High-Order and TVD methods for scalar Equations

1. Introduction

We try to resolve two contradictory requirements:

- we want high order methods
- but without spurious (unphysical) oscillations.

The class of monotone methods do not produce unphysical oscillations, but are at most first order accurate.

We can describe a class of schemes as

$$u_i^{n+1} = H(u_{i-l_L}^n, \dots, u_{i+l_R}^n)$$

$$= \sum_{k=-l_L}^{l_R} b_k u_{i+k}^n$$

For example, the upwind differencing scheme for linear advection,

$$u_i^{n+1} = u_i + c\frac{\Delta t}{\Delta x}(u_{i-1} - u_i)$$

$$= (1 - c\frac{\Delta t}{\Delta x})u_i + c\frac{\Delta t}{\Delta x}u_{i-1}$$

can be expressed as

$$u_i^{n+1} = \sum_{k=-l_L}^{l_R} b_k u_{i+k}^n$$

with $l_L = 1$, $l_R = 0$

$$= b_{-1} u_{i-1}^n + b_0 u_i^n$$

with $b_{i-1} = \frac{c\Delta t}{\Delta x}$ and $b_0 = (1 - \frac{c\Delta t}{\Delta x})$

Definition : A scheme of the form

$$u_i^{n+1} = H\left(u_{i-l_L}^n, \dots, u_{i+l_R}^n\right)$$

$$= \sum_{k=-l_L}^{l_R} b_k u_{i+k}^n$$

is said to be monotone if all coefficients are positive or zero. An alternate definition is

$$\frac{\partial H}{\partial q_k} \geq 0 \qquad \forall k$$

Monotone methods are at most first order accurate, but don't produce unphysical oscillations. One way of resolving the contradiction between linear schemes of high order accuracy and absence of spurious oscillations is by constructing non-linear methods.

Total variation diminishing (TVD) methods are a prominent class of non-linear methods, which we will concern ourselves with.

The theoretical bases of TVD methods are sound for scalar one-dimensional problems only. The one-dimensional scalar theory serves well as a guideline for extending the ideas on a more or less empirical basis for multidimensional problems.

TVD schemes have their roots on the fundamental question of convergence of schemes for non-linear problems. Total Variation Stable schemes have been proven to be convergent. TVD methods are a class of Total Variation Stable schemes and they are based on the requirement that the total variation of the numerical solution be non-increasing in time. In fact, the exact solution of a scalar conservation law poseses this property and therefore TVD numerical methods just mimic a property of the exact solution.
The TVD concept is also very useful in designing schemes.

TVD schemes are intimately linked to traditional artificial viscosity methods. Both TVD and artificial viscosity methods attempt to circumvent Godunov's theorem by constructing schemes of accuracy $p > 1$ such that spurious oscillations near high gradients are eliminated or controlled. But they both effectively resort to the same mechanism: adding artificial viscosity. Artificial Viscosity methods do this explicitly by adding extra terms to the PDEs. TVD methods the artificial viscosity is inherent in the scheme itself, but the way this is activated is rather sophisticated.

2. Basic Properties of Selected Schemes

Given a numerical method, there are four fundamental properties associated with it:

- consistency
- stability
- convergence
- accuracy

2.1 Accuracy

A scheme of the form

$$u_i^{n+1} = \sum_{k=-k_L}^{k_R} b_k \, u_{i+k}^{n}$$

is p-th order accurate in space and time if and only if

$$\sum_{k=-k_L}^{k_R} k^q b_k = (-c)^q, \quad 0 \le q \le p$$

(Roe's theorem)

Proof of Roe's Theorem on Accuracy:

For the solution of the 1D linear advection equation

$$u_t + a u_x = 0$$

where a is a constant wavespeed.

We consider a general scheme of the form

$$u_i^{n+1} = \sum_{\alpha} A_\alpha u_{i+\alpha}^{n}$$

where $u_i = u(i\Delta x, n\Delta t)$, and $\{A_\alpha\}$ is a finite set of constant, nonzero coefficients.

We set

$$u_\alpha^n = B(\alpha \Delta x)^q$$

Without loss of generality, we only consider the origin $(0, 0)$.

Now we perform one time step:

$$u_0^{n+1} = \sum_\alpha A_\alpha B (\alpha \Delta x)^q$$

meanwhile, the exact solution is

$$u_0^{n+1} = u(-a\Delta t, n\Delta t)$$

$$= B(-a\Delta t)^q \qquad \text{since } u_\alpha^n = B(\alpha\Delta x)^q$$

Therefore the solution is exact if and only if

$$\sum_\alpha A_\alpha B (\alpha \Delta x)^q = B(a\Delta t)^q$$

$$\Rightarrow \text{ If } \sum_\alpha A_\alpha = \left(\frac{-a\Delta t}{\Delta x}\right)^q = (-c)^q$$

the solution is exact, and the method reproduces a polynomial of order q exactly, hence the method is q-th order accurate.

Example: Godunov's upwind method

We consider Godunov's upwind method for linear advection:

$$\partial_t u + \partial_x f(u) = 0 \quad , \quad f(u) = au, \; a \geq 0$$

$$\rightarrow \quad u_i^{n+1} = u_i^n + \frac{a\Delta t}{\Delta x}\left(u_{i-1}^n - u_i^n\right)$$

$$= (1 - c)\, u_i^n + c\, u_{i-1} \quad , \quad c = \frac{a\Delta t}{\Delta x}$$

So we have for the expression

$$u_i^{n+1} = \sum_{k=-k_L}^{k_R} b_k \, u_{i-k}^n$$

$$k_L = 1, \quad k_R = 0$$

$$b_{-1} = c \, , \quad b_0 = 1 - c$$

To show the order of accuracy, we verify that for each integer q with $0 \leq q \leq p$ that $\sum k^q b_k = (-c)^q$ with c being the Courant number.

First for $q = 0$:

$$\sum_{k=-1}^{0} k^0 b_k = (-1)^0 \cdot c + 0^0 (1-c)$$

$$= \lim_{x \to 0} (-1)^x c + x^x (1-c)$$

$$= c + 1 - c = 1 = (-c)^0$$

So this checks out.

Now for $q = 1$:

$$\sum_{k=-1}^{0} k^1 b_k = (-1)^1 \cdot c + 0^1 (1-c)$$

$$= -c = (-c)^1$$

This checks out too.

Now for $q = 2$:

$$\sum_{k=-1}^{0} k^2 b_k = (-1)^2 c + 0^2 (1-c)$$

$$= c \neq (-c)^2$$

So it can't be second order accurate.

2.2 Consistency

A scheme is consistent if

$$\sum_{k=k_L}^{k_R} k^q b_k = (-c)^q$$

for $q = 0$.

Or in other words: The method is consistent if it's at least zeroth order accurate.

2.3 Stability

There are multiple methods to determine whether a numerical scheme is stable, like the von Neumann or Lax-Richtmyer analysis. We won't go into details here.

2.4) Convergence

Naturally, the only useful numerical methods are convergent ones. Unfortunately, it is difficult or impossible to prove, theoretically, the convergence of a particular numerical method.

For linear problems, a most useful result is the Lax Equivalence theorem:

> The only convergent schemes are those that are both consistent and stable.

3 WAF type High Order Schemes

The Weighted Average Flux (WAF) approach is a generalisation of the Lax-Wendroff and the Godunov first-order upwind schemes to the non-linear systems of conservation laws. It leads to fully discrete second order accurate schemes.

3.1 The Basic WAF scheme

We work on the approach as applied to a general scalar conservation law

$$u_t + f(u)_x = 0$$

The scheme is based on the explicit conservative formula

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2}\right]$$

The WAF method assumes piece-wise constant data $\{U_i^n\}$, as in the Godunov first order upwind method, i.e.

$$U_i^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t^n)\, dx$$

In the original presentation of the WAF method, the interface flux was defined as an integral average of the flux function $f(u)$, namely

$$f_{i+1/2}^{waf} = \frac{1}{\Delta x} \int_{-1/2\Delta x}^{1/2\Delta x} f\left[U_{i+1/2}, \tfrac{1}{2}\Delta t\right] dx$$

Here the integration goes from the middle of the cell I_i to the middle of the cell I_{i+1}.

The integrand is the physical flux function $f(u)$ in the conservation law (linear or non-linear) evaluated at $U_{i+1/2}(x, \frac{\Delta t}{2})$, where $U_{i+1/2}(x, t)$ is the solution of the Riemann problem with piece-wise constant data (U_i^n, U_{i+1}^n).

Let us develop the full WAF scheme for linear advection.

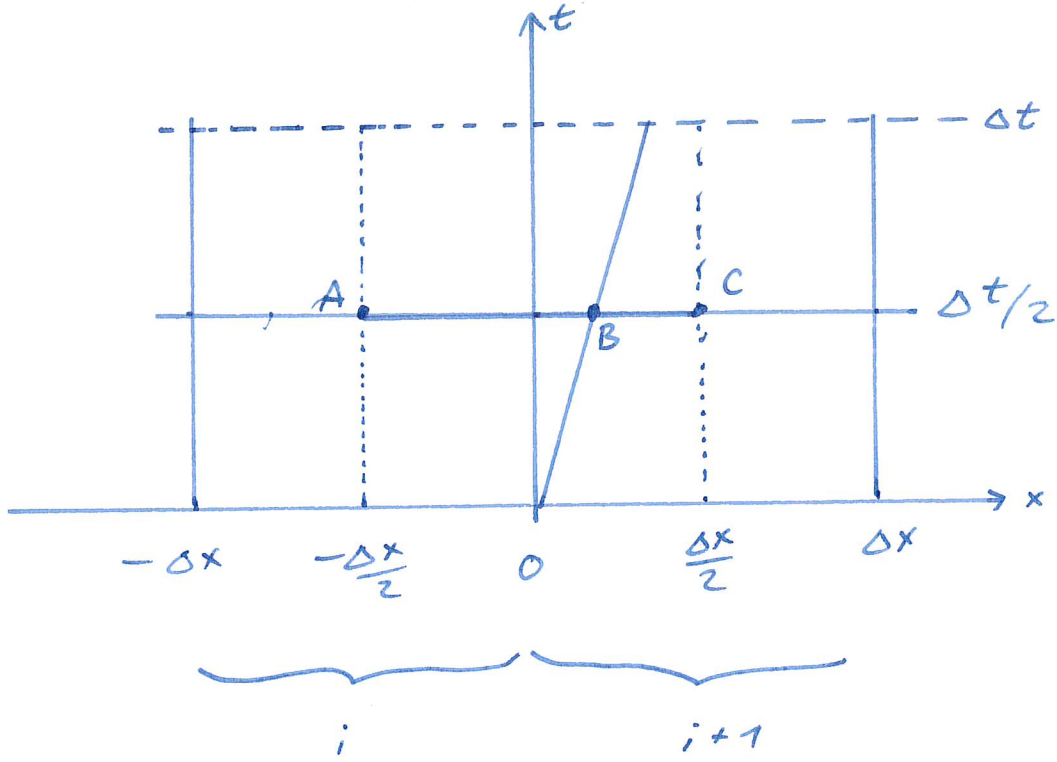
The solution $q_{i+1/2}(x,t)$ of the Riemann problem for the linear advection equation is

$$q_{i+1/2}(x,t) = \begin{cases} q_i^n & \frac{x}{t} \geq a \\ q_{i+1}^n & \frac{x}{t} < a \end{cases}$$

for $a \geq 0$.

The two states in the solution are constant, thus the evaluation of the flux integral becomes trivial.

Suppose the left cell's center is at point A, the right cell's center is at C, and the wave is currently at position B:



Then the integration domain is $[-\frac{\Delta x}{2}, \frac{\Delta x}{2}]$, i.e. the line $\overline{AC}$, which is subdivided into the segments $\overline{AB}$ and $\overline{BC}$ across which the integrals are constant. The lengths of the subdomains are given by

$$\overline{AB} \equiv \beta_1 \Delta x = \frac{\Delta x}{2} + a\frac{\Delta t}{2} = \frac{1}{2}\left(1 + \frac{a\Delta t}{\Delta x}\right)\Delta x$$

$$= \frac{1}{2}(1+c)\Delta x$$

$$\overline{BC} \equiv \beta_2 \Delta x = \frac{\Delta x}{2} - a\frac{\Delta t}{2} = \frac{1}{2}(1-c)\Delta x$$

which gives us

$$f_{i+1/2}^{waf} = \frac{1}{\Delta x} \int_{-\Delta x/2}^{\Delta x_c} f(u_{i+1/2}^{n+1/2})\, dx$$

$$= \frac{1}{\Delta x} \left[\int_A^B f(u_{i+1/2}^{n+1/2})\, dx + \int_B^C f(u_{i+1/2}^{n+1/2})\, dx \right]$$

$$= \frac{1}{\Delta x} \left[\overline{AB}\, f(u_{i+1/2}^{n+1/2}) \Big|_{x=A}^{B} + \overline{BC}\, f(u_{i+1/2}^{n+1/2}) \Big|_{x=B}^{C} \right]$$

$$= \frac{1}{\Delta x} \left[\overline{AB}\, (a u_i^n) + \overline{BC}\, (a u_{i+1}^n) \right]$$

$$= \frac{1}{\Delta x} \left[\frac{1}{2}(1+c)\Delta x (a u_i^n) + \frac{1}{2}(1-c)\Delta x (a u_{i+1}^n) \right]$$

$$= \frac{1}{2}(1+c)(a u_i^n) + \frac{1}{2}(1-c)(a u_{i+1}^n)$$

This numerical flux is identical to that of the Lax-Wendroff (Linear method with downwind slope).

For $a > 0$, the flux is a weighted average of the upwind flux $f_i = a u_i^n$ and downwind flux $f_{i+1} = a u_{i+1}^n$ with weights

$\beta_1 = \frac{1}{2}(1+c)$ and $\beta_2 = \frac{1}{2}(1-c)$,

respectively.

The upwind weight is always larger than the downwind weight and thus the WAF method is upwind biased.

4. MUSCL - Type High-Order methods

The main idea is to modify the piecewise constant data as a first step to achieve higher order of accuracy. This approach has become known as the MUSCL (Monotone Upstream-Centered Scheme for Conservation laws) or Variable Extrapolation approach.

4.1 Data Reconstruction

The simplest way of modifying the piecewise constant data $\{u_i^n\}$ is to replace the constant states u_i^n by piecewise linear functions $\tilde{u}_i^n(x)$. As for the first-order Godunov method, one assumes that u_i^n represents an integral average in cell $I = [x_{i-1/2}, x_{i+1/2}]$ as given by

$$u_i^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t^n)\, dx$$

A piecewise linear local reconstruction of u_i^n is

$$u_i(x) = u_i^n + (x - x_i)\frac{\Delta_i}{\Delta x}, \quad x \in [0, \Delta x]$$

where $\frac{\Delta_i}{\Delta x}$ is a suitably chosen slope of $u_i(x)$ in all I_i. Note that this definition is in local coordinates of the cell, i.e. $x \in [0, \Delta x]$ and the cell center is at $x = \frac{1}{2}\Delta x$.

The values of $u_i(x)$ at the extreme points play a fundamental role. They are given by

$$u_i^L = u_i(0) = u_i^n - \frac{1}{2}\Delta_i$$

$$u_i^R = u_i(\Delta x) = u_i^n + \frac{1}{2}\Delta_i$$

and are usually called boundary extrapolated values.

The integral of $u_i(x)$ in cell I_i is identical to that of u_i^n and thus the reconstruction process retains conservation.

As a consequence of having modified the data, at each interface $i+1/2$ one now may consider the Generalized Riemann Problem

$$\begin{cases} u_t + f(u)_x = 0 \\ u(x,0) = \begin{cases} u_i(x), & x < 0 \\ u_{i+1}(x), & x > 0 \end{cases} \end{cases}$$

to compute an intercell Godunov-Type flux $f_{i+1/2}$. Note that we don't have constant left and right states any longer, but $u_i(x)$ and $u_{i+1}(x)$ instead. The solution no longer contains uniform regions as in the conventional Riemann problem. Wave paths are now curved in $x-t$ space.

For the choice of slope we define

$$\Delta_i = \frac{1}{2}(1+\omega)(u_i^n - u_{i-1}^n) + \frac{1}{2}(1-\omega)(u_{i+1}^n - u_i^n)$$

where $\omega \in [-1, 1]$ is a free parameter.

Some special values for ω:

$\omega = -1$: $\Delta_i = u_{i+1}^n - u_i^n$ downwind slope

$\omega = 0$: $\Delta_i = \frac{u_{i+1}^n - u_{i-1}^n}{2}$ centered slope

$\omega = 1$: $\Delta_i = u_i^n - u_{i-1}^n$ upwind slope

MUSCL data reconstructions that are more accurate are possible, but we won't go into that now.

4.2 The MUSCL-Hancock Method

The MUSCL-Hancock Method (MHM) has three distinct steps to construct fully discrete second-order accurate schemes based on the explicit conservative formula

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2}\right].$$

These are:

1) Data Reconstruction with boundary extrapolated values, i.e.

$$U_i^L = U_i^n - \frac{1}{2}\Delta_i, \qquad U_i^R = U_i^n + \frac{1}{2}\Delta_i$$

2) Evolution of U_i^L, U_i^R by a time $\frac{1}{2}\Delta t$ according to

$$\bar{U}_i^L = U_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(U_i^L) - f(U_i^R)\right]$$

$$\bar{U}_i^R = U_i^R + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(U_i^L) - f(U_i^R)\right]$$

3) Solution of the piece-wise constant data Riemann problem

$$\begin{cases} u_t + f(u)_x = 0 \\ u(x,0) = \begin{cases} \bar{u}_i^R, & x < 0 \\ \bar{u}_{i+1}^L, & x > 0 \end{cases} \end{cases}$$

to find the similarity solution $u_{i+1/2}(x/t)$.

The evolved states $\bar{u}_i^R$ and $\bar{u}_{i+1}^L$ form the piece-wise constant data for a conventional Riemann problem at the cell interface $i+1/2$ with the solution $u_{i+1/2}(x/t)$.

The interface numerical flux $f_{i+1/2}$ is then obtained in exactly the same way as in the Godunov first-order upwind method, i.e.

$$f_{i+1/2} = f(u_{i+1/2}(0))$$

For linear advection, we have

$$f(u) = a u \quad , \quad a \geq 0$$

and thus

$$f(u_i^L) = a u_i^L = a\left(u_i^n - \frac{1}{2}\Delta_i\right)$$

$$f(u_i^R) = a u_i^R = a\left(u_i^n + \frac{1}{2}\Delta_i\right)$$

which we need in step 2) to compute $\bar{u}_i^L$ and $\bar{u}_i^R$:

$$\bar{u}_i^L = u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

$$= u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[a\left(u_i^n - \frac{1}{2}\Delta_i\right) - a\left(u_i^n + \frac{1}{2}\Delta_i\right)\right]$$

$$= u_i^L - \frac{1}{2} a \frac{\Delta t}{\Delta x}\Delta_i$$

$$= \underbrace{u_i^n - \frac{1}{2}\Delta_i}_{= u_i^L} - \frac{1}{2}\underbrace{c}_{a\frac{\Delta t}{\Delta x}}\Delta_i$$

$$= u_i^n - \frac{1}{2}(1+c)\Delta_i$$

$$\bar{u}_i^R = u_i^n + \frac{1}{2}(1-c)\Delta_i$$

Finally, we solve the Riemann problem at the interface $i+1/2$ with the initial data

$$(\bar{u}_i^R, \bar{u}_{i+1}^L)$$

for which the solution is

$$u_{i+1/2}(x/t) = \begin{cases} \bar{u}_i^R = u_i^n + \frac{1}{2}(1-c)\Delta_i, & \frac{x}{t} < a \\ \bar{u}_{i+1}^L = u_{i+1}^n - \frac{1}{2}(1+c)\Delta_i, & \frac{x}{t} > a \end{cases}$$

giving us the flux

$$f_{i+1/2}^{MHM} = \frac{1}{2}(1+\text{sign}(c))f(\bar{u}_i^R) + \frac{1}{2}(1-\text{sign}(c))f(\bar{u}_{i+1}^L)$$

and of course

$$f(u) = au$$

The sign(c) gives us the solution regardless of the advection velocity: with $c = a\frac{\Delta t}{\Delta x}$ and $\Delta t, \Delta x \geq 0$, sign(c) is equal to sign(a), and the factors $\frac{1}{2}(1 \pm \text{sign}(c))$ give 0 or 1, depending on the sign of the advection velocity.

4.3 The Piece-Wise Linear Method

For any conservation law of the form

$$\partial_t u + \partial_x f(u) = 0$$

that is solved by

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2}\right]$$

the Piece-Wise Linear Method (PLM) defines the numerical flux $f_{i+1/2}$ as

$$f_{i+1/2}^{PLM} = \frac{1}{2}\left[f\left(u_{i+1/2}^{grp}(0,0)\right) + f\left(u_{i+1/2}^{grp}(0,\Delta t)\right)\right]$$

where $u_{i+1/2}^{grp}$ is the solution of the Generalized Riemann problem.

The PLM doesn't solve the Generalized Riemann problem directly; Instead, it uses a trapezium rule approximation in time to the integral

$$f_{i+1/2} = \frac{1}{\Delta t}\int_0^{\Delta t} f\left[u_{i+1/2}^{grp}(0,t)\right] dt$$

To determine the fluxes $f(u^{grp}_{i+1/2}(0,0))$ and $f(u^{grp}_{i+1/2}(0,\Delta t))$, we start by

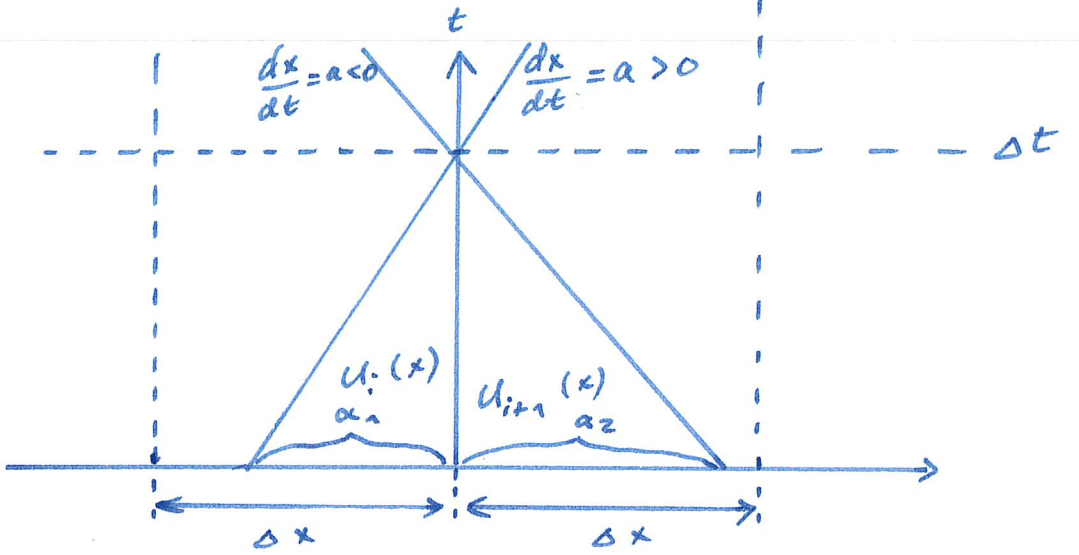
1) Reconstruct data in each cell, and determine the boundary extrapolated values u_L and u_R for each cell.

2) For $f(u^{grp}_{i+1/2}(0,0))$, solve the conventional Riemann problem

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(x,0) = \begin{cases} u_i^R & x < 0 \\ u_{i+1}^L & x > 0 \end{cases} \end{cases}$$

with piece-wise constant data u_L, u_R.

3) For $f(u^{grp}_{i+1/2}(0,\Delta t))$, trace back characteristics from the point $(0,\Delta t)$ to the reconstructed piece-wise linear data



For linear advection with advection velocity $a \geq 0$, in local coordinates (i.e. $u_i(\hat{x}=0) = u_L$, $u_i(\hat{x}=\Delta x) = u_R = u_{i+1/2}$)

$$u_{i+1/2}(\Delta t) = u_i(\hat{x} = \Delta x,\ t = \Delta t)$$

$$= u_i(\hat{x} = \Delta x - a\Delta t,\ t = 0)$$

$$= u_i((1-c)\Delta x,\ 0)$$

and for $a \leq 0$:

$$u_{i+1/2}(\Delta t) = u_{i+1}(\hat{x} = 0,\ t = \Delta t)$$

$$= u_{i+1}(\hat{x} = 0 + |a|\Delta t,\ t = 0)$$

$$= u_{i+1}(c\Delta x,\ t = 0)$$

The choice for u_i in the formula for $u_{i+1/2}(t=\Delta t)$ in the case when $a>0$ and u_{i+1} for $a<0$ was made such that the resulting local coordinate $\hat{x} \in [0, \Delta x]$.
(Above Δx or below 0, it would just mean that we need to go to the neighbouring cell.)

With $u_{i+1/2}(t=\Delta t)$ known, we use it as the solution to the Generalized Riemann problem, i.e.

$$u^{grp}_{i+1/2}(0, \Delta t) = u_{i+1/2}(\Delta t)$$

and we are therefore able to compute the second flux

$$f\left(u^{grp}_{i+1/2}(0, \Delta t)\right)$$

Using the definition for the linear slope reconstruction

$$u_i(x) = u_i + \frac{(\hat{x} - \hat{x}_i)}{\Delta x}\Delta_i \quad , \quad \hat{x}_i = \frac{1}{2}\Delta x$$

we can compute the explicit formula:

$$f(u) = a u$$

For $a > 0$:

$$u_{i+1/2}(0) = u_i^R = u_i^n + \frac{1}{2}\Delta_i$$

$$u_{i+1/2}(\Delta t) = u_i((1-c)\Delta x, 0)$$

$$= u_i^n + \frac{(1-c)\Delta x - 1/2\,\Delta x}{\Delta x}\Delta_i$$

$$= u_i^n + (1/2 - c)\Delta_i$$

Then

$$f_{i+1/2}^{PLM} = \frac{1}{2}\left[f(u_{i+1/2}(0,0)) + f(u_{i+1/2}(0,\Delta t))\right]$$

$$= \frac{1}{2}\left[a(u_i^n + \frac{1}{2}\Delta_i) + a(u_i^n + (\tfrac{1}{2} - c)\Delta_i)\right]$$

$$= \frac{1}{2}\left[2a u_i^n + a(1-c)\Delta_i\right]$$

$$= a\left[u_i^n + \frac{1}{2}(1-c)\Delta_i\right]$$

For $a < 0$:

$$u_{i+1/2}(0,0) = u^L_{i+1} = u^n_{i+1} - \frac{1}{2}\Delta_{i+1}$$

$$u_{i+1/2}(0,\Delta t) = u_{i+1}(c\Delta x, t=0)$$

$$= u^n_{i+1} + \frac{c\Delta x - 1/2\,\Delta x}{\Delta x}\Delta_{i+1}$$

$$= u^n_{i+1} + (c - 1/2)\Delta_{i+1}$$

Then

$$f^{PLM}_{i+1/2} = \frac{1}{2}\left[f(u_{i+1/2}(0,0)) + f(u_{i+1/2}(0,\Delta t))\right]$$

$$= \frac{1}{2}\left[a\left(u^n_{i+1} - \frac{1}{2}\Delta_i\right) + a\left(u^n_{i+1} + (c-1/2)\Delta_{i+1}\right)\right]$$

$$= \frac{1}{2}\left[2a\,u^n_{i+1} + a(c-1)\Delta_{i+1}\right]$$

$$= a\left[u^n_{i+1} - \frac{1}{2}(1-c)\Delta_{i+1}\right]$$

To summarize:

$$f^{PLM}_{i+1/2} = \begin{cases} a\left[u^n_i + \frac{1}{2}(1-c)\Delta_i\right] & a \geq 0 \\ a\left[u^n_{i+1} - \frac{1}{2}(1-c)\Delta_{i+1}\right] & a \leq 0 \end{cases}$$

4.4 The Generalized Riemann Problem Method

The basic ingredient of the GRP method is the solution of the Generalized Riemann Problem

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(x,0) = \begin{cases} u_i(x) & x < 0 \\ u_{i+1}(x), & x > 0 \end{cases} \end{cases}$$

to obtain a Godunov-type numerical flux that yields a second-order accurate scheme.

The GRP method defines a numerical flux as

$$f^{grp}_{i+1/2} = f\left(u^{grp}_{i+1/2}\left(0, \tfrac{1}{2}\Delta t\right)\right)$$

where $u^{grp}_{i+1/2}(x,t)$ is the solution of the Generalized Riemann problem and $u^{grp}_{i+1/2}(0, \frac{1}{2}\Delta t)$ is the mid-point rule approximation in time to the integral

$$f_{i+1/2} = \frac{1}{\Delta t}\int_0^{\Delta t} f\left(u^{grp}_{i+1/2}(0,t)\right)dt$$

As an analytical solution for $u^{grp}_{i+1/2}(0, 1/2\Delta t)$ is difficult at best and impossible at worst, we need another approximation to obtain an expression: We use the Taylor expansion

$$u^{grp}_{i+1/2}(0, 1/2\Delta t) = u^{grp}_{i+1/2}(0,0) + \frac{1}{2}\Delta t \frac{\partial}{\partial t} u^{grp}_{i+1/2}(0,0) + \mathcal{O}(\Delta t^2)$$

The first term, $u^{grp}_{i+1/2}(0,0)$, is the value of $u^{grp}_{i+1/2}(x,t)$ immediately after the interaction of the piece-wise linear states $u_i(x)$, $u_{i+1}(x)$ in the Generalized Riemann Problem. The value is solely determined by the extrapolated boundary values u_i^R, u_{i+1}^L.

To determine the second term, $\frac{1}{2}\Delta t \frac{\partial}{\partial t} u^{grp}_{i+1/2}(0,0)$, then are several possibilities.

Toro suggests a modification of the original GRP method whereby the time derivative in the second term is replaced by a space derivative. Then the required value along the cell interface results from solving an exact Riemann problem for gradients.

Let's do this for linear advection now.

For any p-th order spatial derivative $v = \frac{\partial^p u}{\partial x^p}$, where $u(x,t)$ is a solution to the advection equation

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

we have

$$\frac{\partial v}{\partial t} + a \frac{\partial v}{\partial x} = 0$$

which is easy to show for a smooth function u such that $\frac{\partial^2}{\partial x \partial t} u = \frac{\partial^2}{\partial t \partial x} u$:

$$\partial_x^p [0] = 0 = \partial_x^p \left[\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} \right] = \frac{\partial}{\partial t} \partial_x^p u + a \frac{\partial}{\partial x} \partial_x^p u = \frac{\partial v}{\partial t} + a \frac{\partial v}{\partial x}$$

This means that any spatial gradient of $u(x,t)$ will obey the original PDE.

Hence one may pose Riemann problems for gradients $\frac{\partial^p u}{\partial x^p}$, which if assumed piece-wise constant, lead to conventional Riemann problems.

For the linear advection equation, we have

$$\partial_t u = -a \partial_x u$$

thus the time derivative can be replaced by a space derivative. The space derivative can be found using the above findings and solving the Riemann problem of the $p=1$ derivative:

$$\begin{cases} v_t + a v_x = 0 \\ v(x,0) = \begin{cases} \frac{\Delta_i}{\Delta x}, & x<0 \\ \frac{\Delta_{i+1}}{\Delta x}, & x>0 \end{cases} \end{cases}$$

where I used the fact that $v(x,t) = \frac{\partial u}{\partial x}(x,t)$ and $\frac{\partial u_i}{\partial x} \approx \frac{u_i^R - u_i^L}{\Delta x} = \frac{u_i^n + \frac{1}{2}\Delta_i - (u_i^n - \frac{1}{2}\Delta_i)}{\Delta x} = \frac{\Delta_i}{\Delta x}$ for the expression of the initial conditions.

The solution is simply

$$v(x,t) = \frac{\partial u}{\partial x}(x,t) = \begin{cases} \frac{\Delta_i}{\Delta x}, & \frac{x}{t} < a \\ \frac{\Delta_{i+1}}{\Delta x}, & \frac{x}{t} > a \end{cases}$$

This finally gives us

$$u_{i+1/2}^{grp}(0, \tfrac{1}{2}\Delta t) = u_{i+1/2}^{grp}(0,0) + \tfrac{1}{2}\Delta t \frac{\partial}{\partial t} u_{i+1/2}^{grp}(0,0) + O(\Delta t^2)$$

$$= u_{i+1/2}^{grp}(0,0) - \frac{a}{2}\Delta t \frac{\partial}{\partial x} u_{i+1/2}^{grp}(0,0) + O(\Delta t^2)$$

$$= u_{i+1/2}^{grp} - \frac{1}{2} c \Delta x \frac{\partial}{\partial x} u_{i+1/2}^{grp}$$

and

$$f_{i+1/2}^{grp} = a\left[u_{i+1/2}^{grp} - \frac{1}{2} c \Delta x \frac{\partial}{\partial x} u_{i+1/2}^{grp}(0,0)\right]$$

$$= \begin{cases} a\left[u_i^n + \frac{1}{2}(1-c)\Delta_i\right], & a > 0 \\ a\left[u_{i+1}^n - \frac{1}{2}(1+c)\Delta_{i+1}\right], & a < 0 \end{cases}$$

4.5 Slope-Limiter Centered Schemes

For SLIC schemes, we extend the approach for constructing high-order schemes of any low-order scheme with numerical flux

$$f^{LO}_{i+1/2} = f^{LO}_{i+1/2}(u_L, u_R)$$

where LO stands for "Low Order". The numerical flux for this approach needs to depend on the two states u_L and u_R.

We are interested in low-order schemes that avoid the solution of the Riemann problem.

Similar to the MUSCL-Hancock scheme, we start by reconstructing the boundary extrapolated values.

$$u_i^L = u_i^n - \frac{1}{2}\Delta_i$$

$$u_i^R = u_i^n + \frac{1}{2}\Delta_i$$

And then we evolve them by a time $1/2\,\Delta t$ according to

$$\bar{u}_i^L = u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

$$\bar{u}_i^R = u_i^R + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

But now instead of solving the Riemann problem with data $(\bar{u}_i^R, u_{i+1}^L)$ to find the Godunov first-order upwind flux, we compute a low-order flux with data arguments $(\bar{u}_i^R, \bar{u}_{i+1}^L)$. Thus we have

$$f_{i+1/2}^{slic} = f_{i+1/2}^{LO}(\bar{u}_i^R, \bar{u}_{i+1}^L)$$

A possible choice for the low-order flux is the FORCE (First ORder CEntered scheme) flux, which is given by

$$f_{i+1/2}^{force}(u_L, u_R) = \frac{1}{2}\left[f_{i+1/2}^{RI}(u_L, u_R) + f_{i+1/2}^{LF}(u_L, u_R)\right]$$

where $f^{RI}_{i+1/2}(u_L, u_R)$ is the Richtmyer flux:

$$f^{RI}_{i+1/2}(u_L, u_R) = f\left(u^{n+1/2}_{i+1/2}\right)$$

with $$u^{n+1/2}_{i+1/2} = \frac{1}{2}(u_L + u_R) + \frac{1}{2}\frac{\Delta t}{\Delta x}\left(f(u_L) - f(u_R)\right)$$

and $f^{LF}_{i+1/2}$ is the Lax-Friedrichs scheme flux:

$$f^{LF}_{i+1/2}(u_L, u_R) = \frac{1}{2}\left(f(u_L) + f(u_R)\right) + \frac{1}{2}\frac{\Delta x}{\Delta t}(u_L - u_R)$$

For a smart choice of the flux and the reconstruction, you can even get better than second order accuracy.

To avoid spurious oscillations, slope limiters will be applied, which we will discuss later.

4.6 Other Approaches

4.6.1 Semi-Discrete Schemes

In the semi-discrete approach, or method of lines, one separates space and time discretization processes. First assume some discretization in space, while leaving the problem continuous in time. This results in an ODE in time:

$$\frac{du}{dt} = \frac{1}{\Delta x}\left(f_{i-1/2} - f_{i+1/2}\right)$$

where $f_{i+1/2} = f_{i+1/2}(\{u_i(t)\}, t)$ is an intercell numerical flux.

Any higher order method can be utilized to solve the spatial part of the discretization.

The time discretization results from solving the ODE with any method applicable, even higher order methods like the Runge-Kutta integrators

The separation of time and space discretisation processes in the semi-discrete approach allows enormous flexibility and is well-suited for devising very high order schemes.

4.6.2 Implicit Methods

Implicit methods are conservative schemes of the form

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

where the intercell flux $f_{i+1/2}$ depends on both data values $\{U_i^n\}$ and unknown values $\{U_i^{n+1}\}$.

Implicit methods tend to be expensive in both computing time and memory, but are in theory not restricted to a time step size by stability considerations.

5. Monotone Schemes and Accuracy

5.1 Monotone Schemes

A scheme

$$u_i^{n+1} = H(u_{i-k_L+1}^n, \ldots, u_{i+k_R}^n) \qquad [1]$$

$$= \sum_{k=-k_L+1}^{k_R} b_k u_{i+k}^n \qquad [2]$$

with k_L, k_R two non-negative integers, is said to be monotone if

$$\frac{\partial H}{\partial u_j^n} \geq 0 \quad \forall j \qquad [3]$$

$$\Leftrightarrow \quad b_k \geq 0 \quad \forall j, k \qquad [4]$$

That is, H is a non-decreasing function of each of its arguments.

This definition of a monotone scheme is equivalent to the following property:

if $v_i^n \geq u_i^n \ \forall i$, then $v_i^{n+1} \geq u_i^{n+1} \ \forall i$ [5]

When applied to a scalar non-linear conservation law of the form

$$\partial_t u + \partial_x f(u) = 0$$

a useful property is the following:

Given the data set $\{u_i^n\}$, if the solution set $\{u_i^{n+1}\}$ is obtained with a monotone method, then

$$\max_i \{u_i^{n+1}\} \leq \max_i \{u_i^n\}$$

$$\min_i \{u_i^{n+1}\} \geq \min_i \{u_i^n\}$$

Proof:

Define $v_i^n = \max_j \{u_j^n\} = \text{const} \quad \forall i.$

Then $\partial_t v + \partial_x f(v) = 0 = \partial_t v + \frac{\partial f}{\partial v} \underbrace{\frac{\partial v}{\partial x}}_{=0}$

$\Rightarrow$ $v = \text{const}$ w.r.t. time as well

Since $v_i^n = \max_j \{u_j^n\} \quad \forall i,$ we can write

$$v_i^{n+1} = H(v_{i-k_L+1}, \dots, v_{i+k_R})$$
$$= H(v_i = \max_j \{u_j^n\})$$
$$= b_0 v_i^n$$

With $v_i = \text{const}$ w.r.t. time, $\Rightarrow b_0 = 1$

and $v_i^{n+1} = v_i^n$

Additionally, we use property [5] to get

$$v_i^{n+1} \geq u_i^{n+1} \quad \text{because} \quad v_i^n \geq u_i^n$$

Combining these:

$$v_i^{n+1} = v_i^n \geq u_i^{n+1} \Rightarrow \max_j \{u_j^n\} \geq u_i^{n+1} \; \forall i,j$$

$$\Rightarrow \max_j \{u_j^n\} \geq \max_j \{u_j^{n+1}\}$$

The proof for the $\min_i \{u_i^{n+1}\} \geq \min_i \{u_i^n\}$ follows analogously.

An obvious consequence is revealed when applying these properties recursively:

$$\max_i \{u_i^n\} \leq \max_i \{u_i^{n-1}\} \leq \max_i \{u_i^{n-2}\} \leq \dots \leq \max_i \{u_i^0\}$$

$$\min_i \{u_i^n\} \geq \min_i \{u_i^{n-1}\} \geq \min_i \{u_i^{n-2}\} \geq \dots \geq \min_i \{u_i^0\}$$

⇒ No new extrema are created, and thus spurious oscillations do not appear.

⇒ However, this also means that minima will increase, and maxima will decrease, leading to clipping of extrema, which is a disadvantage of monotone methods.

Another useful consequence is that solutions of monotone schemes satisfy

$$\min_j \{u_j^n\} \le u_i^{n+1} \le \max_j \{u_j^n\}$$

which follows from

$$\min_j\{u_j^n\} \le \underbrace{\min_j \{u_j^{n+1}\}}_{\text{our finding}} \le u_i^{n+1} \le \underbrace{\max_j \{u_i^{n+1}\}}_{\text{our previous finding}} \le \max_j \{u_i^n\}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{by definition of min/max of } u_i^{n+1}}$$

$\Rightarrow$ The solution at any point i is bounded by the minimum and maximum of the data.

[Note that this only holds for scalar conservation laws. The Euler equations aren't scalar laws. Otherwise You'd never have galaxies forming.]

A further theorem applicable to all three-point schemes (i.e. schemes with a 3 point stencil, e.g. u_{i-1}^n, u_i^n u_{i+1}^n) is that if a three-point scheme of the form

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2}\right]$$

for the non-linear conservation law of the form

$$\partial_t u + \partial_x f(u) = 0$$

is monotone, then

$$\frac{\partial}{\partial u_i^n} f_{i+1/2}(u_i^n, u_{i+1}^n) \geq 0 \quad \text{and} \quad \frac{\partial}{\partial u_{i+1}^n} f_{i+1/2}(u_i^n, u_{i+1}^n) \leq 0$$

That is, in a monotone scheme the numerical flux $f_{i+1/2}(u_i^n, u_{i+1}^n)$ is an increasing (i.e. non-decreasing) function of its first argument and a decreasing (i.e. non-increasing) function of its second argument

Proof:

We define

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n)$$

$$\equiv u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2}(u_{i-1}^n, u_i^n) - f_{i+1/2}(u_i^n, u_{i+1}^n)\right]$$

Since the used method is required to be monotone, the following must hold:

$$\frac{\partial H}{\partial u_{i-1}^n} \geq 0$$

$$\Rightarrow \frac{\partial H}{\partial u_{i-1}^n} = \underbrace{\frac{\Delta t}{\Delta x}}_{\geq 0} \frac{\partial f_{i-1/2}}{\partial u_{i-1}^n} \geq 0$$

$$\Rightarrow \frac{\partial f_{i-1/2}}{\partial u_{i-1}^n} \geq 0 \quad \Rightarrow \quad \frac{\partial f_{i+1/2}}{\partial u_i^n} \geq 0$$

also: $$\frac{\partial H}{\partial u_{i+1}^n} = -\frac{\Delta t}{\Delta x} \frac{\partial f_{i+1/2}}{\partial u_{i+1}^n} \geq 0$$

$$\Rightarrow \frac{\partial f_{i+1/2}}{\partial u_{i+1}^n} \leq 0$$

5.2 Monotone Schemes and Godunov's Theorem

Linear (i.e. constant/fixed coefficient) second order accurate schemes to solve the linear advection equation are distinctly better than first-order methods for problems with smooth solutions; however for solutions involving step gradients, such as near a discontinuity, these methods produce spurious oscillations. Monotone methods avoid these oscillations, but have limited accuracy.

Godunov's theorem establishes that the desirable properties of accuracy and monotonicity are, for linear schemes, contradictory requirements. The theorem states that

> There are no monotone, linear schemes for non-linear scalar hyperbolic conservation laws of second order or higher of accuracy.

Proof:

Using Roe's Theorem:

A scheme of the form

$$u_i^{n+1} = H(u_{i-k_L}, \ldots, u_{i+k_R}) = \sum_{k=-k_L}^{k_R} b_k u_{i+k}^n$$

is p-th order accurate in space and time if and only if

$$\sum_{k=-k_L}^{k_R} k^q b_k = (-c)^q, \quad 0 \le q \le p$$

Let us denote the summation

$$s_q = \sum_{k=-k_L}^{k_R} k^q b_k$$

b_k are the constant coefficients of the linear scheme. So for second order accuracy, we require

$$s_0 = 1, \quad s_1 = -c, \quad s_2 = c^2$$

Now we compute and massage S_2:

$$S_2 = \sum_{k=-k_L}^{k_R} k^2 b_k$$

$$= \sum_{k=-k_L}^{k_R} \left([k+c]^2 - 2kc - c^2 \right) b_k$$

$$= \sum_{k=-k_L}^{k_R} (k+c)^2 b_k - 2c \sum_{k=-k_L}^{k_R} k\, b_k - c^2 \sum_{k=-k_L}^{k_R} b_k$$

$$= \sum_{k=-k_L}^{k_R} (k+c)^2 b_k - 2c \underbrace{S_1}_{=-c} - c^2 \underbrace{S_0}_{=+1}$$

$$= \sum_{k=-k_L}^{k_R} (k+c)^2 b_k + 2c^2 - c^2$$

$$= \sum_{k=-k_L}^{k_R} (k+c)^2 b_k + c^2$$

For the method to be second order accurate, we need to have

$$S_2 = c^2$$

In which cases is this condition satisfied?

1) $b_k = 0 \quad \forall k$

But is that even a method?
Nothing happens then, $u_i^{n+1} = 0 \ldots$

2) $c = -k_0$, $\quad b_k = 0 \quad \forall k \neq k_0$

You can only have one Courant number, hence only one coefficient index $k = k_0$ is permissible. Also the Courant number must be an integer.

We assumed a monotone scheme, which demands that $b_k \geq 0 \quad \forall k$.

Hence for $0 < |c| < 1$, we have proved the theorem.

This means that monotone schemes are at most first order accurate, which is too inaccurate to be of practical interest.

In other words, we need to look for a way to circumvent Godunov's theorem. The key to this lies on the assumption made in the theorem that the schemes have fixed coefficients, i.e. are linear schemes.

More concretely, the aim is to find so-called high resolution methods that satisfy the following:

1) The schemes have second or higher order of accuracy in smooth parts of the solution

2) The schemes produce numerical solutions free from spurious oscillations

3) The schemes produce high resolution of discontinuities, i.e. the number of mesh points in the transition zone containing the discontinuous wave is narrow compared to that of the first order monotone method

5.3 Data Compatibility

Definition: Data Compatible Algorithm

A scheme is compatible with a given data set $\{u_i^n\}$ if the solution u_i^{n+1} at each point i, as given by the algorithm, is bounded by the upwind pair (u_{i-s}^n, u_i^n), where $s = \text{sign}(c) = \text{sign}(a)$.

We can rephrase this definition:

$$\min\{u_{i-s}^n, u_i^n\} \le u_i^{n+1} \le \max\{u_{i-s}^n, u_i^n\}$$

The data compatibility condition is equivalent to requiring

$$0 \le \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i^n} \le 1$$

Proof:

a) Let $u_{i-s}^n \leq u_i^n$

then

$$u_{i-s}^n \leq u_i^{n+1} \leq u_i^n$$

$$\Rightarrow u_{i-s}^n - u_i^n \leq u_i^{n+1} - u_i^n \leq 0$$

$$\Rightarrow 1 \geq \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i^n} \geq 0$$

switch of signs because $u_{i-s}^n - u_i^n < 0$!

b) analogously for $u_{i-s}^n \geq u_i^n$

Where does this data compatibility come from?

Well, it's mimicking a property of conservation laws of the form

$$\partial_t u + \partial_x f(u) = 0$$

where $u = u(x,t)$ and f is a convex flux function, i.e. $\frac{d^2 f}{du^2} > 0$.

Suppose you have a function $V(x,t)$ and a function $u(x,t)$, both of which satisfy the conservation law equation. Then if

$$V_0(x) = V(x, t=0) \geq u_0(x) = u(x, t=0) \quad \forall x,$$

$V(x,t) \geq u(x,t)$ for all t as well.

To demonstrate this property, let us find the analytical solution to $u(x,t)$ from the conservation law

$$\frac{\partial u}{\partial t} + \lambda(u) \frac{\partial u}{\partial x} = 0$$

$$\rightarrow \frac{1}{\lambda(u)} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

Let $u(x,t) \equiv f(x)\, g(t)$

Then

$$\frac{1}{\lambda} f \frac{\partial g}{\partial t} + g \frac{\partial f}{\partial x} = 0$$

$$\Rightarrow \frac{1}{\lambda} \frac{1}{g} \cdot \frac{\partial g}{\partial t} + \frac{1}{f} \frac{\partial f}{\partial x} = 0$$

$$\Rightarrow \frac{1}{\lambda} \frac{1}{g} \frac{\partial g}{\partial t} = -\frac{1}{f} \frac{\partial f}{\partial x} \equiv -\alpha \quad \text{const.}$$

Solving individually:

$$\frac{\partial f}{f} = +\alpha\, dx$$

$$f = C_1 e^{\alpha x}$$

and

$$\frac{\partial g}{g} = -\lambda(u)\,\alpha\, dt$$

$$g = C_2 e^{-\alpha \int_0^t \lambda\, dt}$$

giving

$$u = f \cdot g = C e^{\alpha\left(x - \int_0^t \lambda\, dt\right)}$$

we also have

$$u(x, t=0) = u_0 = C e^{\alpha\left(x - \underbrace{\int_0^0 \lambda\, dt}_{=0}\right)}$$

$$= C e^{\alpha x}$$

and we see that we can write

$$u(x, t) = u\left(x = x - \int_0^t \lambda\, dt,\ t=0\right) =$$

$$= u_0\left(x - \int_0^t \lambda\, dt\right)$$

Now back to our assumption:

Let $v_0 \geq u_0 \quad \forall x$

Then $v(x,t) = v_0(x - \int \lambda dt) \geq u_0(x - \int \lambda dt) =$

$= u(x,t)$

$\Rightarrow$ If $v_0 \geq u_0 \quad \forall x$, then

$v(x,t) \geq u(x,t) \quad \forall x$

Note: This requires that $v_0 \geq u_0$ for all x, not at every x. Big difference.

So the idea of data compatibility relies on the boundedness of the problem.

What if it is violated?

Well, then you're allowing new minima or maxima to form, since your solution is not bounded by the previous minima or maxima, and you get oscillations.

The interesting point is that when you start checking second order methods, you will find that not all methods are compatible with all possible data, but only for a certain range of values for r_i.

with

$$r_i = \frac{u_i^n - u_{i-1}^n}{u_{i+1}^n - u_i^n}$$

such that

$$0 \leq \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i} \leq 1$$

So how about we use different methods for different values of r_i such that always a compatible method is used?

In other words, pick your method based on the values that you currently are working on. Changing methods means that we have variable coefficients, and picking the method based on r_i means that the choice is a function of the local state. So we arrived at non-linear methods!

6. Total Variation Diminishing Methods

6.1 The Total Variation

It can be proved that Total Variation Stable methods are convergent. A subclass of total Variation Stable methods are those whose total variation doesn't increase in time; Those are commonly refered to as total variation diminishing methods.

Given a function $u = u(x)$, the total variation of u is defined as

$$TV(u) = \lim_{\delta \to 0} \sup \frac{1}{\delta} \int_{-\infty}^{\infty} |u(x+\delta) - u(x)| \, dx$$

if $u(x)$ is smooth, then the definition is equivalent to

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

For convergence purposes, it suffices to define the total variation of $u(x,t)$ at fixed times $t = t^n$. If $u^n = \{u_i^n\}$ is a mesh function, then the total variation is defined as

$$TV(u^n) = \sum_{i=-\infty}^{\infty} |u_{i+1}^n - u_i^n|$$

In order for $TV(u^n)$ to be finit, one must assume $u_i^n = 0$ or $u_i^n = \text{const}$ as $i \to \pm\infty$.

A fundamental property of the exact solution of the IVP for the non-linear scalar conservation law, provided $u(x,0)$ has bounded total variation, is

1) No new local extrema in x may be formed

2) The value of a local minimum doesn't decrease, and the value of a local maximum doesn't increase.

From this it follows that the $TV(u(t))$ is a decreasing function of time, that is

$$TV(u(t_2)) \leq TV(u(t_1)) \quad \forall\, t_2 \geq t_1$$

6.2 TVD and Monotonicity Preserving Schemes

Consider a numerical scheme of the form

$$U_i^{n+1} = H(U_{i-k_L+1}^n, \dots, U_{i+k_R}^n)$$

with k_L, k_R two non-negative integers, to solve a scalar conservation law.

Definition: A scheme $U_i^{n+1} = H(U_{i-k_L+1}^n, \dots, U_{i+k_R}^n)$ is said to be a Total Variation Diminishing scheme, if

$$TV(U^{n+1}) \le TV(U^n) \quad \forall n$$

Definition: Schemes of the same form as above are said to be Monotonicity Preserving Schemes if whenever the data $\{U_i^n\}$ is monotone the solution set $\{U_i^{n+1}\}$ is monotone in the same sense. That is if $\{U_i^n\}$ is monotone increasing, so is $\{U_i^{n+1}\}$, and if $\{U_i^n\}$ is monotone decreasing, so is $\{U_i^{n+1}\}$.

In general, the set S_{mon} of monotone schemes is contained in the set S_{TVD} of TVD schemes and this in turn is contained in the set S_{mpr} of monotonicity preserving schemes, that is

$$S_{mon} \subseteq S_{TVD} \subseteq S_{mpr}$$

we leave out the proof of this theorem.

A linear scheme as applied to the linear advection equation is Monotonicity Preserving if and only if the coefficients b_k are non-negative, i.e. $b_k \geq 0 \; \forall k$, when

$$u_i^{n+1} = \sum_{k=-k_l}^{k_R} b_k u_{i+k}$$

This is also the condition for a scheme to be monotone, hence monotone schemes are monotonicity preserving in the case of linear advection, and thus also TVD.

Now consider the class of non-linear schemes

$$U_i^{n+1} = U_i^n - C_{i-1/2}\,\Delta U_{i-1/2} + D_{i+1/2}\,\Delta U_{i+1/2}$$

with

$$\Delta U_{i-1/2} = U_i - U_{i-1}$$

$$\Delta U_{i+1/2} = U_{i+1} - U_i$$

For any scheme of this form to solve a (linear) hyperbolic scalar conservation law, a sufficient condition for the scheme to be TVD is that the coefficients satisfy

$$C_{i+1/2} \geq 0$$

$$D_{i+1/2} \geq 0$$

$$0 \leq C_{i+1/2} + D_{i+1/2} \leq 1$$

Proof: We apply the scheme to two consecutive cells i and $i+1$:

$$u_i^{n+1} = u_i^n - C_{i-1/2}(u_i^n - u_{i-1}^n) + D_{i+1/2}(u_{i+1}^n - u_i^n)$$

$$u_{i+1}^{n+1} = u_{i+1}^n - C_{i+1/2}(u_{i+1}^n - u_i^n) + D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)$$

Then

$$u_{i+1}^{n+1} - u_i^{n+1} = u_{i+1}^n - C_{i+1/2}(u_{i+1}^n - u_i^n) + D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)$$
$$- u_i^n + C_{i+1/2}(u_i^n - u_{i-1}^n) - D_{i+1/2}(u_{i+1}^n - u_i^n)$$

$$= u_{i+1}^n - C_{i+1/2}u_{i+1}^n + C_{i+1/2}u_i^n + D_{i+3/2}u_{i+2}^n - D_{i+3/2}u_{i+1}^n$$
$$- u_i^n + C_{i-1/2}u_i^n - C_{i-1/2}u_{i-1}^n - D_{i+1/2}u_{i+1}^n + D_{i+1/2}u_i^n$$

$$= u_{i+1}^n(1 - C_{i+1/2} - D_{i+1/2}) + u_i^n(-1 + C_{i+1/2} + C_{i-1/2} + D_{i+$$
$$+ u_{i-1}(-C_{i-1/2}) + D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)$$

$$= (u_{i+1}^n - u_i^n)(1 - C_{i+1/2} - D_{i+1/2}) +$$
$$+ (u_i - u_{i-1})C_{i-1/2} + D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)$$

Now take the absolute value:

$$|u_{i+1}^{n+1} - u_i^{n+1}| = |(u_{i+1}^n - u_i^n)(1 - C_{i+1/2} - D_{i+1/2}) +$$

$$+ (u_i - u_{i-1}) C_{i-1/2} + D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)|$$

$$\leq |(u_{i+1}^n - u_i^n)(1 - C_{i+1/2} - D_{i+1/2})| +$$

$$|(u_i - u_{i-1}) C_{i-1/2}| + |D_{i+3/2}(u_{i+2}^n - u_{i+1}^n)|$$

$$\leq |u_{i+1}^n - u_i^n||1 - C_{i+1/2} - D_{i+1/2}| + |C_{i-1/2}||u_i - u_{i-1}| +$$

$$+ |D_{i+3/2}||u_{i+2}^n - u_{i+1}^n|$$

Now apply the sufficient conditions:

$$C_{i+1/2} \geq 0, \quad D_{i+1/2} \geq 0, \text{ and } 1 - C_{i+1/2} - D_{i+1/2} \geq 0$$

Then

$$|u_{i+1}^{n+1} - u_i^{n+1}| \leq |u_{i+1}^n - u_i^n| - C_{i+1/2}|u_{i+1}^n - u_i^n| -$$

$$- D_{i+1/2}|u_{i+1}^n - u_i^n| + C_{i-1/2}|u_i - u_{i-1}| +$$

$$+ D_{i+3/2}|u_{i+2}^n - u_{i+1}^n|$$

$$= |u_{i+1}^n - u_i^n| +$$

$$+ C_{i-1/2}|u_i^n - u_{i-1}^n| - C_{i+1/2}|u_{i+1}^n - u_i^n| +$$

$$+ D_{i+3/2}|u_{i+2}^n - u_{i+1}^n| - D_{i+\frac{1}{2}}|u_{i+1}^n - u_i^n|$$

Finally, let us sum over all i:

$$\sum_i |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_i |u_{i+1}^n - u_i^n| +$$

$$+ \sum_i C_{i-1/2} |u_i^n - u_{i-1}^n| - \sum_i C_{i+1/2} |u_{i+1}^n - u_i^n| +$$

$$+ \sum_i D_{i+3/2} |u_{i+2}^n - u_{i+1}^n| - \sum D_{i+1/2} |u_{i+1}^n - u_i^n|$$

$$= \sum_i |u_{i+1}^n - u_i^n|$$

Here we use for the final step that when summing over all i, the second and third row cancel out. Remember that we require $|u_{i+1} - u_i| \to 0$ for $i \to \pm\infty$, so by walking over all i it cancels out nicely.

So we arrive at the sought result

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

by assuming the sufficient conditions above.

7. Flux Limiter Methods

7.1 TVD version of the WAF method

The big idea is to utilize variable coefficients in the method based on the local situation, i.e. the $\{u_i^n\}$, such that the method becomes TVD.

For convenience, we rewrite the WAF flux

$$f_{i+1/2} = \frac{1}{2}(1+c)(au_i^n) + \frac{1}{2}(1-c)(au_{i+1}^n)$$

as

$$f_{i+1/2} = \frac{1}{2}(1+\phi)(au_i^n) + \frac{1}{2}(1-\phi)(au_{i+1}^n)$$

for $a \geq 0$.

Familiar methods result from particular choices of ϕ:

$\phi = |c|$: Lax – Wendroff method

$\phi = s \equiv sign(a) = sign(c)$: Godunov 1st order upwind

$\phi = -s$: Downwind method, unconditionally unstable

The purpose is to find appropriate ranges for ϕ as a function of some data-dependent variables that produce a TVD version of the WAF scheme.

A first upper and lower bound comes from the extreme cases $\phi = s$ and $\phi = -s$, giving us

$$-1 \leq \phi \leq 1$$

Let us write the fluxes explicitly:

$$f_{i+1/2} = \frac{1}{2}(1+\phi_{i+1/2})(a u_i^n) + \frac{1}{2}(1-\phi_{i+1/2})(a u_{i+1}^n)$$

$$f_{i-1/2} = \frac{1}{2}(1+\phi_{i-1/2})(a u_{i-1}^n) + \frac{1}{2}(1-\phi_{i-1/2})(a u_i^n)$$

Now let us apply the data compatibility requirement

$$0 \leq \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i^n} \leq 1$$

where $s = \text{sign}(a) = \text{sign}(c)$

which is physically the requirement that u_i^{n+1} is bounded by the previously present states u_i^n, u_{i-s}^n.

This is a stronger constraint than Harten's theorem, which states that a scheme of the form

$$U_i^{n+1} = U_i^n - C_{i-1/2}\,\Delta U_{i-1/2} + D_{i+1/2}\,\Delta U_{i+1/2}$$

with

$$C_{i+1/2} \geq 0$$

$$D_{i+1/2} \geq 0$$

$$0 \leq C_{i+1/2} + D_{i+1/2} \leq 1$$

will be TVD.

Let us show that the data compatibility is a stricter constraint:

Let $s = +1$, and let us omit the indices $i \pm 1/2$ on the coefficients C and D.

Then the Harten scheme is written as

$$U_i^{n+1} = U_i^n - C(U_i^n - U_{i-s}^n) + D(U_{i+s}^n - U_i^n)$$

$$= U_i^n + C(U_{i-s}^n - U_i^n) + D(U_{i+s}^n - U_i^n)$$

$$\Rightarrow \quad \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i^n} = C + D \frac{u_{i+1}^n - u_i^n}{u_{i-s}^n - u_i^n}$$

$$= C + D \frac{u_{i+1}^n - u_i^n}{u_{i-s}^n - u_i^n} + D - D$$

$$= C + D + D\left(\frac{u_{i+1}^n - u_i^n}{u_{i-s}^n - u_i^n} - 1 \right)$$

$$= C + D + D\left(\frac{u_{i+1}^n - u_{i-s}^n}{u_{i-s}^n - u_i^n} \right)$$

Using the condition

$$0 \leq C + D \leq 1$$

we obtain

$$D \frac{u_{i+1}^n - u_{i-s}^n}{u_{i-s}^n - u_i^n} \leq \frac{u_i^{n+1} - u_i^n}{u_{i-s}^n - u_i^n} \leq 1 + D \frac{u_{i+1}^n - u_{i-s}^n}{u_{i-s}^n - u_i}$$

Comparing with the data compatibility condition

$$0 \leq \frac{U_i^{n+1} - U_i^n}{U_{i-s}^n - U_i^n} \leq 1$$

we see that the values are both restrained over equally sized intervals, but the interval is not fixed for the Harten condition, hence the data compatibility condition is stricter.

Using the explicit expression for the fluxes, the full expression for the WAF scheme is given by

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

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

$$= u_i^n + \frac{c}{2}\left[u_{i-1}^n + \phi_{i-1/2} u_{i-1}^n + u_i^n - \phi_{i-1/2} u_i^n - \right.$$
$$\left. - u_i^n - \phi_{i+1/2} u_i^n - u_{i+1}^n + \phi_{i+1/2} u_{i+1}^n\right]$$

$$= u_i^n + \frac{c}{2}\left[(u_{i-1}^n - u_i^n) - \phi_{i-1/2}(u_i^n - u_{i-1}^n) - \right.$$
$$\left. - (u_{i+1}^n - u_i^n) + \phi_{i+1/2}(u_{i+1}^n - u_i^n)\right]$$

$$= u_i^n + \frac{c}{2}\left[(u_{i-1}^n - u_i^n)(1+\phi_{i-1/2}) + (u_{i+1}^n - u_i^n)(\phi_{i+1/2} - 1)\right]$$

Then

$$u_i^{n+1} - u_i^n = \frac{c}{2}\left[(u_{i-1}^n - u_i^n)(1+\phi_{i-1/2}) + (u_{i+1}^n - u_i^n)(\phi_{i+1/2} - 1)\right]$$

and

$$\frac{u_i^{n+1} - u_i^n}{u_{i-1}^n - u_i^n} = \frac{c}{2}\left[(1+\phi_{i-1/2}) + \frac{u_{i+1}^n - u_i^n}{u_{i-1}^n - u_i^n}(\phi_{i+1/2} - 1)\right]$$

$$= \frac{c}{2}\left[1 + \phi_{i-1/2} + \frac{u_{i+1}^n - u_i^n}{u_i^n - u_{i-1}^n}\left(1 - \phi_{i+1/2}\right)\right]$$

$$= \frac{c}{2}\left[\frac{1}{r_{i+1/2}}\left(1 - \phi_{i+1/2}\right) + \phi_{i-1/2} + 1\right]$$

with $$r_{i+1/2} \equiv \frac{u_i^n - u_{i-1}^n}{u_{i+1}^n - u_i^n}$$

Now let us apply

$$0 \le \frac{u_i^{n+1} - u_i^n}{u_{i-1}^n - u_i^n} =$$

$$= \frac{c}{2}\left[\frac{1}{r_{i+1/2}}\left(1 - \phi_{i+1/2}\right) + \phi_{i-1/2} + 1\right]$$

$$\Rightarrow -1 \le \frac{1}{r_{i+1/2}}\left(1 - \phi_{i+1/2}\right) + \phi_{i-1/2}$$

And for

$$\frac{c}{2}\left[\frac{1}{r_{i+1/2}}(1-\phi_{i+1/2})+\phi_{i-1/2}+1\right]=\frac{u_i^{n+1}-u_i^n}{u_{i-1}^n-u_i^n}\leq 1$$

$$\Rightarrow \frac{1}{r_{i+1/2}}(1-\phi_{i+1/2})+\phi_{i-1/2}\leq \frac{2}{c}-1=\frac{2-c}{c}$$

Finally giving us

$$-1\leq \frac{1}{r_{i+1/2}}(1-\phi_{i+1/2})+\phi_{i-1/2}\leq \frac{2-c}{c}$$

We obtain this condition for both positive and negative advection velocities a, where for $a\leq 0$ we have

$$r_{i+1/2}=\frac{u_{i+2}^n-u_{i+1}^n}{u_{i+1}^n-u_i^n}$$

Put together, we see that

$$r_{i+1/2} = \begin{cases} \dfrac{u_i^n - u_{i-1}^n}{u_{i+1}^n - u_i^n} & a > 0 \\ \dfrac{u_{i+2}^n - u_{i+1}^n}{u_{i+1}^n - u_i^n} & a < 0 \end{cases}$$

and $r_{i+1/2}$ is always the ratio of the upwind change to the local change.

$r_{i+1/2}$ will now be regarded as a flow parameter that will cause ϕ to adjust to the local conditions on the data.

What is left to do is to find a suitable range of values for $\phi_{i+1/2}$ as a function of $r_{i+1/2}$ and $|c|$ such that the scheme will be TVD.

To this end we select two inequalities, one for $\phi_{i+1/2}$ and one for $\phi_{i-1/2}$, so that both inequalities

$$-1 \leq \frac{1}{r_{i+1/2}} \left(1 - \phi_{i+1/2}\right) + \phi_{i-1/2} \leq \frac{2 - |c|}{|c|}$$

are satisfied.

[the $|c|$ enters the inequality so that it will be valid for both positive and negative wave speeds a, and only the definition of $r_{i+1/2}$ changes.]

Let us introduce L into the inequalities:

Let

[redacted]

$$L \leq \phi_{i-1/2} \leq 1$$

such that when we add these inequalities to

$$-1-L \leq \frac{1}{r_{i+1/2}}\left(1-\phi_{i+1/2}\right) \leq \frac{2(1-|c|)}{|c|}$$

$\phi_{i-1/2} \leq 1$ comes from Harten's condition, where we have $0 \leq C_{i-1/2} = \frac{1}{2}(1+\phi_{i-1/2}) \leq 1$ for the extreme case $D_{i+1/2} = 0$
(The second condition actually is $0 \leq C+D \leq 1$)

Now let's study this inequality in detail, and start with the LHS:

$$-(1+L) \leq \frac{1}{r}(1-\phi)$$

- if $r > 0$:

$$-(1+L)\,r \leq (1-\phi)$$

$$\Rightarrow \phi \leq 1+(1+L)\,r \equiv \phi_L(r)$$

- if $r < 0$:

$$-(1+L)\,r \geq (1-\phi)$$

$$\Rightarrow \phi \geq 1+(1+L)\,r \equiv \phi_L(r)$$

For the right inequality:

$$\frac{1}{r}(1-\phi) \leq \frac{2(1-|c|)}{|c|}$$

- if $r > 0$:

$$\phi \geq 1 - \frac{2(1-|c|)}{|c|} r \equiv \phi_R(r)$$

- if $r < 0$:

$$\phi \leq 1 - \frac{2(1-|c|)}{|c|} r \equiv \phi_R(r)$$

To sum up all our conditions:

1) Global boundaries:

$$L \leq \phi$$

$$-1 \leq \phi \leq 1$$

2) $r > 0$:

$$1 - \frac{2(1-|c|)}{|c|} r \leq \phi \leq 1 + (1+L)r$$

3) $r < 0$:

$$1 + (1+L)r \leq \phi \leq 1 - \frac{2(1-|c|)}{|c|} r$$

If these conditions are met, the method will be TVD.

We impose however one more condition:

$$\phi(r=1) = |c|$$

This ensures second order accuracy for values of r close to 1, i.e. when the upwind change is comparable to the local change, i.e. when the local situation is smooth.

Respecting these conditions, we can let creativity flow and find a multitude of flux limiters, like

Superbee: $\phi = \begin{cases} 1 & \text{if } r \le 0 \\ 1-2(1-|c|)r & \text{if } 0 \le r \le 1/2 \\ |c| & \text{if } \frac{1}{2} \le r \le 1 \\ 1-(1-|c|)r & \text{if } 1 \le r \le 2 \\ 2|c|-1 & \text{if } r \ge 2 \end{cases}$

minmod: $\phi = \begin{cases} 1 & \text{if } r \le 0 \\ 1-(1-|c|)r & \text{if } 0 \le r \le 1 \\ |c| & \text{if } r \ge 1 \end{cases}$

van Leer: $\phi = \begin{cases} 1 & \text{if } r \le 0 \\ 1 - \frac{2(1-|c|)r}{1+r} & \text{if } r \ge 0 \end{cases}$

and many others.

▽ Note however that these expressions only hold for the WAF scheme ▽

7.2 The General Flux Limiter Approach

A general flux-limiter approach can be used to construct high-order TVD schemes. It requires a high-order flux $f^{HI}_{i+1/2}$ associated with a scheme of accuracy greater than one and a low-order flux $f^{LO}_{i+1/2}$ associated with a monotone first-order scheme.

For a model conservation law

$$\partial_t u + \partial_x f(u) = 0$$

as solved by

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2}\right]$$

we define a high-order TVD flux as

$$f_{i+1/2}^{TVD} = f_{i+1/2}^{LO} + \phi_{i+1/2}\left[f_{i+1/2}^{HI} - f_{i+1/2}^{LO}\right]$$

where $\phi_{i+1/2}$ is a flux limiter function yet to be determined.

Suppose

$$f_{i+1/2}^{LO} = \alpha_0\, a u_i^n + \alpha_1\, a u_{i+1}^n$$

$$f_{i+1/2}^{HI} = \beta_0\, a u_i^n + \beta_1\, a u_{i+1}^n$$

different choices for α and β give us various familiar methods.

This gives as the expression

$$f_{i+1/2}^{TVD} = f_{i+1/2}^{LO} + \phi_{i+1/2}\left[f_{i+1/2}^{HI} - f_{i+1/2}^{LO}\right]$$

$$= \alpha_0\, a u_i^n + \alpha_1\, a u_{i+1}^n + \phi_{i+1/2}\Big[\beta_0\, a u_i^n + \beta_1\, a u_{i+1}^n - \alpha_0\, a u_i^n - \alpha_1\, a u_{i+1}^n\Big]$$

$$= \left[\alpha_0 + \phi_{i+1/2}(\beta_0 - \alpha_0)\right] a u_i^n + \left[\alpha_1 + \phi_{i+1/2}(\beta_1 - \alpha_1)\right] a u_{i+1}^n$$

and finally

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\left[f_{i-1/2} - f_{i+1/2} \right]$$

$$= u_i^n + \frac{\Delta t}{\Delta x}\Big[\left[\alpha_0 + \phi_{i-1/2}(\beta_0 - \alpha_0)\right] a\, u_{i-1}^n +$$

$$+ \left[\alpha_1 + \phi_{i-1/2}(\beta_1 - \alpha_1)\right] a\, u_i^n -$$

$$- \left[\alpha_0 + \phi_{i+1/2}(\beta_0 - \alpha_0)\right] a\, u_i^n -$$

$$- \left[\alpha_1 + \phi_{i+1/2}(\beta_1 - \alpha_1)\right] a\, u_{i+1}^n \Big]$$

$$= u_i^n + C\Big[\left(\alpha_0 + \phi_{i-1/2}(\beta_0 - \alpha_0)\right)(u_{i-1}^n - u_i^n) +$$

$$\left(\alpha_1 + \phi_{i+1/2}(\beta_1 - \alpha_1)\right)(u_i^n - u_{i+1}^n) \Big]$$

Comparing with the Harten way of writing a scheme:

$$u_i^{n+1} = u_i^n - C(u_i - u_{i-1}) + D(u_{i+1} - u_i)$$

shows

$$C = c\left[\alpha_0 + (\beta_0 - \alpha_0)\,\phi_{i-1/2}\right]$$

$$D = -c\left[\alpha_1 + (\beta_1 - \alpha_1)\,\phi_{i+1/2}\right]$$

Note: I cheated a bit with the indices $i \pm 1/2$ for the flux limiter function ϕ as it doesn't really matter at this point.

7.2.1 TVD Upwind Flux Limiter Schemes

How to construct a TVD Upwind Flux limiter Scheme:

- Pick a low-order scheme and a high order scheme to set the α and β coefficients
- Use Harten's TVD condition to set inequalities
- Impose a lower and upper boundary for the flux limiter, ψ_S and ψ_T, depending on what methods you want to be able to reproduce in limiting cases, i.e. what boundary schemes you want.
- Using these inequalities, find the TVD regions for which $\phi(r)$ results in a TVD method
- Pick something such that $\phi(r)$ is in the TVD region for all r

8. Slope Limiter Methods

The MUSCL approach allows for construction of higher order schemes by replacing the assumption of piecewise linear states $\{u_i^n\}$ with higher order interpolations, e.g. a piecewise linear description:

$$u_i^n(x) = u_i^n + \frac{x - x_i}{\Delta x} \Delta_i$$

where $x \in [0, \Delta x]$ and $x_i = \frac{1}{2}\Delta x$ in local cell coordinates.

Now we construct non-linear versions of these schemes by replacing the slopes Δ_i in the data reconstruction step by limited slopes $\bar{\Delta}_i$ according to some TVD constraints.

8.1 TVD Conditions

Apparently restricting the boundary extrapolated values u_i^L, u_i^R to satisfy

$$\min\{u_{i-1}^n, u_i^n\} \le u_i^L \le \max\{u_{i-1}^n, u_i^n\}$$

$$\min\{u_i^n, u_{i+1}^n\} \le u_i^R \le \max\{u_i^n, u_{i+1}^n\}$$

does not lead to useful results.

Instead, we impose a restriction on the evolved boundary extrapolated values $\bar{u}_i^L$, $\bar{u}_i^R$. Recall that many MUSCL-type schemes try to avoid the arising generalized Riemann problem, and one of the ways of doing that is to first evolve the states as e.g.

$$\bar{u}_i^L = u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

$$\bar{u}_i^R = u_i^R + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

Theorem: If the evolved boundary extrapolated values $\{\bar{u}_k^{L,R}\}$ satisfy

$$\begin{cases} \min\{u_{i-1}^n, u_i^n\} \le \bar{u}_i^L \le \max\{u_{i-1}^n, u_i^n\} \quad \forall i \\ \min\{u_i^n, u_{i+1}^n\} \le \bar{u}_i^R \le \max\{u_i^n, u_i^{nn}\} \quad \forall i \end{cases}$$

then for any monotone scheme of the form

$$v_i^{n+1} = \sum_{k=-1}^{1} b_k v_{i+k}^n$$

applied to $\{\bar{u}_k^{L,R}\}$ one has

$$\min_k \{u_k^n\}_{k=i-2}^{i+2} \le u_i^{n+1} \le \max_k \{u_k^n\}_{k=i-1}^{i+2}$$

Hence the corresponding scheme is TVD.

Proof: Emitted.

8.2 Construction of TVD slopes

Now we construct limited slopes $\bar{\Delta}_i$ to replace Δ_i.

If the limited slopes $\bar{\Delta}_i$ are chosen according to

$$\bar{\Delta}_i = \frac{1}{2}\left[\text{sign}(\Delta_{i-1/2}) + \text{sign}(\Delta_{i+1/2})\right] \times \left[\min\left(\beta_{i-1/2}|\Delta_{i-1/2}|, \beta_{i+1/2}|\Delta_{i+1/2}|\right)\right]$$

with $\beta_{i-1/2} = \frac{2}{1+c}$, $\beta_{i+1/2} = \frac{2}{1-c}$

then the resulting MUSCL scheme is TVD.

Proof:

a) The TVD condition is

$$\min\{u_{i-1}^n, u_i^n\} \le \bar{u}_i^L \le \max\{u_{i-1}^n, u_i^n\} \quad \forall i$$

b) We have

$$\bar{u}_i^L = u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}\left[f(u_i^L) - f(u_i^R)\right]$$

$$= u_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x} a\left[u_i^L - u_i^R\right]$$

$$= u_i^n - \frac{1}{2}\bar{\Delta}_i + 1/2\,\frac{\Delta t}{\Delta x} a\left[u_i^n - \frac{1}{2}\bar{\Delta}_i - u_i^n - \frac{1}{2}\bar{\Delta}_i\right]$$

$$= u_i^n - \frac{1}{2}(1+c)\bar{\Delta}_i$$

Then from a) we get when we put b) into it:

Assume $u_{i-1}^u \leq u_i^u$

Then

$$u_{i-1}^u \leq \bar{u}_i^L = u_i^u - \frac{1}{2}(1+c)\bar{\delta}_i \leq u_i^u$$

First inequality:

$$u_{i-1}^u \leq u_i^u - \frac{1}{2}(1+c)\bar{\delta}_i$$

$$u_{i-1}^u - u_i^u = -\Delta u_{i-1/2} \leq -\frac{1}{2}(1+c)\bar{\delta}_i$$

$$\frac{2}{1+c}\Delta u_{i-1/2} \geq \bar{\delta}_i$$

Second inequality:

$$u_i^u - \frac{1}{2}(1+c)\bar{\delta}_i \leq u_i^u$$

$$-\frac{1}{2}(1+c)\bar{\delta}_i \leq u_i^u - u_i^u = 0$$

$$\bar{\delta}_i \geq 0$$

$$\Rightarrow \quad 0 \leq \bar{\delta}_i \leq \frac{2}{1+c}\Delta u_{i-1/2} \quad \text{for} \quad \Delta u_{i-1/2} \geq 0$$

i.e. $u_i \geq u_{i-1}$

We can analogously obtain

$$\frac{2}{1+c}\Delta_{i-1/2} \le \bar{\Delta}_i \le 0 \quad \text{for } \Delta_{i-1/2} \le 0$$

and for the pair (u_{i+1}, u_i), we get

$$0 \le \bar{\Delta}_i \le \frac{2}{1-c}\Delta u_{i+1/2} \quad \Delta u_{i+1/2} \ge 0$$

$$\frac{2}{1-c}\Delta u_{i+1/2} \le \bar{\Delta}_i \le 0 \quad \Delta u_{i+1/2} \le 0$$

So we have 4 cases:

$\Delta u_{i+1/2} \ge 0$, $\Delta u_{i-1/2} \le 0$:

$$\bar{\Delta}_i \le 0, \quad \bar{\Delta}_i \ge 0 \quad \Rightarrow \bar{\Delta}_i = 0$$

$\Delta u_{i+1/2} \le 0$, $\Delta u_{i-1/2} \ge 0$:

$$\bar{\Delta}_i \ge 0, \quad \bar{\Delta}_i \le 0 \quad \Rightarrow \bar{\Delta}_i = 0$$

$\Rightarrow$ If $\text{sign}(\Delta u_{i+1/2}) \neq \text{sign}(\Delta_{i-1/2})$, $\bar{\Delta}_i = 0$

we can write that as

$$\frac{1}{2}\left(\text{sign}(\Delta u_{i+1/2}) + \text{sign}(\Delta u_{i-1/2})\right)$$

and add it as a multiplication factor that gives 0 if the signs are different and 1 if the signs are the same.

The following two cases remain:

$\Delta u_{i+1/2} \geq 0$, $\Delta u_{i-1/2} \geq 0$:

$$\bar{\sigma}_i \leq \frac{2}{1+c}\Delta u_{i+1/2}, \quad \bar{\sigma}_i \leq \frac{2}{1+c}\Delta u_{i-1/2}$$

$$\Rightarrow \bar{\sigma}_i \leq \min\left\{\frac{2}{1-c}\Delta u_{i+1/2}, \frac{2}{1+c}\Delta u_{i-1/2}\right\}$$

$\Delta u_{i+1/2} \leq 0$, $\Delta u_{i-1/2} \leq 0$:

$$\bar{\sigma}_i \geq \frac{2}{1-c}\Delta u_{i+1/2}, \quad \bar{\sigma}_i \geq \frac{2}{1+c}\Delta u_{i-1/2}$$

$$\Rightarrow \bar{\sigma}_i \leq \frac{2}{1-c}|\Delta u_{i+1/2}|, \quad \bar{\sigma}_i \leq \frac{2}{1+c}|\Delta u_{i-1/2}|$$

Since $\Delta u_{i\pm 1/2} \leq 0$

$$\Rightarrow \bar{\sigma}_i \leq \min\left\{\frac{2}{1-c}|\Delta u_{i+1/2}|, \frac{2}{1+c}\Delta u_{i-1/2}|\right\}$$

$\Rightarrow$ We can combine both conditions into

$$\bar{\sigma}_i \leq \min\left\{\frac{2}{1-c}|\Delta u_{i+1/2}|, \frac{2}{1+c}|\Delta u_{i-1/2}|\right\}$$

And finally, if we replace $\leq$ with $=$, we arrive at

$$\bar{\sigma}_i = \frac{1}{2}\left(\mathrm{sign}(\Delta u_{i+1/2}) + \mathrm{sign}(\Delta u_{i-1/2})\right) \times \min\left\{\beta_{i+1/2}|\Delta u_{i+1/2}|, \beta_{i-1/2}|\Delta u_{i-1/2}|\right\}$$

8.3 Slope Limiters

Using the TVD analysis that we just completed, we can construct upwind and centered slope limiter methods.

Let us express

$$\bar{\Delta}_i = \xi_i \Delta_i$$

with

$$\Delta_i = \frac{1}{2}(1+\omega)\Delta u_{i-1/2} + \frac{1}{2}(1-\omega)\Delta u_{i+1/2}$$

Recall that in the derivation of the TVD condition for $\bar{\Delta}_i$ we used

$$\Delta_{i+1/2} = u_{i+1}^n - u_i^n = \Delta u_{i+1/2}$$

but are using a more general case now; Let us re-write the condition properly first.

We obtain a TVD region for $\xi(r)$ as follows:

Let $r = \dfrac{\Delta u_{i-1/2}}{\Delta u_{i+1/2}}$

and let us express the TVD condition in units of Δi:

First factor:

$$\frac{1}{2}\left(\text{sign}\left(\Delta u_{i+1/2}\right) + \text{sign}\left(\Delta u_{i-1/2}\right)\right)$$

is equivalent to

$$\xi(r) = 0 \quad \text{for} \quad r < 0$$

Second part: Assume $\beta_{i+1/2}|\Delta u_{i+1/2}| > \beta_{i-1/2}|\Delta u_{i-1/2}|$

Then

$$\bar{\Delta}_i = \xi(r)\Delta_i = \beta_{i-1/2}\left|\Delta u_{i-1/2}\right|$$

$$\Rightarrow \xi(r) = \frac{\beta_{i-1/2}\,\Delta u_{i-1/2}}{\Delta i}$$

$$= \frac{\beta_{i-1/2}\,\Delta u_{i-1/2}}{\frac{1}{2}(1+\omega)\Delta u_{i-1/2} + \frac{1}{2}(1-\omega)\Delta u_{i+1/2}}$$

$$= \frac{2\beta_{i-1/2}}{1 + \omega + (1-\omega)\,1/r}$$

$$= \frac{2\beta_{i-1/2}\,r}{1-\omega + (1+\omega)r}$$

Now suppose we have

$$\beta_{i+1/2}\,|\Delta u_{i+1/2}| < \beta_{i-1/2}\,|\Delta u_{i-1/2}|$$

Then

$$\xi(r) = \frac{\beta_{i+1/2}\,\Delta u_{i+1/2}}{\Delta i}$$

$$= \frac{2\beta_{i+1/2}\,\Delta u_{i+1/2}}{(1+\omega)\Delta u_{i-1/2} + (1-\omega)\Delta u_{i+1/2}}$$

$$= \frac{2\,\beta_{i+1/2}}{1-\omega + (1+\omega)r}$$

So finally we may express the TVD condition as

$$\bar{\Delta}_i = \xi(r)\,\Delta_i$$

$$\Delta_i = \frac{1}{2}(1+\omega)\Delta u_{i-1/2} + \frac{1}{2}(1-\omega)\Delta u_{i+1/2}$$

$$\xi(r) = \begin{cases} 0 & \text{for } r \le 0 \\ \min\{\xi_L(r), \xi_R(r)\} & \text{for } r \ge 0 \end{cases}$$

$$\xi_L(r) = \frac{2\beta_{i-1/2}\, r}{1-\omega+(1+\omega)r}$$

$$\xi_R(r) = \frac{2\beta_{i+1/2}}{1-\omega+(1+\omega)r}$$

$$r = \frac{\Delta u_{i-1/2}}{\Delta u_{i+1/2}}$$

With the TVD region defined this way, we can find slope limiters that follow these rules. Examples are

$$\text{Superbee:} \quad \xi(r) = \begin{cases} 0 & r \le 0 \\ 2r & 0 \le r \le 1/2 \\ 1 & 1/2 \le r \le 1 \\ \min\{r, \xi_R(r), 2\} & r \ge 1 \end{cases}$$

$$\text{van Leer:} \quad \xi(r) = \begin{cases} 0 & r \le 0 \\ \min\left\{\frac{2r}{1+r}, \xi_R(r)\right\} & r \ge 0 \end{cases}$$

$$\text{minbee:} \quad \xi(r) = \begin{cases} 0 & r \le 0 \\ r & 0 \le r \le 1 \\ \min\{1, \xi_R(r)\} & r \ge 1 \end{cases}$$

8.4 Limited Slopes Obtained From Flux Limiters

For the model linear advection equation, we can establish a connection between upwind-based flux limiter schemes and MUSCL-type schemes.

This can be accomplished by selecting limited slopes $\bar{D}_i$ in the reconstruction step to reproduce conventional upwind flux-limiters $\psi_{i+1/2}$.

Let us construct the TVD flux using

$$f_{i+1/2}^{TVD} = f_{i+1/2}^{LO} + \psi_{i+1/2}\left[f_{i+1/2}^{HI} - f_{i+1/2}^{LO}\right]$$

and use the upwind Godunov low order flux:

$$f^{LO}_{i+1/2} = \frac{1}{2}(1+s)a u_i^n + \frac{1}{2}(1-s)a u_{i+1}^n$$

$$s = \text{sign}(a)$$

and the Lax-Wendroff flux for the high order flux:

$$f^{HI}_{i+1/2} = \frac{1}{2}(1+c)a u_i^n + \frac{1}{2}(1-c)a u_{i+1}^n$$

Then we have

$$f^{HI}_{i+1/2} - f^{LO}_{i+1/2} = \frac{a}{2}\left[u_i^n(1+c-1-s) + u_{i+1}^n(1-c-1+s)\right]$$

$$= \frac{a}{2}\left[u_i^n(c-s) + (s-c)u_{i+1}^n\right]$$

$$= \frac{a}{2}(s-c)(u_{i+1}^n - u_i^n)$$

$$= \frac{a}{2}(s-c)\,\Delta u_{i+1/2}$$

So the TVD flux for the scheme is

$$f_{i+1/2} = \begin{cases} a u_i^n + \psi_{i+1/2} \frac{1}{2}(1-c)\Delta u_{i+1/2} & a > 0 \\ a u_{i+1}^n - \psi_{i+1/2} \frac{1}{2}(1+c)\Delta u_{i+1/2} & a < 0 \end{cases}$$

Now recall that an upwind based slope limiter method has the intercell flux

$$f_{i+1/2} = \begin{cases} a u_i^n + \frac{1}{2}(1-c) a \bar{\Delta}_i & a > 0 \\ a u_{i+1} - \frac{1}{2}(1+c) a \bar{\Delta}_{i+1} & a < 0 \end{cases}$$

So we can immediately see that

$$\psi_{i+1/2} \Delta u_{i+1/2} = \begin{cases} \bar{\Delta}_i & a > 0 \\ \bar{\Delta}_{i+1} & a < 0 \end{cases}$$