sharper resolution of the discontinuity in the next time step than we would obtain with the minmod slopes.

One choice of limiter that gives the reconstruction of Fig. 9 (b), while still giving second order accuracy for smooth solutions, is the so-called superbee limiter introduced by Roe

$$\sigma_i^n = \text{maxmod}\left(\sigma_i^{(1)}, \sigma_i^{(2)}\right) \tag{5.53}$$

where

$$\sigma_i^{(1)} = \text{minmod}\left(\left(\frac{Q_i^n - Q_{i-1}^n}{\Delta x}\right), 2\left(\frac{Q_{i+1}^n - Q_i^n}{\Delta x}\right)\right),$$

$$\sigma_i^{(2)} = \text{minmod}\left(2\left(\frac{Q_i^n - Q_{i-1}^n}{\Delta x}\right), \left(\frac{Q_{i+1}^n - Q_i^n}{\Delta x}\right)\right).$$

Each one-sided slope is compared with twice the opposite one-sided slope. Then the maxmod function in (5.53) selects the argument with larger modulus. In regions where the solution is smooth this will tend to return the larger of the two one-sided slopes, but will still be giving an approximation to q_x, and hence we expect second-order accuracy. The superbee limiter is also TVD in general.

Fig. 10 (b) shows the same test problem as before but with the superbee method. The discontinuity stays considerably sharper. On the other hand, we see a tendency of the smooth hump to become steeper and squared off. This is sometimes a problem with superbee – by choosing the larger of the neighboring slopes it tends to steepen smooth transitions near inflection points.

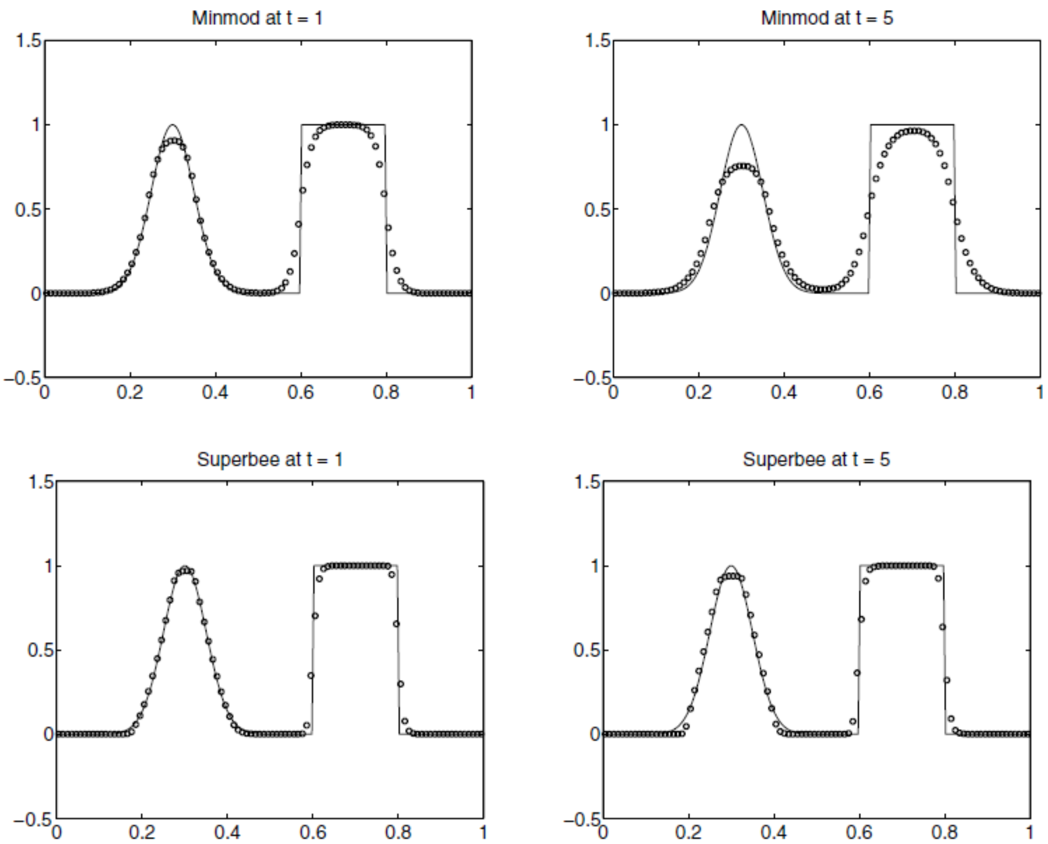
5.8 Flux Formulation with Piecewise Linear Reconstruction

The slope-limiter methods described above can be written as flux-differencing methods of the form (5.34):

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

The updating formulas derived above can be manipulated algebraically to determine what the numerical flux function must be. Alternatively, we can derive the numerical flux by computing the exact flux through the interface $x_{i-1/2}$ using the piecewise linear solution $\tilde{q}^n(x,t)$, by integrating $\bar{u}\tilde{q}^n(x_{i-1/2},t)$ in time from t_n to t_{n+1}. For the advection equation this is easy to do and we find that

$$F_{i-1/2}^n \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(q(x_{i-1/2},t))dt = \bar{u}Q_{i-1}^n + \frac{1}{2}\bar{u}(\Delta x - \bar{u}\Delta t)\sigma_{i-1}^n.$$

(a) 

(b)

Figure 10: Tests on the advection equation with different high-resolution methods, as in Fig 7: (a) minmod limiter, (b) superbee limiter. Courtesy of LeVeque R. J. 2002.

Using this in the flux-differencing formula (5.34) gives

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right) - \frac{1}{2}\frac{\bar{u}\Delta t}{\Delta x}(\Delta x - \bar{u}\Delta t)(\sigma_i^n - \sigma_{i-1}^n).$$

If we also consider the case $\bar{u} < 0$, then we will find that in general the numerical flux for a slope-limiter method is

$$F_{i-1/2}^n = \begin{cases} \bar{u}Q_{i-1}^n + \frac{1}{2}\bar{u}(\Delta x - \bar{u}\Delta t)\sigma_{i-1}^n & \text{if} \quad \bar{u} \geq 0 \\ \bar{u}Q_i^n - \frac{1}{2}\bar{u}(\Delta x + \bar{u}\Delta t)\sigma_i^n & \text{if} \quad \bar{u} \leq 0 \end{cases} \tag{5.54}$$

or

$$F^n_{i-1/2} = \begin{cases} \bar{u}Q^n_{i-1} + \frac{1}{2}|\bar{u}|(\Delta x - |\bar{u}|\Delta t)\sigma^n_{i-1} & \text{if} \quad \bar{u} \geq 0 \\ \bar{u}Q^n_i + \frac{1}{2}|\bar{u}|(\Delta x - |\bar{u}|\Delta t)\sigma^n_i & \text{if} \quad \bar{u} \leq 0 \end{cases} \tag{5.55}$$

Rather than associating a slope σ^n_i with the ith cell, the idea of writing the method in terms of fluxes between cells suggests that we should instead associate our approximation to q_x with the cell interface $x_{i-1/2}$ where $F^n_{i-1/2}$ is defined. Across the interface $x_{i-1/2}$ we have a jump

$$\Delta Q^n_{i-1/2} = Q^n_i - Q^n_{i-1} \tag{5.56}$$

and this jump divided by Δx gives an approximation to q_x . This suggests that we write the flux (5.55) as

$$F^n_{i-1/2} = \bar{u}^- Q^n_i + \bar{u}^+ Q^n_{i-1} + \frac{1}{2}|\bar{u}| \left(1 - \frac{|\bar{u}|\Delta t}{\Delta x}\right) \delta^n_{i-1/2} \tag{5.57}$$

where $\bar{u}^- = \min(\bar{u}, 0)$, $\bar{u}^+ = \max(\bar{u}, 0)$, and

$$\delta^n_{i-1/2} = \text{a limited version of} \ \ \Delta Q^n_{i-1/2}. \tag{5.58}$$

If $\delta^n_{i-1/2}$ is the jump $\Delta Q^n_{i-1/2}$ itself, then (5.57) gives the Lax-Wendroff method. From the form (5.57), we see that the Lax-Wendroff flux can be interpreted as a modification to the upwind flux (5.36). By limiting this modification we obtain a different form of the high-resolution methods.

5.9 Flux Limiters

From the above discussion it is natural to view the Lax-Wendroff method as the basic second-order method based on piecewise linear reconstruction, since defining the jump $\delta^n_{i-1/2}$ in (5.58) in the most obvious way as $\Delta Q^n_{i-1/2}$ at the interface $x_{i-1/2}$ results in that method. Other second-order methods have fluxes of the form (5.57) with different choices of $\delta^n_{i-1/2}$. The slope-limiter methods can then be reinterpreted as flux-limiter methods by choosing $\delta^n_{i-1/2}$ to be a limited version of $\Delta Q^n_{i-1/2}$ (5.58). In general we will set

$$\delta^n_{i-1/2} = \phi(\theta^n_{i-1/2})\Delta Q^n_{i-1/2} \tag{5.59}$$

where

$$\theta^n_{i-1/2} = \frac{\Delta Q^n_{I-1/2}}{\Delta Q^n_{i-1/2}}. \tag{5.60}$$

The index I here is used to represent the interface on the upwind side of $x_{i-1/2}$:

$$I = \begin{cases} i-1 & \text{if } \ \bar{u} > 0, \\ i+1 & \text{if } \ \bar{u} < 0. \end{cases} \tag{5.61}$$

The ratio $\theta^n_{i-1/2}$ can be thought of as a measure of the smoothness of the data near $x_{i-1/2}$. Where the data is smooth we expect $\theta^n_{i-1/2} \approx 1$ (except at extrema). Near a discontinuity we expect that $\theta^n_{i-1/2}$ may be far from 1

The function $\phi(\theta)$ is the flux-limiter function, whose value depends on the smoothness. Setting $\phi(\theta) = 1$ for all θ gives the Lax-Wendroff method, while setting $\phi(\theta) = 0$ gives upwind. More generally we might want to devise a limiter function ϕ that has values near 1 for $\theta \approx 1$, but that reduces (or perhaps increases) the slope where the data is not smooth.

There are many other ways one might choose to measure the smoothness of the data besides the variable θ defined in (5.60). However, the framework proposed above results in very simple formulas for the function ϕ corresponding to many standard methods, including all the methods discussed so far.

In particular, note the nice feature that choosing

$$\phi(\theta) = \theta$$

results in the Beam-Warming method. We also find that Fromms method can be obtained by choosing

$$\phi(\theta) = \frac{1}{2}(1 + \theta).$$

In summary,

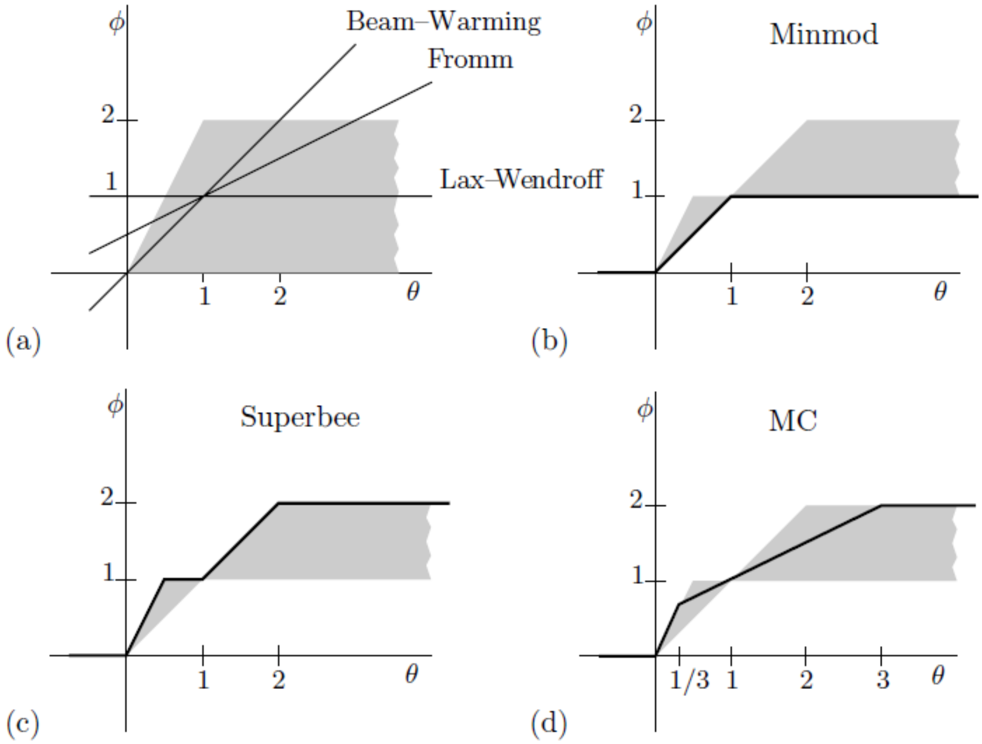


Figure 11: Limiter functions $\phi(\theta)$. (a) The shaded regions shows where function values must lie for the method to be TVD. The second-order linear methods have functions $\phi(\theta)$ that leave this region. (b) The shaded region is the Sweby region of second-order TVD methods. The minmod limiter lies along the lower boundary. (c) The superbee limiter lies along the upper boundary. (d) The MC limiter is smooth at $\phi = 1$. Courtesy of LeVeque R. J. 2002.

Linear methods:

$$\begin{aligned} \text{upwind}: &\quad \phi(\theta) = 0, \\ \text{Lax}-\text{Wendroff}: &\quad \phi(\theta) = 1, \\ \text{Beam}-\text{Warming}: &\quad \phi(\theta) = \theta, \\ \text{Fromm}: &\quad \tfrac{1}{2}(1+\theta). \end{aligned} \tag{5.62}$$

High-resolution limiters (TVD methods):

$$
\begin{aligned}
\text{minmod :} \quad & \phi(\theta) = \text{minmod}(1, \theta), \\
\text{superbee :} \quad & \phi(\theta) = \max(0, \min(1, 2\theta), \min(2, \theta), \\
\text{MC :} \quad & \phi(\theta) = \max(0, \min(1+\theta)/2, 2, 2\theta) \\
\text{van Leer :} \quad & \phi(\theta) = \tfrac{\theta+|\theta|}{1+|\theta|}.
\end{aligned} \tag{5.63}
$$

A wide variety of other limiters have also been proposed in the literature. The dispersive nature of the LaxWendroff method also causes a slight shift in the location of the smooth hump, a phase error, that is visible in Fig. 7, particularly at the later time $t = 5$. Another advantage of using limiters is that this phase error can be essentially eliminated. Fig. 12 shows a computational example where the initial data consists of a wave packet, a high-frequency signal modulated by a Gaussian. With a dispersive method such a packet will typically propagate at an incorrect speed corresponding to the numerical group velocity of the method. The LaxWendroff method is clearly quite dispersive. The high-resolution method shown in Fig. 12(c) performs much better. There is some dissipation of the wave, but much less than with the upwind method.

5.10 TVD Limiters

For simple limiters such as minmod, it is clear from the derivation as a slope limiter that the resulting method is TVD, since it is easy to check that (5.48) is satisfied. For more complicated limiters we would like to have an algebraic proof that the resulting method is TVD. A fundamental tool in this direction is the following theorem of Harten, which can be used to derive explicit algebraic conditions on the function ϕ required for a TVD method.

Theorem (Harten)[2] Consider a general method of the form

$$
Q_i^{n+1} = Q_i^n - C_{i-1}^n \left(Q_i^n - Q_{i-1}^n\right) + D_i^n \left(Q_{i+1}^n - Q_i^n\right)
$$

over one time step, where the coefficents C_{i-1}^n and D_i^n are arbitrary values (which in particular may depend on values of Q^n in some way, i.e., the method may be nonlinear. Then

$$
TV(Q^{n+1}) \leq TV(Q^{n+1}) \ ,
$$

[2]Page 116, R. J. LeVeque 2002.

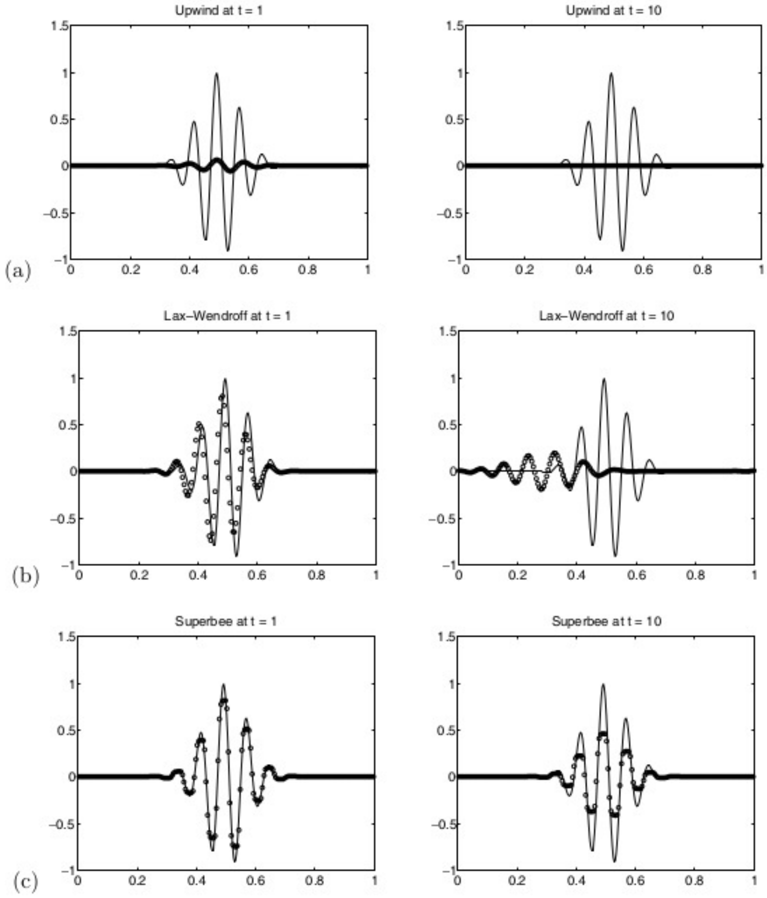


Figure 12: Tests on the advection equation with different methods on a wave packet. Results at time $t = 1$ and $t = 10$ are shown, corresponding to 1 and 10 revolutions through the domain in which the equation $q_t + q_x = 0$ is solved with periodic boundary conditions. Courtesy of LeVeque R. J. 2002.

provided the following conditions are satisfied:

$$C^n_{i-1} \geq 0 \quad \text{for every} \quad i \ ,$$

$$D^n_i \geq 0 \quad \text{for every} \quad i \ ,$$

$$C^n_i + D^n_i \leq 1 \quad \text{for every} \quad i \ .$$

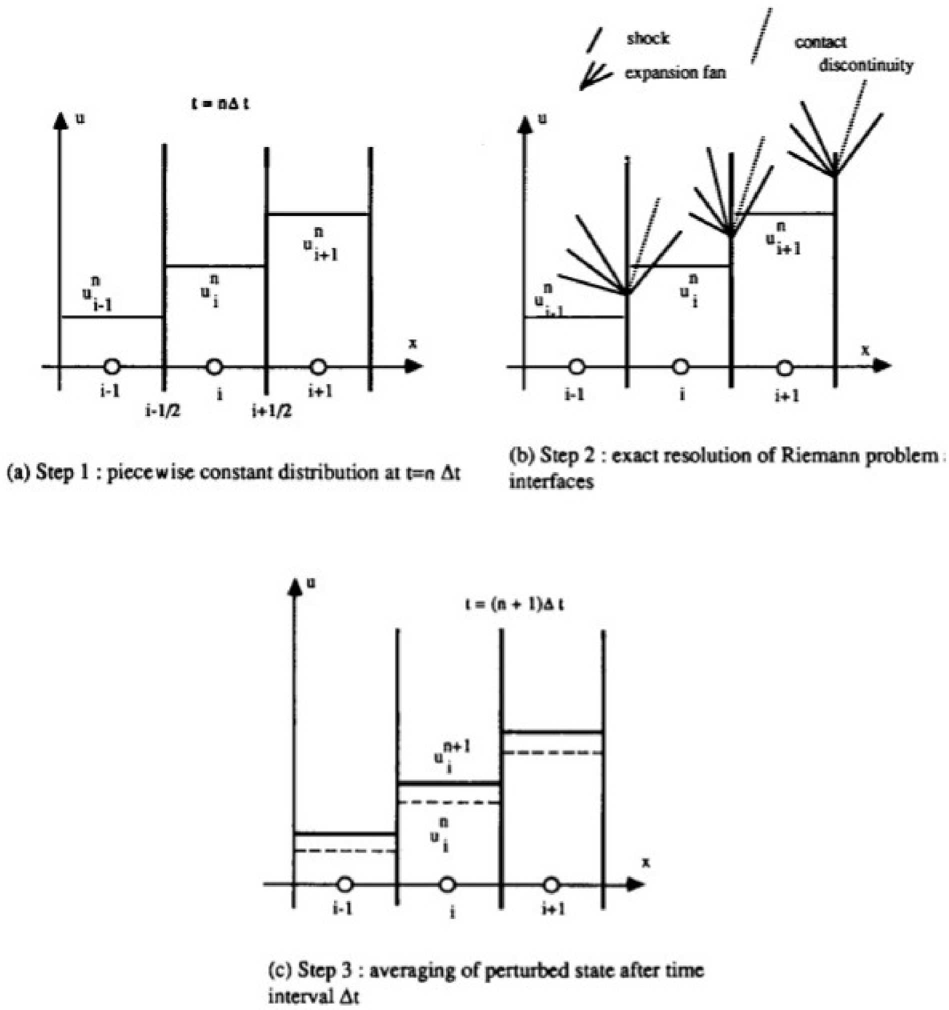


Figure 13: The three basic steps of Godunov's method. Courtesy of Hirsch C. 1988-1990.

6 Godunovs Method for Nonlinear Euler equations

Godunovs method (the REA Algorithm) can be easily generalized to the one-dimensional nonlinear Euler equations (see Fig. 13).