1. The Scalar Conservation Law

1.1 Introduction and smooth solution

In this text we consider the initial value problem

$$\begin{aligned} &u_t + f(u)_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = u_0(x) \end{aligned} \tag{1.1}$$

where the function $u(t,x)$ is the unknown and $f(u)$ and $u_0(x)$ are given functions.

It is a generalization of the hyperbolic problem

$$\begin{aligned} &u_t + au_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = u_0(x) \end{aligned} \tag{1.2}$$

with which the reader is supposed to be familiar. Problem (1.2) is usually analyzed using Fourier series. Since problem (1.1) is in general non linear, Fourier methods can not be used.

The choice $f(u) = u^2/2$ yields the inviscid Burger's equation, an equation interesting because of its resemblance to the equations of fluid dynamics. It is widely used as a model problem.

The equation $u_t + f(u)_x = 0$ is called a *conservation law.* By integrating over $-\infty < x < \infty$ one gets

$$\frac{d}{dt}\int_{-\infty}^{\infty} u(x,t)\,dx = 0$$

assuming that $f(u)$ vanishes as $|x| \to \infty$. Thus the name derives from the fact that the integral of u is conserved in time.

The function $f(u)$ is called *flux function.* By integrating over $a < x < b$ one gets

$$\frac{d}{dt}\int_{a}^{b} u(x,t)\,dx = f(u(t,a)) - f(u(t,b)) \tag{1.3}$$

which can be given the interpretation that the integral of u over a finite interval can change due to in- or outflow at the boundaries $x = a$ and $x = b$.

If we carry out the x differentiation we get

$$u_t + a(u)u_x = 0$$

where $a(u) = f'(u)$. In the same way as for problem (1.2), we can make the definition

Definition 1.1. *The characteristics are the curves in the x-t plane defined by*

$$dx(t)/dt = a(u(t,x(t))) \tag{1.4}$$

We have a theorem similar to the one for the linear case.

Theorem 1.2. *If the solution $u(t, x)$ is differentiable, it is constant along the characteristics.*

Proof: The chain rule is used to evaluate the derivative of u along a characteristic curve

$$\frac{du(t, x(t))}{dt} = u_t + \frac{dx(t)}{dt} u_x = u_t + a(u) u_x = 0$$

using (1.4). The derivative is zero and the solution constant.

The theorem and (1.4) implies that the characteristics are straight lines. The following theorem further shows that there are many similarities between (1.1) and (1.2).

Theorem 1.3. *The solution, u, to problem (1.1) satisfies*

$$u = u_0(x - a(u)t) \tag{1.5}$$

if it is differentiable.

Proof: Insert (1.5) into the PDE and use the chain rule. The result from doing this is

$$(1 + u_0'(x - a(u)t)a'(u)t)(u_t + a(u)u_x) = 0$$

We differentiate (1.5) with respect to time and obtain

$$u_t = u_0'(x - a(u)t)(-a'(u)tu_t - a(u))$$

Solve for u_t

$$u_t = -\frac{u_0' a}{1 + u_0' a' t}$$

Since we assume that u has continous derivative, the denominator $1 + u_0' a' t$ must be different from zero, and thus the factor multiplying $u_t + a(u)u_x$ can be divided out and the proof is complete.

If the above non linear algebraic equation has a unique solution, a very efficient solution procedure for problem (1.1) is to solve (1.5) by Newton's method.

1.2 Non smoothness, Jump condition

The major difference between the linear and the non linear equations is that for the latter, the solution in the class of continuous functions may fail to exist after a finite time, no matter how smooth the initial data are. We give three examples to show how this failure occurs.

Example 1.1 (Geometric description of smoothness failure)

$$\begin{aligned} &u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = \sin x \end{aligned}$$

By differentiation $a(u) = u$ and thus the slope of the characteristics are u. Initially in the point $x = \pi/2$, the slope and the solution are 1 and in the point $x = 3\pi/2$ the slope and function are -1.

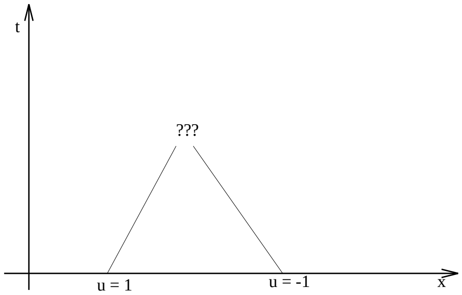


Figure 1.1. Values are transported along the characteristics

The value 1 is transported to the right and the value -1 to the left, at some time they will meet, thereby causing a failure of smoothness in the solution.

Example 1.2 (Algebraic description of smoothness failure) Consider (1.5). By implicit differentiation with respect to t we get

$$u_t = u_0'(x - a(u)t)(-a'(u)tu_t - a(u))$$

Solve for u_t

$$u_t = -\frac{u_0' a}{1 + u_0' a' t} \tag{1.6}$$

if $a'u_0'$ is < 0 at some point, we see from (1.6) that there will be a blow up of the derivative at $t = -1/(u_0'a')$.

Example 1.2 shows that under certain conditions, such as e.g. $a'(u) > 0$ and $u_0'(x) > 0$, a smooth solution does exist.

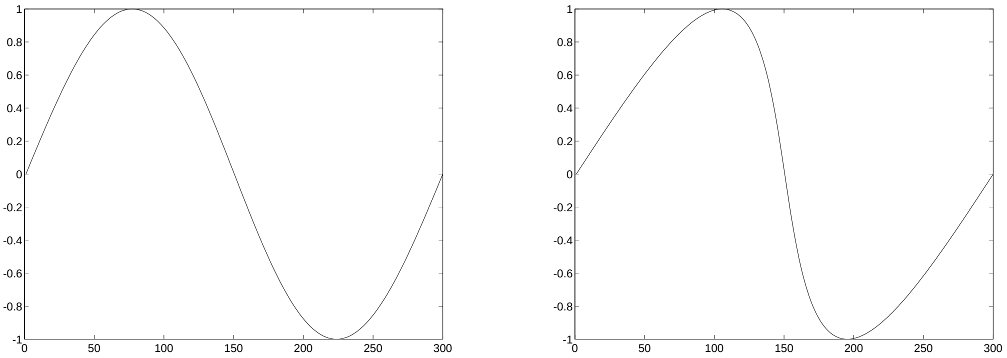
Example 1.3 (Dynamic description of smoothness failure) The same problem as in example 1.1 is considered

$$\begin{aligned} &u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = \sin x \end{aligned}$$

The differential equation can be written

$$u_t + uu_x = 0$$

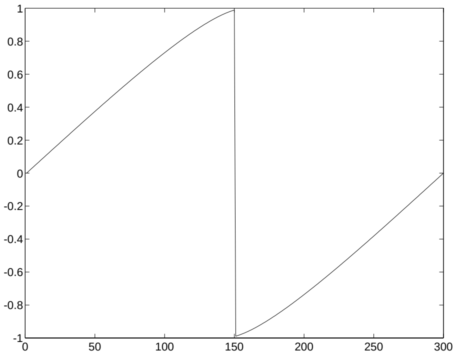
and u can, in analogy with the linear hyperbolic equation, be interpreted as the speed with which the initial data propagates. For the sine wave below, the maxima travels to the right with speed 1 and the minima to the left with speed -1. This causes a gradual sharpening of the gradients with time,



a) at t = 0 b) at t=0.1

Fig. 1.2. A solution to Burgers' equation.

and finally the waves break into discontinuities.



c) at t = 0.5

The examples shows the necessity to extend the solutions into the class of functions with discontinuities. The partial differential equation does not make sense for non differentiable functions. We can however interpret the derivatives in the sense of distributions. More specifically this means that the equation is multiplied by a smooth test function, $\varphi \in C_o^\infty(\mathbf{R}^+ \times \mathbf{R})$, and then integrated in time and space. Integration by parts afterwards moves the derivatives to the smooth test functions. Doing this yields

$$\int_0^\infty \int_{-\infty}^\infty \varphi_t u + \varphi_x f(u)\,dx\,dt + \int_{-\infty}^\infty \varphi(0,x)u_0(0,x)\,dx = 0 \tag{1.7}$$

The boundary terms at $t, |x| = \infty$ does not contribute, since φ is assumed to have compact support. We define

Definition 1.4. *A weak solution to (1.1) is a function $u(t,x)$ satisfying (1.7) for all smooth test functions $\varphi \in C_0^\infty$.*

In the specific case of one discontinuity, separating two smooth parts of the solution we can use the conservation property of the original problem (1.1) to obtain the following theorem.

Theorem 1.5. *(Rankine-Hugoniot) Assume that a discontinuity is moving with speed s and that the value of u to the left of the jump is u_L and to the right u_R. The the following holds*

$$s(u_L - u_R) = f(u_L) - f(u_R)$$

Proof: Use the integrated form (1.3)

$$\frac{d}{dt}\int_a^b u\,dx = f(u(t,a)) - f(u(t,b)) \tag{1.8}$$

assume there is one discontinuity moving on the curve $x(t)$ and that the solution is smooth otherwise. Separate (1.8) into smooth parts

$$\frac{d}{dt}(\int_a^{x(t)} u\,dx + \int_{x(t)}^b u\,dx) = f(u(t,a)) - f(u(t,b))$$

The differentiation can now be carried out, giving

$$\int_a^{x(t)} u_t\,dx + u(t,x(t)-)x'(t) + \int_{x(t)}^b u_t\,dx - u(t,x(t)+)x'(t) = f(u(t,a)) - f(u(t,b))$$

Now use $u_t = -f_x$ in the integrals. Performing the integration gives

$$\begin{aligned} f(u(t,a)) - f(u(t,x(t)-)) + u(t,x(t)-)x'(t) + f(u(t,x(t)+)) - \\ f(u(t,b)) - u(t,x(t)+)x'(t) = f(u(t,a)) - f(u(t,b)) \end{aligned}$$

The desired result is obtained by rearranging this expression, and using the notations $u(t,x(t)-) = u_L$, $u(t,x(t)+) = u_R$, $x'(t) = s$.

1.3. Uniqueness, Entropy condition

When we extend the class of admissible solution from the differentiable functions to non differentiable functions, we unfortunately loose uniqueness. The extended class of functions is too large.

We therefore impose an extra condition the so called *entropy condition* which tells us, in case of multiple solutions, which solution is the correct one. The name derives from application to gas dynamics, in which case there is only one solution satisfying the physically correct condition of entropy decrease.

As we will see later, entropy conditions are important when we study numerical methods, since some convergent numerical methods does not converge to the solution singled out by the entropy condition.

The theory is considerably simplified if the flux function is convex ($f''(u) > 0$). Therefore we start with that case. The typical example of non uniqueness is the following

Example 1.4 Two possible solutions to the problem

$$u_t + (u^2/2)_x = 0 \qquad -\infty < x < \infty \;\; 0 < t$$
$$u(0,x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

are

$$u_1(t,x) = \begin{cases} 0 & x < t/2 \\ 1 & x \geq t/2 \end{cases}$$

The jump is moving with the speed $s = 1/2$ obtained from the Rankine-Hugoniot condition, and

$$u_2(t,x) = \begin{cases} 0 & x < 0 \\ x/t & 0 \leq x \leq t \\ 1 & x > t \end{cases}$$

The second solution is a so called expansion wave (or rarefaction wave). It is easy to see that these functions solves the problem, by inserting them into the differential equation. By looking at the characteristics in the $x - t$ plane, we get the following picture of the solution 1 in the example above

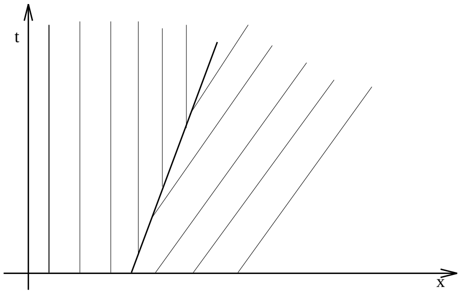


Fig. 1.3. Diverging characteristics.

This solution is not a good one for the following reasons

1. Sensitivity to perturbations. A small disturbance in the discontinuity will propagate out into the solution and affect the smooth parts.

2. There are characteristics emanating from the discontinuity. We would like the solution to be determined by the initial data. Consequently, if at some time t we trace a characteristics backwards we should end at some point at the time zero. This is not true for this solution.

   For the following example point one and two are resolved in a satisfactory way.

   Example 1.5 The problem

$$u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t$$
$$u(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x > 0 \end{cases}$$

has a solution

$$u(t,x) = \begin{cases} 1 & x < t/2 \\ 0 & x \geq t/2 \end{cases} .$$

The jump is moving with the speed $s = 1/2$ obtained from the Rankine-Hugoniot condition. The characteristics are pointing into the jump

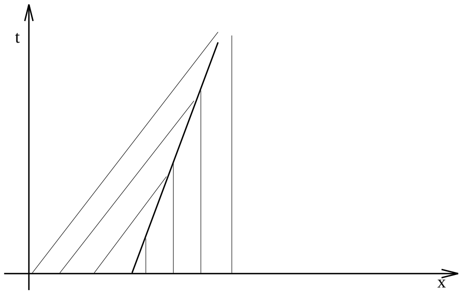


Fig. 1.4. Converging characteristics.

Here a small disturbance in the jump will immediately disappear into the discontinuity, and at a given time, we can always follow a characteristic backwards to time zero.

Example 1.5 gives a motivation for the following definition.

Definition 1.6. *A discontinuity with left state u_L and right state u_R, moving with speed s for a conservation law with convex flux function is entropy satisfying if*

$$f'(u_L) > s > f'(u_R) \tag{1.9}$$

This means that the characteristics are going towards the discontinuity as time increases.

An entropy satisfying discontinuity is also called a *shock*. The significance of the above definition can be seen in the following theorem

Theorem 1.7. *The initial value problem (1.1) with convex flux function and arbitrary integrable initial data has a unique weak solution in the class of functions satisfying (1.9) across all jumps.*

Proof: The proof of existence uses the exact solution formula which we describe in the next section. We refer to [17] for the details, and the uniqueness.

For the non convex conservation law the condition (1.9) has to be satisfied for all u between u_L and u_R. The flux function could look like in fig. 1.5.

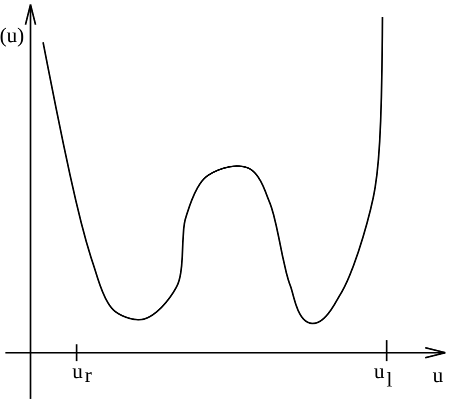


Fig. 1.5. Non convex flux function.

Here a jump between u_L and u_R, satisfies condition (1.9), but is still not the correct solution. It has turned out that it is necessary to require the following entropy condition for a general non convex conservation law

$$\frac{f(u_L)-f(u)}{u_L-u} \geq \frac{f(u_R)-f(u_L)}{u_R-u_L} \quad \text{all } u \in [u_L,u_R] \text{ or } [u_R,u_L] \tag{1.10}$$

It is important that all values u between u_L and u_R are involved. Intuitively we can understand (1.10) as requiring the characteristics to go into the shock for the entire family of shocks between u_L and u, $u \in [u_L, u_R]$. Geometrically (1.10) can be interpreted as the graph $u - f(u)$ must lie below the chord between $(u_L, f(u_L))$ and $(u_R, f(u_R))$ if $u_L > u_R$, and above if $u_L < u_R$. (1.10) can be derived from the inviscid limit of the problem

$$u_t + f(u)_x = \epsilon u_{xx} \tag{1.11}$$

where ϵ is a positive parameter. (1.11) has a unique smooth solution. The physically relevant solution of (1.1) is defined as the solution of (1.11) as $\epsilon \to 0$. We give a derivation of (1.10) later in this section.

There is a result similar to theorem 1.7 for the entropy condition (1.10).

Theorem 1.8. *The initial value problem (1.1) with arbitrary integrable initial data has a unique weak solution in the class of functions satisfying (1.10) across all jumps.*

Proof: Not given here. We refer to [16].

An example of a conservation law with non convex flux function is the so called Buckley-Leverett equation

$$u_t + (\frac{u^2}{u^2+(1-u)^2/4})_x = 0$$

which occurs in the theory of flow through porous media.

There is an alternative way of getting entropy conditions which we now describe. First the equation

$$u_t + f(u)_x = \epsilon u_{xx}$$

and $E'(u)$ are multiplied. $E(u)$ is a strictly convex ($E''(u) > 0$), differentiable function, which we will call the entropy function.

$$E'u_t + E'f(u)_x = \epsilon E'(u)u_{xx}$$

Define $F'(u) = E'(u)f'(u)$, the equation takes a form similar to the original one

$$E(u)_t + F(u)_x = \epsilon E'(u)u_{xx}.$$

Using the identity

$$E(u)_{xx} = E''(u)(u_x)^2 + E'(u)u_{xx}$$

we rewrite the viscosity term and get

$$E(u)_t + F(u)_x = \epsilon(E(u)_{xx} - E''(u)(u_x)^2) \leq \epsilon E(u)_{xx}$$

where the last inequality follows from the fact that $E(u)$ is convex. Let now $\epsilon \to 0$ and we arrive at

$$E(u)_t + F(u)_x \leq 0 \tag{1.12}$$

where the inequality should be understood to be valid in the sense of distributions.

Thus we have showed that if $u(t,x)$ is a solution to the original conservation law, obtained as the vanishing viscosity limit solution of (1.11), the additional inequality (1.12) is valid. As previously mentioned, the vanishing viscosity solution is the physically relevant one which we want an entropy condition to choose for us. As an entropy condition we take (1.12), or across jump discontinuities

$$s(E(u_R) - E(u_L)) - (F(u_R) - F(u_L)) \geq 0 \tag{1.13}$$

which follows from (1.12) by calculations similar to the proof of theorem 1.5. We have now three different entropy conditions, (1.9), (1.10) and (1.13), we finish by investigating the relationship between them. By using the definition $E'f' = F'$ it is easy to prove the identity

$$s(E(u_R) - E(u_L)) - (F(u_R) - F(u_L)) = \int_{u_L}^{u_R} E''(u)(su_L - f(u_L) - (su - f(u)))\,du$$

The function inside the integral is familiar, using the definition of s, the shock speed, we can rewrite entropy condition (1.10) as

$$\begin{array}{ll} su - f(u) \leq su_L - f(u_L) & u_L < u_R \\ su - f(u) \geq su_L - f(u_L) & u_L > u_R \end{array}$$

thus we immediately get (1.10)⇒ (1.13) from

$$\int_{u_L}^{u_R} E''(u)(su_L - f(u_L) - (su - f(u)))\,du \geq 0.$$

and $E''(u) > 0$. For the implication in other direction it is necessary to assume that (1.13) is valid for all convex $E(u)$, or at least a class sufficiently large to assure that

$$\int E''(u)g(u)\,du \geq 0 \Rightarrow g(u) \geq 0.$$

One example of such a class is given in exercise 5. In the special case $f(u)$ convex the sign of $su_L - f(u_L) - (su - f(u))$ does not change over the interval $[u_L, u_R]$, and one convex entropy function is sufficient. Summary:

(1.10) $\Rightarrow$ (1.13) for any convex entropy function.
(1.13) for a "large" class of entropy functions$\Rightarrow$(1.10).
(1.13) with one entropy function $\Leftrightarrow$ (1.9).
(1.10)$\Rightarrow$ (1.9),
(1.9) $\Rightarrow$ (1.10) if $f(u)$ convex.

Here the last two implications are easily shown and left as an exercise

1.3 Exact solution formulas

For reference we here give some analytic solution formulas without proving them. The equation

$$u_t + (u^2/2)_x = \epsilon u_{xx}$$

can be solved exactly [15], the formula is not given here. A similar result has been obtained for the problem

$$\begin{aligned} &u_t + f(u)_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = u_0(x) \end{aligned} \tag{1.14}$$

with $f(u)$ convex. The solution at a fixed point (t, x) is obtained from

Theorem 1.9. *The solution to (1.14) is given by*

$$u(t,x) = b((x-y)/t) \tag{1.15}$$

where $b(u)$ is the inverse function of $f'(u)$, (which exists since $f(u)$ is convex) and y is the value which minimizes ((t, x) are still kept fixed)

$$G(x,y,t) = \int_{-\infty}^{y} u_0(s)\,ds + th((x-y)/t).$$

Here $h(u)$ is a function determined from $h'(u) = b(u)$, and $h(f'(0)) = 0$.

We refer to [17] for a derivation of the formulas.

The problem with piecewise constant initial data, will be of importance to some of the numerical methods encountered later on. In the scalar case it is possible to solve the problem

$$\begin{aligned} &u_t + (f(u))_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = \begin{cases} u_L & x < 0 \\ u_R & x > 0 \end{cases} \end{aligned}$$

analytically for any differentiable flux function $f(u)$. u_L and u_R are constants. First one proves that the solution only depends on x/t. Let the solution be $u(t,x) = u(x/t) = u(\zeta)$. The following formulas then give a closed expression for $u(\zeta)$.

$$\begin{aligned} u(\zeta) &= -\frac{d}{d\zeta}(\min_{w\in[u_L,u_R]}(f(w)-\zeta w)) \qquad u_L < u_R \\ u(\zeta) &= -\frac{d}{d\zeta}(\max_{w\in[u_R,u_L]}(f(w)-\zeta w)) \qquad u_L > u_R \end{aligned} \tag{1.16}$$

The differentiation is made in the sense of distributions. We refer to [20] for a derivation of these formulas.

Exercises

1. In [17] the following entropy condition is given

$$\begin{aligned} f(\alpha u_R + (1-\alpha)u_L) &\le \alpha f(u_R) + (1-\alpha)f(u_L) \quad u_R < u_L \\ f(\alpha u_R + (1-\alpha)u_L) &\ge \alpha f(u_R) + (1-\alpha)f(u_L) \quad u_R > u_L \end{aligned}$$

all $\alpha \in [0,1]$. Show that this entropy condition is equivalent to (1.10). What is geometrical interpretation of the above entropy condition ?

2. The formula (1.5) can not be used to solve the problems

$$\begin{aligned} &u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \ \ 0 < t \\ &u(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x > 0 \end{cases} \end{aligned}$$

and

$$\begin{aligned} &u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \ \ 0 < t \\ &u(0,x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases} \end{aligned} .$$

Try to use it anyway, to investigate how the formula fails. Use then formula (1.15) in the last section to obtain a correct solution.

3. Consider the problem

$$\begin{aligned} &u_t + (f(u))_x = 0 \quad -\infty < x < \infty \ \ 0 < t \\ &u(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x > 0 \end{cases} \end{aligned} .$$

with $f(u) = 1.1u^4 - 2u^3 + u^2$. One possible solution is

$$u(t,x) = \begin{cases} 1 & x < 0.1t \\ 0 & x > 0.1t \end{cases} .$$

Show that this solution satisfies the entropy condition (1.9) but not (1.10).

4. Solve the problem

$$\begin{aligned} &u_t + (u^3/3)_x = 0 \qquad -\infty < x < \infty \;\; 0 < t \\ &u(0,x) = \begin{cases} 1 & x < 0 \\ -1 & x > 0 \end{cases} \end{aligned} \qquad .$$

using the exact formula (1.16).

5. Show that the entropy condition

$$E(u)_t + F(u)_x \leq 0$$

for all

$$\begin{aligned} E(u) &= \begin{cases} u - c & u \geq c \\ 0 & u < c \end{cases} \\ F(u) &= \begin{cases} f(u) - f(c) & u \geq c \\ 0 & u < c \end{cases} \end{aligned}$$

with c a real constant, implies the entropy condition (1.10). Thus instead of requiring (1.13) for all convex $E(u)$, the subclass above can be used.

2. Numerical Methods for the Scalar Conservation Law

2.1 Notations

We will describe some numerical methods applied to a one dimensional problem using a uniform grid. This is for clarity of exposition, the changes required for more space dimensions and curvilinear grids are straightforward.

We consider a discretization of the x axis

$$x_j \quad j = \dots, -2, -1, 0, 1, 2, \dots$$

The uniform spacing is $\Delta x = x_{j+1} - x_j$. We divide the time into time levels $t_0 = 0, t_1, t_2, \dots$. The time step $\Delta t = t_{n+1} - t_n$ will be constant.

We here avoid boundaries by considering the problem on the entire domain $-\infty < x < \infty$. The analysis below could have been done, using periodicity instead, as is usually done in the linear case. In practical computations it is, of course, not possible to use an infinite number of grid points. Thus, in order to verify numerically the results below, it is necessary to use a periodic problem.

The following notations will be used

$$\begin{aligned}
u_j^n &= \text{ The numerical solution at the point } (t_n, x_j) \\
D_+ u_j &= (u_{j+1} - u_j)/\Delta x \\
\Delta_+ u_j &= u_{j+1} - u_j \\
D_- u_j &= D_+ u_{j-1} \\
\Delta_- u_j &= \Delta_+ u_{j-1} \\
D_0 u_j &= \frac{1}{2}(D_+ + D_-)u_j \\
\Delta_0 u_j &= \frac{1}{2}(\Delta_+ + \Delta_-)u_j
\end{aligned}$$

The operators D_+ and D_- approximates $\partial/\partial x$ to first order accuracy, D_0 gives second order accuracy. We write this as

$$\begin{aligned}
D_+ u_j &= u_x(x_j) + O(\Delta x) \\
D_0 u_j &= u_x(x_j) + O(\Delta x^2)
\end{aligned}$$

Thus e=$O(\Delta x^p)$ denotes a quantity which goes to zero with the same rate as Δx^p when Δx goes to zero i.e.

$$0 < C_1 \leq \lim_{\Delta x \to 0} |e|/\Delta x^p \leq C_2$$

with C_1 and C_2 positive constants.

2.2 Definitions and General Results

We now consider the method

$$u_j^{(1)} = u_j^n - \Delta t D_- f(u_j^n)$$
$$u_j^{(2)} = u_j^{(1)} - \Delta t D_+ f(u_j^{(1)})$$
$$u_j^{n+1} = (u_j^{(2)} + u_j^n)/2$$

which approximates the scalar conservation law

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

to second order accuracy in both time and space. The method is popular in computational aerodynamics where it is known as MacCormack's scheme. We use this scheme to demonstrate how the numerical solution can misbehave when the solution to the partial differential equation is discontinuous.

Example 2.1 The solution to the problem

$$u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t$$
$$u(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases}$$

is a translation of the initial step function with velocity 1/2 (from the Rankine Hugoniot condition). The solution satisfies the entropy condition. Below the solution obtained using MacCormack's scheme is displayed. The solid line is the exact solution, the circles are the numerical solution.

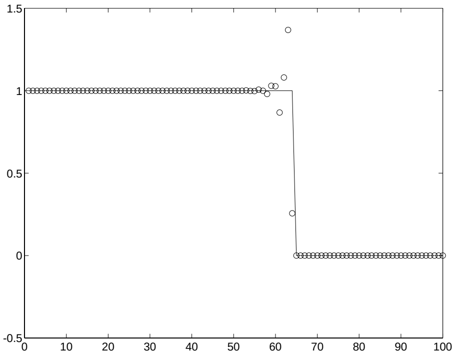


Fig.2.1. MacCormack

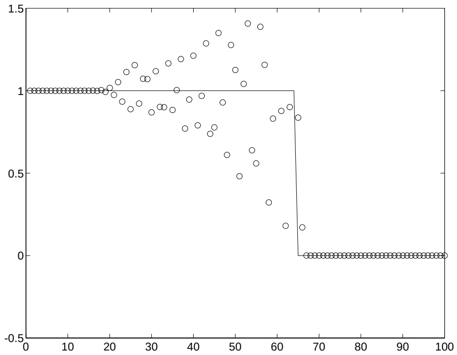


Fig.2.2. Leapfrog

The scheme does not behave well near the shock. The oscillations around the shock are related to the well known Gibb's phenomenon in Fourier analysis. There is a small amount of numerical viscosity in this scheme, which keeps the oscillations near the shock. With a scheme, like leapfrog, which only has dispersive errors and no numerical damping, the oscillations spread out all over the computational domain. This text describes how difference schemes which gives a solution without these erroneous oscillations can be designed.

Example 2.2 The problem

$$u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t$$
$$u(0,x) = \begin{cases} -1 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

has the following solution

$$u(t,x) = \begin{cases} -1 & x < -t \\ x/t & -t \leq x \leq t \\ 1 & x > t \end{cases}$$

However using MacCormack's scheme, we instead get the solution

$$u(t,x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases}$$

which also is a weak solution to the problem, but which does not satisfy the entropy condition. The result is plotted in fig. 2.3..

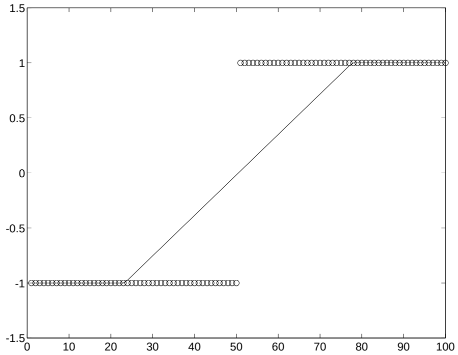


Fig.2.3. MacCormack fails to produce entropy solution.

The consistency with the conservation law does not guarantee that a scheme picks up the entropy satisfying solution. We will try to find difference schemes were such a guarantee is available.

One usual standard form for difference approximations to the conservation law is the *conservative form*,

$$u_j^{n+1} = u_j^n - \lambda(h(u_{j-q+1}^n, u_{j-q+2}^n, \ldots, u_{j+p+1}^n) - h(u_{j-q}^n, u_{j-q+1}^n, \ldots, u_{j+p}^n)) \qquad (2.1)$$

The notation $\lambda = \Delta t/\Delta x$ is used. The function $h(u_{j-q}^n, \ldots, u_{j+p}^n)$ is called the *numerical flux function*. By Taylor expansion one can show that consistency with the conservation law requires

$$f(u) = h(u, u, \ldots, u).$$

Here we mean consistency in the sense that a smooth solution inserted into the difference formula gives a truncation error proportional to $\Delta t(\Delta t^p + \Delta x^q)$, with $p > 0, q > 0$. The conservative form implies that (if $u_j^n \to 0$ as $j \to \pm\infty$)

$$\sum_{j=-\infty}^{\infty} u_j^{n+1} = \sum_{j=-\infty}^{\infty} u_j^n,$$

the discrete counterpart of (1.3) holds.

We often write $h^n_{j-1/2} = h(u^n_{j-q}, u^n_{j-q+1}, \dots, u^n_{j+p})$, and thus the conservative approximation becomes

$$\frac{u^{n+1}_j - u^n_j}{\Delta t} + \frac{h^n_{j+1/2} - h^n_{j-1/2}}{\Delta x} = 0.$$

It is possible to invent schemes that are consistent with the differential equation, but not on conservative form. For such a scheme one can obtain solution were the shocks move with incorrect speed. Consistency in the usual sense does not take in to account the discontinuities, and therefore not the Rankine-Hugoniot condition. Loosely speaking, we can say that the conservative form means consistency with the Rankine-Hugoniot condition. The following theorem states this more precisely

Theorem 2.1. *If u^n_j is computed with a consistent difference approximation on conservative form and $u^n_j \to u(t,x)$ as $\Delta t, \Delta x \to 0$ in $L^1_{loc}(\mathbf{R}^+, \mathbf{R})$ then $u(t,x)$ is a weak solution to the conservation law.*

Proof: We write the scheme (2.1)

$$\frac{u^{n+1}_j - u^n_j}{\Delta t} + \frac{h^n_{j+1/2} - h^n_{j-1/2}}{\Delta x} = 0.$$

Multiply with a test function $\varphi \in C^\infty_0(\mathbf{R}^+, \mathbf{R})$, and sum over n and j,

$$\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \varphi^n_j \frac{u^{n+1}_j - u^n_j}{\Delta t} + \varphi^n_j \frac{h^n_{j+1/2} - h^n_{j-1/2}}{\Delta x} = 0.$$

Now, do partial summation using the rule $\sum_{j=c}^{d} a_j \Delta_+ b_j = -\sum_{j=c+1}^{d} b_j \Delta_- a_j - a_c b_c + a_d b_{d+1}$. We get

$$-\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} (u^{n+1}_j \frac{\varphi^{n+1}_j - \varphi^n_j}{\Delta t} + h^n_{j+1/2} \frac{\varphi^n_{j+1} - \varphi^n_j}{\Delta x}) - \sum_{j=-\infty}^{\infty} \frac{u^0_j \varphi^0_j}{\Delta t} = 0$$

All boundary terms except the one at $t = 0$ disappears, since φ is compactly supported. Multiply $\Delta t \Delta x$ and use the assumption that $u^n_j \to u$. The sum will converge towards the integral

$$-\int_0^\infty \int_{-\infty}^{\infty} u\varphi_t + f(u)\varphi_x \, dx \, dt - \int_{-\infty}^{\infty} u_0 \varphi \, dx = 0$$

here the consistency $h(u, u, \dots, u) = f(u)$ is used. Thus, by definition 1.4, the limit function u is a weak solution of the conservation law.

Remark: The theorem is an if-then statement (implication in one direction). It is possible to have non-conservative form, but still get a solution with correctly moving shocks. An example is the approximation $u_j D_0 u_j$ on non conservative form to $(u^2/2)_x$ is in fact conservative in the sense that $\sum_{j=-\infty}^{\infty} u_j D_0 u_j = 0$ if $u_j \to 0$ as $j \to \pm\infty$.

Example 2.1 showed that there might be problems near the shocks in certain difference methods. We now turn to the problem of characterize a good numerical solution without oscillations around the shock.

The most popular measure for oscillations is

Definition 2.2. *A difference method is called total variation decreasing (TVD) if it produces a solution satisfying*

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^n|$$

for all $n \geq 0$.

We will sometimes use the notation $TV(u^n) = \sum_{j=-\infty}^{\infty} |u_{j+1}^n - u_j^n|$. Originally the concept was called total variation non-increasing (TVNI), but TVD has become the standard term. We give an example to clarify the meaning of the definition.

Example 2.3 Consider the problem $u_t + (u^2/2)_x = 0$ with initial data

$$u_j^0 = 1 \quad j < 0$$
$$u_j^0 = 0 \quad j \geq 0$$

Approximate with the Lax-Wendroff scheme at some CFL number. After one time step the solution is

$$u_j^1 = 1 \quad j \leq -1$$
$$u_0^1 = 1.34$$
$$u_1^1 = 0.23$$
$$u_j^1 = 0 \quad j > 1$$

thus the scheme produced a small overshoot. The variation at t_0 was =1. The variation at t_1 is $\ldots + 0 + 0.34 + 1.11 + 0.23 + 0 + \ldots = 1.68$. The overshoot shows as an increase in the total variation. Thus the Lax-Wendroff scheme is not TVD.

It is natural to require TVD, since the solution to the continuous problem $u(t, x)$ satisfies

$$\frac{d}{dt} \int_{-\infty}^{\infty} |\frac{\partial u}{\partial x}| \, dx \leq 0.$$

It has turned out that the TVD criterion is sometimes too restrictive. We will later on in some cases replace it with

Definition 2.3. *A difference method is called essentially non oscillatory (ENO) if it produces a solution satisfying*

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^n| + O(\Delta x^p)$$

for all $n \geq 0$ and some $p \geq 1$.

Example 2.2 shows that the numerical solution can fail to satisfy the entropy condition. The theorem below provides one way to investigate whether a scheme is entropy satisfying or not.

Theorem 2.4. *If a difference method produces a solution, which also satisfies the discrete entropy condition*

$$(E(u_j^{n+1}) - E(u_j^n))/\Delta t + (H(u_{j-q+1}^n, \ldots, u_{j+p+1}^n) - H(u_{j-q}^n, \ldots, u_{j+p}^n))/\Delta x \leq 0$$

with $H(u_{j-q}, \ldots, u_{j+p})$ a numerical entropy flux consistent with the entropy flux of the differential equation,

$$H(u, \ldots, u) = F(u)$$
$$F'(u) = E'(u)f'(u)$$

then if u_j^n converges the limit will satisfy

$$E(u)_t + F(u)_x \leq 0$$

Proof: Similar to the proof of theorem 2.1.

There is an important class of schemes satisfying a discrete entropy condition.

Definition 2.5. *The difference scheme*

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n)$$

is called an E scheme if

$$\begin{cases} h_{j+1/2} \leq f(u) & \text{all } u \in [u_j, u_{j+1}] \text{ if } u_j < u_{j+1} \\ h_{j+1/2} \geq f(u) & \text{all } u \in [u_{j+1}, u_j] \text{ if } u_{j+1} < u_j \end{cases}$$

This definition is made because of the following theorem

Theorem 2.6. *An E scheme satisfies the semi discrete entropy condition*

$$\frac{dE(u_j(t))}{dt} + D_+ H_{j-1/2} \leq 0$$

for all convex $E(u)$. The numerical entropy flux is given by

$$H_{j-1/2} = F(u_j) + E'(u_j)(h_{j-1/2} - f(u_j))$$

Proof: Start from the difference method

$$du_j(t)/dt = -(h_{j+1/2} - h_{j-1/2})/\Delta x$$

Multiply by $E'(u_j)$, where $E(u)$ is a convex function, so that we get the first term in the semi discrete entropy condition

$$\Delta x \frac{dE(u_j)}{dt} = -E'(u_j)(h_{j+1/2} - h_{j-1/2})$$

where we also multiplied by Δx. Introduce the entropy flux $F'(u) = E'(u)f'(u)$ by

$$\Delta x \frac{dE(u_j)}{dt} + F(u_{j+1}) - F(u_j) = -E'(u_j)(h_{j+1/2} - h_{j-1/2}) + F(u_{j+1}) - F(u_j) \quad (2.2)$$

We can write

$$F(u_{j+1}) - F(u_j) = \int_{u_j}^{u_{j+1}} F'(u)\,du =$$
$$\int_{u_j}^{u_{j+1}} E'(u)f'(u)\,du = [E'f]_{u_j}^{u_{j+1}} - \int_{u_j}^{u_{j+1}} E''(u)f(u)\,du$$

Using this expression for $\Delta_+ F(u_j)$ in the right hand side of (2.2) yields

$$\Delta x \frac{dE(u_j)}{dt} + F(u_{j+1}) - F(u_j) = -E'(u_j)(h_{j+1/2} - h_{j-1/2}) +$$
$$E'(u_{j+1})f(u_{j+1}) - E'(u_j)f(u_j) - \int_{u_j}^{u_{j+1}} E''(u)f(u)\,du$$

Add and subtract $E'(u_{j+1})h_{j+1/2}$ to the right hand side

$$\Delta x \frac{dE(u_j)}{dt} + F(u_{j+1}) - F(u_j) = -E'(u_{j+1})(h_{j+1/2} - f(u_{j+1})) +$$
$$E'(u_j)(h_{j-1/2} - f(u_j)) + (E'(u_{j+1}) - E'(u_j))h_{j+1/2} - \int_{u_j}^{u_{j+1}} E''(u)f(u)\,du$$

which can be written

$$\Delta x \frac{dE(u_j)}{dt} + F(u_{j+1}) - F(u_j) + \Delta_+(E'(u_j)(h_{j-1/2} - f(u_j))) =$$
$$\int_{u_j}^{u_{j+1}} E''(u)\,du h_{j+1/2} - \int_{u_j}^{u_{j+1}} E''(u)f(u)\,du$$

and thus by defining $H_{j-1/2} = F(u_j) + E'(u_j)(h_{j-1/2} - f(u_j))$ the result follows from

$$\Delta x \frac{dE(u_j)}{dt} + \Delta_+ H_{j-1/2} = \int_{u_j}^{u_{j+1}} E''(u)(h_{j+1/2} - f(u))\,du$$

The convexity of $E(u)$ means that $E''(u) > 0$, thus the right hand side is non positive if $h_{j+1/2}$ satisfies the requirements in the theorem.

Remark: It is possible to prove the Entropy condition for the time discretized approximation (forward Euler) as well. We gave the proof for the semi discrete approximation, because it gives a clear picture of how definition 2.5. enters into it, and because the proof is considerably simpler than the proof for the fully discrete case.

It is possible to prove that an E scheme has at most order of accuracy one.

Note that in the definition of an E scheme, a statement is made about *all* values of u between u_{j+1} and u_j. The theory in chapter one makes it probable that for non convex flux functions it is necessary to have information of how the flux function behaves between the grid points. We give an example to clarify this statement.

Example 2.4 The problem

$$u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t$$
$$u(0,x) = \begin{cases} 1 & x < 0 \\ -1 & x \geq 0 \end{cases}$$

has solution

$$u(t,x) = \begin{cases} 1 & x < 0 \\ -1 & x \geq 0 \end{cases} .$$

Assume that a difference method is given which gives the steady solution profile

$$\begin{aligned} u_j^n &= 1 \quad j \leq -1 \\ u_0^n &= 0.8 \\ u_1^n &= -0.8 \\ u_j^n &= -1 \quad j \geq 2 \end{aligned}$$

for all n. Make a deformation of the flux function as in fig. 2.4. below.

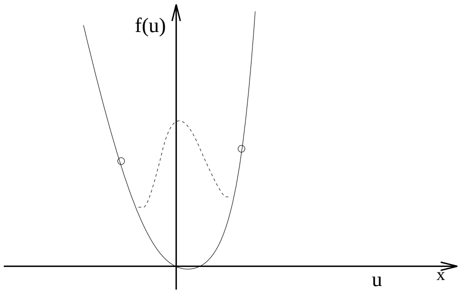


Fig.2.4. Deformed flux function

The steady shock does not satisfy the entropy condition for the deformed flux function (cf. chapter 1). The deformed flux coincides with $f(u) = u^2/2$ for $|u| > 0.8$, and a scheme which only relies on flux values at the grid points, does not have sufficient information to distinguish between the deformed flux and the quadratic one.

As an example we now give two classes of difference methods, monotone schemes and three point schemes, where the TVD and entropy properties have been worked out.

2.3 Monotone Schemes

The TVD and ENO properties are usually difficult to investigate for a given scheme. We therefore start analysing the subclass of monotone schemes, which is easy to distinguish. We write an explicit difference method in general as

$$u_j^{n+1} = G(u_{j-q}^n, \ldots, u_{j+p+1}^n). \tag{2.3}$$

With this notation we introduce the class of monotone schemes

Definition 2.7. *The scheme (2.3) is monotone if the function G is an increasing function of all its arguments, i.e.*

$$\frac{\partial G(u_{-q}, \ldots, u_{p+1})}{\partial u_i} \geq 0 \qquad -q \leq i \leq p+1$$

We let G_i denote the partial derivative of G with respect to its ith argument. Note that it is implicitly assumed that G is a differentiable function.

Theorem 2.8. *Monotone conservative schemes are TVD.*

Proof:

$$\sum_{j=-\infty}^{\infty} |u_{j+1}^{n+1} - u_j^{n+1}| = \sum_{j=-\infty}^{\infty} |G(u_{j-q+1}^n, \ldots, u_{j+p+2}^n) - G(u_{j-q}^n, \ldots, u_{j+p+1}^n)| =$$

$$\sum_{j=-\infty}^{\infty} |G(u_{j-q}^n + \Delta_+ u_{j-q}^n, \ldots, u_{j+p+1}^n + \Delta_+ u_{j+p+1}^n) - G(u_{j-q}^n, \ldots, u_{j+p+1}^n)| =$$

$$\sum_{j=-\infty}^{\infty} | \sum_{k=-q}^{p+1} \int_0^1 G_k(u_{j-q}^n + \theta\Delta_+ u_{j-q}^n, \ldots, u_{j+p+1}^n + \theta\Delta_+ u_{j+p+1}^n)\Delta_+ u_{j+k}^n \, d\theta| \leq$$

Use the monitonicity and the triangle inequality

$$\sum_{j=-\infty}^{\infty} \sum_{k=-q}^{p+1} \int_0^1 G_k(u_{j-q}^n + \theta\Delta_+ u_{j-q}^n, \ldots, u_{j+p+1}^n + \theta\Delta_+ u_{j+p+1}^n)|\Delta_+ u_{j+k}^n| \, d\theta =$$

Change index from j to $m = j + k$

$$\sum_{m=-\infty}^{\infty} \sum_{k=-q}^{p+1}$$

$$\int_0^1 G_k(u_{m-k-q}^n + \theta\Delta_+ u_{m-k-q}^n, \ldots, u_{m-k+p+1}^n + \theta\Delta_+ u_{m-k+p+1}^n) \, d\theta |\Delta_+ u_m^n|$$

The result follows from the fact that

$$\sum_{k=-q}^{p+1} G_k(v_{m-k-q}, \ldots, v_{m-k+p+1}) = 1 \tag{2.4}$$

since we then get from above

$$\sum_{j=-\infty}^{\infty} |u_{j+1}^{n+1} - u_j^{n+1}| \leq$$
$$\sum_{m=-\infty}^{\infty} \sum_{k=-q}^{p+1}$$
$$\int_0^1 G_k(u_{m-k-q}^n + \theta\Delta_+ u_{m-k-q}^n, \ldots, u_{m-k+p+1}^n + \theta\Delta_+ u_{m-k+p+1}^n)\, d\theta |\Delta_+ u_m^n| =$$
$$\sum_{m=-\infty}^{\infty} \int_0^1 d\theta |\Delta_+ u_m^n| = \sum_{m=-\infty}^{\infty} |u_{m+1}^n - u_m^n|$$

It remains to prove (2.4). To make the formulas simpler, we do this only for the three point scheme

$$G(v_{m-1}, v_m, v_{m+1}) = v_m - \lambda(h(v_m, v_{m+1}) - h(v_{m-1}, v_m))$$

where $h(u, v)$ is the numerical flux function. If we denote the derivative of h with respect to its first argument h_1 and let h_2 be the derivative with respect to the second argument, we get

$$\begin{aligned}
G_{-1}(v_{m-1}, v_m, v_{m+1}) &= \lambda h_1(v_{m-1}, v_m)\\
G_0(v_{m-1}, v_m, v_{m+1}) &= 1 - \lambda(h_1(v_m, v_{m+1}) - h_2(v_{m-1}, v_m))\\
G_1(v_{m-1}, v_m, v_{m+1}) &= -\lambda h_2(v_m, v_{m+1})
\end{aligned}$$

For the three point scheme (2.4) becomes

$$\begin{aligned}
&G_{-1}(v_m, v_{m+1}, v_{m+2}) + G_0(v_{m-1}, v_m, v_{m+1}) + G_1(v_{m-2}, v_{m-1}, v_m) =\\
&\lambda h_1(v_m, v_{m+1}) + 1 - \lambda(h_1(v_m, v_{m+1}) - h_2(v_{m-1}, v_m)) - \lambda h_2(v_{m-1}, v_m) = 1
\end{aligned}$$

and the proof is complete. The general (2.4) follows similarly by converting to derivatives of the numerical flux function.

Theorem 2.9. *Except for one trivial case, monotone schemes are at most first order accurate.*

To prove this we first state the following theorem, which does not use the monotonicity of the scheme, and thus holds in general for all first order schemes on conservative form.

Theorem 2.10. *The truncation error of the method*

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n)$$

is

$$\tau_j^n = -\Delta t^2 (q(u)u_x)_x + \Delta t O(\Delta t^2 + \Delta x^2)$$

where

$$q(u) = (\frac{1}{\lambda^2}\sum k^2 G_k(u,\ldots,u) - f'(u)^2)/2$$

and we denote $u = u(t_n, x_j)$.

Proof: By definition the truncation error τ_j^n is

$$\begin{aligned}
&\tau_j^n = u(t_{n+1}, x_j) - G(u(t_n, x_{j-q}), \ldots, u(t_n, x_{j+p+1})) = \\
&u + \Delta t u_t + \frac{\Delta t^2}{2} u_{tt} - (G(u,\ldots,u) + \sum_{k=-q}^{p+1} G_k(u,\ldots,u)(u(t_n, x_{j+k}) - u) + \\
&\frac{1}{2}\sum_{k=-q}^{p+1}\sum_{m=-q}^{p+1} G_{km}(u,\ldots,u)(u(t_n, x_{j+k}) - u)(u(t_n, x_{j+m}) - u) + \\
&O(\Delta t(\Delta t^2 + \Delta x^2)))
\end{aligned}$$

where we use the notation $u = u(t_n, x_j)$. We have here Taylor expanded the difference scheme in u. Next we expand the functions u in x and arrive at (modulo second order terms)

$$\begin{aligned}
&\tau_j^n = u + \Delta t u_t + \frac{\Delta t^2}{2} u_{tt} - \\
&(u + \Delta x u_x \sum_{k=-q}^{p+1} k G_k + \frac{\Delta x^2}{2} u_{xx} \sum_{k=-q}^{p+1} k^2 G_k + \frac{\Delta x^2}{2}(u_x)^2 \sum_{k=-q}^{p+1}\sum_{m=-q}^{p+1} km G_{km})
\end{aligned}$$

We use G to denote $G(u,\ldots,u)$, and similarly for the derivatives of G. Now add and subtract the expression

$$\frac{\Delta x^2}{2}(u_x)^2 \sum_{k=-q}^{p+1}\sum_{m=-q}^{p+1} k^2 G_{km} \tag{2.5}$$

The reason for this is that the last term together with (2.5) becomes

$$\frac{\Delta x^2}{2}(u_x)^2 \sum_{k=-q}^{p+1}\sum_{m=-q}^{p+1} (km - k^2) G_{km} = 0 \tag{2.6}$$

We omit the proof that this sum is zero for the moment. If we accept (2.6) as a fact, we get

$$\begin{aligned}
\tau_j^n =& \Delta t u_t + \frac{\Delta t^2}{2} u_{tt} - (\Delta x u_x \sum_{k=-q}^{p+1} k G_k + \\
&\frac{\Delta x^2}{2} u_{xx} \sum_{k=-q}^{p+1} k^2 G_k + \frac{\Delta x^2}{2}(u_x)^2 \sum_{k=-q}^{p+1}\sum_{m=-q}^{p+1} k^2 G_{km})
\end{aligned}$$

Now use

$$\sum_{k=-q}^{p+1} kG_k(u,\ldots,u) = -\lambda f'(u) \tag{2.7}$$

to eliminate the zero order terms. We omit the proof of (2.7) as well. The first order terms remains

$$\tau_j^n = \frac{\Delta t^2}{2}u_{tt} - (\frac{\Delta x^2}{2}u_{xx}\sum_{k=-q}^{p+1} k^2G_k + \frac{\Delta x^2}{2}u_x\sum_{k=-q}^{p+1}(k^2G_k)_x) =$$

$$\frac{\Delta t^2}{2}u_{tt} - \frac{\Delta x^2}{2}(u_x\sum_{k=-q}^{p+1} k^2G_k)_x$$

Finally we remove the time derivatives by substituting u_{tt} with $((f')^2u_x)_x$. This can be done because

$$\begin{aligned}&u_{tt} = -f_{xt} = -(f'(u)u_x)_t = \\ &-(f''(u)u_tu_x + f'(u)u_{xt}) = f''(u)f_xu_x + f'f_{xx} = \\ &(f'f_x)_x = ((f')^2u_x)_x\end{aligned}$$

and the truncation error becomes

$$\tau_j^n = \frac{\Delta t^2}{2}((((f')^2 - \frac{1}{\lambda^2}\sum_{k=-q}^{p+1} k^2G_k)u_x)_x$$

which is what we wanted to prove. It remains to prove (2.6) and (2.7). That can be done by writing out the conservative form of the method and let the derivatives of G instead become derivatives of the numerical flux function h. It is a straightforward calculation similar to the one that was done in the last part of theorem 2.8., and we do not give it here.

Finally we give the proof of theorem 2.9. By (2.7)

$$\lambda^2(f'(u))^2 = (\sum_{k=-q}^{p+1} kG_k)^2 = (\sum_{k=-q}^{p+1} k\sqrt{G_k}\sqrt{G_k})^2$$

were the monotonicity is used to split G_k into square roots. The Cauchy-Schwartz inequality gives

$$\lambda^2(f'(u))^2 \leq \sum_{k=-q}^{p+1} k^2G_k \sum_{k=-q}^{p+1} G_k = \sum_{k=-q}^{p+1} k^2G_k$$

by writing out the conservative form, it is easy to see that $\sum_{k=-q}^{p+1} G_k = 1$. It follows that

$$q(u) = (\frac{1}{\lambda^2}\sum k^2G_k(u,\ldots,u) - f'(u)^2)/2 \leq 0$$

where $q(u)$ is the function defined in theorem 2.10.,

$$\tau_j^n = -\Delta t^2 (q(u) u_x)_x.$$

From Cauchy-Schwartz, we know that strict inequality ($<$) holds except if $kG_k = const. G_k \Rightarrow$ the method is a pure translation. This is the trivial case mentioned in theorem 2.9. We conclude that strict inequality holds, except in this case, and therefore the truncation error does not vanish. The accuracy is one.

Theorem 2.11. *Monotone schemes satisfy the discrete entropy condition*

$$(E(u_j^{n+1}) - E(u_j^n))/\Delta t + (H_{j+1/2}^n - H_{j-1/2}^n)/\Delta x \leq 0$$

for the class of entropies $E(u) = |u - c|$ *all* $c \in \mathbf{R}$, *and where the numerical entropy flux,* $H_{j+1/2}^n$ *is consistent with the entropy flux*

$$F(u) = sign(u - c)(f(u) - f(c))$$

Proof: Introduce the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Define the numerical entropy flux as

$$\begin{aligned} H_{j-1/2}^n = H(u_{j-q}^n, \ldots, u_{j+p}^n) = \\ h(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p}^n) - h(c \wedge u_{j-q}^n, \ldots, c \wedge u_{j+p}^n) \end{aligned}$$

where h is the numerical flux of the scheme. With this numerical entropy flux, we obtain

$$\begin{aligned} E(u_j^n) - \lambda \Delta_+ H_{j-1/2}^n = |u_j^n - c| - \\ \lambda \Delta_+ h(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p}^n) - \lambda \Delta_+ h(c \wedge u_{j-q}^n, \ldots, c \wedge u_{j+p}^n) \end{aligned}$$

and since $|u - c| = u \vee c - u \wedge c$, we arrive at

$$\begin{aligned} E(u_j^n) - \lambda \Delta_+ H_{j-1/2}^n &= u_j^n \vee c - u_j^n \wedge c - \lambda \Delta_+ H_{j-1/2}^n \\ &= G(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p+1}^n) - G(c \wedge u_{j-q}^n, \ldots, c \wedge u_{j+p+1}^n) \end{aligned} \tag{2.8}$$

From the monotonicity of the method we get

$$\begin{aligned} u_j^{n+1} &= G(u_{j-q}^n, \ldots, u_{j+p+1}^n) \leq G(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p+1}^n) \\ c &= G(c, \ldots, c) \leq G(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p+1}^n) \end{aligned}$$

and thus

$$u_j^{n+1} \vee c \leq G(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p+1}^n)$$

similarly we see that

$$-(u_j^{n+1} \wedge c) \leq -G(c \wedge u_{j-q}^n, \ldots, c \wedge u_{j+p+1}^n)$$

Finally,

$$E(u_j^{n+1}) = |u_j^{n+1} - c| = u^{n+1} \vee c - u_j^{n+1} \wedge c \leq$$
$$G(c \vee u_{j-q}^n, \ldots, c \vee u_{j+p+1}^n) - G(c \wedge u_{j-q}^n, \ldots, c \wedge u_{j+p+1}^n)$$
$$= |u_j^n - c| - \lambda \Delta_+ H_{j-1/2}^n$$

where (2.8) was used in the last equality, This is the desired entropy inequality.

Remark: The class of entropy functions in the previous theorem is sufficiently large to assure that the entropy condition (1.10) will hold for the limit solution.

2.4 Three point schemes

For three point schemes ($h_{j+1/2} = h(u_{j+1}, u_j)$), there is a complete characterization of TVD schemes in terms of the numerical viscosity coefficient.

We first state the theorem on which all proofs that a scheme is TVD is based. To apply this theorem, it is necessary to write the difference method differently. The incremental form or *I-form* of the difference approximation to (1.1) is

$$u_j^{n+1} = u_j^n + C_{j+1/2}\Delta_+ u_j^n - D_{j-1/2}\Delta_- u_j^n$$

from the I-form the TVD property can be obtained through

Theorem 2.12. *If*

$$C_{j+1/2} \geq 0 \quad D_{j+1/2} \geq 0 \qquad C_{j+1/2} + D_{j+1/2} \leq 1$$

then the method is TVD.

Proof: Apply Δ_+ to both sides of the I-form and sum over j.

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| = \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^n + \Delta_+(C_{j+1/2}\Delta_+ u_j^n) - \Delta_+(D_{j-1/2}\Delta_- u_j^n)|$$

rearranging terms gives

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| =$$
$$\sum_{j=-\infty}^{\infty} |C_{j+3/2}\Delta_+ u_{j+1}^n + (1 - C_{j+1/2} - D_{j+1/2})\Delta_+ u_j^n + D_{j-1/2}\Delta_+ u_{j-1}^n|$$

Apply the triangle inequality on the right hand side

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} |C_{j+3/2}\Delta_+ u_{j+1}^n| +$$
$$\sum_{j=-\infty}^{\infty} |(1 - C_{j+1/2} - D_{j+1/2})\Delta_+ u_j^n| + \sum_{j=-\infty}^{\infty} |D_{j-1/2}\Delta_+ u_{j-1}^n|$$

Now use the assumption that $C_{j+1/2}, D_{j+1/2}$ are positive and that the sum $C_{j+1/2} + D_{j+1/2} \leq 1$.

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} C_{j+3/2}|\Delta_+ u_{j+1}^n| +$$
$$\sum_{j=-\infty}^{\infty} (1 - C_{j+1/2} - D_{j+1/2})|\Delta_+ u_j^n| + \sum_{j=-\infty}^{\infty} D_{j-1/2}|\Delta_+ u_{j-1}^n|$$

Finally shift the indices in the first and third sum on the right hand side

$$\sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} C_{j+1/2}|\Delta_+ u_j^n| +$$
$$\sum_{j=-\infty}^{\infty} (1 - C_{j+1/2} - D_{j+1/2})|\Delta_+ u_j^n| + \sum_{j=-\infty}^{\infty} D_{j+1/2}|\Delta_+ u_j^n| = \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^n|$$

and the TVD property is proved.

Remark: The condition $C_{j+1/2} + D_{j+1/2} \leq 1$ corresponds to the CFL condition in the linear case, and is not required for a semi discrete method of lines approximation.

We now introduce the viscosity form or *Q-form* of the difference approximation to (1.1) as

$$u_j^{n+1} = u_j^n - \Delta t D_0 f(u_j^n) + \frac{1}{2}\Delta_+(Q_{j-1/2}\Delta_- u_j^n) \qquad (2.9)$$

where Q is the *numerical viscosity coefficient.* A three point scheme is uniquely defined through its coefficient of numerical viscosity as can be seen from the conversion formulas at the end of this section. Thus there is only one degree of freedom in chosing a three point scheme.

If we rewrite (2.9) on conservative form, the numerical flux function becomes

$$h_{j+1/2} = \frac{1}{2}(f(u_{j+1}) + f(u_j)) - \frac{1}{2\lambda}Q_{j+1/2}(u_{j+1} - u_j). \qquad (2.10)$$

This is seen by inserting this numerical flux function into the conservative (or *C-form*) and rearranging terms.

If we apply theorem 2.12. to the Q-form we get the following characterization of three point TVD schemes.

Theorem 2.13. *A three point scheme is TVD if and only if the numerical viscosity coefficient satisfies*

$$\lambda|a_{j+1/2}| \leq Q_{j+1/2} \leq 1$$

where $a_{j+1/2}$ is the local wave speed

$$a_{j+1/2} = \begin{cases} \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} & u_j \neq u_{j+1} \\ f'(u_j) & u_j = u_{j+1} \end{cases}$$

Proof. Starting from the conservative form, we add and subtract $f(u_j)$, and get

$$u_j^{n+1} = u_j^n - \lambda(\frac{h_{j+1/2}^n - f(u_j^n)}{u_{j+1}^n - u_j^n}(u_{j+1}^n - u_j^n) - \frac{h_{j-1/2}^n - f(u_j^n)}{u_j^n - u_{j-1}^n}(u_j^n - u_{j-1}^n))$$

Thus we can identify

$$C_{j+1/2} = -\lambda\frac{h_{j+1/2}^n - f(u_j^n)}{u_{j+1}^n - u_j^n}$$
$$D_{j-1/2} = -\lambda\frac{h_{j-1/2}^n - f(u_j^n)}{u_j^n - u_{j-1}^n}$$

Insert the expression (2.10) for $h_{j+1/2}$ into these formulas

$$C_{j+1/2} = -\lambda\frac{f(u_{j+1}^n) - f(u_j^n)}{2(u_{j+1}^n - u_j^n)} + Q_{j+1/2}/2 = \frac{1}{2}(Q_{j+1/2} - \lambda a_{j+1/2})$$
$$D_{j-1/2} = -\lambda\frac{f(u_{j-1}^n) - f(u_j^n)}{2(u_j^n - u_{j-1}^n)} + Q_{j-1/2}/2 = \frac{1}{2}(Q_{j-1/2} + \lambda a_{j-1/2})$$

The positivity of $C_{j+1/2}$ and $D_{j+1/2}$ means that

$$Q_{j+1/2} \geq \lambda a_{j+1/2} \quad \text{and}$$
$$Q_{j+1/2} \geq -\lambda a_{j+1/2}$$

which is equivalent to the lower limit in the theorem

$$Q_{j+1/2} \geq \lambda |a_{j+1/2}|$$

The condition $C_{j+1/2} + D_{j+1/2} \leq 1$ becomes

$$Q_{j+1/2} \leq 1$$

and the theorem follows.

The quantity $a_{j+1/2}$ is important and will be used throughout this text. The second order Lax-Wendroff scheme has $Q_{j+1/2} = \lambda^2 a_{j+1/2}^2$, and is clearly outside the TVD region. Thus we get the following result

Corollary 2.14. *Three point TVD schemes are at most first order accurate. The situation can be viewed in fig. 2.5.*

This result and the corresponding result for monotone schemes might seem depressing. First order schemes are not accurate enough to be of use in practice. However, higher order methods are developed using first order methods as building blocks. This is the motivation for the study of first order methods.

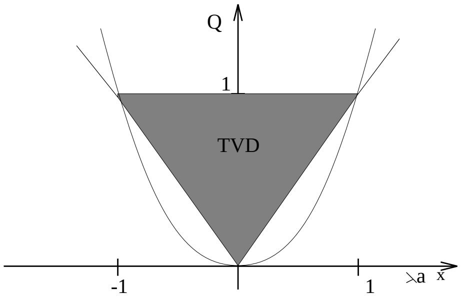


Fig.2.5. TVD domain for numerical viscosity

For reference we conclude this section by a listing of the three different standard forms to write an approximation to (1.1), and formulas for converting between them. The conservative form (C-form), the incremental form (I-form) and the viscosity form (Q-form).

Q-form to C-form

$$h_{j+1/2} = \frac{1}{2}(f(u_j) + f(u_{j+1})) - \frac{1}{2\lambda} Q_{j+1/2} \Delta_+ u_j$$

C-form to Q-form

$$Q_{j+1/2} = \lambda \frac{f(u_j) + f(u_{j+1}) - 2h_{j+1/2}}{\Delta_+ u_j}$$

Q-form to I-form

$$C_{j+1/2} = \frac{1}{2}(Q_{j+1/2} - \lambda a_{j+1/2})$$

$$D_{j+1/2} = \frac{1}{2}(Q_{j+1/2} + \lambda a_{j+1/2})$$

I-form to Q-form

$$Q_{j+1/2} = C_{j+1/2} + D_{j+1/2}$$

C-form to I-form

$$C_{j+1/2} = -\lambda \frac{h^n_{j+1/2} - f(u^n_j)}{u^n_{j+1} - u^n_j}$$

$$D_{j+1/2} = -\lambda \frac{h^n_{j+1/2} - f(u^n_{j+1})}{u^n_{j+1} - u^n_j}$$

I-form to C-form

$$h_{j+1/2} = f(u_j) - \frac{1}{\lambda} C_{j+1/2} \Delta_+ u_j = f(u_{j+1}) - \frac{1}{\lambda} D_{j+1/2} \Delta_+ u_j$$

2.5 Some Schemes

Here we give some examples of three point approximations, which can be analyzed using the theorems in section 2.4. The schemes are important in their own rights, some of them will come up later in versions of higher order accuracy and in extension to nonlinear systems of conservation laws.

Example 2.5 *The upwind scheme.* This scheme is the lower TVD limit in theorem 2.13, i.e.

$$Q_{j+1/2} = \lambda |a_{j+1/2}|.$$

Writing out the conservative form the scheme becomes

$$h_{j+1/2} = \frac{1}{2}(f(u_{j+1})+f(u_j))-\frac{1}{2}|\frac{f(u_{j+1})-f(u_j)}{u_{j+1}-u_j}|(u_{j+1}-u_j) = \begin{cases} f(u_{j+1}) & a_{j+1/2} < 0 \\ f(u_j) & a_{j+1/2} > 0 \end{cases}$$

and we can see the reason why this is called the upwind scheme. The scheme takes the flux value from the direction of the characteristics. For the linear equation $u_t + au_x = 0$ the wave speed $a_{j+1/2} = a$ is constant and the scheme becomes

$$\begin{aligned} u_j^{n+1} &= u_j^n - a\Delta t D_+ u_j^n \quad a < 0 \\ u_j^{n+1} &= u_j^n - a\Delta t D_- u_j^n \quad a > 0 \end{aligned}$$

or with $a^+ = \max(0,a)$ and $a^- = \min(0,a)$,

$$u_j^{n+1} = u_j^n - \Delta t(a^- D_+ u_j^n + a^+ D_- u_j^n)$$

By considering the example

$$\begin{aligned} & u_t + (u^2/2)_x = 0 \quad -\infty < x < \infty \;\; 0 < t \\ & u(0,x) = \begin{cases} -1 & x < 0 \\ 1 & x \geq 0 \end{cases} \end{aligned}$$

it is easy to see that the upwind scheme does not satisfy the entropy condition. The scheme does not contain enough viscosity to break the expansion shock into an expansion wave. The scheme is attractive because it has the least possible viscosity to suppress oscillations.

Example 2.6. *The Lax-Friedrichs scheme.* At the other end of the TVD interval in theorem 2.13 we find the Lax-Friedrichs scheme, which has the viscosity

$$Q_{j+1/2} = 1$$

and numerical flux function

$$h_{j+1/2} = \frac{1}{2}(f(u_{j+1}) + f(u_j)) - \frac{1}{2\lambda}(u_{j+1} - u_j)$$

This scheme is extremely diffusive, and smears shocks enormously. The advantage of the scheme is its simplicity, and the fact that the numerical flux function is infinitely differentiable with respect to its arguments. This is of importance for steady state computations when, in Newton type methods, the Jacobian of the scheme is required. It is also a requirement when the formal order of accuracy is derived.

Example 2.7. *The Godunov scheme.* This scheme has the viscosity coefficient

$$Q_{j+1/2} = \lambda \max_{(u-u_j)(u-u_{j+1})\leq 0} \frac{f(u_{j+1}) + f(u_j) - 2f(u)}{u_{j+1} - u_j} \tag{2.11}$$

The scheme was originally derived for the Euler equations in gas dynamics, where it was constructed as solving a Riemann problem locally between each two grid points. This derivation will be given later. Here we can instead explain the Godunov scheme as the lower limit in the definition of the E schemes

$$h_{j+1/2} = \begin{cases} \min_{u_j<u<u_{j+1}} f(u) & \text{if } u_j < u_{j+1} \\ \max_{u_{j+1}<u<u_j} f(u) & \text{if } u_j > u_{j+1} \end{cases}$$

From this definition it is straightforward to derive the expression (2.11) for the viscosity coefficient. The Godunov scheme is the E scheme with smallest coefficient of viscosity. It is also a TVD scheme.

Example 2.8. *The Engquist-Osher scheme.* The E-O scheme was designed with the intent of improving the upwind scheme with respect to entropy and convergence to steady state. The viscosity coefficient takes into account all values between u_{j+1} and u_j by integrating over this interval,

$$Q_{j+1/2} = \frac{\lambda}{u_{j+1} - u_j} \int_{u_j}^{u_{j+1}} |f'(u)|\, du$$

If f' does not change sign between u_j and u_{j+1}, then we see that the viscosity is equal to the viscosity of the upwind scheme. The advantages with this method is that it is an E scheme and that the numerical flux is a C^1 function of its arguments, making it suitable for steady state computations. The scheme is TVD.

Example 2.9. *The Lax-Wendroff scheme.* The only choice of viscosity that gives a second order accurate approximation (both in space and time) is the Lax-Wendroff scheme

$$Q_{j+1/2} = \lambda^2 a_{j+1/2}^2$$

The scheme is not TVD, but it is important because of its optimality. (The only three point second order scheme). Later when we discuss second order TVD schemes, the Lax-Wendroff scheme will play an important role.

The Godunov, E-O and the upwind schemes coincide if the flux function derivative $f'(u)$ does not change sign between u_j and u_{j+1}. The *sonic points* are the u values for which $f'(u) = 0$. It is usually around the sonic points the entropy condition is hard to satisfy. There have been suggested a number of fixes for the upwind scheme to satisfy the entropy condition. One which is commonly used in computational fluid dynamics (CFD) is the choice

$$Q_{j+1/2} = \begin{cases} \lambda|a_{j+1/2}| & \text{if } \lambda|a_{j+1/2}| > 2\epsilon \\ (\lambda a_{j+1/2})^2/(4\epsilon) + \epsilon & \text{if } \lambda|a_{j+1/2}| \leq 2\epsilon \end{cases}$$

The viscosity is prevented from going to zero when $|a_{j+1/2}| = 0$, and the viscosity becomes a C^1 function of u_j, u_{j+1}. The disadvantage is that we now have a parameter to tune.

As a summary the schemes are plotted as function of increasing viscosity in fig.2.6.

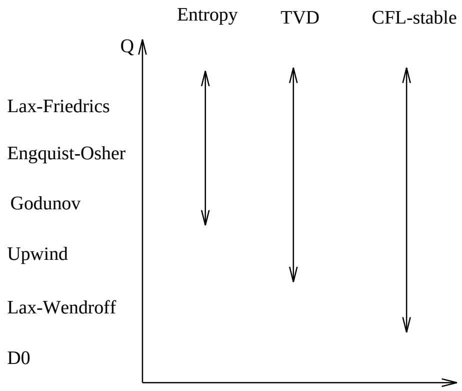


Fig.2.6. Properties of methods as function of the viscosity

2.6. Two space dimensions

In two space dimensions we approximate the conservation law

$$u_t + f_1(u)_x + f_2(u)_y = 0$$

on some domain, by the explicit difference method

$$u^{n+1}_{i,j} = u^n_{i,j} - \lambda_x \Delta_{+i} h^n_{i-1/2,j} - \lambda_y \Delta_{+j} g^n_{i,j-1/2}$$

where $\lambda_x = \Delta t/\Delta x$, $\lambda_y = \Delta t/\Delta y$. $h^n_{i-1/2,j} = h(u^n_{i-q,j}, \dots, u^n_{i+p,j})$ is a flux consistent with the flux f_1, and similarly for $g^n_{i,j-1/2}$. We can choose $h^n_{i-1/2,j}$ as a one dimensional flux formula described in this chapter. Note however that with this straightforward generalization the Lax-Wendroff scheme will not maintain second order accuracy in two dimensions. In the case of Lax-Wendroff, it is better to use operator splitting, then second order accuracy can be kept.

For two space dimensions, it is possible to prove that TVD, in the sense that

$$\sum_{i,j} \Delta y |u^n_{i+1,j} - u^n_{i,j}| + \Delta x |u^n_{i,j+1} - u^n_{i,j}|$$

is decreasing, implies overall first order accuracy. First order is too restrictive. Instead we write the scheme as

$$u^{n+1}_{i,j} = u^n_{i,j} + A_{i+1/2,j}\Delta_{+i}u^n_{i,j} - B_{i-1/2,j}\Delta_{-i}u^n_{i,j} + C_{i,j+1/2}\Delta_{+j}u^n_{i,j} - D_{i,j-1/2}\Delta_{-j}u^n_{i,j}$$

and take

$$\begin{array}{l} A_{i+1/2,j} \geq 0 \;\; B_{i+1/2,j} \geq 0 \;\; C_{i,j+1/2} \geq 0 \;\; D_{i,j+1/2} \geq 0 \\ A_{i+1/2,j} + B_{i+1/2,j} + C_{i,j+1/2} + D_{i,j+1/2} \leq 1 \end{array} \qquad (2.12)$$

as a criterion for a scheme with good properties with respect to shocks. Unlike the one dimensional case, (2.12) does not imply TVD, and thus allows for second order accurate schemes in two space dimensions.

Exercises

1. Show that the Engquist-Osher scheme, the Godunov scheme and the upwind scheme coincides when applied to the linear problem

$$u_t + au_x = 0$$

2. Determine the smallest constant d that makes the Lax-Wendroff scheme with added viscosity

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda\Delta_+ h_{j-1/2} + d\Delta_+\Delta_- u_j^n \\ h_{j+1/2} &= \frac{1}{2}(f(u_j^n) + f(u_{j+1}^n)) - \frac{1}{2\lambda}(\lambda a_{j+1/2})^2\Delta_+ u_j^n \end{aligned}$$

TVD. Determine the cfl stability condition for the resulting TVD scheme. Does the scheme satisfy an entropy condition ?

3. Assume the initial data

$$u_j^0 = \begin{cases} u_L & j < 0 \\ u_R & j \geq 0 \end{cases}$$

are given to the general three point scheme

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n) \\ h_{j+1/2}^n &= \frac{1}{2}(f(u_j^n) + f(u_{j+1}^n)) - \frac{1}{2\lambda}Q(u_{j+1}^n - u_j^n) \end{aligned}$$

Determine conditions on Q such that

$$TV(u^1) \leq TV(u^0)$$

Compare with theorem 2.13.

4. Show that the method

$$u_j^{n+1} = u_j^n - \begin{cases} \Delta t D_+ f(u_j^n) & \text{if } f'(u_j^n) < 0 \\ \Delta t D_- f(u_j^n) & \text{if } f'(u_j^n) > 0 \end{cases}$$

is not conservative.

3. Second order accurate TVD methods

3.1 Limitations of Accuracy

Before starting to describe second order schemes for shock computations we give some necessary conditions for such schemes. We saw in the previous chapter that three point TVD schemes are at most first order accurate. Thus a second order TVD scheme on C-form

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n)$$

must involve more than three points on the time level t_n.

In fact second order accuracy everywhere is not compatible with the TVD constraint.

Theorem 3.1. *At smooth extrema which are not sonic points a TVD scheme is first order accurate.*

Proof: Write the method on I-form

$$u_j^{n+1} = u_j^n + C_{j+1/2}\Delta_+ u_j^n - D_{j-1/2}\Delta_- u_j^n$$

we consider a general explicit scheme, and thus

$$C_{j+1/2} = C(u_{j-q+1}^n, \dots, u_{j+p+1}^n) \quad D_{j-1/2} = D(u_{j-q}^n, \dots, u_{j+p}^n)$$

In all accuracy investigations, it is necessary to assume that the solution u_j is smooth, to allow for Taylor expansion. The truncation error is expanded as

$$\tau_j^n = -u(x_j, t_{n+1}) + u(x_j, t_n) + C_{j+1/2}\Delta_+ u_j^n - D_{j-1/2}\Delta_- u_j^n =$$
$$-\Delta t u_t - \frac{\Delta t^2}{2} u_{tt} + (C + \sum_{k=-q}^{p} C_k((k+1)\Delta x u_x + O(\Delta x^2)))(\Delta x u_x +$$
$$\frac{\Delta x^2}{2} u_{xx} + O(\Delta x^3)) - (D + \sum_{k=-q}^{p} D_k(k\Delta x u_x + O(\Delta x^2)))(\Delta x u_x -$$
$$\frac{\Delta x^2}{2} u_{xx} + O(\Delta x^3))$$

where we use the notation $C = C(u, \dots, u)$ and C_k is the derivative of C with respect to its kth argument, evaluated at $u, \dots, u$, and where u is $u(t_n, x_j)$. Simplify the expression

$$\tau_j^n = -\Delta t u_t + (C - D)\Delta x u_x - \frac{\Delta t^2}{2} u_{tt} +$$
$$(C + D)\frac{\Delta x^2}{2} u_{xx} + \sum_{k=-q}^{p} C_k(k+1)\Delta x^2 (u_x)^2 - \sum_{k=-q}^{p} D_k k \Delta x^2 (u_x)^2 + O(\Delta x^3)$$

Consistency yields

$$C - D = -\lambda f'(u) \tag{3.1}$$

At a smooth extreme point $u_x = 0$, and the condition for second order accuracy there becomes

$$\lambda^2 u_{tt} = (C + D) u_{xx}$$

but

$$u_{tt} = (f'(u)^2 u_x)_x = f'(u)^2 u_{xx} + f'(u)_x^2 u_x$$

and at extrema the condition for second order second order accuracy thus becomes

$$C + D = (\lambda f'(u))^2 \tag{3.2}$$

Solving (3.1) and (3.2) for C and D gives

$$\begin{aligned} 2C &= (\lambda f'(u))^2 - \lambda f'(u) \\ 2D &= (\lambda f'(u))^2 + \lambda f'(u) \end{aligned} \tag{3.3}$$

If $f'(u) \neq 0$, the CFL condition $\lambda f'(u) < 1$ implies that not both C and D can be non negative. For TVD it is necessary that $C_{j+1/2}$ and $D_{j+1/2}$ are non negative. $C = C_{j+1/2} + O(\Delta x)$, but (3.3) means that one of C, D are negative of order one, $(= -|O(1)|)$. Thus if $f'(u) \neq 0$ we have proved the impossibility to satisfy the TVD condition given in theorem 2.12. If on the other hand $f'(u) = 0$, the above argument is not true. This is the exception "u non sonic" mentioned in the theorem.

This result can be interpreted geometrically as clipping of extrema displayed in the figures 3.1 and 3.2. In order to maintain TVD, the maximum can not be placed on the exact solution curve at $t + \Delta t$, since this would correspond to an increase in the variation.

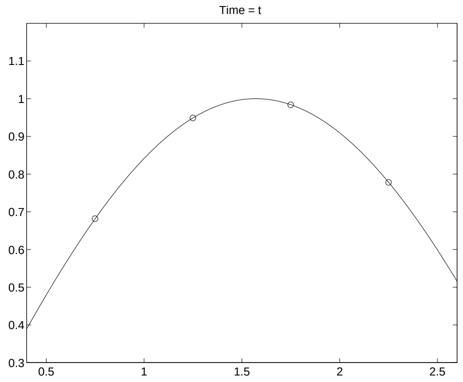


Fig.3.1. Time t

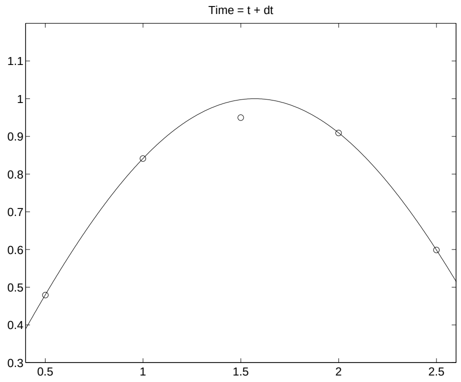


Fig.3.2. Translated to time $t + \Delta t$

It is also necessary to consider non linear methods, as seen in the following theorem

Theorem 3.2. *A linear difference approximation*

$$u_j^{n+1} = \sum_{k=-q}^{p} a_k u_{j+k}^n$$

which is TVD, is at most first order accurate.

Proof: Consider the function

$$u_j^n = \begin{cases} 1 & j \leq 0 \\ 0 & j > 0 \end{cases} .$$

Then $TV(u^n) = 1$. We evaluate the variation after one time step

$$TV(u^{n+1}) = \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{n+1}| = \sum_{j=-\infty}^{\infty} | \sum_{k=-q}^{p} a_k \Delta_+ u_{j+k}^n | = \sum_{k=-q}^{p} |a_k|$$

The scheme is consistent if

$$\sum_{k=-q}^{p} a_k = 1 \qquad (3.4)$$

if $a_k < 0$ for some k, then (3.4) gives

$$TV(u^{n+1}) = \sum_{k=-q}^{p} |a_k| > 1$$

and the method is not TVD. Thus $a_k \geq 0$ and the result follows from theorem 2.9 since a_k positive means that the scheme is monotone.

Note that the theorem does not state which partial differential equation we approximate. Second order TVD schemes (away from smooth extrema) must be non linear even when applied to the linear partial differential equation $u_t + au_x = 0$.

There exist a large number of second order TVD methods. They have in common that they all degenerate to first order accuracy at smooth extrema, and are non linear schemes. We distinguish two main classes of methods

1. Equation simultaneously discretized in time and space. These schemes are TVD modifications of the Lax-Wendroff scheme

$$u_j^{n+1} = u_j^n - \Delta t D_0 f(u_j^n) + \frac{\Delta t^2}{2} D_+(a_{j-1/2}^2 D_- u_j^n)$$

2. Spatially second order semi discrete approximations, which leaves the time discretization as a separate choice. These schemes are TVD modifications of the method

$$\frac{du_j}{dt} = -D_0 f(u_j^n)$$

What method to use depends on the specific application. A general guideline can be given based on the unmodified schemes. The class 1 is suited for time dependent calculations, while methods in class 2 are better for finding a stationary solution, since the spatial discretization does not depend on a time step. The generalization of the Lax-Wendroff scheme to more space dimensions than one is somewhat complicated, but operator splitting dimension by dimension can be used. For the semi-discrete methods, the two and three dimensional cases are straightforward.

We describe second order TVD schemes based on the Lax-Wendroff method in sections 3.2, 3.3 and 3.4. Sections 3.5 and 3.6 deal with semi discrete methods.

3.2 The Modified Flux Method

For the linear problem $u_t + au_x = 0$, one can show that the highest order of accuracy for a three point scheme is two, and that the Lax-Wendroff scheme is the only scheme, which has this optimal property. This scheme is not TVD. We show below how it is possible to modify the viscosity of the Lax-Wendroff scheme so that it becomes a TVD viscosity wherever necessary, i.e. in the neighborhood of shocks.

The method is sometimes named the modified flux method, due to the following interpretation. If $h_{j+1/2}$ is the numerical flux of a first order TVD scheme then

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{h_{j+1/2}^n-h_{j-1/2}^n}{\Delta x}=u_t+f(u)_x+O(\Delta x).$$

If $h_{j+1/2}^{LW}$ is the numerical flux of the Lax-Wendroff scheme then

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{h_{j+1/2}^{nLW}-h_{j-1/2}^{nLW}}{\Delta x}=u_t+f(u)_x+O(\Delta x^2),$$

and therefore

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{h_{j+1/2}^n-h_{j-1/2}^n}{\Delta x}=u_t+f(u)_x+ \\ \frac{h_{j+1/2}^n-h_{j+1/2}^{nLW}-(h_{j-1/2}^n-h_{j-1/2}^{nLW})}{\Delta x}+O(\Delta x^2).$$

From this formula we immediately obtain

Lemma 3.3. *If*

$$h_{j+1/2}-h_{j+1/2}^{LW}=O(\Delta x^2)$$

and if the leading error term in the $O(\Delta x^2)$ is smooth, then the method using the flux $h_{j+1/2}$ is second order accurate.

Proof: The assumption of smooth error term give

$$\frac{\Delta_+ h_{j-1/2}}{\Delta x}=\frac{\Delta_+ h_{j-1/2}^{LW}}{\Delta x}+\frac{\Delta_+ O(\Delta x^2)}{\Delta x}=\frac{\Delta_+ h_{j-1/2}^{LW}}{\Delta x}+O(\Delta x^2)$$

Thus one power of Δx is lost from dividing by Δx and one power gained by taking the difference.

We will apply a first order TVD method to a problem with the modified flux function

$$g_j = f_j + \frac{1}{\lambda}b_j$$

where b_j is some quantity resembling the difference

$$\lambda(h_{j+1/2}^{LW}-h_{j+1/2})=\frac{1}{2}(Q_{j+1/2}-Q_{j+1/2}^{LW})\Delta_+ u_j^n$$

Thus we let $Q^{LW}_{j+1/2}$ denote the numerical viscosity of the Lax-Wendroff method, and $Q_{j+1/2}$ the viscosity of the first order TVD method.

Apply now a first order TVD scheme to the problem with flux function g_j, the modified numerical flux becomes

$$h^M_{j+1/2} = \frac{1}{2}(f_{j+1} + f_j) - \frac{1}{2\lambda}(Q(g)_{j+1/2}\Delta_+ u^n_j - (b_{j+1} + b_j)) \tag{3.5}$$

where we write $Q(g)_{j+1/2}$ to stress that the viscosity is evaluated using the flux g_j. Theorem 2.13. is used to find the condition for TVD

$$|\lambda a_{j+1/2} + \frac{b_{j+1} - b_j}{u_{j+1} - u_j}| \leq Q(g)_{j+1/2} \leq 1 \tag{3.6}$$

By comparison with the Lax-Wendroff flux we get the following condition for second order accuracy

$$Q(g)_{j+1/2} - \frac{b_{j+1} + b_j}{u_{j+1} - u_j} = Q^{LW}_{j+1/2} + O(\Delta x) \tag{3.7}$$

We have two problems here, the first is how to determine the flux modification b_j so that (3.6) and (3.7) can be satisfied. A second problem arises if b_j is only known at the grid points, but our TVD scheme requires flux values intermediate between grid points when $Q(g)$ is to be evaluated, such as e.g. the Godunov or the Engquist-Osher schemes.

Introduce the notation

$$d_{j+1/2} = \frac{1}{2}(Q_{j+1/2} - Q^{LW}_{j+1/2})\Delta_+ u^n_j$$

Define

$$b_j = \begin{cases} 0 & \text{if } \Delta_+ u^n_j \Delta_- u^n_j < 0 \\ \text{sign}(\Delta_+ u^n_j)\min(|d_{j+1/2}|, |d_{j-1/2}|) & \text{otherwise} \end{cases} \tag{3.8}$$

Note that $b_j = 0$ at extrema, and thus that no modification will be made there. The accuracy at extrema is first order in accordance with theorem 3.1. Note also that because of theorem 2.13 $Q_{j+1/2} - Q^{LW}_{j+1/2} > 0$, and thus that $d_{j+1/2}$ and $\Delta_+ u^n_j$ have the same sign.

Theorem 3.4. *If b_j is given by (3.8) then the scheme with numerical flux (3.5) is second order accurate away from extrema if the viscosity coefficient, $Q(g)_{j+1/2}$ satisfies*

$$|Q(g)_{j+1/2} - Q_{j+1/2}| = O(\Delta x) \tag{3.9}$$

Proof: Assume that the solution is a smooth function. Away from extrema $b_j = d_{j+1/2}$ or $d_{j-1/2}$ and similarly for b_{j+1}, but

$$d_{j+1/2} = d_{j-1/2} + O(\Delta x^2)$$

so that

$$\frac{b_{j+1} + b_j}{\Delta_+ u^n_j} = \frac{2d_{j+1/2} + O(\Delta x^2)}{\Delta_+ u^n_j} = 2d_{j+1/2}/\Delta_+ u^n_j + O(\Delta x)$$

and thus, by the definition of $d_{j+1/2}$,

$$Q(g)_{j+1/2} - \frac{b_{j+1}+b_j}{\Delta_+ u_j^n} = Q(g)_{j+1/2} - Q_{j+1/2} + Q^{LW}_{j+1/2} + O(\Delta x)$$

The condition for second order accuracy (3.7) is satisfied if

$$Q(g)_{j+1/2} - Q_{j+1/2} = O(\Delta x)$$

This theorem had been very easy to prove if we had defined $b_j = d_{j+1/2}$ always, the more complicated definition of b_j is made to make it possible to prove the TVD property of the method, which we now proceed to do. First we define the modified viscosity as

$$Q(g)_{j+1/2} = Q_{j+1/2} + |\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| \tag{3.10}$$

Thus we use an upwind approximation to the modified part of the flux. With this viscosity we prove

Theorem 3.5. *The scheme defined by the numerical flux (3.5) and with b_j given by (3.8) and $Q(g)_{j+1/2}$ by (3.10) is TVD and second order accurate away from extrema under the cfl condition $Q_{j+1/2} \leq \frac{2}{3}$*

Proof: The lower part of the TVD inequality (3.6) is immediate from the triangle inequality

$$|\lambda a_{j+1/2} + \frac{b_{j+1}-b_j}{u_{j+1}-u_j}| \leq Q_{j+1/2} + |\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| = Q(g)_{j+1/2}$$

and where we use that $Q_{j+1/2}$ is the viscosity of a TVD scheme. The upper limit

$$Q_{j+1/2} + |\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| \leq 1$$

is shown using the fact that b_j and b_{j+1} always have the same sign and thus that

$$|\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| \leq \frac{\max(|b_j|,|b_{j+1}|)}{|\Delta_+ u^n_j|} \leq \frac{1}{2}(Q_{j+1/2} - Q^{LW}_{j+1/2})$$

so that

$$Q(g)_{j+1/2} = Q_{j+1/2} + |\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| \leq \frac{3}{2}Q_{j+1/2} - \frac{1}{2}Q^{LW}_{j+1/2} \leq \frac{3}{2}Q_{j+1/2}$$

The upper inequality is satisfied under the cfl condition $Q_{j+1/2} \leq \frac{2}{3}$. Second order accuracy follows directly from theorem 3.4 by the observation that

$$|\frac{b_{j+1}-b_j}{u^n_{j+1}-u^n_j}| = O(\Delta x)$$

And the theorem has been proved. Note that in the special case of an upwind approximation $Q_{j+1/2} = \lambda|a_{j+1/2}|$ the cfl condition can be relaxed, because then $Q^{LW} = Q^2$, and the upper TVD inequality becomes

$$\frac{3}{2}Q_{j+1/2} - \frac{1}{2}Q^2_{j+1/2} = \frac{1}{2}(Q_{j+1/2}+1) - \frac{1}{2}(1-Q_{j+1/2})^2 \le \frac{1}{2}(Q_{j+1/2}+1)$$

which is ≤ 1 if $\lambda|a_{j+1/2}| \le 1$.

Instead of defining $Q(g)_{j+1/2}$ through (3.10) we could have extended b_j (defined by (3.8)) to be defined for all u by a piecewise linear interpolation. It is then possible to prove that any first order three point TVD scheme applied to the flux function $f + \frac{1}{\lambda}b$ will lead to a $Q(g)_{j+1/2}$ which satisfies the requirements above for second order accuracy away from extrema and TVD, under a cfl condition similar to the one above.

The scheme using (3.8), (3.10) can be rewritten as

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(f_{j+1}^n - f_{j-1}^n) + \frac{1}{2}\Delta_+(Q^n_{j-1/2}\Delta_-u_j^n) -$$
$$\frac{1}{2}(b_{j+1} - b_{j-1}) + \frac{1}{2}\Delta_+(|\frac{\Delta_-b_j}{\Delta_-u_j^n}|\Delta_-u_j^n)$$

i.e. to convert a first order TVD scheme to a second order one, we can add the extra terms

$$-\frac{1}{2}(b_{j+1} - b_{j-1}) + \frac{1}{2}\Delta_+(|\frac{\Delta_-b_j}{\Delta_-u_j^n}|\Delta_-u_j^n)$$

without changing the original scheme. This makes the modification easy to implement into a computer program where the first order method is available. The correction term is sometimes called antidiffusive flux, since it is consistent with the equation $u_t = -cu_{xx}$ with $c > 0$.

It is easy to see that the method has a five point stencil, and that it is a non linear method when applied to the linear equation $u_t + au_x = 0$.

3.3 The Weighted Upwind-Lax-Wendroff Method

We next described another class of methods, based on the same idea of switching to the Lax-Wendroff method whenever possible due to the TVD constraint. This second class of methods have all numerical flux functions which can be written as a weighted average of the upwind method and the Lax-Wendroff method,

$$h_{j+1/2} = (1 - w_{j+1/2})h^{UPW}_{j+1/2} + w_{j+1/2}h^{LW}_{j+1/2}.$$

Any first order TVD method can be used instead of the upwind flux, $h^{UPW}_{j+1/2}$. The idea is to have $w_{j+1/2} \approx 1$, when the solution is smooth, and $w_{j+1/2} \approx 0$ near discontinuities. Note that the methods in the previous section can not be written in this way, due to the non-linear dependence of $Q(g)_{j+1/2}$ on the modified flux.

For this class of methods, the known results about TVD have mainly been worked out for the linear problem $u_t + au_x = 0$. For this problem we obtain the numerical flux

$$h_{j+1/2} = a(u_{j+1} + u_j)/2 - \frac{1}{2}\lambda|a|\Delta_+ u_j + \frac{1}{2}(\lambda|a| - (\lambda a)^2)w_{j+1/2}\Delta_+ u_j \qquad (3.11)$$

Example of weight functions are

$$w_{j+1/2} = \begin{cases} \phi(r_j) & \text{if } a > 0 \\ \phi(1/r_{j+1}) & \text{if } a < 0 \end{cases} \qquad (3.12)$$

or

$$w_{j+1/2} = \phi(r_j) + \phi(1/r_{j+1}) - 1 \qquad (3.13)$$

Where we define

$$r_j = \frac{\Delta_- u_j}{\Delta_+ u_j}$$

as a measure of the smoothness of u_j. When u_j is smooth, and does not have an extreme point, $r_j = 1 + O(\Delta x)$.

The function $\phi(r)$ is called *limiter*. We require that $\phi(1) = 1$, which implies that

$$\phi(r_j) = 1 + O(\Delta x) \quad \phi(1/r_j) = 1 + O(\Delta x)$$

and consequently

$$h_{j+1/2} = h^{LW}_{j+1/2} + (1 - w_{j+1/2})(h^{UPW}_{j+1/2} - h^{LW}_{j+1/2}) = h^{LW}_{j+1/2} + O(\Delta x)O(\Delta x)$$

at smooth non-extreme points for the weight functions (3.12), (3,13). According to lemma 3.3, $\phi(1) = 1$ thus guarantees second order of accuracy. The TVD property is investigated in the next theorem.

Theorem 3.6. *The method with numerical flux (3.11), and limiter (3.12), approximating $u_t + au_x = 0$ is TVD if $\phi(r)$ satisfies*

$$0 \le \phi(r) \le 2 \quad 0 \le \phi(r)/r \le 2$$

Proof: Assume that $a > 0$. The proof for $a < 0$ is similar. We will apply theorem 2.12, and begin therefore by writing the method using (3.11), (3.12) as

$$u_j^{n+1} = u_j^n - \lambda a \Delta_- u_j^n - \frac{\lambda a - (\lambda a)^2}{2}(\phi(r_j)\Delta_+ u_j^n - \phi(r_{j-1})\Delta_- u_j^n)$$

With the definitions

$$C_{j+1/2} = 0$$
$$D_{j-1/2} = \lambda a + \frac{\lambda a - (\lambda a)^2}{2}(\frac{1}{r_j}\phi(r_j) - \phi(r_{j-1}))$$

we can write the method as

$$u_j^{n+1} = u_j^n + C_{j+1/2}\Delta_+ u_j^n - D_{j-1/2}\Delta_- u_j^n.$$

Assuming the cfl condition $\lambda a \leq 1$, we see that the TVD condition $0 \leq D_{j-1/2} \leq 1$ is satisfied if

$$-2 \leq \frac{1}{r_j}\phi(r_j) - \phi(r_{j-1}) \leq 2.$$

This condition is true if e.g.

$$0 \leq \phi(r) \leq 2 \qquad 0 \leq \phi(r)/r \leq 2$$

Example of a function satisfying the conditions on $\phi(r)$ in theorem 3.6 is

$$\phi(r) = \begin{cases} \frac{2r}{r+1} & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases}$$

There is a special terminology for this class of methods. The scheme with limiter (3.12) is called an *upwind* TVD scheme, and the scheme with the limiter (3.13) a *symmetric* TVD scheme, thus indicating whether the upwind direction is required in the computation of the weight function. Note that in both cases the upwind direction is required when computing the flux $h_{j+1/2}^{UPW}$. The symmetric TVD scheme is simpler than the upwind TVD scheme, but we pay for the simlicity because the TVD analysis for the case (3.13) (exercise 3) will give more restrictive conditions on ϕ.

3.4 The Flux Corrected Transport Method

The methods described in sections 3.2 and 3.3 can abstractly be written

$$u^{n+1} = L(u^n) + M(u^n)$$

where L is the first order TVD scheme, and M is the modification such that the resulting scheme is TVD and such that $L + M$ is the Lax-Wendroff scheme whenever possible due to the TVD constraint.

We now turn to another method based on the same idea of modifying the Lax-Wendroff scheme, but instead on the form

$$\begin{aligned} u^* &= L(u^n) \\ u^{n+1} &= u^* + M(u^*) \end{aligned} \tag{3.14}$$

where L is a first order TVD scheme and M is a modification such that $L(u^n)+M(L(u^n))$ is TVD and the Lax-Wendroff scheme whenever possible. We thus implement the second order modification as a corrector step to the TVD predictor. This method is known as the flux corrected transport method (FCT). We thus use the predictor step

$$u_j^* = u_j^n - \lambda\Delta_- h_{j+1/2}^n$$

where $h_{j+1/2}^n$ is the numerical flux of a first order TVD method. The corrector step is

$$u_j^{n+1} = u_j^* - (b_{j+1/2} - b_{j-1/2}) \tag{3.15}$$

where

$$b_{j+1/2} = \begin{cases} 0 & \text{if } \Delta_+ u_j^* \Delta_- u_j^* < 0 \text{ or } \Delta_+ u_{j+1}^* \Delta_- u_{j+1}^* < 0 \\ s\min(\frac{1}{2}|\Delta_- u_j^*|, d_{j+1/2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_+ u_{j+1}^*|) & \text{otherwise} \end{cases}. \tag{3.16}$$

Here $s = \text{sign}(\Delta_+ u_j^*)$ and $d_{j+1/2} = \frac{1}{2}(Q_{j+1/2} - Q_{j+1/2}^{LW})$, where $Q_{j+1/2}$ is the numerical viscosity of the first order predictor, and $Q_{j+1/2}^{LW}$ is the numerical viscosity of the Lax-Wendroff method.

Again we can see that no change is made at extrema, and thus that the accuracy is only first order there. The easiest way to understand the formula above is through the proof of the following theorem.

Theorem 3.7. *The FCT method (3.14) where L is a first order TVD scheme and M is given by (3.15), (3.16) is TVD and second order accurate away from extrema.*

Proof: To prove TVD define

$$f_j = \begin{cases} 0 & \text{if } \Delta_+ u_j^* \Delta_- u_j^* < 0 \\ s\min(\frac{1}{2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_- u_j^*|) & \text{otherwise} \end{cases}$$

with $s = \text{sign}(\Delta_+ u_j^*)$. Write the corrector as

$$u_j^{n+1} = u_j^* + C_{j+1/2}\Delta_+ u_j^* - D_{j-1/2}\Delta_- u_j^*$$

with

$$C_{j+1/2} = \frac{-b_{j+1/2} + f_j}{\Delta_+ u_j^*} \quad D_{j-1/2} = \frac{-b_{j-1/2} + f_j}{\Delta_- u_j^*}$$

and then use theorem 2.12 to show that $TV(u^{n+1}) \leq TV(u^*)$. TVD of the total method follows since the predictor assures that $TV(u^*) \leq TV(u^n)$. At extrema $b_{j+1/2} = 0$ and C, D are obviously non negative. Assume that $\Delta_+ u_j^* \Delta_- u_j^* > 0$. We then have

$$\begin{aligned} C_{j+1/2} = & \frac{sign(\Delta_+ u_j^*)}{\Delta_+ u_j^*}(\min(\frac{1}{2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_- u_j^*|) - \\ & \min(\frac{1}{2}|\Delta_+ u_{j+1}^*|, d_{j+1/2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_- u_j^*|)) \geq 0 \end{aligned}$$

since $0 \leq d_{j+1/2} \leq 1/2$. Similarly for $D_{j+1/2}$ we have

$$\begin{aligned} D_{j+1/2} = & \frac{sign(\Delta_+ u_j^*)}{\Delta_+ u_j^*}(\min(\frac{1}{2}|\Delta_+ u_{j+1}^*|, \frac{1}{2}|\Delta_+ u_j^*|) - \\ & \min(\frac{1}{2}|\Delta_+ u_{j+1}^*|, d_{j+1/2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_- u_j^*|)) \geq 0 \end{aligned}$$

Finally we have to prove that $C_{j+1/2} + D_{j+1/2} \leq 1$. This follows from

$$C_{j+1/2} + D_{j+1/2} = \frac{f_{j+1} + f_j - 2b_{j+1/2}}{\Delta_+ u_j^*} \leq \frac{\frac{1}{2}|\Delta_+ u_j^*| + \frac{1}{2}|\Delta_+ u_j^*| - 2|b_{j+1/2}|}{|\Delta_+ u_j^*|} \leq 1$$

Next we prove second order accuracy. Assume that u_j is smooth, and that there are no local extrema. Then

$$\begin{aligned} b_{j+1/2} &= s \min(\frac{1}{2}|\Delta_+ u_j^* + O(\Delta x^2)|, d_{j+1/2}|\Delta_+ u_j^*|, \frac{1}{2}|\Delta_+ u_j^* + O(\Delta x^2)|) \\ &= d_{j+1/2}\Delta_+ u_j^* + O(\Delta x^2) \end{aligned}$$

and, since $u_j^* = u_j^n + O(\Delta x)$,

$$\begin{aligned} b_{j+1/2} = \frac{1}{2}(Q_{j+1/2} - Q_{j+1/2}^{LW})\Delta_+ u_j^* + O(\Delta x^2) = \\ \frac{1}{2}(Q_{j+1/2} - Q_{j+1/2}^{LW})(\Delta_+ u_j^n + O(\Delta x^2)) + O(\Delta x^2) = \\ \frac{1}{2}(Q_{j+1/2} - Q_{j+1/2}^{LW})\Delta_+ u_j^n + O(\Delta x^2) \end{aligned}$$

Thus for the total flux of the FCT method we have

$$\begin{aligned} h_{j+1/2}^n + \frac{1}{\lambda}b_{j+1/2} = \frac{1}{2}(f_{j+1}^n + f_j^n) - \frac{1}{2\lambda}Q_{j+1/2}\Delta_+ u_j^n + \\ \frac{1}{2\lambda}(Q_{j+1/2} - Q_{j+1/2}^{LW})\Delta_+ u_j^n + O(\Delta x^2) = \quad . \\ h_{j+1/2}^{LW} + O(\Delta x^2) \end{aligned}$$

The scheme is Lax-Wendroff up to truncation error, and thus second order accurate (see lemma 3.3).

Remark: The method of artificial compression (ACM) is a method on the form (3.14), (3.15), but with

$$b_{j+1/2} = \begin{cases} 0 \quad \text{if } \Delta_+ u_j^* \Delta_- u_j^* < 0 \text{ or } \Delta_+ u_{j+1}^* \Delta_- u_{j+1}^* < 0 \\ \operatorname{sign}(\Delta_+ u_j^*) \min(|\Delta_- u_j^*|, |\Delta_+ u_j^*|, |\Delta_+ u_{j+1}^*|) \text{ otherwise} \end{cases}$$

this correction sharpens discontinuities and can be made TVD with some changes, but is not in general second order accurate (not even away from extrema).

Originally, FCT was defined using the scheme in exercise 2.2 as predictor. This gives $d_{j+1/2} = \frac{1}{8}$, a constant, and the computation of the antidiffusive flux in the corrector step becomes very simple. Furthermore, FCT was defined using the corrector flux

$$b_{j+1/2} = \begin{cases} 0 \quad \text{if } \Delta_+ u_j^* \Delta_- u_j^* < 0 \text{ or } \Delta_+ u_{j+1}^* \Delta_- u_{j+1}^* < 0 \\ s \min(|\Delta_- u_j^*|, d_{j+1/2}|\Delta_+ u_j^*|, |\Delta_+ u_{j+1}^*|) \text{ otherwise} \end{cases} . \tag{3.17}$$

which in general does not lead to a TVD method.

We have here modified the flux (3.17) with factors $\frac{1}{2}$ in some places, to make the total method TVD for arbitrary TVD predictors. Alternatively a more restrictive CFL condition could have been imposed on the corrector step, e.g. $\lambda \leq 1/2$.

Example 3.1 An example to show that (3.17) can increase the variation. Take the monotone function

$$u_1^* = 0 \;\; u_2^* = 1 \;\; u_3^* = 2 \;\; u_4^* = 2.1 \;\; u_5^* = 3 \;\; u_6^* = 4$$

Using (3.14) with $d = \frac{1}{8}$ gives

$$b_{1+1/2} = 1/8 \;\; b_{2+1/2} = 1/10 \;\; b_{3+1/2} = 1/80 \;\; b_{4+1/2} = 1/10 \;\; b_{5+1/2} = 1/8$$

and finally

$$u_1^{n+1} = 0 \;\; u_2^{n+1} = 1.025 \;\; u_3^{n+1} = 2.0875 \;\; u_4^{n+1} = 2.0125 \;\; u_5^{n+1} = 2.975 \;\; u_6^{n+1} = 0$$

A maximum and a minimum have been introduced, which leads to an increase in variation.

3.5 Semi Discrete Inner TVD Schemes

The semi-discrete methods are divided into two different groups, the inner schemes which are the analogue of

$$\frac{du_j(t)}{dt} = -\frac{1}{\Delta x}\Delta_- f(\frac{1}{2}(u_{j+1}+u_j))$$

and the outer schemes which are the analogue of

$$\frac{du_j(t)}{dt} = -\frac{1}{\Delta x}\Delta_- \frac{1}{2}(f(u_{j+1})+f(u_j))$$

Before starting the description, we state the semi discrete version of theorem 2.12. A semi discrete method is TVD if

$$TV(u(t_2)) \leq TV(u(t_1)) \quad \text{all } \; t_2 > t_1$$

and the theorem is

Theorem 3.8. *The method*

$$\frac{du_j}{dt} = C_{j+1/2}\Delta_+ u_j - D_{j-1/2}\Delta_- u_j$$

is TVD if

$$C_{j+1/2} \geq 0 \quad D_{j+1/2} \geq 0$$

Proof: Is left to the reader.

In addition to this TVD condition, we will also require

$$C_{j+1/2}\Delta x \leq A \qquad D_{j+1/2}\Delta x \leq A \tag{3.18}$$

where A is a constant. This because if the problem is discretized in time with an explicit method one gets the third condition in theorem 2.12 (or a similar condition if another method than forward Euler is used in time)

$$\Delta t(C_{j+1/2} + D_{j+1/2}) \leq 1$$

which can be satisfied for a cfl condition $\lambda \leq 1/(2A)$ if (3.18) hold.

We start with a description of the inner TVD schemes. Assume

$$h_{j+1/2} = h(u_{j+1}, u_j)$$

is a numerical flux of a three point first order TVD scheme. This is an approximation to the flux in the intermediate point $x_{j+1/2}$. As a more accurate approximation of this flux we instead take

$$h_{j+1/2} = h(u^R_{j+1/2}, u^L_{j+1/2}) \tag{3.19}$$

and use

$$\frac{du_j(t)}{dt} = -\frac{1}{\Delta x}\Delta_- h_{j+1/2} \tag{3.20}$$

where $u^R_{j+1/2}$ and $u^L_{j+1/2}$ are approximations from the right and from the left to the value of u at the point $x_{j+1/2}$.

One way to interpret this is that a piecewise linear interpolation of the values u_j is made

$$u = u_j + s_j(x - x_j)/\Delta x \quad x_{j-1/2} < x < x_{j+1/2}$$

we then take

$$\begin{aligned} u^R_{j-1/2} &= u_j - s_j/2 \\ u^L_{j+1/2} &= u_j + s_j/2 \end{aligned} \tag{3.21}$$

The slopes have to be constructed so that they meet the requirements for second order accuracy and TVD. We will here follow a more general outline, and allow $u^R_{j+1/2}$ and $u^L_{j+1/2}$ to be any values, not necessarily obtained from piecewise linear interpolation.

Remark: The inner scheme with piecewise linear interpolation is sometimes referred to as "the MUSCL scheme".

The condition for second order accuracy can be seen from

Theorem 3.9. *If*

$$\begin{aligned} u^R_{j+1/2} - u_{j+1/2} &= O(\Delta x^2) \\ u^L_{j+1/2} - u_{j+1/2} &= O(\Delta x^2) \end{aligned}$$

where $u_{j+1/2} = (u_{j+1} + u_j)/2$ and the numerical flux function is Lipschitz continuous then the approximation (3.20) is second order accurate in space.

Proof: The numerical flux $f(\frac{1}{2}(u_{j+1}+u_j))$ leads to a second order accurate method. We prove the theorem by showing

$$h_{j+1/2} - f(\frac{1}{2}(u_{j+1} + u_j)) = O(\Delta x^2)$$

Begin by using the consistency $f(u) = h(u, u)$, then use the Lipschitz condition

$$\begin{aligned} &h(u^R_{j+1/2}, u^L_{j+1/2}) - f(u_{j+1/2}) = \\ &h(u^R_{j+1/2}, u^L_{j+1/2}) - h(u_{j+1/2}, u_{j+1/2}) \leq \\ &L_1(u^R_{j+1/2} - u_{j+1/2}) + L_2(u^L_{j+1/2} - u_{j+1/2}) \end{aligned}$$

We can see that $h_{j+1/2} - f(u_{j+1/2}) = O(\Delta x^2)$ and thus the order is two. In the same way as in lemma 3.3, it is necessary that the leading term in the $O(\Delta x^2)$ is smooth. This will not always be the case near extreme points in the methods described below.

We give two sets of conditions for TVD, the first is

Theorem 3.10. *If the scheme with numerical flux $h(u_{j+1}, u_j)$ is TVD, then the approximation using the numerical flux $h(u^R_{j+1/2}, u^L_{j+1/2})$ is TVD if*

$$\begin{aligned}
\frac{u^R_{j+1/2} - u^L_{j+1/2}}{u_{j+1} - u_j} &\geq 0 \\
\frac{u^R_{j-1/2} - u^L_{j+1/2}}{u_{j+1} - u_j} &\leq 0 \\
\frac{u^R_{j-1/2} - u^L_{j+1/2}}{u_j - u_{j-1}} &\leq 0
\end{aligned} \tag{3.22}$$

Proof: Write $-\Delta_+ h(u^R_{j-1/2}, u^L_{j-1/2})$ as

$$\begin{aligned}
&-\left(h(u^R_{j+1/2}, u^L_{j+1/2}) - f(u^L_{j+1/2})\right) + \left(h(u^R_{j-1/2}, u^L_{j+1/2}) - f(u^L_{j+1/2})\right) \\
&-\left(h(u^R_{j-1/2}, u^L_{j+1/2}) - f(u^R_{j-1/2})\right) + \left(h(u^R_{j-1/2}, u^L_{j-1/2}) - f(u^R_{j-1/2})\right)
\end{aligned}$$

From the theory of first order schemes, we know that

$$\begin{aligned}
C^{(1)}_{j+1/2} &= -\lambda \frac{h(u_{j+1}, u_j) - f(u_j)}{u_{j+1} - u_j} \geq 0 \\
D^{(1)}_{j-1/2} &= -\lambda \frac{h(u_j, u_{j-1}) - f(u_j)}{u_j - u_{j-1}} \geq 0
\end{aligned} \tag{3.23}$$

which can be used by writing

$$-\Delta_+ h(u^R_{j-1/2}, u^L_{j-1/2}) = C_{j+1/2}\Delta_+ u_j - D_{j-1/2}\Delta_- u_j$$

with

$$\begin{aligned}
C_{j+1/2} = &-\frac{h(u^R_{j+1/2}, u^L_{j+1/2}) - f(u^L_{j+1/2})}{u^R_{j+1/2} - u^L_{j+1/2}} \frac{u^R_{j+1/2} - u^L_{j+1/2}}{u_{j+1} - u_j} + \\
&\frac{h(u^R_{j-1/2}, u^L_{j+1/2}) - f(u^L_{j+1/2})}{u^R_{j-1/2} - u^L_{j+1/2}} \frac{u^R_{j-1/2} - u^L_{j+1/2}}{u_{j+1} - u_j} \\
D_{j-1/2} = &\frac{h(u^R_{j-1/2}, u^L_{j+1/2}) - f(u^R_{j-1/2})}{u^R_{j-1/2} - u^L_{j+1/2}} \frac{u^R_{j-1/2} - u^L_{j+1/2}}{u_j - u_{j-1}} - \\
&\frac{h(u^R_{j-1/2}, u^L_{j-1/2}) - f(u^R_{j-1/2})}{u^R_{j-1/2} - u^L_{j-1/2}} \frac{u^R_{j-1/2} - u^L_{j-1/2}}{u_j - u_{j-1}}
\end{aligned}$$

thus by using (3.23), we find that $C_{j+1/2} \geq 0, D_{j+1/2} \geq 0$ if (3.22) holds.

We can obtain less restrictive TVD conditions if we add assumptions about the first order numerical flux. One example of this is

Theorem 3.11. *If the scheme with numerical flux $h(u_{j+1}, u_j)$ is monotone, then the approximation using the numerical flux $h(u^R_{j+1/2}, u^L_{j+1/2})$ is TVD if*

$$\begin{aligned} \frac{u^R_{j+1/2} - u^R_{j-1/2}}{u_{j+1} - u_j} &\geq 0 \\ \frac{u^L_{j+1/2} - u^L_{j-1/2}}{u_j - u_{j-1}} &\geq 0 \end{aligned} \tag{3.24}$$

Proof: Write $-\Delta_+ h(u^R_{j-1/2}, u^L_{j-1/2})$ as

$$\begin{aligned} &-(h(u^R_{j+1/2}, u^L_{j+1/2}) - h(u^R_{j-1/2}, u^L_{j+1/2})) - \\ &\qquad (h(u^R_{j-1/2}, u^L_{j+1/2}) - h(u^R_{j-1/2}, u^L_{j-1/2})) = \\ &- \int_0^1 h_1(u^R_{j-1/2} + \theta\Delta_+ u^R_{j-1/2}, u^L_{j+1/2})\, d\theta \Delta_+ u^R_{j-1/2} - \\ &\qquad \int_0^1 h_2(u^R_{j-1/2}, u^L_{j-1/2} + \theta\Delta_+ u^L_{j-1/2})\, d\theta \Delta_+ u^L_{j-1/2} \end{aligned}$$

where h_1 and h_2 are the derivatives of h with respect to its first and second argument respectively. Since the scheme

$$u^{n+1}_j = u^n_j - \lambda(h(u^n_{j+1}, u^n_j) - h(u^n_j, u^n_{j-1}))$$

is assumed to be monotone, $h_1 < 0$ and $h_2 > 0$. Thus by taking

$$\begin{aligned} C_{j+1/2} &= - \int_0^1 h_1(u^R_{j-1/2} + \theta\Delta_+ u^R_{j-1/2}, u^L_{j+1/2})\, d\theta \frac{u^R_{j+1/2} - u^R_{j-1/2}}{u_{j+1} - u_j} \\ D_{j-1/2} &= \int_0^1 h_2(u^R_{j-1/2}, u^L_{j-1/2} + \theta\Delta_+ u^L_{j-1/2})\, d\theta \frac{u^L_{j+1/2} - u^L_{j-1/2}}{u_j - u_{j-1}} \end{aligned}$$

TVD follows from theorem 2.12 if (3.24) holds.

We are now ready to describe how to do the piecewise linear interpolation. Apply theorem 3.10 to the right and left values (3.21). The resulting inequalities are

$$\begin{aligned} 1 - \frac{1}{2}\frac{s_{j+1} + s_j}{\Delta_+ u_j} &\geq 0 \\ \frac{s_j}{\Delta_+ u_j} &\geq 0\,. \\ \frac{s_j}{\Delta_- u_j} &\geq 0 \end{aligned}$$

If $\Delta_+ u_j \Delta_- u_j < 0$, it is necessary that $s_j = 0$, i.e. the usual degeneracy to first order accuracy at extrema. Condition for second order accuracy is obtained from theorem 3.9

$$u_{j+1} - \frac{s_{j+1}}{2} = \frac{1}{2}(u_{j+1} + u_j) + O(\Delta x^2)$$
$$u_j + \frac{s_j}{2} = \frac{1}{2}(u_{j+1} + u_j) + O(\Delta x^2)$$

which is equivalent to

$$s_j = \Delta_+ u_j + O(\Delta x^2)$$

Since $\Delta_- u_j = \Delta_+ u_j + O(\Delta x^2)$, the choice

$$s_j = \begin{cases} 0 & \text{if } \Delta_+ u_j \Delta_- u_j < 0 \\ \operatorname{sign}(\Delta_+ u_j) \min(|\Delta_+ u_j|, |\Delta_- u_j|) & \text{otherwise} \end{cases}$$

leads to a second order TVD scheme. The function above is called the *minmod* function, and we write

$$s_j = \operatorname{minmod}(\Delta_+ u_j, \Delta_- u_j)$$

This is the only example we give of a choice of slopes satisfying the requirements in theorem 3.10. Instead we now turn to theorem 3.11. The TVD requirements there gives more freedom of choice.

Apply theorem 3.11 to the right and left values (3.21). The resulting inequalities are

$$\frac{u_{j+1} - s_{j+1}/2 - u_j + s_j/2}{\Delta_+ u_j} \geq 0$$
$$\frac{u_j + s_j/2 - u_{j-1} - s_{j-1}/2}{\Delta_- u_j} \geq 0$$

which simplifies to

$$1 - \frac{s_{j+1} - s_j}{2\Delta_+ u_j} \geq 0$$
$$1 + \frac{s_{j+1} - s_j}{2\Delta_+ u_j} \geq 0$$

i.e.

$$|s_{j+1} - s_j| \leq 2|\Delta_+ u_j| \tag{3.25}$$

By taking

$$s_j = \begin{cases} 0 & \Delta_+ u_j \Delta_- u_j \leq 0 \\ B(\Delta_+ u_j, \Delta_- u_j) & \text{otherwise} \end{cases}$$

where B is a function which has the same sign as its arguments, s_j and s_{j+1} will always have the same sign. (3.25) is then satisfied if

$$\max(|s_j|, |s_{j+1}|) \leq 2|\Delta_+ u_j|$$

which holds if $B(x, y)$ is such that

$$|B(x, y)| \leq 2\min(|x|, |y|). \tag{3.26}$$

This follows because (3.26) implies that

$$\max(|s_j|, |s_{j+1}|) \leq 2\max(\min(|\Delta_+ u_j|, |\Delta_- u_j|), \min(|\Delta_+ u_{j+1}|, |\Delta_+ u_j|)) \leq 2|\Delta_+ u_j|$$

We now give some examples of functions, B, that are sometimes used in computations. The condition for second order accuracy

$$s_j = \Delta_+ u_j + O(\Delta x^2)$$

is translated into

$$B(\Delta_+ u_j, \Delta_+ u_j + O(\Delta x^2)) = \Delta_+ u_j + O(\Delta x^2)$$

This is satisfied if

$$B(x, x) = x$$

and B is Lipschitz continuous, which can easily be checked to hold for the examples below.

Example 3.2 *minmod* slope limiter. Take

$$B(x, y) = \text{sign}(x)\min(|x|, |y|)$$

(3.26) is clearly satisfied. This is the slope limiter already encountered in connection with theorem 3.10.

Example 3.3 *van Leer's* slope limiter. This is the function

$$B(x, y) = \frac{2xy}{x + y}$$

An advantage is that this is a smooth function of its arguments. (3.26) follows from

$$\begin{aligned} |B(x, y)| &\leq \frac{2|x|}{1 + |y/x|} \leq 2|x| \quad |x| \leq |y| \\ |B(x, y)| &\leq \frac{2|y|}{1 + |x/y|} \leq 2|y| \quad |y| \leq |x| \end{aligned}$$

Example 3.4 *Superbee* slope limiter. This is the function

$$B(x, y) = \begin{cases} \text{sign}(x)\max(|x|, |y|) & \text{if } x/2 \leq y \leq 2x \\ 2\text{sign}(x)\min(|x|, |y|) & \text{if } x/2 > y \text{ or } y > 2x \end{cases}$$

This is a slope limiter which gives high compression, (3.26) is easy to verify.

Example 3.5 *van Albada's* slope limiter.

$$B(x, y) = \frac{x^2 y + y^2 x}{x^2 + y^2}$$

another smooth function which satisfies the requirements for second order TVD. The reader is asked to verify (3.26). This limiter is not set equal to zero if $xy < 0$. In computer programs we use the modification

$$B(x,y) = \frac{(x^2+\epsilon^2)y + (y^2+\epsilon^2)x}{2\epsilon^2 + x^2 + y^2}$$

with ϵ a small constant, to avoid difficulties when $x = y = 0$.

Note also that the function $B(x,y) = (x+y)/2$ leads to a centered difference method, which does not satisfy the TVD conditions.

These were all examples of slope limiters. We now generalize these limiters into limiter functions, which in general can not be interpreted as piecewise linear interpolation. Instead the values $u^R_{j-1/2}, u^L_{j+1/2}$ can be considered as interpolated from the right and the left respectively. We take

$$\begin{aligned} u^R_{j-1/2} &= u_j - \frac{1}{2}\psi(\frac{1}{r_j})\Delta_+ u_j \\ u^L_{j+1/2} &= u_j + \frac{1}{2}\psi(r_j)\Delta_- u_j \end{aligned} \tag{3.27}$$

where r_j is defined as

$$r_j = \frac{\Delta_+ u_j}{\Delta_- u_j}$$

The function $\psi(r)$ is called a *limiter.* If $\psi \equiv 1$, the interpolation interpretation becomes clear. This is a generalization of the piecewise linear interpolation, assume that

$$\psi(r) = r\psi(1/r) \tag{3.28}$$

Compare with (3.21), and it is clear that (3.27) corresponds to taking

$$s_j = \psi(r_j)\Delta_- u_j$$

We will now consider (3.27), without requiring (3.28). In this way a wider class of TVD methods can be treated.

Apply theorem 3.11 to (3.27)

$$\frac{u_{j+1} - \frac{1}{2}\psi(\frac{1}{r_{j+1}})\Delta_+ u_{j+1} - (u_j - \frac{1}{2}\psi(\frac{1}{r_j})\Delta_+ u_j)}{\Delta_+ u_j} \geq 0$$
$$\frac{u_j + \frac{1}{2}\psi(r_j)\Delta_- u_j - (u_{j-1} + \frac{1}{2}\psi(r_{j-1})\Delta_- u_{j-1})}{\Delta_- u_j} \geq 0$$

which simplifies to

$$1 - \frac{1}{2}r_{j+1}\psi(\frac{1}{r_{j+1}}) + \frac{1}{2}\psi(\frac{1}{r_j}) \geq 0$$
$$1 + \frac{1}{2}\psi(r_{j+1}) - \frac{1}{2r_j}\psi(r_j) \geq 0$$

Thus if ψ is such that

$$1+\frac{1}{2}\psi(s)-\frac{1}{2r}\psi(r)\geq 0 \qquad \text{all } r,s$$

the method is TVD. By inspecting the I-form in the proof of theorem 3.11, we see that the boundedness (3.18) means that the this expression is bounded from above by a constant, i.e.

$$0\leq 1+\frac{1}{2}\psi(s)-\frac{1}{2r}\psi(r)\leq A \qquad \text{all } r,s \tag{3.29}$$

Theorem 3.12. *If the limiter function ψ is Lipschitz continuous and the following holds for all r*

$$\begin{aligned}&\psi(1)=1\\ &m\leq\psi(r)\leq M\\ &M+2-2A\leq\frac{\psi(r)}{r}\leq 2+m\end{aligned}$$

for some constants $m>-1, M, A$, then the second order semi discrete method, obtained by putting (3.27) into a first order numerical flux function is second order accurate and TVD if the first order flux corresponds to a monotone scheme.

Proof: We have seen above that (3.29) implies TVD. Estimate the expression in (3.29) using the given bounds,

$$1+\frac{1}{2}\psi(s)-\frac{1}{2r}\psi(r)\geq 1+\frac{1}{2}m-\frac{1}{2}(2+m)=0$$

We obtain the upper bound similarly,

$$1+\frac{1}{2}\psi(s)-\frac{1}{2r}\psi(r)\leq 1+\frac{1}{2}M-\frac{1}{2}(M+2-2A)=A$$

(3.29) holds and the method is TVD. Second order accuracy follows if $\psi(1)=1$ and ψ is Lipschitz in a neighborhood of 1. If $u_x\neq 0$, then $r_j=1+O(\Delta x)$ and $\psi(r_j)=1+O(\Delta x)$, thus

$$\begin{aligned}u^R_{j-1/2}&=u_j-\frac{1}{2}\psi(\frac{1}{r_j})\Delta_+u_j=u_j-\frac{1}{2}\Delta_+u_j+O(\Delta x^2)=\\ &\frac{1}{2}(u_j+u_{j-1})-\frac{1}{2}\Delta_+\Delta_-u_j+O(\Delta x^2)=\frac{1}{2}(u_j+u_{j-1})+O(\Delta x^2)\end{aligned}$$

and similarly for $u^L_{j+1/2}$. Theorem 3.9 gives second order accuracy. The condition $m>-1$ means that the point $(1,\psi(1))$ is inside the TVD region. Since $\psi(r)$ is bounded, we have in general $\lim_{r\to\pm\infty}\psi(r)/r=0$, which means that zero must be an allowed value for $\psi(r)/r$. This is no problem since A can be chosen large enough so that the lower bound is negative.

The TVD domain is outlined in the figure 3.3. together with a shaded curve indicating a limiter function inside the TVD domain.

We now give some examples of commonly used limiter functions. The reader is asked to verify that the functions satisfy the conditions for TVD and second order accuracy given in theorem 3.12.

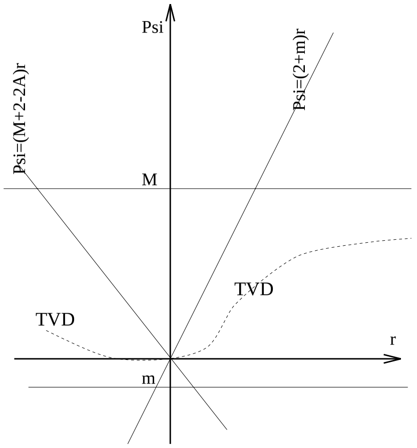


Fig.3.3. TVD region.

Example 3.6 The slope limiters given previously can be converted to limiter functions with the following results. The minmod limiter becomes

$$\psi(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ \min(1,r) & \text{otherwise} \end{cases}$$

van Leer's limiter becomes

$$\psi(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ \frac{2r}{r+1} & \text{otherwise} \end{cases}$$

The superbee limiter becomes

$$\psi(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ \max(\min(2r,1),\min(r,2)) & \text{otherwise} \end{cases}$$

and finally the van Albada's limiter becomes

$$\psi(r) = \frac{r^2+r}{r^2+1}$$

which one usually does not put equal to zero for r negative. It is inside the TVD region, also for $r < 0$.

Now we give some limiters for which $\psi(r)/r \neq \psi(1/r)$, and thus which can not be interpreted as piecewise linear interpolation.

Example 3.7 A generalization of the minmod limiter is

$$\psi(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ \min(c,r) & \text{otherwise} \end{cases}$$

where c is a constant, $1 \leq c \leq 2$.

Example 3.8 *2/3-limiters.* These are limiters with the additional property

$$\psi'(1) = 2/3$$

In this case one can show that the scheme is third order accurate away from extrema. Note that $\psi(r)/r = \psi(1/r)$ implies that $\psi'(1) = 1/2$, so that it is impossible to interpret these limiters as piecewise linear reconstruction. Some examples of limiters in this class are

$$\psi(r) = \frac{4r^2 + 2r}{3(r^2+1)}$$

or

$$\psi(r) = \frac{2r^2 + r}{2r^2 - r + 2}$$

Of course, these 2/3-limiters satisfy $\psi(1) = 1$ and are inside the TVD region in figure 3.3.

We finally show in figures 3.4 – 3.7 some of the functions in the examples above.

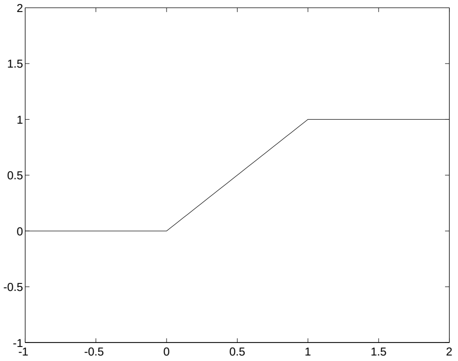


Fig.3.4. Minmod limiter function

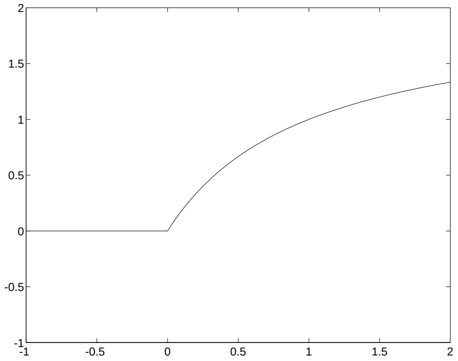


Fig.3.5. van Leer limiter function

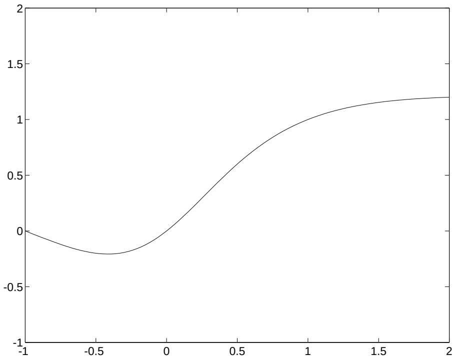


Fig.3.6. van Albada limiter function

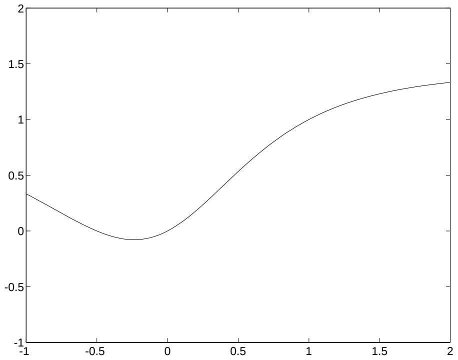


Fig.3.7. 2/3-limiter function

3.6 Semi Discrete Outer TVD Schemes

The last part of this chapter is devoted to the outer TVD semi discrete schemes. Begin with the centered difference approximation

$$\frac{du_j}{dt} = -\frac{1}{\Delta x}\Delta_+ \frac{1}{2}(f(u_j) + f(u_{j-1}))$$

introduce the numerical flux $h_{j+1/2}$ from a first order TVD scheme and rewrite the centered approximation as

$$\frac{du_j}{dt} = -\frac{1}{\Delta x}\Delta_+(h_{j-1/2} + \frac{1}{2}(f(u_j) - h_{j-1/2}) + \frac{1}{2}(f(u_{j-1}) - h_{j-1/2}))$$

we can interpret the numerical flux of this method as a first order one, with two correction terms. To make the approximation TVD it is necessary to sometimes switch off the correction terms. We do this using a limiter function (flux limiter) similar to what was done for the inner scheme. The numerical flux for the outer second order TVD schemes is

$$h^{(2)}_{j+1/2} = h_{j+1/2} + \frac{1}{2}\psi(r_j^+)(f(u_{j+1}) - h_{j+1/2}) + \frac{1}{2}\psi(r_{j+1}^-)(f(u_j) - h_{j+1/2}) \qquad (3.30)$$

where we use

$$r_j^+ = \frac{f(u_j) - h_{j-1/2}}{f(u_{j+1}) - h_{j+1/2}} \qquad r_j^- = \frac{f(u_j) - h_{j+1/2}}{f(u_{j-1}) - h_{j-1/2}}$$

these quantities and the way the limiter functions ψ depend on them are defined in such a way that a reasonably simple condition on $\psi(r)$ is obtained from the TVD requirement. The formulas above are easiest understood through the proof of the following theorem.

Theorem 3.13. *If the limiter function $\psi(r)$ satisfies*

$$0 \le 1 - \frac{1}{2}\psi(r) + \frac{1}{2s}\psi(s) \le A \quad \text{all } r, s \qquad (3.31)$$

for a constant A, then the outer semi discrete method is TVD.

Proof: We write the method with the numerical flux (3.30) on I-form

$$\begin{aligned}-\Delta_+ h^{(2)}_{j-1/2} = -(h_{j+1/2} - f_j + \frac{1}{2}(f_{j+1} - h_{j+1/2})\psi(r_j^+) + \\ \frac{1}{2}(f_j - h_{j+1/2})\psi(r_{j+1}^-) - h_{j-1/2} + f_j - \\ \frac{1}{2}(f_j - h_{j-1/2})\psi(r_{j-1}^+) - \frac{1}{2}(f_{j-1} - h_{j-1/2})\psi(r_j^-))\end{aligned}$$

Thus it is possible to define

$$C_{j+1/2} = -\frac{h_{j+1/2} - f_j}{u_{j+1} - u_j}(1 - \frac{1}{2}\psi(r_{j+1}^-) + \frac{1}{2r_j^-}\psi(r_j^-))$$

$$D_{j-1/2} = -\frac{h_{j-1/2} - f_j}{u_j - u_{j-1}}(1 - \frac{1}{2}\psi(r_{j-1}^+) + \frac{1}{2r_j^+}\psi(r_j^+))$$

from which we see that if the inequality

$$1 - \frac{1}{2}\psi(r) + \frac{1}{2s}\psi(s) \geq 0 \qquad \text{all } r, s$$

is true, then $C_{j+1/2}, D_{j+1/2}$ are non negative, and TVD follows from theorem 3.8. Finally, the boundedness (3.18) gives the upper bound.

Finally we further investigate condition (3.31) to obtain a theorem similar to theorem 3.12.

Theorem 3.14. *If the limiter function ψ is Lipschitz continuous and the following holds for all r*

$$\begin{aligned} &\psi(1) = 1 \\ &m \leq \psi(r) \leq M \\ &M - 2 \leq \frac{\psi(r)}{r} \leq 2A - 2 + m \end{aligned}$$

for some constants $m, M < 2, A$, then the second order outer semi discrete method, using the flux (3.30), where $h_{j+1/2}$ corresponds to a first order TVD scheme is second order accurate and TVD.

Proof: We verify the condition (3.31). Use the given bounds to obtain the lower limit

$$1 - \frac{1}{2}\psi(r) + \frac{1}{2s}\psi(s) \geq 1 - \frac{1}{2}M + \frac{1}{2}(M - 2) \geq 0$$

and the upper limit

$$1 - \frac{1}{2}\psi(r) + \frac{1}{2s}\psi(s) \leq 1 - \frac{1}{2}m + \frac{1}{2}(2A - 2 + m) \leq A$$

and TVD follows from theorem 3.13. Second order accuracy follows if $\psi(1) = 1$ and ψ is Lipschitz in a neighborhood of 1. If $f_x \neq 0$, then $r_j = 1 + O(\Delta x)$ and $\psi(r_j) = 1 + O(\Delta x)$, thus the numerical flux of the method becomes

$$\begin{gathered} h_{j+1/2} + \frac{1}{2}(1 + O(\Delta x))(f(u_{j+1}) - h_{j+1/2}) + \frac{1}{2}(1 + O(\Delta x))(f(u_j) - h_{j+1/2}) = \\ \frac{1}{2}(f(u_{j+1}) + f(u_j)) + O(\Delta x^2) \end{gathered}$$

where we use that the flux difference $f(u_{j+1}) - h_{j+1/2} = O(\Delta x)$ and similar for the other flux difference. Thus the numerical flux is equal to the flux of a second order scheme up to truncation error. The condition $M < 2$ means that zero is included in the interval, which is necessary by the same argument as was given in the proof of theorem 3.12. Since A can be chosen large enough, the point $(1, \psi(1))$ can be included into the TVD region.

The TVD region for the outer limiters is thus similar to the TVD region for the inner limiters, shown in figure 3.3.

In fact most of the examples of limiter functions given in this chapter satisfy both the requirements in theorem 3.12 and 3.14. The condition such a limiter has to satisfy is obtained if we combine the two TVD regions.

Theorem 3.15. *If $\psi(r)$ satisfies*

$$\begin{aligned}&\psi(1) = 1\\&m \leq \psi(r) \leq M\\&M-2 \leq \frac{\psi(r)}{r} \leq 2+m\end{aligned}$$

with $M < 2, m > -1$, then $\psi(r)$ will give a second order TVD scheme if it is used either as an inner or an outer limiter.

Remark: By choosing the constant $A = 2$, the TVD conditions for inner and outer limiters coincide.

Let us finally compare the formula (3.30) with a simpler weighted upwind-centered method, which we define in analogy with the methods in section 3.3, as having the following numerical flux function

$$h^{(2)}_{j+1/2} = (1 - w_{j+1/2})h_{j+1/2} + w_{j+1/2}h^c_{j+1/2}.$$

The centered flux is $h^c_{j+1/2} = (f_{j+1} + f_j)/2$, and $h_{j+1/2}$ is the numerical flux of a first order TVD method. With this method we can retain the simplicity of the weighted TVD methods, and at the same time avoid the difficulties at steady state and multi dimensional computations associated with the Lax-Wendroff method.

If we write out the formula above, we obtain

$$h^{(2)}_{j+1/2} = h_{j+1/2} + w_{j+1/2}((f(u_{j+1}) + f(u_j))/2 - h_{j+1/2})$$

and we see that this is a simplification of formula (3.30) where we lump together the two correction terms, so that they are multiplied with the same weight. We could e.g. define

$$w_{j+1/2} = \psi(r_j) + \psi(1/r_{j+1}) - 1$$

with

$$r_j = \frac{Q_{j-1/2}\Delta_- u_j}{Q_{j+1/2}\Delta_- u_j} = \frac{(f_j + f_{j-1})/2 - h_{j-1/2}}{(f_{j+1} + f_j)/2 - h_{j+1/2}}$$

It is possible to show TVD for this method, under conditions on ψ, similar to the previous analysis in this chapter.

Time discretization will be discussed in section 4. For the moment, we recommend the Runge-Kutta method

$$\begin{aligned}&u^1 = u^n - \lambda\Delta_+ h^n_{j-1/2}\\&u^2 = u^1 - \lambda\Delta_+ h^1_{j-1/2}\\&u^{n+1} = (u^n + u^2)/2\end{aligned}$$

to be used for second order accuracy in time. It can be proved that if the semi discrete problem is TVD, the fully discretized problem is TVD too, if the method above is used.

Exercises

1. Give conditions on s_j such that the piecewise linear interpolation

$$u(x) = u_j + s_j(x - x_j)/\Delta x \qquad x_{j-1/2} \le x \le x_{j+1/2}$$

does not increase the variation, i.e.

$$\begin{aligned} u_j + s_j/2 &\le u_{j+1} - s_{j+1}/2 && \text{if u increasing} \\ u_j + s_j/2 &\ge u_{j+1} - s_{j+1}/2 && \text{if u decreasing} \end{aligned}$$

2. The 2/3-limiters in example 3.8 can be viewed as a modification of the difference scheme (the κ-scheme)

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

for a certain value of κ. Here it is applied to the problem $u_t + au_x = 0$, $a > 0$. Investigate how the spatial accuracy depends on the parameter κ.

3. Give conditions on $\phi(r)$ such that the method with numerical flux (3.11) and limiter (3.13) is TVD when applied to the problem $u_t + au_x = 0$, $a > 0$.

4. Higher Order of Accuracy

4.1 Point values and cell averages

In this section we will not strictly enforce the TVD constraint. As we have seen, TVD leads to restrictions in the accuracy. It is, however, necessary to use some of the ideas in the previous sections to make the increase in variation "as small as possible". A pure centered difference approach is not sufficient as can be seen from the following experiment. We solve $u_t + (u^2/2)_x = 0$, with a step function as initial data. The scheme

$$\frac{du_j}{dt} = -D_0(I - \frac{\Delta x^2}{6} D_+ D_-) f(u_j) - d\Delta x^3 (D_+ D_-)^2 u_j$$

is used, discretized in time using a fourth order Runge-Kutta method. The spatial discretization is a fourth order accurate centered difference together with a fourth order artificial dissipation term. There is a viscosity parameter d left to tune the method for shocks. The result in fig. 4.1 shows the solution after shocks have formed for some different values of the viscosity d. In the first picture the viscosity parameter is too small to give substantial damping of oscillations. in the second picture, $d = 0.2$ which was the best value according to subjective judgement by looking at the results. In the last picture $d = 0.235$, which turned out to be the largest possible viscosity due to the CFL constraint. The dissipation operator was not taken into account in the CFL condition.

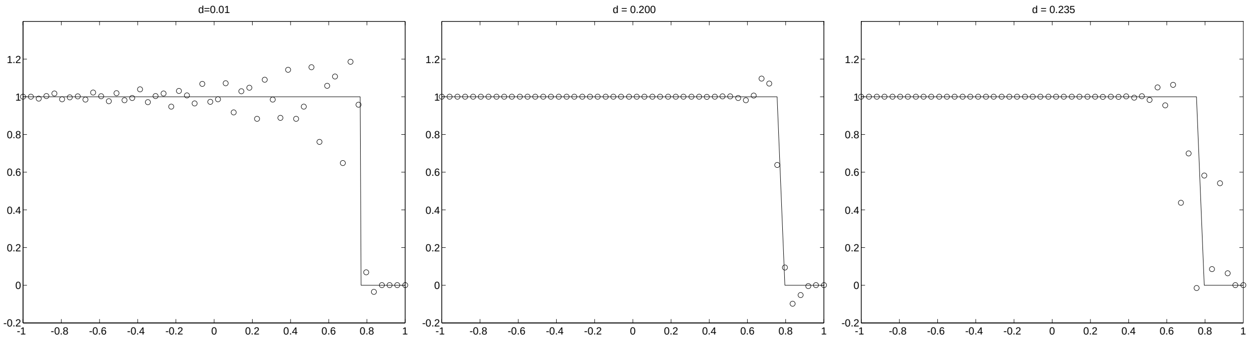


Fig. 4.1. Fourth order solution of Burger's equation.

Note that the results are not particularly good, not even after tuning the viscosity.

The higher order methods described in this chapter will give good results for this problem. We must however issue a warning that the theory for higher order non oscillatory schemes is not well developed.

This far we have not made any distinction between cell averages and point values. Consider the grid

$$x_j \quad j = \ldots, -1, 0, 1, \ldots$$

with $\Delta x = x_j - x_{j-1} =$ constant. The numerical approximation u_j^n at (t_n, x_j) can be thought of as an approximation to the point value $u(t_n, x_j)$. Alternatively, we introduce the *cells*, c_j as

$$c_j = \{x | x_{j-1/2} \leq x \leq x_{j+1/2}\}$$

where $x_{j+1/2} = (x_j + x_{j+1})/2$, and view u_j^n as an approximation to the cell average

$$\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(t_n, x)\,dx.$$

The situation is depicted in fig. 4.2.

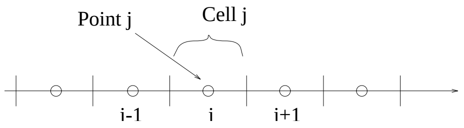


Fig. 4.2. Grid cells and grid points.

The distinction between these two views is not important for methods with accuracy ≤ 2, since

$$u(t_n, x_j) = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(t_n, x)\,dx + O(\Delta x^2)$$

In this section, however, we will treat higher order of accuracy than two. We first analyze semi discrete methods, and save the time discretization until the last section. The *cell average* based higher order schemes are the generalization of the inner schemes described in the previous chapter. The schemes starts from the following exact formula for the cell average. Integrate

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

with respect to x over one cell at t. The result is

$$\frac{d}{dt}\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(t, x)\,dx + \frac{f(u(t, x_{j+1/2})) - f(u(t, x_{j-1/2}))}{\Delta x} = 0 \qquad (4.1)$$

Compare this with the numerical approximation

$$\frac{du_j}{dt} + \frac{h_{j+1/2} - h_{j-1/2}}{\Delta x} = 0 \qquad (4.2)$$

If the numerical flux approximates the flux of the exact solution at the cell interface

$$h_{j+1/2} = f(u(t, x_{j+1/2})) + O(\Delta x^p)$$

then (4.2) is a p th order approximation to the PDE in terms of its *cell averages*.

One usual way to find higher order approximations is to make a piecewise polynomial approximation, $L(x)$ of $u(t_n, x)$ from the given cell averages u_j^n. Inside each cell $u(t_n, x)$ is approximated by a polynomial, and at the cell interfaces, $x_{j+1/2}$, there may be jumps. From this piecewise polynomial the numerical flux is obtained as $h(u_{j+1/2}^R, u_{j+1/2}^L)$, where $h(u_{j+1}, u_j)$ is the numerical flux of a first order TVD method and the end values are

$$u_{j+1/2}^R = \lim_{x \to x_{j+1/2}+} L(x)$$

$$u_{j+1/2}^L = \lim_{x \to x_{j+1/2}-} L(x)$$

Fig. 4.3 below shows a piecewise parabolic approximation.

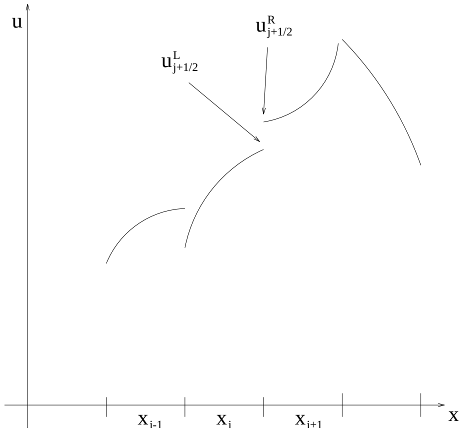


Fig. 4.3. Piecewise parabolic reconstruction.

The *point value* based higher order methods starts from the observation that if

$$f(u(x_j)) = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} F(x)\,dx$$

for some function $F(x)$, then

$$f(u(x_j))_x = \frac{F(x_{j+1/2}) - F(x_{j-1/2})}{\Delta x}$$

and thus if the numerical flux satisfies

$$h_{j+1/2} = F(x_{j+1/2}) + O(\Delta x^p)$$

the scheme (4.2) is p th order accurate in terms of point values. The function $F(x)$ can be obtained by interpolation of the grid function

$$G_{j+1/2} = \int_a^{x_{j+1/2}} F(x)\,dx = \sum_{k=a}^{j} f(u_k)\Delta x$$

and then taking the derivative of the interpolation polynomial, $F(x) = dG(x)/dx$. The point based algorithm is much easier to generalize to more than one space dimension.

In section 4.2. we show some different ways to do the piecewise polynomial reconstruction. When the time discretization is made, extra care has to be taken to get the same high order of accuracy as for the spatial approximation. This is the topic of section 4.4.

4.2 Inner interpolation gives a cell average scheme

There are three ingredients in an inner high order scheme

1. A First order numerical flux.
2. A piecewise polynomial interpolation to find $u^L_{j+1/2}$, and $u^R_{j+1/2}$.
3. A time discretization.

The topic of this section is 2., polynomial interpolation. We will consider the problem of finding the values of the solution at the cell interfaces, $u^R_{j+1/2}$, $u^L_{j+1/2}$, $j = \ldots, -1, 0, 1, \ldots$ from given cell averages. This is done by piecewise polynomial interpolation, and in such a way that the variation of the interpolant is is as small as possible. Strictly speaking, this is not an interpolation problem, since the function is given as cell averages, while an interpolation problem requires the function at certain points. The term *reconstruction* is therefore used to denote the process of finding an approximation to a function whose cell averages are given.

One method in this class is the so called piecewise parabolic method (PPM). It contains a reconstruction step using parabolic polynomials. The reconstruction algorithm contains limiters to ensure monotonicity. We here give the algorithm without details, just to give the reader an understanding for the complexity of the PPM reconstruction step.

(a) Define the primitive function $V_{j+1/2} = \sum_{k=a}^{j} u_k \Delta x$.
(b) Interpolate $V_{j+1/2}$ using piecewise quartic polynomial.
(c) Define $u^L_{j+1/2} = u^R_{j+1/2} = dV(x_{j+1/2})/dx$.
(d) Modify the left and right values obtained in (c), so that they both are between u_j and u_{j+1}.
(e) If the parabola in cell j (parabola through $u^R_{j-1/2}$, $u^L_{j+1/2}$ and satisfying $\int_{c_j} u\,dx = \Delta x u_j$) has an extreme point inside the cell, modify it such that it becomes monotone inside the cell.
(f) If a cell is inside a discontinuity replace the parabola with a line, which gives a steeper shock representation than the original parabola. As discontinuity detector the following conditions are used

```
  if  $\Delta_+\Delta_- u_{j-1}\Delta_+\Delta_- u_{j+1} < 0$
 and  $|\Delta_+\Delta_-\Delta_+ u_j| \geq M_1$
 and  $\Delta_+\Delta_-\Delta_+ u_j \Delta_+ u_j < 0$
 and  $|\Delta_0 u_j| \geq M_2$
then  there is a discontinuity in cell $j$.
```

We next describe another method for obtaining high order interpolation of discontinuous functions. The essentially non oscillatory (ENO) interpolation is a systematic way to incrementally increase the accuracy to any order by adding points to the interpolation polynomial from the left or from the right, depending on in which direction the function is least oscillatory.

The process is described using Newton's form of the interpolation polynomial. Assume that the function $g(x)$ is known at the points $x_j, j = \ldots, -1, 0, 1, \ldots$. Define the divided differences $[x_i, \dots, x_{i+r}]g$ recursively by

$$[x_i]g = g(x_i)$$

$$[x_i, \dots, x_{i+r}]g = \frac{[x_{i+1}, \dots, x_{i+r}]g - [x_i, \dots, x_{i+r-1}]g}{x_{i+r} - x_i}$$

Newton's polynomial interpolating g at the points $x_1, \dots, x_n$ is then given by

$$P^n(x) = \sum_{i=1}^{n} (x - x_1)(x - x_2) \dots (x - x_{i-1})[x_1, \dots, x_i]g$$

where $(x - x_i) \dots (x - x_j) = 1$ if $i > j$. This form is convinient, since if we want to add another point to the interpolation problem, we can immediately update the interpolation polynomial using the formula

$$P^{n+1}(x) = P^n(x) + (x - x_1) \dots (x - x_n)[x_1, \dots, x_{n+1}]g$$

Proof of the above statements and description of various interpolation procedures can be found in any textbook on approximation theory.

We now give an algorithm for constructing a piecewise N degree polynomial continuous interpolant $L(x)$ from the given grid function u_j, with

$$L(x_j) = u_j$$

and which does introduce as small amount of oscillations as possible. Start by defining the linear polynomial

$$L^1(x) = u_j + (x - x_j)(u_{j+1} - u_j)/\Delta x \quad x_j \le x < x_{j+1}$$

and the indices

$$k^1_{min} = j$$
$$k^1_{max} = j + 1$$

to bookkeep the stencil width. The interpolation proceeds recursively as follows. Define the divided differences

$$a_p = [x_{k^{p-1}_{min}}, \dots, x_{k^{p-1}_{max}+1}]u$$
$$b_p = [x_{k^{p-1}_{min}-1}, \dots, x_{k^{p-1}_{max}}]u$$

where thus we add one point to the right for a_p and one point to the left for b_p. Next use the smallest difference to update the polynomial.

if $|a_p| < |b_p|$ then

$$L^p(x) = L^{p-1}(x) + a_p \prod_{k=k^{p-1}_{min}}^{k^{p-1}_{max}} (x - x_k)$$
$$k^p_{max} = k^{p-1}_{max} + 1$$
$$k^p_{min} = k^{p-1}_{min}$$

else

$$L^p(x) = L^{p-1}(x) + b_p \prod_{k=k_{min}^{p-1}}^{k_{max}^{p-1}} (x - x_k)$$

$$k_{max}^p = k_{max}^{p-1}$$
$$k_{min}^p = k_{min}^{p-1} - 1$$

Thus $L^p(x)$ is a degree p polynomial which interpolates $u(x)$ and which is constructed from the smallest possible divided differences.

We next show how this interpolation algorithm can be used to solve the reconstruction problem. There are two ways to do this.

1. Reconstruction by primitive function (RP).
2. Reconstruction by deconvolution (RD).

In the first method (RP), we observe that the primitive function

$$U(x_{j+1/2}) = \int_{-\infty}^{x_{j+1/2}} u(t_n, x)\, dx = \sum_{k=-\infty}^{j} u_k^n \Delta x$$

is known at the points $x_{j+1/2}$. The function $U(x)$ is interpolated using the ENO interpolation algorithm above. The interpolation polynomial, $L(x)$, is differentiated to get the approximation to $u(t_n, x)$. Thus the left and right values required in the numerical flux are

$$u_{j+1/2}^L = \frac{dL(x_{j+1/2}-)}{dx}$$
$$u_{j+1/2}^R = \frac{dL(x_{j+1/2}+)}{dx}$$

$L(x)$ is continuous, but the derivatives may have different values from the left and from the right at the break points $x_{j+1/2}$. In this way the reconstructed function becomes piecewise continuous.

To describe the second method (RD), we first note that

$$\overline{u}(x) = \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} u(y)\, dy = \int_{-1/2}^{1/2} u(x + s\Delta x)\, ds \qquad (4.3)$$

where thus $\overline{u}(x)$ is the cell average. We interpolate the given cell averages, using the ENO interpolation algorithm above, to get an approximation to $\overline{u}(x)$, and then find the approximation to $u(x)$ by inverting ("deconvolute") (4.3).

To invert (4.3) use Taylor expansion

$$\overline{u}(x_j) = \int_{-1/2}^{1/2} \sum_{\nu=0}^{N-1} \frac{s\Delta x^\nu}{\nu!} \frac{d^\nu u}{dx^\nu}(x_j)\, dx + O(\Delta x^N) =$$
$$\sum_{\nu=0}^{N-1} \frac{1}{\nu!} \Delta x^\nu \frac{d^\nu u}{dx^\nu}(x_j) \int_{-1/2}^{1/2} s^\nu\, ds + O(\Delta x^N)$$

All the derivatives $d^\nu u/dx^\nu(x_j)$ are unknowns but there is only one equation. To introduce more equations it is necessary to consider the derivatives of $\overline{u}(x)$. Similarly as above one gets

$$\frac{d^k\overline{u}(x_j)}{dx^k} = \sum_{\nu=0}^{N-k-1} \frac{1}{\nu!}\Delta x^\nu \frac{d^{\nu+k}u}{dx^{\nu+k}}(x_j)\int_{-1/2}^{1/2} s^\nu \, ds + O(\Delta x^{N-k})$$

for $k = 1, \ldots, N-1$. In this way N equations are obtained for the N unknowns $d^\nu u/dx^\nu(x_j), \nu = 0, \ldots, N-1$.

The reconstruction polynomial is then defined as

$$L^{N-1}(x) = \sum_{\nu=0}^{N-1} \frac{(x-x_j)^\nu}{\nu!}\frac{d^\nu u}{dx^\nu}(x_j) \quad x_{j-1/2} < x < x_{j+1/2} \tag{4.5}$$

In summary the algorithm becomes

1. Use the ENO interpolation algorithm to interpolate the cell averages u_j. The result is a polynomial of degree N, $Q^N(x)$, piecewise differentiable with breakpoints at x_j.
2. Evaluate the derivatives $dQ^N(x_j)/dx$. Since x_j are break points extra care has to be taken. We define
$$\frac{dQ^N(x_j)}{dx} = minmod(\frac{dQ^N(x_j-)}{dx}, \frac{dQ^N(x_j+)}{dx})$$
and similarly for higher order derivatives.
3. Solve the upper triangular linear system of equations
$$\frac{d^kQ^N(x_j)}{dx^k} = \sum_{\nu=0}^{N-k-1} \frac{1}{\nu!}\Delta x^\nu \frac{d^{k+\nu}u}{dx^{k+\nu}}(x_j)\int_{-1/2}^{1/2} s^\nu \, ds + O(\Delta x^N) \quad k = 0, \ldots, N-1$$
to get the derivatives $d^\nu u/dx^\nu(x_j)$.
4. Define (4.5) as the piecewise polynomial reconstruction.

Example 4.1. We derive the second order ENO scheme through RP. Second order means doing piecewise linear reconstruction. Thus the primitive function has to be interpolated using degree 2 polynomials. We obtain

$$U(x) = U_{j+1/2} + (x - x_{j+1/2})u_j + \frac{1}{2\Delta x}(x - x_{j-1/2})(x - x_{j+1/2})m(\Delta_- u_j, \Delta_+ u_j)$$
$$x_{j-1/2} < x < x_{j+1/2}$$

where

$$m(x,y) = \begin{cases} x & \text{if } |x| \leq |y| \\ y & \text{if } |x| \geq |y| \end{cases}$$

The linear approximation inside cell j becomes

$$\frac{dU}{dx} = u_j + \frac{x - x_j}{\Delta x} m(\Delta_- u_j, \Delta_+ u_j)$$

this is a scheme on the form treated in chapter 3 (see p.50), with

$$s_j = m(\Delta_+ u_j, \Delta_- u_j)$$

the TVD condition (3.25) is satisfied. Thus this is a TVD scheme which and consequently it degenerates to first order at extrema.

In general one can prove that the RP ENO scheme using degree N polynomials has truncation error $O(\Delta x^N)$, except at points where any of the first $N-1$ derivatives disappears, there the truncation error is $O(\Delta x^{N-1})$. It is also possible to prove that the truncation error for RD ENO method using degree N polynomials is $O(\Delta x^N)$ always.

4.3 Outer interpolation gives a point value scheme

This scheme is based on interpolation of the numerical fluxes. To achieve the desired order of accuracy, it is necessary that the interpolated fluxes have a sufficient amount of derivatives. This is a very important point which somewhat restricts the possible choices of first order numerical flux to build the method from.

Assume that the point values

$$u_j \quad j = \dots, -2, -1, 0, 1, 2, \dots$$

are known. The idea of this method was outlined in section 4.1. We form the interpolant of

$$H_{j+1/2} = \Delta x \sum_{k=a}^{j} f(u_k)$$

by using the ENO interpolation algorithm, and then take the numerical flux as

$$h_{j+1/2} = \frac{dH(x_{j+1/2})}{dx}.$$

The interpolation is made piecewise polynomial with break points x_j. This direct approach have to be modified somewhat. If we carry out the above scheme we get

$$H^1(x) = \begin{cases} H_{j+1/2} + (x - x_{j+1/2}) f_j & \text{if } |f_j| < |f_{j+1}| \\ H_{j+1/2} + (x - x_{j+1/2}) f_{j+1} & \text{if } |f_{j+1}| < |f_j| \end{cases} \qquad x_j < x < x_{j+1}$$

which leads to

$$h_{j+1/2} = \begin{cases} f_j & \text{if } |f_j| < |f_{j+1}| \\ f_{j+1} & \text{if } |f_{j+1}| < |f_j| \end{cases}$$

if the order of accuracy is chosen =1. Although this flux is consistent, the resulting method is not TVD (Exercise 2). It is crucial that the first order approximation is TVD. From numerical experiments, it is possible to verify that this method is not non oscillatory no matter how high the accuracy of the interpolant. Instead we make the first order version of this method TVD, by taking

$$H^1(x) = \begin{cases} H_{j+1/2} + (x - x_{j+1/2}) f_j & \text{if } a_{j+1/2} \geq 0 \\ H_{j+1/2} + (x - x_{j+1/2}) f_{j+1} & \text{if } a_{j+1/2} < 0 \end{cases} \qquad x_j < x < x_{j+1}$$

the first order method is then the upwind scheme. Continuing the interpolation to higher order leads to a non oscillatory high order scheme, but the method does not satisfy an entropy condition.

We obtain a more general way of choosing the starting first order polynomial if we consider a first order TVD flux $h_{j+1/2}$ and split it as

$$h_{j+1/2} = f_j^+ + f_{j+1}^-$$

where f^+ corresponds to positive wave speeds and f^- to negative wave speeds. As an example the Engquist-Osher scheme (section 2.5) can be written on this form with

$$f^+(u) = \begin{cases} f(u) & \text{if } f'(u) > 0 \\ 0 & \text{if } f'(u) < 0 \end{cases} \qquad f^-(u) = \begin{cases} 0 & \text{if } f'(u) > 0 \\ f(u) & \text{if } f'(u) < 0 \end{cases}$$

Another example is the Lax-Friedrichs scheme, where

$$\begin{aligned} f^+(u) &= (f(u) + \frac{1}{\lambda}u)/2 \\ f^-(u) &= (f(u) - \frac{1}{\lambda}u)/2 \end{aligned}$$

or the modified Lax-Friedrichs scheme

$$\begin{aligned} f^+(u) &= (f(u) + \alpha u)/2 \\ f^-(u) &= (f(u) - \alpha u)/2 \end{aligned}$$

with $\alpha = \max |f'(u)|$.

We define the starting polynomials

$$\left.\begin{aligned} H_-^1(x) &= H_{j+1/2} + (x - x_{j+1/2})f_{j+1}^- \\ H_+^1(x) &= H_{j+1/2} + (x - x_{j+1/2})f_j^+ \end{aligned}\right. \qquad x_j < x < x_{j+1}$$

and then continue the ENO interpolation of f^+ and f^- respectively through the points x_j to arbitrary order of accuracy, p. Finally

$$h_{j+1/2} = \frac{dH_+^p(x_{j+1/2})}{dx} + \frac{dH_-^p(x_{j+1/2})}{dx}$$

The truncation error for this method will involve differences of the functions f^+ and f^-. Thus to achieve the expected accuracy it is necessary to have $f^+, f^- \in C^p$, p large enough. Because of this, the scheme has mostly been used together with the C^∞ Lax-Friedrichs numerical flux, or the modified Lax-Friedrichs numerical flux. However the Lax-Friedrichs scheme does not always give sufficient shock resolution. Although the higher order versions, obtained as described above, performs much better than the first order Lax-Friedrichs, there is still need for first order TVD methods giving better shock resolution than Lax-Friedrichs and having more derivatives than the upwind or the Engquist-Osher schemes, to be used as building blocks for this method.

We conclude with some remarks about two space dimensions. For the problem

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

the method described in this section can be applied separately in the x- and y- directions to approximate $\partial/\partial x$ and $\partial/\partial y$ respectively (see section 2.6). There are no extra complications. For the cell centered scheme, the two dimensional generalization of formula (4.1) gives an integral around the cell boundary. This integral is required to p th order accuracy, which can be done by a numerical quadrature formula. If e.g., $p = 4$ this means using two values on each cell side. Thus for each cell, we need a two dimensional reconstruction, which is a non trivial problem in its own right, and then we have 8 flux evaluations to make, two on each side. The cell centered scheme quickly becomes more computationally expensive than the point centered scheme.

4.4 Time discretization

The easiest way to obtain a high order time discretization is to use a Runge-Kutta method. However it has been observed that e.g., the classical fourth order Runge-Kutta method can cause large amount of oscillations in the solution although the space discretization is made TVD. Therefore, we have to be extra careful about how to design Runge-Kutta schemes.

We consider the semi discrete approximation

$$\frac{du}{dt} = L(u)$$

to the problem

$$u_t = -f(u)_x$$

where we know that the forward Euler approximation

$$u^{n+1} = u^n + \Delta t L(u^n)$$

leads to a TVD or ENO method. The semi discrete TVD methods treated previously can all be written

$$\frac{du_j}{dt} = C_{j+1/2}\Delta_+ u_j - D_{j-1/2}\Delta_- u_j$$

with non negative $C_{j+1/2}, D_{j+1/2}$. From theorem 2.12 it follows that the forward Euler time discretization is TVD under the CFL constraint $\Delta t(C_{j+1/2} + D_{j+1/2}) \leq 1$, all j. Thus it is not too restrictive to assume TVD for the forward Euler time discretization. The idea of TVD Runge-Kutta methods is to write the scheme as a convex combination of forward Euler steps. One general form for explicit m stage Runge-Kutta methods is

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(i)} &= u^{(0)} + \Delta t \sum_{k=0}^{i-1} c_{ik} L(u^{(k)}) \quad i = 1, 2, \ldots, m \\ u^{n+1} &= u^{(m)} \end{aligned} \tag{4.5}$$

For each stage, the weights $\alpha_k^{(i)}$, $k = 0, \ldots, i-1$, satisfying

$$\alpha_k^{(i)} \geq 0$$
$$\sum_{k=0}^{i-1} \alpha_k^{(i)} = 1$$

are introduced. Then (4.5) can be rewritten

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(i)} &= \sum_{k=0}^{i-1} \alpha_k^{(i)} u^{(k)} + \beta_k^{(i)} \Delta t L(u^{(k)}) \quad i = 1, 2, \ldots, m \\ u^{n+1} &= u^{(m)} \end{aligned} \tag{4.6}$$

with $\beta_k^{(i)} = c_{ik} - \sum_{s=k+1}^{i-1} c_{sk} \alpha_s^{(i)}$. By writing

$$u^{(i)} = \sum_{k=0}^{i-1} \alpha_k^{(i)} (u^{(k)} + \frac{\beta_k^{(i)} \Delta t}{\alpha_k^{(i)}} L(u^{(k)})) \quad i = 1, 2, \ldots, m$$

it is easy to prove

Theorem 4.1. *If $\beta_k^{(i)} > 0$ and the method*

$$u^{n+1} = u^n + \Delta t L(u^n)$$

is TVD under the CFL condition $\lambda \leq \lambda_0$, ($\lambda = \Delta t / \Delta x$), then the method (4.6) is TVD under the CFL condition

$$\lambda \leq \lambda_0 \min_{i,k} \frac{\alpha_k^{(i)}}{\beta_k^{(i)}} \tag{4.7}$$

Proof: For each stage it holds

$$\begin{aligned} TV(u^{(i)}) = \sum_{j=-\infty}^{\infty} |\Delta_+ u_j^{(i)}| &\leq \sum_{j=-\infty}^{\infty} \sum_{k=0}^{i-1} \alpha_k^{(i)} |\Delta_+ (u_j^{(k)} + \frac{\beta_k^{(i)}}{\alpha_k^{(i)}} \Delta t L(u_j^{(k)}))| \\ &\leq \sum_{k=0}^{i-1} \alpha_k^{(i)} TV(u^{(k)}) \end{aligned}$$

where we used that $\alpha_k^{(i)} > 0$ and that the forward Euler parts in the sum above are TVD under the constraint (4.7). Use induction by assuming that $TV(u^{(k)}) \leq TV(u^{(0)})$ for all $k \leq i-1$. This is certainly true for $i = 1$. The inequality above gives

$$TV(u^{(i)}) \leq (\sum_{k=0}^{i-1} \alpha_k^{(i)}) TV(u^{(0)}) = TV(u^{(0)})$$

Thus we have proved $TV(u^{(m)}) \leq TV(u^{(0)})$, which is the TVD condition $TV(u^{n+1}) \leq TV(u^n)$.

If, however, some $\beta_k^{(i)} < 0$ then the step

$$u^{(k)} + \frac{\beta_k^{(i)} \Delta t}{\alpha_k^{(i)}} L(u^{(k)})$$

corresponds to a reversal of time, and we replace the operator $L(u)$ with an operator $\hat{L}(u)$ which is such that $-\hat{L}(u)$ approximates the problem

$$u_t = f(u)_x$$

in a TVD (or ENO) fashion. We give an example to show how the operator $\hat{L}(u)$ is derived.

Example 4.3. We approximate

$$u_t = u_x$$

using the stable TVD approximation

$$u^{n+1} = u^n + \Delta t D_+ u^n$$

Thus, using the formulas above, $L(u) = \Delta t D_+$. The operator $\hat{L}$ is obtained from approximating

$$u_t = -u_x$$

using the stable TVD approximation

$$u^{n+1} = u^n - \Delta t D_- u^n$$

From this we find that $-\hat{L}(u) = -\Delta t D_-$, and thus

$$\hat{L}(u) = \Delta t D_-$$

The CFL condition for the case of negative $\beta_k^{(i)}$ with $\hat{L}(u)$ replacing $L(u)$ is obtained from using the absolute value in (4.7).

To derive some particular Runge-Kutta methods, we start from (4.6) and investigate, by Taylor expansion, the possible methods. We give one example.

Example 4.4 Require second order accuracy and $m = 2$. The method is

$$\begin{aligned}
u^{(0)} &= u^n \\
u^{(1)} &= u^{(0)} + \Delta t \beta_0^{(1)} L(u^{(0)}) \\
u^{(2)} &= \alpha_0^{(2)} u^{(0)} + \Delta t \beta_0^{(2)} L(u^{(0)}) + \alpha_1^{(2)} u^{(1)} + \Delta t \beta_1^{(2)} L(u^{(1)}) \\
u^{n+1} &= u^{(2)}
\end{aligned}$$

where we will choose $\alpha_k^{(i)} \geq 0$. To get the accuracy, take an exact solution to $u_t = L(u)$, and insert it into the method. The truncation error in time is (dropping all terms of $O(\Delta t^3)$)

$$\tau = u(t+\Delta t) - (\alpha_0^{(2)} u + \Delta t \beta_0^{(2)} L(u) + \alpha_1^{(2)} (u + \Delta t \beta_0^{(1)} L(u)) + \\ \Delta t \beta_1^{(2)} L(u + \Delta t \beta_0^{(1)} L(u)))$$

where we use u to denote the exact solution $u(t)$. Taylor expansion gives

$$\tau = (1 - \alpha_0^{(2)} - \alpha_1^{(2)}) u + (1 - \beta_0^{(2)} - \beta_1^{(2)} - \alpha_1^{(2)} \beta_0^{(1)}) \Delta t u_t + \\ \frac{\Delta t^2}{2} u_{tt} - \Delta t^2 \beta_1^{(2)} \beta_0^{(1)} L(u) L\prime(u)$$

The observation $u_{tt} = (L(u))_t = L'(u) u_t = L'(u) L(u)$ gives the conditions for second order accuracy

$$\alpha_0^{(2)} + \alpha_1^{(2)} = 1$$
$$\beta_0^{(2)} = 1 - \frac{1}{2\beta_0^{(1)}} - \alpha_1^{(2)} \beta_0^{(1)}$$
$$\beta_1^{(2)} = \frac{1}{2\beta_0^{(1)}}$$

The factors

$$\frac{1}{|\beta_0^{(1)}|} \quad \frac{\alpha_0^{(2)}}{|\beta_0^{(2)}|} \quad \frac{\alpha_1^{(2)}}{|\beta_1^{(2)}|}$$

comes into the CFL condition. If they all can be made ≥ 1, this Runge-Kutta scheme will be non oscillatory under the same CFL condition as the forward Euler scheme is non oscillatory. If furthermore we can choose all β non negative, we can avoid the operator $\hat{L}(u)$. We first try $\beta_0^{(1)} = 1$, which gives $\beta_1^{(2)} = 1/2$ and then to keep the next CFL factor =1, we take $\alpha_1^{(2)} = 1/2$. This leads to $\alpha_0^{(2)} = 1/2$ and $\beta_0^{(2)} = 0$. We have obtained the method

$$u^{(1)} = u^{(0)} + \Delta t L(u^{(0)})$$
$$u^{(2)} = \frac{1}{2}(u^{(0)} + u^{(1)} + \Delta t L(u^{(1)}))$$

which is the same as the method given on page 58. This method gives second order in time and retains the non oscillatory features from the semi discrete approximation. It does not require $\hat{L}(u)$, which saves programming effort.

In a similar way we can derive the third order TVD Runge-Kutta method

$$u^{(1)} = u^{(0)} + \Delta t L(u^{(0)})$$
$$u^{(2)} = \frac{3}{4} u^{(0)} + \frac{1}{4} u^{(1)} + \frac{\Delta t}{4} L(u^{(1)})$$
$$u^{(3)} = \frac{1}{3} u^{(0)} + \frac{2}{3} u^{(2)} + \frac{2\Delta t}{3} L(u^{(2)})$$

This method has CFL factor one, and is thus stable under the CFL condition obtained from the forward Euler discretization. For higher accuracy than three, no method not involving $\hat{L}(u)$ is known. It is recommended that a known high order Runge-Kutta method is written on the form (4.6), and then $L(u)$ is replaced by $\hat{L}(u)$ wherever $\beta < 0$.

The Runge-Kutta approach seems to be the simplest way to do time discretization and we recommend it. There are however other ways. We conclude with a brief discussion of these alternative time discretizations. One example is a Lax-Wendroff type of method. It relies on the formula

$$u^{n+1} = u^n + \Delta t (u^n)_t + \frac{\Delta t^2}{2}(u^n)_{tt} + \dots$$

The time derivatives are then replaced by spatial ones.

$$\begin{aligned} u_t &= -f(u)_x \\ u_{tt} &= (f'(u)f(u)_x)_x \\ &\dots \end{aligned}$$

The spatial derivatives are approximated using the ENO scheme. The procedure becomes very complicated for higher order of accuracy than 2 in time. It is not suited for steady state computations, but the stencil will not be as wide as for the Runge-Kutta schemes.

Another, more one dimensional method, can be given by a direct integration of the conservation law in the $x-t$ plane. This method is derived for schemes based on cell averages. Consider the conservation law

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

integrate around one cell in the $x-t$ plane as depicted in fig. 4.4,

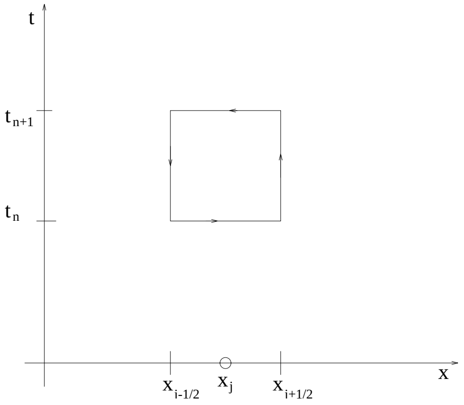


Fig. 4.4. Path of integration in the $x-t$ plane.

and use Green's formula to obtain

$$\oint_C u\,dx - f(u)\,dt = 0$$

Evaluate this integral directly

$$\int_{x_{j-1/2}}^{x_{j+1/2}} u(t_n,x)\,dx - \int_{t_n}^{t_{n+1}} f(u(t,x_{j+1/2}))\,dt - \int_{x_{j-1/2}}^{x_{j+1/2}} u(t_{n+1},x)\,dx + \int_{t_n}^{t_{n+1}} f(u(t,x_{j-1/2}))\,dt = 0$$

Letting u_j^n denote the cell average, the above integral becomes

$$u_j^{n+1} = u_j^n - \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} (f(u(t,x_{j+1/2})) - f(u(t,x_{j-1/2})))\,dt.$$

Note that no approximation has yet been made. Assume that the cell averages u_j^n are known, and that $L(t_n,x)$ is the function obtained by piecewise polynomial reconstruction from the cell averages. As numerical approximation to the PDE we take

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}(h_{j+1/2}^n - h_{j-1/2}^n)$$

Here, $h_{j+1/2}^n$ approximates

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(v(t,x_{j+1/2}))\,dt$$

with $v(t,x)$ the solution to the PDE for $t > t_n$ using $L(t_n,x)$ as initial data. For first order accuracy we can approximate

$$\int_{t_n}^{t_{n+1}} f(v(t,x_{j+1/2}))\,dt \approx \Delta t f(v(t_n,x_{j+1/2})) = \Delta t h(u_{j+1}^n, u_j^n)$$

The reconstructed function has break points at $x_{j+1/2}$, and therefore $f(v(t_n,x_{j+1/2}))$ is not uniquely defined. We use the semi discrete flux $h_{j+1/2}$ to be the flux at $(t_n,x_{j+1/2})$. This leads to the usual forward Euler method, which is only first order accurate in time. For higher order accuracy the flux at other points may be needed e.g., second order accuracy can be obtained from the approximation

$$\int_{t_n}^{t_{n+1}} f(v(t,x_{j+1/2}))\,dt \approx \frac{\Delta t}{2}(f(v(t_n,x_{j+1/2})) + f(v(t_{n+1},x_{j+1/2}))) = \frac{\Delta t}{2}(h(u_{j+1/2}^{nR}, u_{j+1/2}^{nL}) + f(v(t_{n+1},x_{j+1/2})))$$

where the value $v(t_{n+1},x_{j+1/2})$ is found by tracing the characteristic through the point $(t_{n+1},x_{j+1/2})$ backward to $t = t_n$, where the reconstructed function is known, and $h(u_{j+1/2}^{nR}, u_{j+1/2}^{nL})$ is the semi discrete flux function evaluated at the known time t_n.

There is an abundance of methods based on this idea. Another example is when one half step using the first order scheme is taken to obtain a value at $(t_{n+1/2},x_{j+1/2})$

and then the approximation

$$\int_{t_n}^{t_{n+1}} f(v(t, x_{j+1/2}))\,dt \approx \Delta t f(v(t_{n+1/2}, x_{j+1/2}))$$

is used to get a second order accurate scheme.

Exercises

1. The second order cell based RD ENO scheme has second order accuracy everywhere and is consequently not TVD. Write this scheme on slope limiter form (as on p.50), and thus derive the ENO limiter function

$$\begin{aligned} B(\Delta_+ u_j, \Delta_- u_j) = minmod(\Delta_+ u_j - \frac{1}{2} m(\Delta_+\Delta_- u_{j+1}, \Delta_+\Delta_- u_j),\\ \Delta_- u_j + \frac{1}{2} m(\Delta_+\Delta_- u_j, \Delta_+\Delta_- u_{j-1})) \end{aligned}$$

   with

$$m(x, y) = \begin{cases} x & \text{if } |x| < |y| \\ y & \text{if } |y| < |x| \end{cases}$$

2. Show that the method, using the numerical flux

$$h_{j+1/2} = \begin{cases} f(u_j) & \text{if } |f(u_j)| < |f(u_{j+1})| \\ f(u_{j+1}) & \text{if } |f(u_{j+1})| \leq |f(u_j)| \end{cases}$$

   is not a TVD method.

3. Derive the limiter function $\psi(r)$ corresponding to the slope limiter in example 4.1,

$$m(x, y) = \begin{cases} x & \text{if } |x| < |y| \\ y & \text{if } |y| < |x| \end{cases}$$

   and show that it can fit inside the TVD domain given in fig. 3.3, chapter 3.

5. Systems of conservation laws

5.1 Linear systems

When we apply the methods for scalar problems to systems of hyperbolic partial differential equations, one of the most important facts is that the methods have to be applied to the characteristic variables. We here give an example of a linear system to illustrate this.

Example 5.1. Consider the system

$$\begin{pmatrix} u \\ v \end{pmatrix}_t + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.1}$$

The PDE can be decoupled into two independent scalar problems by a diagonalizing transformation. It is easy to verify that the eigenvectors and eigenvalues of the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

are

$$\lambda_1 = 1 \;\; \mathbf{r}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \lambda_2 = -1 \;\; \mathbf{r}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

We introduce the diagonalizing matrix

$$R = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and multiply (5.1) by R^{-1}. The result is

$$\begin{aligned} w_t + w_x &= 0 \\ z_t - z_x &= 0 \end{aligned} \tag{5.2}$$

where the transformed variables are defined as

$$\begin{pmatrix} w \\ z \end{pmatrix} = R^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{2} \begin{pmatrix} u+v \\ u-v \end{pmatrix} \tag{5.3}$$

Next, we consider this PDE on $-\infty < x < \infty$, $t > 0$, and give the initial data

$$w(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \qquad z(0,x) = \begin{cases} -1 & x < 0 \\ 0 & x \geq 0 \end{cases}$$

which, by (5.3), corresponds to

$$u(0,x) = 0 \qquad v(0,x) = \begin{cases} 2 & x < 0 \\ 0 & x \geq 0 \end{cases}$$

The solution for $t > 0$ is

$$w(t,x) = \begin{cases} 1 & x - t < 0 \\ 0 & x - t \geq 0 \end{cases} \qquad z(t,x) = \begin{cases} -1 & x + t < 0 \\ 0 & x + t \geq 0 \end{cases}$$

as easily found from the diagonal form (5.2). In the variables $(u\ v)$ this corresponds to

$$u(t,x) = \begin{cases} 0 & x < -t \\ 1 & -t \le x \le t \\ 0 & x \ge t \end{cases} \qquad v(t,x) = \begin{cases} 2 & x < -t \\ 1 & -t \le x \le t \\ 0 & x \ge t \end{cases} .$$

The solution is depicted in Figures 5.1 and 5.2 below.

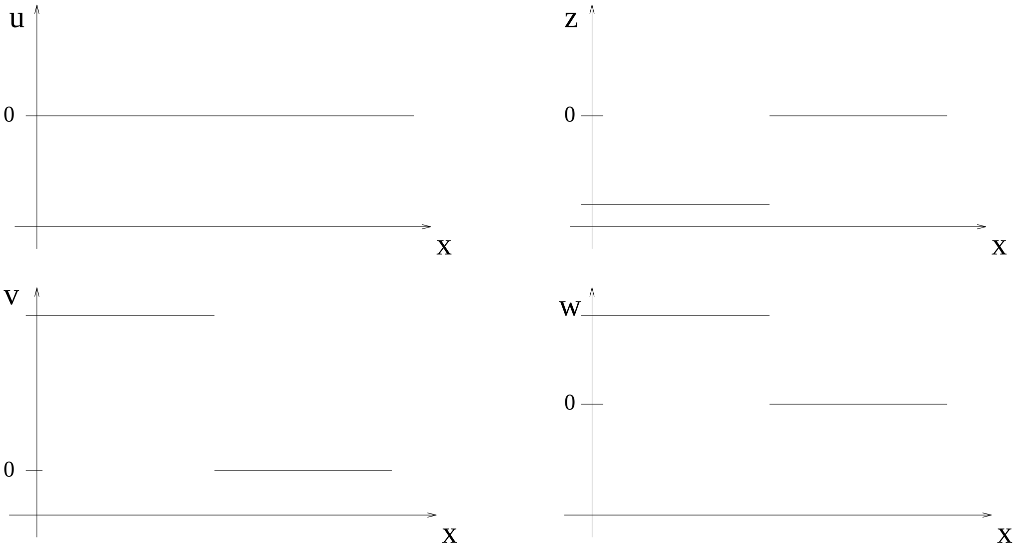


Fig. 5.1a. Original variables. Fig. 5.1b. Characteristic variables.

Initial data.

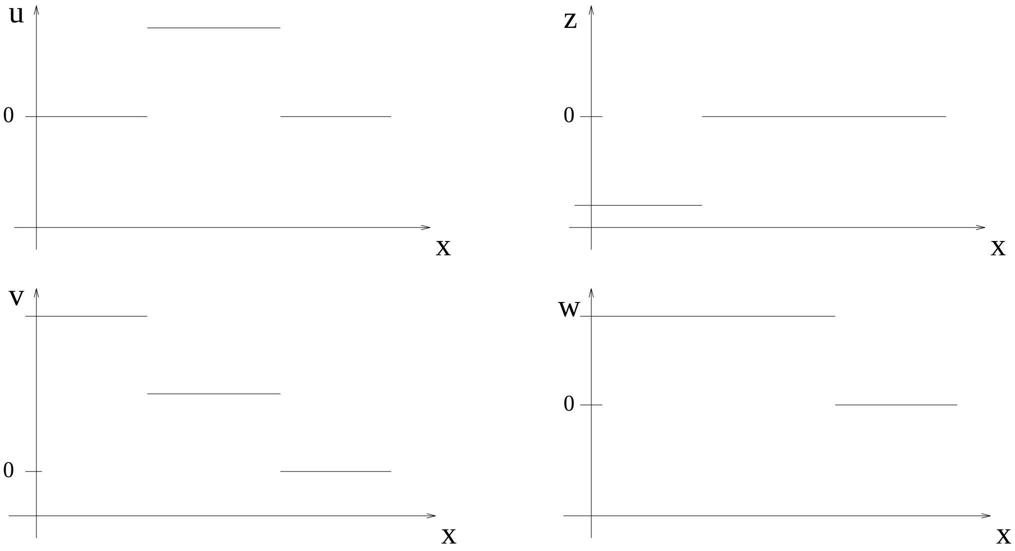


Fig. 5.2a. Original variables. Fig. 5.2b. Characteristic variables.

Solution at time > 0.

The point with this example is that the variable u is zero at $t = 0$, but immediately develops a square pulse at $t > 0$. Thus there is no TVD property in the variable u, and therefore it is not reasonable to use a TVD scheme componentwise in $(u\ v)$. A TVD method has to be applied in the characteristic variables $(w\ z)$.

In general we consider linear systems

$$\mathbf{u}_t + A\mathbf{u}_x = \mathbf{0}$$

where A is a diagonalizable matrix. Then we can perform the transformation

$$R^{-1}\mathbf{u}_t + R^{-1}ARR^{-1}\mathbf{u}_x = \mathbf{0}$$

where R is the matrix of eigenvectors of A. Introducing the characteristic variable

$$\mathbf{v} = R^{-1}\mathbf{u},$$

we obtain the decoupled system

$$\mathbf{v}_t + \Lambda\mathbf{v}_x = \mathbf{0} \tag{5.4}$$

where Λ is the diagonal matrix consisting of the eigenvalues of A. We thus have a set of m independent scalar equations, $(v_k)_t + \lambda_k(v_k)_x = 0$ which have solutions $v_{k0}(x - \lambda_k t)$ for given initial data $v_{k0}(x)$.

We use the decoupling of the linear system to solve the problem

$$\begin{aligned} \mathbf{u}_t + A\mathbf{u}_x &= \mathbf{0} \\ \mathbf{u}(x,0) &= \begin{cases} \mathbf{u}_L & \text{if } x < 0 \\ \mathbf{u}_R & \text{if } x > 0 \end{cases} \end{aligned}$$

where $\mathbf{u}_L$ and $\mathbf{u}_R$ are two constant states. A hyperbolic partial differential equation with the initial data consisting of two constant states is called a *Riemann problem.* According to the discussion above, the solution can be written

$$\mathbf{u}(x,t) = \sum_{k=1}^{m} v_{k0}(x - \lambda_k t)\mathbf{r}_k,$$

where now all the functions $v_{k0}(x)$ are step functions with a jump at $x = 0$. $\mathbf{r}_k$ are the eigenvectors of A i.e., the columns of R. We assume that the eigenvalues are enumerated in increasing order $\lambda_1 < \lambda_2 < \ldots < \lambda_m$. Let us denote

$$v_{k0}(x) = \begin{cases} v_{kL} & x < 0 \\ v_{kR} & x > 0 \end{cases}.$$

The solution is thus piecewise constant, with changes when $x - \lambda_k t$ changes sign for some k. From this observation the solution formula

$$\mathbf{u}(x,t) = \sum_{k=1}^{q} v_{kR}\mathbf{r}_k + \sum_{k=q+1}^{m} v_{kL}\mathbf{r}_k = \mathbf{u}_L + \sum_{k=1}^{q}(v_{kR} - v_{kL})\mathbf{r}_k \quad \lambda_q < x/t < \lambda_{q+1}$$

follows easily. The solution is thus constant on wedges in the $x - t$ plane, as seen in Fig. 5.3.

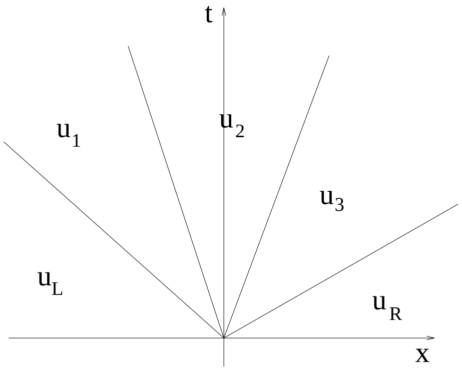


Fig. 5.3. Solution of the linear Riemann problem in the $x-t$ plane, $m=4$.

As seen above, the states inside the wedges are given by

$$\mathbf{u}_q = \mathbf{u}_L + \sum_{k=1}^{q}(v_{kR} - v_{kL})\mathbf{r}_k$$

with $\mathbf{u}_L = \mathbf{u}_0, \mathbf{u}_R = \mathbf{u}_m$.

5.2 Non linear systems

If the coefficient matrix is diagonalizable, a linear system can be decoupled into a number of independent scalar problems. This is however not true for a non linear system, the diagonalizing transformation R is now a function of $\mathbf{u}(x,t)$ and we can not use a relation like $(R\mathbf{v})_t = R\mathbf{v}_t$ which was essential in deriving (5.4). The non linear system is more complicated than a collection of scalar non linear problems.

We consider the equation

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}$$

where the solution vector is $\mathbf{u} = (u_1(x,t), \dots, u_m(x,t))^T$. The Jacobian matrix of the flux function is denoted $A(\mathbf{u}) = \partial\mathbf{f}/\partial\mathbf{u}$. The eigenvalues of $A(\mathbf{u})$, $\lambda_i(\mathbf{u})$ are assumed to be real, distinct, and ordered in increasing order

$$\lambda_1(\mathbf{u}) < \lambda_2(\mathbf{u}) < \dots \lambda_m(\mathbf{u})$$

The corresponding eigenvectors are denoted $\mathbf{r}_1(\mathbf{u}), \dots, \mathbf{r}_m(\mathbf{u})$.

We generalize the convexity condition $f''(u) \neq 0$ to systems as follows.

Definition 5.1. *The kth field is genuinely non linear if $\mathbf{r}_k^T \nabla_u \lambda_k(\mathbf{u}) \neq 0$ for all $\mathbf{u}$.*

If a scalar problem is linear then $f''(u) = 0$. This condition is generalized to

Definition 5.2. *The kth field is linearly degenerate if $\mathbf{r}_k^T \nabla_u \lambda_k(\mathbf{u}) = 0$ for all $\mathbf{u}$.*

Here ∇_u is the gradient operator with respect to $\mathbf{u}$,

$$\nabla_u a = (\frac{\partial a}{\partial u_1}, \dots, \frac{\partial a}{\partial u_m}).$$

It is not hard to verify that definition 5.1 and 5.2 degenerates to the convexity and the linearity conditions respectively in the scalar case $m = 1$.

We will here discuss three types of solutions.

1. Shocks.
2. Rarefaction waves.
3. Contact discontinuities.

In section 5.3 we will show how these three types of solutions can be pieced together to form a solution of the Riemann problem for the non linear system. For the scalar equation we have seen a shock solution in example 1.5 and an expansion wave solution in u_2 in example 1.4.

We first describe shock solutions. These satisfy the Rankine-Hugoniot condition,

$$s(\mathbf{u}_L - \mathbf{u}_R) = \mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R) \tag{5.5}$$

which is derived in the same way as for the scalar problem. We also require an entropy condition. Since we are dealing with the generalization of the convex conservation law, we will look for an entropy condition which generalizes the condition (1.9) i.e., the characteristics should point into the shock.

Definition 5.3. *Let k be a genuinely non linear field. A k-shock is a discontinuity satisfying (5.5) and for which it holds*

$$\lambda_k(\mathbf{u}_L) > s > \lambda_k(\mathbf{u}_R)$$
$$\lambda_{k-1}(\mathbf{u}_L) < s < \lambda_{k+1}(\mathbf{u}_R)$$

The meaning of this definition is first that the shock is in the kth characteristic variable, and second that the number of undetermined quantities at the shock (i.e., the number of characteristics pointing out from the shock) is equal to the number of equations given by (5.5). If we consider the shock as a boundary we see that definition 5.3 means that the characteristics $1, \dots, k-1$ are inflow quantities into the region on the left of the shock. The characteristics $k+1, \dots, m$ are inflow quantities into the region on the right of the shock. Thus there are $m-1$ inflow variables which we must specify. Eliminating s from (5.5) gives $m-1$ equations, thus the number of equations and unknown are equal.

Assume that $\mathbf{u}_L$ is given, we investigate which states $\mathbf{u}_R$ can be connected to $\mathbf{u}_L$ through a shock wave. (5.5) is a system of m equations for the $m+1$ unknowns, $\mathbf{u}_R$, s. We expect to find a one parameter family of solutions $\mathbf{u}_R$. Furthermore, it is natural to have one such family of solutions for each eigenvalue, corresponding in the linear case to placing the discontinuity in any of the m characteristic variables $v_i, i = 1, \dots, m$. These intuitive ideas are stated in the following theorem. The proof is not given here. See e.g., [18] for a proof.

Theorem 5.4. *Assume that the kth field is genuinely non linear. The set of states $\mathbf{u}_R$ near $\mathbf{u}_L$ which can be connected to $\mathbf{u}_L$ through a k-shock form a smooth one parameter family $\mathbf{u}_R = \mathbf{u}(p), -p_0 \leq p \leq 0$. $\mathbf{u}_R(0) = \mathbf{u}_L$. The shock speed, s is also a smooth function of p.*

Formally we could use (5.5) to obtain a shock solution for $p > 0$ as well, but it turns out that the entropy condition is not satisfied for p positive. The situation is similar to the scalar equation, where the entropy condition imposes the restriction that shocks only can jump downwards (see examples 1.4 and 1.5).

We next investigate the rarefaction wave solutions. A rarefaction wave centered at x=0 is a solution which only depends on x/t i.e., $\mathbf{u}(x,t)=\mathbf{b}(x/t)$. Inserting this ansatz into the equation gives

$$-\frac{x}{t^2}\mathbf{b}'+\frac{1}{t}A(\mathbf{b})\mathbf{b}'=\mathbf{0}$$

We denote $\xi=x/t$ and $\mathbf{b}'=d\mathbf{b}/d\xi$ and we thus have

$$(A(\mathbf{b})-\xi)\mathbf{b}'=\mathbf{0}.$$

The solution is given in terms of eigenvalues and eigenvectors

$$\xi=\lambda(\mathbf{b}(\xi)) \qquad \mathbf{b}'=c\mathbf{r}(\mathbf{b})$$

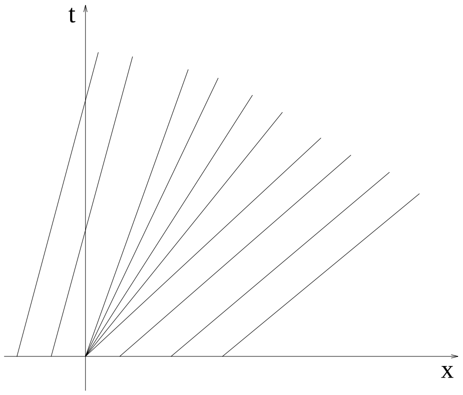
c is a constant. Here it is possible to use genuine non linearity to show that $c=1$. For a given state $\mathbf{u}_L$, we thus can solve the ordinary differential equation

$$\begin{aligned} &\mathbf{b}'(\xi)=\mathbf{r}(\mathbf{b}(\xi)) \quad \xi_0\le\xi\le\xi_0+p \\ &\xi_0=\lambda(\mathbf{b}(\xi_0)) \end{aligned} \tag{5.6}$$

to some final point ξ_0+p, where p is a sufficiently small parameter value. The state $\mathbf{u}_R=\mathbf{b}(\xi_0+p)$ is in this way connected to the state $\mathbf{u}_L=\mathbf{b}(\xi_0)$ through a k-rarefation wave. From the above computations we obtain the following theorem.

Theorem 5.5. *Assume that the kth field is genuinely non linear. The set of states* $\mathbf{u}_R$ *near* $\mathbf{u}_L$ *which can be connected to* $\mathbf{u}_L$ *through a k-rarefaction wave form a smooth one parameter family* $\mathbf{u}_R=\mathbf{u}(p), 0\le p\le p_0$. $\mathbf{u}_R(0)=\mathbf{u}_L$.

Fig. 5.4a shows an example of the k-characteristics for a rarefaction wave and Fig. 5.4b shows one example of a component of the solution $\mathbf{u}(x/t)$ at a fixed time.



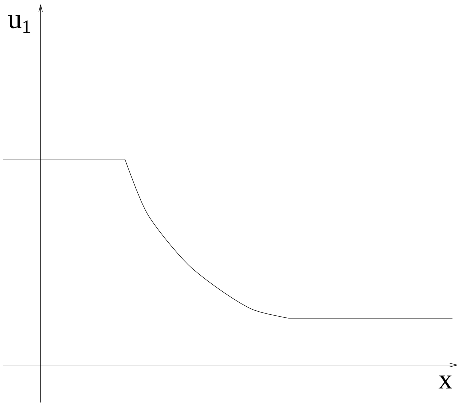


Fig. 5.4a. Characteristics in one field.

Fig. 5.4b. Solution at a time $t>0$.

We summarize the shock and rarefaction cases above as follows.

Theorem 5.6. *Assume that the kth field is genuinely non linear. Given the state $\mathbf{u}_L$, there is a one parameter family of states $\mathbf{u}_R = \mathbf{u}(p), -p_0 \le p \le p_0$ which can be connected to $\mathbf{u}_L$ through a k-shock ($p \le 0$) or a k-rarefaction wave ($p \ge 0$). $\mathbf{u}(p)$ is twice continuously differentiable.*

The differentiability is proved by expanding the function $\mathbf{u}(p)$ around $p = 0$, and can be found in e.g., [18]. For the example $m = 2$, the situation is displayed in Fig. 5.5. The curves show where it is possible to place $\mathbf{u}_R$ in order to connect it to the given state $\mathbf{u}_L$ through a shock or a rarefaction wave.

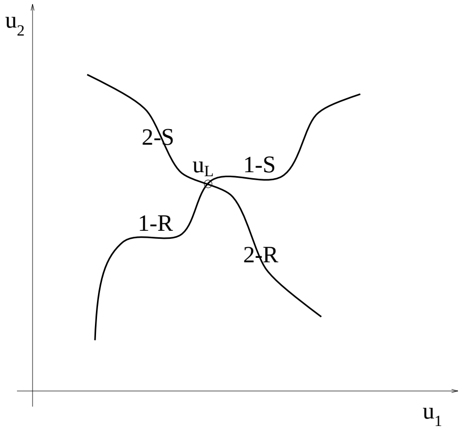


Fig. 5.5. Phase plane. 2-S = 2-shock. 1-R = 1-rarefaction.

Next we define the Riemann invariants. They are quantities which are constant on rarefaction waves, and can be

Definition 5.7. *A k-Riemann invariant is a smooth scalar function $w(u_1, \dots, u_m)$, such that*

$$\mathbf{r}_k^T \nabla_u w = 0$$

i.e., the gradient of w is perpendicular to the kth eigenvector of A.

Theorem 5.8. *There exist $m-1$ k-Riemann invariants with linearly independent gradients.*

Proof: The vector field

$$\mathbf{r}_k^T \nabla_u = \sum_{i=1}^{m} r_i(\mathbf{u}) \frac{\partial}{\partial u_i}$$

can by a coordinate transformation $\mathbf{v} = \mathbf{u}(\mathbf{v})$ be written

$$\frac{\partial}{\partial v_1}$$

and we choose

$$w_1(\mathbf{v}) = v_2 \quad w_2(\mathbf{v}) = v_3 \quad \dots \quad w_{m-1}(\mathbf{v}) = v_m.$$

The functions $w_i, i = 1, \ldots, m-1$ will then satisfy

$$\frac{\partial w_i}{\partial v_1} = 0$$

and have linearly independent gradients. Transforming back yields functions $w_i(\mathbf{u})$ with the desired properties.

Riemann invariants are used for computing the states across a rarefaction wave. The useful property is given in the following theorem.

Theorem 5.9. *The k-Riemann invariants are constant on a k-rarefaction wave.*

Proof: We have seen above that on a k-rarefaction wave, the solution is a function of $\xi = x/t$ and satisfies

$$\mathbf{u}'(\xi) = \mathbf{r}_k(\mathbf{u}(\xi))$$

Let w be a k-Riemann invariant. We obtain

$$\frac{dw}{d\xi} = \sum_i \frac{du_i}{d\xi}\frac{\partial w}{\partial u_i} = \mathbf{u}'(\xi)^T \nabla_u w = \mathbf{r}_k(\mathbf{u})^T \nabla_u w = 0$$

by the definition of Riemann invariant. Thus $dw/d\xi = 0$ and w is constant on the k-rarefaction wave.

Theorem 5.9 gives the following equations for two states connected by a k-rarefaction wave

$$w_i(\mathbf{u}_L) = w_i(\mathbf{u}_R) \quad i = 1, \ldots, m-1 \tag{5.7}$$

where w_i are the k-Riemann invariants. For a rarefaction wave, these relations can be used similarly as (5.5) is used for a shock e.g. to determine the state $\mathbf{u}_R$ from a given $\mathbf{u}_L$.

Let us finally investigate the linearly degenerate fields. Assume that the kth field is linearly degenerate, and define the curve $\mathbf{u}(p)$ through

$$\frac{d\mathbf{u}(p)}{dp} = \mathbf{r}_k(\mathbf{u}(p)) \tag{5.8}$$

The kth eigenvalue is constant on this curve, since the linear degeneracy gives

$$\frac{d\lambda_k(\mathbf{u})}{dp} = \frac{d\mathbf{u}}{dp}\nabla_u \lambda_k = \mathbf{r}_k(\mathbf{u})^T \nabla_u \lambda_k = 0.$$

Theorem 5.10. *Assume that the kth field is linearly degenerate. The states on the curve (5.8) can all be connected to $\mathbf{u}_L$ through a discontinuity moving with speed $s = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}(p))$.*

Proof: Define the function

$$\mathbf{G}(\mathbf{u}(p)) = \mathbf{f}(\mathbf{u}(p)) - s\mathbf{u}(p)$$

which appears in the Rankine-Hugoniot condition. Differentiate this function with respect to p to obtain

$$\frac{d\mathbf{G}}{dp} = (A(\mathbf{u}(p)) - s)\frac{d\mathbf{u}}{dp}$$

which is zero due to (5.8), and the definition of s. Thus

$$\mathbf{f}(\mathbf{u}(p)) - s\mathbf{u}(p) = const. = \mathbf{f}(\mathbf{u}_L) - s\mathbf{u}_L$$

and the Rankine-Hugoniot condition is satisfied.

These discontinuities are called *contact discontinuities.* The kth characteristics are parallel to the discontinuity. This wave have many similarities with the solution of a linear problem, $u_t + au_x = 0$ with a single discontinuity as initial data. The discontinuity is propagating along the characteristics with the wave speed a. There is no entropy condition for a linearly degenerate field, as in the linear equation the solution in the weak sense is unique.

For systems which do not consist of linearly degenerate and genuinely nonlinear fields, the entropy condition in definition 5.3 has to be replaced with something more general. The situation is similar to the scalar equation when the entropy condition (1.9) is not enough for non-convex conservation laws. We can define the more general entropy condition for systems in the same way as for the scalar equation i.e., in terms of a class of entropy functions $E(\mathbf{u})$, which satisfy an inequality

$$E(\mathbf{u})_t + F(\mathbf{u})_x \leq 0$$

with $\nabla_u F^T = \nabla_u E^T A(\mathbf{u})$.

The formulas are is similar to the formulas for the scalar case.

We will apply the numerical methods to the equations of gas dynamics, which consist of genuinely non linear and linearly degenerate fields. Thus definition 5.3 will be sufficient, and we do not here develop more general entropy conditions.

5.3. The Riemann problem for non linear hyperbolic systems

We here solve the Riemann problem

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}$$
$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_L & \text{if } x < 0 \\ \mathbf{u}_R & \text{if } x > 0 \end{cases}$$

where $\mathbf{u}_L$ and $\mathbf{u}_R$ are two constant states. The solution of this problem will sometimes be used in numerical methods where we solve a Riemann problem locally between the grid points.

We assume that all characteristic fields are genuinely non linear, and that $\mathbf{u}_L$ and $\mathbf{u}_R$ are sufficiently close, such that we can apply the parametrization in theorem 5.6. The solution is similar to the solution for the linear equation, in the sense that it consists of $m+1$ constant states separated by shocks or rarefaction waves.

To construct the solution, we connect $\mathbf{u}_L$ to a new state $\mathbf{u}_1$ by a 1-wave (shock or rarefaction). We write this as

$$\mathbf{u}_1 = \mathbf{u}(p_1, \mathbf{u}_L).$$

Next the state $\mathbf{u}_1$ is connected to the a state $\mathbf{u}_2$ by a 2-wave.

$$\mathbf{u}_2 = \mathbf{u}(p_2, \mathbf{u}_1) = \mathbf{u}(p_1, p_2, \mathbf{u}_L)$$

We continue this to the mth state, given as a function of m parameters and the left state,

$$\mathbf{u}_m = \mathbf{u}(p_1, p_2, \ldots, p_m, \mathbf{u}_L)$$

By requiring

$$\mathbf{u}_R = \mathbf{u}_m$$

we obtain the system of m algebraic equations

$$\mathbf{u}_R = \mathbf{u}(p_1, p_2, \ldots, p_m, \mathbf{u}_L) \tag{5.9}$$

for the m unknown $p_1, \ldots, p_m$. If $\mathbf{u}_L$ and $\mathbf{u}_R$ are close enough, it follows from the inverse mapping theorem that we can always solve (5.9). To see this it is necessary to check that the Jacobian of the mapping

$$f(p_1, p_2, \ldots, p_m) = -\mathbf{u}_R + \mathbf{u}(p_1, p_2, \ldots, p_m, \mathbf{u}_L)$$

is non-singular at $(0, \ldots, 0)$. From the rarefaction wave solutions (5.6), we see that

$$\mathbf{u}(p) = \mathbf{u}_L + p\mathbf{r}_k + O(p^2)$$

which by the smoothness at $p = 0$ (theorem 5.6) must hold for p both positive and negative. For the Riemann problem we thus obtain

$$\begin{aligned}
&\mathbf{u}_1 = \mathbf{u}_L + p_1\mathbf{r}_1 + O(p_1^2) \\
&\mathbf{u}_2 = \mathbf{u}_1 + p_2\mathbf{r}_2 + O(p_2^2) = \mathbf{u}_L + p_1\mathbf{r}_1 + p_2\mathbf{r}_2 + O((p_1+p_2)^2) \\
&\ldots \\
&\mathbf{u}_R = \mathbf{u}_L + p_1\mathbf{r}_1 + p_2\mathbf{r}_2 + \ldots + p_m\mathbf{r}_m + O((p_1+p_2+\ldots+p_m)^2)
\end{aligned}$$

which shows that the Jacobian of the mapping at $(0, \dots, 0)$ is R, the matrix of eigenvectors. This matrix is non singular by the hyperbolicity. Thus the inverse mapping theorem applies and we have a solution of the type shown in Fig. 5.6.

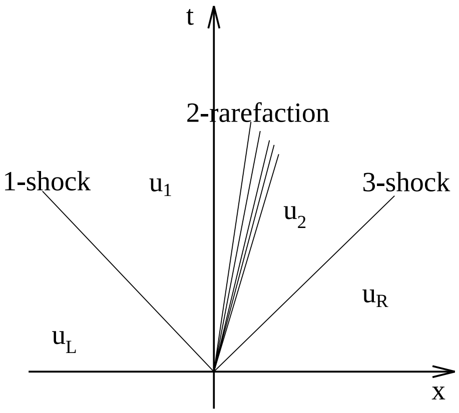


Fig. 5.6. Example wave structure in the solution of the Riemann problem.

In Fig. 5.7a we give an example of a solution for $m = 2$ in the phase plane. Fig. 5.7b indicates the wave structure in the $x - t$ plane, and Fig. 5.7c shows the corresponding solution in variable $u_1(x, t)$ as function of x for a given time.

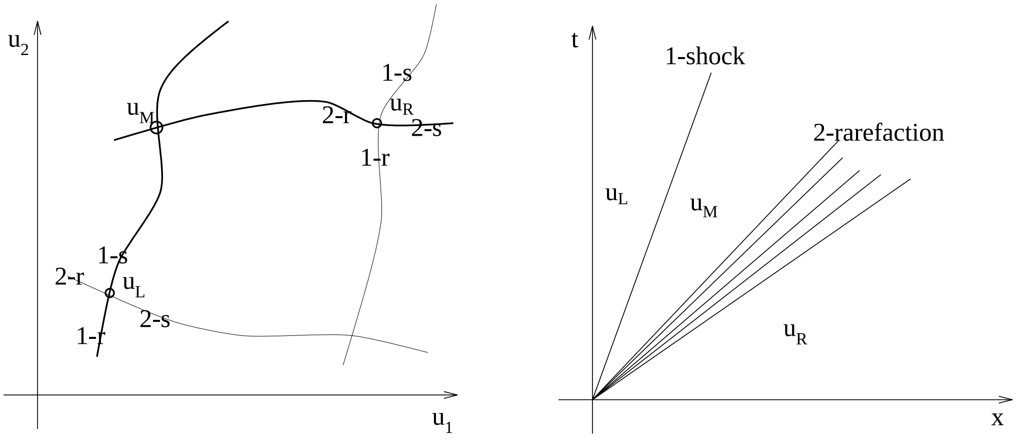


Fig. 5.7a. Example of a phase plane plot. Fig. 5.7b. Corresponding wave structure.

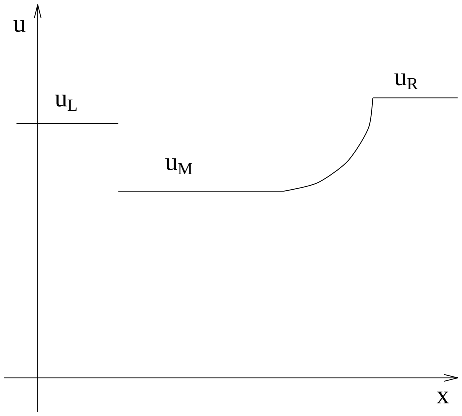


Fig. 5.7c. Solution at a time $t > 0$ for the waves in Fig. 5.7b.

5.4. Existence of solution

We saw in section 5.3 that there exist a solution to the Riemann problem if the states $\mathbf{u}_L$ and $\mathbf{u}_R$ are sufficiently close. The only known result on existence for the problem

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}$$
$$\mathbf{u}(0,x) = \mathbf{u}_0(x)$$

has been proved under a similar assumption, namely that the initial data have sufficiently small variation.

Theorem 5.10. *Assume a non linear hyperbolic system is given, with all the fields genuinely non linear. There are constants δ, K such that if the variation of the initial data is sufficiently small in the sense that*

$$||\mathbf{u}_0 - \mathbf{c}||_\infty + TV(\mathbf{v}_0) \le \delta$$

for some constant state $\mathbf{c}$, then a weak solution $\mathbf{u}(x,t)$ exist, and is such that

$$TV(\mathbf{u}(t,.)) \le K\, TV(\mathbf{u}_0)$$

It is not known whether this solution is unique or satisfies the entropy condition. The theorem is proved by showing convergence of the random choice difference method. The method is interesting because of the convergence properties, but is in practical cases outperformed by most other methods. Therefore, we describe the method here and not in the chapter on difference methods for systems.

The method is defined on a staggered grid, the approximation is u^0_{2j} at t_0 and u^1_{2j-1} at t_1 e.t.c., see Fig. 5.8.

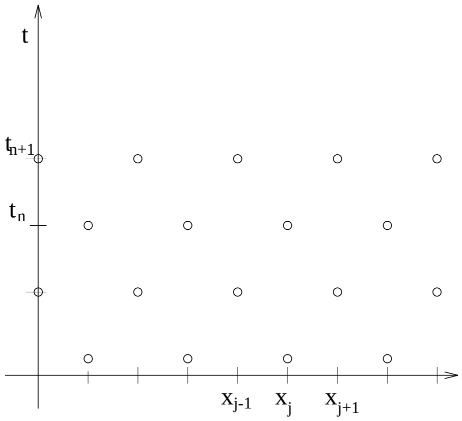


Fig. 5.8. Staggered grid.

Given a solution $\ldots, u^n_{j-2}, u^n_j, u^n_{j+2}, \ldots$, the random choice method consists of the following steps to determine the solution at u^{n+1}_{j+1}

1. Solve the Riemann problem at x_{j+1} with u^n_j as left state and u^n_{j+2} as right state.
2. Let the new time level t_{n+1} be such that no waves from (t_n, x_{j+1}) is outside the interval $[x_j, x_{j+2}]$ at t_{n+1}. This corresponds to a CFL condition.

3. Choose a point $(t_{n+1}, x(\theta)) = (t_{n+1}, x_j + \theta(x_{j+2} - x_j))$ where θ is a random number in $[0, 1]$. The same random number is used for all cells.
4. Define the new value, u_{j+1}^{n+1}, as the value of the solution of the Riemann problem at this random point.

The same procedure is then repeated to get u_j^{n+2} from $u_{j-1}^{n+1}, u_{j+1}^{n+1}$ etc.

In Fig. 5.9, we give a picture of the local Riemann problems in the $x - t$ plane as obtained by this algorithm.

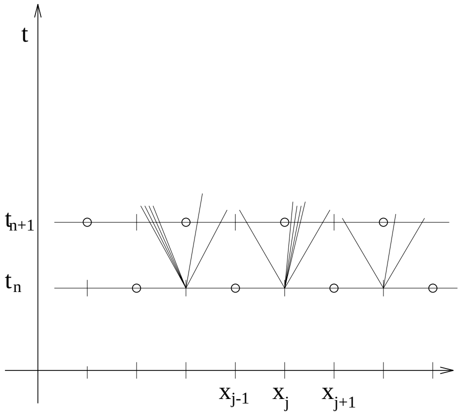


Fig. 5.9. Riemann problems are solved locally at cell interfaces.

The main advantage of the random choice method is that all grid values are obtained as solutions of local Riemann problems. Thus no new intermediate values are introduced in shocks, which in some applications can be of value. Because of the local Riemann problems, control of the variation can be achieved by using estimates for the solutions of the Riemann problem. We do not give the proof of theorem 5.10 here. It is technically complicated, but does not rely on any advanced mathematical concepts.

6. Numerical methods for systems of conservation laws

6.1. Simple waves in gas dynamics

We will consider the generalization of the first order schemes in chapter 2 to systems of equations. For the special case of the gas dynamics equations

$$\begin{pmatrix} \rho \\ m \\ e \end{pmatrix}_t + \begin{pmatrix} m \\ \rho u^2 + p \\ (e+p)u \end{pmatrix}_x = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{6.1}$$

specific formulas will be given. In (6.1) $m = \rho u$ is the momentum, ρ the density, u the velocity, p the pressure and e the total energy of an inviscid fluid. An additional relation to link p to the other variables ρ, m, e is obtained by assuming the perfect gas law

$$p = (\gamma - 1)(e - \frac{1}{2}\rho u^2),$$

where γ is a constant specific for the fluid in question. For air one usually takes $\gamma = 1.4$.

Since there is not sufficient theory available to derive a systematic treatment, the ideas for systems are based on the TVD ideas for scalar equations. The methods for systems are derived in a heuristic way. Thus this chapter can only describe “how to” derive methods for system and not “if” or “why” the methods will give correct answers.

We will give first order accurate methods. Similar to the scalar case second order methods can be derived from the first order ones by piecewise linear interpolation.

In the random choice method and the Godunov method, we have to solve a Riemann problem exactly. In section 6.2 we show how to do this for (6.1) through an iterative procedure. We begin by giving some formulas for simple wave solutions of (6.1) i.e., shocks, rarefaction waves, and contact discontinuities. We noted in chapter 5 that the eigenvalues and eigenvectors are important for the wave structure. Thus we begin by finding these quantites for (6.1).

Theorem 6.1. *For the gas dynamics equations (6.1), where*

$$p = (\gamma - 1)(e - \frac{1}{2}\rho u^2),$$

the eigenvalues and eigenvectors of the Jacobian $\partial \mathbf{f}/\partial \mathbf{u}$ *are*

$$\lambda_1 = u - c \quad \lambda_2 = u \quad \lambda_3 = u + c$$

$$\mathbf{r}_1 = \begin{pmatrix} 1 \\ u - c \\ h - uc \end{pmatrix} \quad \mathbf{r}_2 = \begin{pmatrix} 1 \\ u \\ u^2/2 \end{pmatrix} \quad \mathbf{r}_3 = \begin{pmatrix} 1 \\ u + c \\ h + uc \end{pmatrix}$$

where the sound of speed, c, and the enthalpy h are defined by

$$c = \sqrt{\frac{\gamma p}{\rho}} \qquad h = \frac{e+p}{\rho}$$

Proof: Straightforward calculations, not given here.

The formulas for the eigenvectors and eigenvalues enables us to verify that

$$\mathbf{r}_1^T\nabla\lambda_1 = -(\gamma+1)\frac{c}{2\rho} \neq 0 \qquad \mathbf{r}_2^T\nabla\lambda_2 \equiv 0 \qquad \mathbf{r}_3^T\nabla\lambda_3 = (\gamma+1)\frac{c}{2\rho} \neq 0.$$

For example

$$\mathbf{r}_2^T\nabla\lambda_2 = \frac{\partial u}{\partial \rho} + u\frac{\partial u}{\partial m} + u^2/2\frac{\partial u}{\partial e} =$$
$$\frac{-m}{\rho^2} + \frac{u}{\rho} + 0 = 0$$

Thus the 1 and 3 fields are genuinely non linear and can cause a shock wave or a rarefaction wave to appear in the solution. The 2 field is linearly degenerate and can only give rise to contact discontinuities.

The jump condition is

$$\begin{aligned} s[\rho] &= [m] \\ s[m] &= [\rho u^2 + p] \\ s[e] &= [u(e+p)] \end{aligned} \tag{6.2}$$

where we use the notation $[q] = q_R - q_L$. The jump condition can be rewritten in the following form

$$[\rho v] = 0 \tag{6.3a}$$
$$[\rho v^2 + p] = 0 \tag{6.3b}$$
$$v_L[\frac{2}{\gamma-1}c^2 + v^2] = 0 \tag{6.3c}$$

where we have defined $v = u - s$ as the speed relative to the shock wave. The derivation of (6.3) from (6.2) is somewhat tedious and we omit it here. The form (6.3) is easier to use in proving some of the theorems below.

Finally to obtain conditions which connect states separated by a rarefaction wave, we need the Riemann invariants. From the definition of the Riemann invariant w_k,

$$\mathbf{r}_k^T\nabla w_k = 0,$$

we get by straightforward calculations the following results. For the 1-field

$$w_1^{(1)} = u + \frac{2}{\gamma-1}c \quad w_1^{(2)} = p\rho^{-\gamma}, \tag{6.4a}$$

the 2-field

$$w_2^{(1)} = u \quad w_2^{(2)} = p \tag{6.4b}$$

and finally the 3-field

$$w_3^{(1)} = u - \frac{2}{\gamma-1}c \quad w_3^{(2)} = p\rho^{-\gamma} \tag{6.4c}$$

Thus there are two Riemann invariants for each characteristic field. As seen in the previous chapter the Riemann invariants are constant on rarefaction waves, so that e.g., for two states $\mathbf{u}_L, \mathbf{u}_R$, separated by a 1 rarefaction

$$\begin{aligned} u_L + \frac{2}{\gamma-1}c_L &= u_R + \frac{2}{\gamma-1}c_R \\ p_L\rho_L^{-\gamma} &= p_R\rho_R^{-\gamma} \end{aligned}$$

and similarly for the other fields.

The conditions that connects two states are different if the separating wave is a shock wave or a rarefaction wave. It is therefore necessary to distinguish between these two cases when solving the Riemann problem. One useful criterion is derived in the following theorem.

Theorem 6.2. *For 1-waves*

$$\begin{aligned} p_L < p_R &\quad \text{for shocks} \\ p_L > p_R &\quad \text{for rarefaction waves} \end{aligned}$$

for 3-waves

$$\begin{aligned} p_L > p_R &\quad \text{for shocks} \\ p_L < p_R &\quad \text{for rarefaction waves} \end{aligned}$$

and for the contact discontinuity

$$p_L = p_R$$

Proof: The conditions for shocks are derived from the jump condition (6.3) and the entropy inequalities. We prove the theorem for the 1-waves. First assume a 1-shock. The entropy condition according to definition 5.3 is

$$u_L - c_L > s > u_R - c_R, \qquad s < u_R$$

from which

$$v_L > c_L, \qquad 0 < v_R < c_R$$

follows. Note that v_L and v_R are positive. (6.3c) gives

$$\frac{\gamma+1}{\gamma-1}c_L^2 < \frac{2c_L^2}{\gamma-1} + v_L^2 = \frac{2c_R^2}{\gamma-1} + v_R^2 < \frac{\gamma+1}{\gamma-1}c_R^2$$

and hence

$$c_L < c_R.$$

Use (6.3c) again to obtain

$$0 < \frac{c_R^2}{\gamma-1} - \frac{c_L^2}{\gamma-1} = \frac{1}{2}(v_L^2 - v_R^2)$$

Since v_L, v_R are positive we find that

$$v_L > v_R$$

From (6.3b) we obtain the pressure difference

$$p_L - p_R = \rho_R v_R^2 - \rho_L v_L^2$$

which by (6.3a) is

$$p_L - p_R = \rho_L v_L (v_R - v_L) < 0$$

The result $p_L < p_R$ have been obtained. Next consider a 1 rarefaction wave. We have the inequality

$$u_L - c_L < u_R - c_R$$

which states that the head of the wave travels faster than its tail, see Fig. 6.1 below. The slope of the 1- characteristics is $u_L - c_L$ to the left of the wave and $u_R - c_R$ to the right.

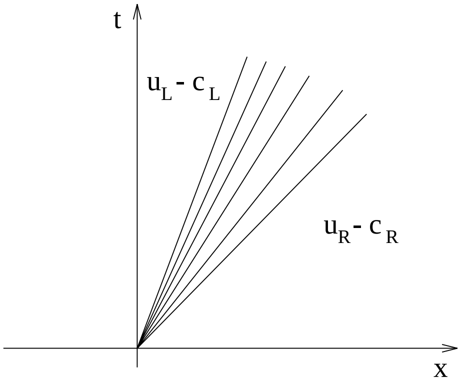


Fig. 6.1. 1-rarefaction wave for (6.1).

From the 1 Riemann invariant (6.4a) we obtain

$$p_R \rho_R^{-\gamma} = p_L \rho_L^{-\gamma}$$
$$\frac{p_R}{p_L} = \left(\frac{\rho_R}{\rho_L}\right)^{\gamma}$$

We use the definition $c^2 = \gamma p/\rho$ to eliminate ρ, with the result

$$\frac{p_R}{p_L} = \left(\frac{c_R}{c_L}\right)^{\frac{2\gamma}{\gamma-1}} \tag{6.5}$$

The second 1 Riemann invariant gives

$$u_L - c_L + \frac{\gamma+1}{\gamma-1} c_L = u_L + \frac{2}{\gamma-1} c_L = u_R + \frac{2}{\gamma-1} c_R =$$
$$u_R - c_R + \frac{\gamma+1}{\gamma-1} c_R > u_L - c_L + \frac{\gamma+1}{\gamma-1} c_R$$

and hence

$$c_L > c_R$$

(6.5) finally gives the result

$$p_L > p_R$$

The proof for the 3 waves is similar and we omit it. For the contact discontinuity, p is a Riemann invariant and according to chapter 5, does not change across it.

6.2. The Riemann problem in gas dynamics

We are now ready to solve the Riemann problem in gas dynamics. Assume that the initial data

$$\mathbf{u}(0,x) = \begin{cases} \mathbf{u}_L & x < 0 \\ \mathbf{u}_R & x > 0 \end{cases}$$

are given. We want to solve (6.1) for these data forwards in time. The solution consists of a 1-wave a contact discontinuity and a 3-wave as seen in Fig. 6.2 below.

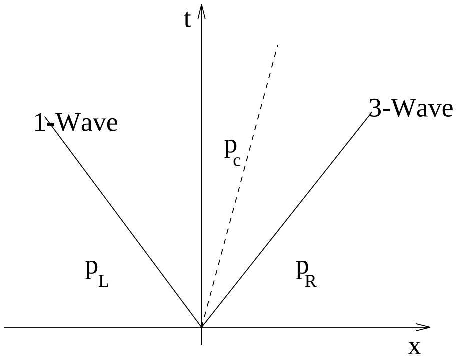


Fig. 6.2. General wave structure for (6.1).

The 1-wave and the 3-wave can be either a shock or a rarefaction wave. The 2-wave is always a contact discontinuity. The pressure does not change across the contact discontinuity, and therefore it should be easier to find an equation for the single unknown pressure than for a quantity which changes across the contact and thus has two unknown states. We begin with finding the intermediate pressure p_c from an iterative process. The rest of the state variables can then be found from direct formulas.

Theorem 6.3. *Define*

$$M_L = -\frac{p_L - p_c}{u_L - u_c} \qquad M_R = \frac{p_R - p_c}{u_R - u_c}$$

then the formulas

$$M_L = \sqrt{\rho_L p_L}\,\phi(p_c/p_L)$$
$$M_R = \sqrt{\rho_R p_R}\,\phi(p_c/p_R)$$

are valid. The function $\phi(x)$ is defined as

$$\phi(x) = \begin{cases} (\frac{\gamma+1}{2}x + \frac{\gamma-1}{2})^{1/2} & \text{if } x \geq 1 \\ \frac{\gamma-1}{2\sqrt{\gamma}}\frac{1-x}{1-x^{(\gamma-1)/2\gamma}} & \text{if } x < 1 \end{cases}$$

Proof: We prove the theorem for the left wave, the proof for the right wave is similar and leads to the same conclusion. First assume that the left wave is a shock. The jump conditions (6.3 a,b) gives

$$p_L - p_c = \rho_L v_L (v_c - v_L).$$

since $v = u - s$ this means that

$$M_L = \rho_L v_L \tag{6.6}$$

Solve (6.3a) for ρ_c and insert into (6.3b,c). We obtain

$$\begin{aligned} &p_L + \rho_L v_L^2 = p_c + \rho_L v_L v_c \\ &\frac{\gamma p_L}{(\gamma-1)\rho_L} + \frac{1}{2}v_L^2 = \frac{\gamma p_c v_c}{(\gamma-1)\rho_L v_L} + \frac{1}{2}v_c^2 \end{aligned}$$

Next solve the first equation above for v_c and insert it into the second. After some simplifying algebra the result is

$$\frac{\rho_L v_L^2}{p_L} = \frac{\gamma-1}{2} + \frac{\gamma+1}{2}\frac{p_c}{p_L}$$

Using (6.6), one gets

$$\frac{M_L^2}{\rho_L p_L} = \frac{\gamma-1}{2} + \frac{\gamma+1}{2}\frac{p_c}{p_L}$$

From theorem 6.2 $p_c/p_L > 1$, since the left wave must be a 1 wave. After taking the squareroot the positive sign should be chosen, since by (6.6) M_L is positive (v_L positive can be seen from the proof of theorem 6.2). Thus we have proved that

$$M_L = \sqrt{\rho_L p_L}\phi(p_L/p_c).$$

if $p_L/p_c > 1$.

Next assume that the left wave is a rarefaction wave. Then the first Riemann invariant for the 1 wave gives

$$\begin{aligned} &u_L + \frac{2}{\gamma-1}c_L = u_c + \frac{2}{\gamma-1}c_c \\ &\Leftrightarrow \\ &\frac{u_L - u_c}{c_L} = \frac{\gamma-1}{2}(\frac{c_c}{c_L} - 1) \\ &\Leftrightarrow \\ &-\frac{p_L - p_c}{c_L M_L} = \frac{2}{\gamma-1}(\frac{c_c}{c_L} - 1) \end{aligned} \tag{6.7}$$

The second Riemann invariant gives

$$p_L \rho_L^{-\gamma} = p_c \rho_c^{-\gamma}.$$

By using the definition $c^2 = \gamma p/\rho$ to eliminate ρ, we can rewrite this as

$$\frac{c_c}{c_L} = \left(\frac{p_c}{p_L}\right)^{\frac{\gamma-1}{2\gamma}}$$

Inserting into (6.7) yields

$$-\frac{p_L - p_c}{c_L M_L} = \frac{2}{\gamma - 1}(\left(\frac{p_c}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} - 1)$$

Finally use $c^2 = \gamma p/\rho$ to eliminate c_L from the left hand side. We know from theorem 6.2 that $p_c/p_L < 1$, and thus the obtained result

$$\frac{\sqrt{\rho_L p_L}}{\sqrt{\gamma} M_L}(p_c/p_L - 1) = \frac{2}{\gamma - 1}(\left(\frac{p_c}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} - 1)$$

is easily seen to be equivalent to

$$M_L = \sqrt{\rho_L p_L}\phi(p_c/p_L).$$

The derivation of the expressions for M_R is analogous and not given here.

The different cases $x < 1$ and $x \geq 1$ corresponds to rarefaction waves and shock waves respectively, as seen from theorem 6.2. Note that the computation of

$$\frac{1-x}{1-x^p}$$

becomes numerically ill conditioned for x close to one. It is therefore good practice in a computer program to replace this function by the polynomial approximation

$$\frac{1}{p} + \frac{p-1}{2p}(1-x)$$

if $1 - \epsilon < x < 1$, where ϵ is a small number and depends on the machine precision used.

Eliminating u_c from the definition of M_L and M_R gives finally the formula

$$p_c = (u_L - u_R + p_R/M_R + p_L/M_L)/(1/M_R + 1/M_L).$$

The iterative method for finding p_c is defined as

$$p_c^* = (u_L - u_R + p_R/M_R^k + p_L/M_L^k)/(1/M_R^k + 1/M_L^k) \qquad (6.8a)$$

$$p_c^{k+1} = \max(p_c^*, \epsilon_1) \qquad (6.8b)$$

where ϵ_1 is introduced to prevent negative pressure during the iteration. Theorem 6.3 is used to evaluate $M_L^k = \sqrt{\rho_L p_L}\phi(p_c^k/p_L)$ and $M_R^k = \sqrt{\rho_R p_R}\phi(p_c^k/p_R)$. The initial guess, $p_c^0 = (p_R + p_L)/2$ has turned out to work well in computer programs. Convergence is usually fast, but for strong rarefactions degradation in convergence rate has been observed. If no convergence is achieved after a fixed number of iterations, we replace (6.8b) by

$$p_c^{k+1} = \alpha \max(\epsilon_1, p_c^*) + (1-\alpha)p_c^k$$

where $\alpha = 1/2$. If there is still convergence problems, we reduce α further.

After p_c is found, we compute

$$u_c = u_R - (p_R - p_c)/M_R$$

Theorem 6.2. gives complete information about the wave configuration. If $p_c/p_L < 1$ the 1-wave is a rarefaction, otherwise a shock and if $p_c/p_R < 1$ the 3-wave is a rarefaction, otherwise a shock. The contact discontinuity lies between the 1 wave and the 3 wave, and propagates with velocity u_c.

For each point (x, t) in which we want to compute the solution, we make tests to decide whether the point is

a) To the left of the 1- wave.
b) Inside the 1-wave if it is a rarefaction.
c) To the right of the 1 wave but to the left of the contact discontinuity.
d) To the right of the contact discontinuity but to the left of the 3-wave.
e) Inside the 3-wave, if it is a rarefaction.
f) To the right of the 3 wave.

The jump condition, or the invariance of the Riemann invariants over the rarefaction waves, gives formulas for the intermediate quantities. Because we know the intermediate pressure p_c it turns out that there is no need to solve any equations, but all required quantities are found from direct formulas. A fortran program which solves the Riemann problem in gas dynamics is supplied in the appendix.

The formulas to determine $\mathbf{u}(x, t)$ will be different in each of the different cases. The computer program will thus contain a certain amount of formulas, but the execution time will be reasonable, since only one branch of the alternatives is actually executed. On a vector computer the situation becomes more troublesome, since there are difficulties in making `IF` statements vectorize.

We have now constructed a solution of the Riemann problem. By further analysis of the solution procedure it is possible to prove

Theorem 6.4. *There is a unique solution of the Riemann problem for the gas dynamics equations (6.1) if*

$$u_R - u_L < \frac{2}{\gamma - 1}(c_L + c_R) \tag{6.9}$$

When (6.9) is not satisfied, there will be a vacuum present in the solution and the intermediate state will therefore not be well defined.

6.3. The Godunov, Roe, and Osher methods

In this section we give a description of three of the best shock capturing methods for systems of conservation laws. First the Godunov scheme is described, since its main feature is the solution of a Riemann problem, most of the description has already been made in section 6.2. This scheme is important since other methods are often thought of as its simplification. However, it is not necessary to make this interpretation.

Second we give the generalization of the upwind scheme to system, known as Roe's method. Finally the Engquist-Osher scheme for systems is described, it is usually called Osher's method.

6.3.1. Godunov's method. Godunov's method has many features in common with the random choice method. The following algorithm describes the method.

1. We start from given $\mathbf{u}_j^n$ the numerical solution at time t_n. The solution is defined for all x by piecewise constant interpolation
$$\mathbf{u}^n(x) = \mathbf{u}_j^n \quad x_{j-1/2} < x < x_{j+1/2}.$$
2. We then solve the Riemann problems at all break points $x_{j+1/2}$. The next time level t_{n+1} is made small enough such that no waves from two different Riemann problems interact. This gives a CFL condition $\Delta t/\Delta x <$ const. Let
$$\mathbf{w}((x - x_{j+1/2})/(t - t_n), \mathbf{u}_j, \mathbf{u}_{j+1})$$
denote the solution of the Riemann problem at $(x_{j+1/2}, t_n)$.
3. The new solution is defined as the average over cell j of the solution of the Riemann problems in 2 i.e.,
$$\begin{aligned}\mathbf{u}_j^{n+1} =& \frac{1}{\Delta x}\Big(\int_{x_{j-1/2}}^{x_j} \mathbf{w}((x - x_{j-1/2})/(t_{n+1} - t_n), \mathbf{u}_{j-1}, \mathbf{u}_j)\, dx \\ &+ \int_{x_j}^{x_{j+1/2}} \mathbf{w}((x - x_{j+1/2})/(t_{n+1} - t_n), \mathbf{u}_j, \mathbf{u}_{j+1})\, dx\Big)\end{aligned}$$

We write this algorithm on conservative form,
$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda\Delta_+\mathbf{h}_{j-1/2}^n,$$

by, taking a contour integral around one cell in the $x - t$ plane. Since the solution in $t_n \le t < t_{n+1}$ satisfies the PDE exactly the integral is zero, and we get

$$\begin{aligned}&\int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{u}(t_n, x)\, dx - \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{w}(0, \mathbf{u}_{j+1}, \mathbf{u}_j))\, dt - \\ &\qquad \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{u}(t_{n+1}, x)\, dx + \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{w}(0, \mathbf{u}_j, \mathbf{u}_{j-1}))\, dt = 0\end{aligned} \tag{6.10}$$

Since we have defined $\mathbf{u}_j^{n+1}$ as the cell average of the solution, and since the integrals in time have time independent integrands, we can rewrite (6.10) as a difference approximation on conservative form with

$$\mathbf{h}_{j+1/2}^n = \mathbf{f}(\mathbf{w}(0, \mathbf{u}_{j+1}, \mathbf{u}_j)).$$

Thus Godunov's method is implemented using the solution procedure in section 6.2. to solve one Riemann problem in each grid point. This solution is evaluated at $x = 0$, and the flux function is evaluated with the solution as argument.

It has often been argued that in Godunov's method a large amount of work is spent to solve a Riemann problem exactly. The information thus computed is mostly wasted since it is only used to form an average.

In this context the generalization of the upwind and the Engquist-Osher schemes to systems, which we now proceed to describe, can be viewed as a Godunov scheme with a simplified solution of the Riemann problem. For each of the scalar methods there are usually a number of different generalizations to systems, some complicated and some very simplified.

6.3.2. Roe's method. We now describe the upwind scheme. For a scalar conservation law, this is the method with numerical flux

$$h^n_{j+1/2} = \frac{1}{2}(f_{j+1} + f_j) - \frac{1}{2}|a_{j+1/2}|(u^n_{j+1} - u^n_j)$$

This scheme is generalized using the eigenvalues of a jacobian matrix as wave speeds. A matrix

$$A_{j+1/2} = A(\mathbf{u}_j, \mathbf{u}_{j+1})$$

with $A(\mathbf{u}, \mathbf{u}) = A(\mathbf{u}) = \partial \mathbf{f}/\partial \mathbf{u}$ is defined, and the scheme becomes

$$\mathbf{h}^n_{j+1/2} = \frac{1}{2}(\mathbf{f}_{j+1} + \mathbf{f}_j) - \frac{1}{2}|A_{j+1/2}|(\mathbf{u}^n_{j+1} - \mathbf{u}^n_j)$$

where the absolute value of the matrix is defined as

$$|A| = R \begin{pmatrix} |\lambda_1| & 0 & \dots & 0 \\ 0 & |\lambda_2| & \dots & 0 \\ 0 & \dots & \ddots & 0 \\ 0 & \dots & 0 & |\lambda_m| \end{pmatrix} R^{-1}.$$

Here λ_j are the eigenvalues and R is the matrix with the eigenvectors as columns. We can see this as a local diagonalization of the system. The matrix can be chosen as $A_{j+1/2} = A((\mathbf{u}_{j+1} + \mathbf{u}_j)/2)$, but the best result seems to be obtained by using a matrix which satisfies the straightforward generalization of a division in the scalar case,

$$\mathbf{f}(\mathbf{u}_{j+1}) - \mathbf{f}(\mathbf{u}_j) = A_{j+1/2}(\mathbf{u}_{j+1} - \mathbf{u}_j).$$

P. Roe has showed how to construct such matrices, the upwind scheme is therefore sometimes called Roe's method. The Roe matrix for the gas dynamics equations is found by evaluating the jacobian at a weighted average,

$$A_{j+1/2} = A(m(\mathbf{u}_{j+1}, \mathbf{u}_j)),$$

where $m(u, v)$ is the weighting procedure described below. The mean value density, velocity and enthalpy is computed using the weights

$$w_1 = \frac{\sqrt{\rho_j}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}} \quad w_2 = \frac{\sqrt{\rho_{j+1}}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}}$$

Thus we compute

$$\begin{aligned}
\rho &= w_1 \rho_j + w_2 \rho_{j+1} \\
u &= w_1 u_j + w_2 u_{j+1} \\
h &= w_1 h_j + w_2 h_{j+1} \\
c^2 &= (\gamma - 1)(h - \frac{1}{2} u^2)
\end{aligned}$$

from which the eigenvalues and eigenvectors of $A_{j+1/2}$ are found using theorem 6.1. In practice the term

$$|A| \Delta_+ \mathbf{u}_j = R |\Lambda| R^{-1} \Delta_+ \mathbf{u}_j$$

is evaluated as

$$\sum_{k=1}^{m} |\lambda_k| \alpha_k \mathbf{r}_k$$

where α_k is solution of the linear system of equations

$$R\alpha = \Delta_+ \mathbf{u}_j.$$

For the Euler equations, α_k can be derived analytically with the following result

$$\begin{aligned}
\alpha_2 &= \frac{\gamma - 1}{c^2}((h - u^2)\Delta_+ \rho_j + u \Delta_+ m_j - \Delta_+ e_j) \\
\alpha_1 &= \frac{1}{2}(\Delta_+ \rho_j - \alpha_2 + \frac{u \Delta_+ \rho_j - \Delta_+ m_j}{c}) \\
\alpha_3 &= \frac{1}{2}(\Delta_+ \rho_j - \alpha_2 - \frac{u \Delta_+ \rho_j - \Delta_+ m_j}{c})
\end{aligned}$$

It is assumed that the matrix R contains the eigenvectors of the jacobian $A(\mathbf{u})$. The quantities without index thus belong to the state in which the jacobian is evaluated.

6.3.3. Osher's method. The Engquist-Osher scheme for systems is usually called the Osher scheme. The numerical flux is

$$\mathbf{h}^n_{j+1/2} = \frac{1}{2}(\mathbf{f}_{j+1} + \mathbf{f}_j) - \frac{1}{2} \int_{\mathbf{u}_j}^{\mathbf{u}_{j+1}} |A(\mathbf{u})| \, d\mathbf{u}$$

The integral of the absolute value of the jacobian,

$$\int_{\mathbf{u}_j}^{\mathbf{u}_{j+1}} |A(\mathbf{u})| \, d\mathbf{u} \tag{6.11}$$

is not path independent, and thus we have to describe an integration path in order to define the method. Osher chose a path which follows the eigenvectors. If $\mathbf{u}(s), 0 \leq s \leq 1$

is a parametrization of the integration path then the following formulas describes the curve.

$$\frac{d\mathbf{u}}{ds} = \mathbf{r}_1 \quad 0 \le s \le s_1$$
$$\frac{d\mathbf{u}}{ds} = \mathbf{r}_2 \quad s_1 \le s \le s_2$$
$$\ldots$$
$$\frac{d\mathbf{u}}{ds} = \mathbf{r}_m \quad s_{m-1} \le s \le 1$$

The first step in the algorithm consists of determining the points $\mathbf{u}(s_k)$. The Riemann invariants, w_k are constant on path k, because

$$\frac{dw_k(\mathbf{u})}{ds} = \nabla w_k^T \frac{d\mathbf{u}}{ds} = \nabla w_k^T \mathbf{r}_k = 0.$$

Thus we have $m-1$ relations

$$w_k^{(j)}(\mathbf{u}_k) = w_k^{(j)}(\mathbf{u}_{k+1}), \quad j = 1, \ldots, m-1 \tag{6.12}$$

for each subpath $s_k < s < s_{k+1}$. The total number of equations is $m(m-1)$. The unknowns are the intermediate states $\mathbf{u}_k$, $k = 1, \ldots, m-1$. Since each state is a vector of m components, the total number of unknowns are also $m(m-1)$. We begin with solving the non linear system of equations (6.12) for the unknowns $\mathbf{u}_k$. Second, we evaluate the integral

$$\int_{\mathbf{u}_j}^{\mathbf{u}_{j+1}} |A|\, d\mathbf{u} = \sum_{k=1}^{m} \int_{\mathbf{u_{k-1}}}^{\mathbf{u}_k} |A|\, d\mathbf{u}$$

where each subpath integral is evaluated using the formula

$$\int_{\mathbf{u}_{k-1}}^{\mathbf{u}_k} |A|\, d\mathbf{u} = \int_{s_{k-1}}^{s_k} |A| \mathbf{r}_k\, ds = \int_{s_{k-1}}^{s_k} |\lambda_k| \mathbf{r}_k\, ds.$$

The subpath integral over $[s_{k-1}, s_k]$ is further divided into pieces where λ_k has constant sign. Without absolute value the integrals are easy to evaluate. As an example, assume that λ_k is positive on $[s_{k-1}, s_*]$ and negative on $[s_*, s_k]$, where $s_* \in [s_{k-1}, s_k]$. Define $\mathbf{u}_* = \mathbf{u}(s_*)$. The subpath integral becomes

$$\int_{s_{k-1}}^{s_*} \lambda_k \mathbf{r}_k\, ds - \int_{s_*}^{s_k} \lambda_k \mathbf{r}_k\, ds = \int_{\mathbf{u}_{k-1}}^{\mathbf{u}_*} A\, d\mathbf{u} - \int_{\mathbf{u}_*}^{\mathbf{u}_k} A\, d\mathbf{u} = -\mathbf{f}(\mathbf{u}_k) - \mathbf{f}(\mathbf{u}_{k-1}) + 2\mathbf{f}(\mathbf{u}_*).$$

In a computer program, the integral (6.11) is determined by adding or subtracting a number of terms $\mathbf{f}(\mathbf{u}_c)$, where $\mathbf{u}_c$ are points of changing subpath or points where λ_k changes sign on a subpath.

We next describe how to implement this method for the Euler equations of gas dynamics. In one space dimension $m = 3$ and we have the following path of integration for the integral (6.11)

$$\begin{aligned}\frac{d\mathbf{u}}{ds} &= \mathbf{r}_1 \qquad 0 \le s \le s_1\\ \frac{d\mathbf{u}}{ds} &= \mathbf{r}_2 \qquad s_1 \le s \le s_2\\ \frac{d\mathbf{u}}{ds} &= \mathbf{r}_3 \qquad s_2 \le s \le 1\end{aligned}$$

The following notation has become standard

$$\begin{aligned}\mathbf{u}_j &= \mathbf{u}(0) = \mathbf{u}_0\\ \mathbf{u}(s_1) &= \mathbf{u}_{1/3}\\ \mathbf{u}(s_2) &= \mathbf{u}_{2/3}\\ \mathbf{u}_{j+1} &= \mathbf{u}(1) = \mathbf{u}_1\end{aligned}$$

We start by determining the intermediate states $\mathbf{u}_{1/3}$ and $\mathbf{u}_{2/3}$. The Riemann invariants (6.4) leads to the following system of equations

$$\begin{aligned}&u_0 + \frac{2}{\gamma-1}c_0 = u_{1/3} + \frac{2}{\gamma-1}c_{1/3}\\ &p_0\rho_0^{-\gamma} = p_{1/3}\rho_{1/3}^{-\gamma}\\ &u_{1/3} = u_{2/3}\\ &p_{1/3} = p_{2/3}\\ &u_1 - \frac{2}{\gamma-1}c_1 = u_{2/3} - \frac{2}{\gamma-1}c_{2/3}\\ &p_1\rho_1^{-\gamma} = p_{2/3}\rho_{2/3}^{-\gamma}\end{aligned}$$

First define

$$\alpha = \left(\frac{p_0\rho_0^{-\gamma}}{p_1\rho_1^{-\gamma}}\right)^{1/2\gamma} = \left(\frac{p_{1/3}\rho_{1/3}^{-\gamma}}{p_{2/3}\rho_{2/3}^{-\gamma}}\right)^{1/2\gamma} = \left(\frac{\rho_{2/3}}{\rho_{1/3}}\right)^{1/2}$$

where we use $p_{1/3} = p_{2/3}$. From the definition $c^2 = \gamma p/\rho$ it follows that

$$\alpha = \frac{c_{1/3}}{c_{2/3}}$$

This equation together with

$$\begin{aligned}u_{1/3} &= u_0 + \frac{2}{\gamma-1}(c_0 - c_{1/3})\\ u_{2/3} &= u_1 - \frac{2}{\gamma-1}(c_1 - c_{2/3})\end{aligned} \tag{6.13}$$

leads to the following formula for $u_{1/3}$ (since we know that $u_{2/3} = u_{1/3}$)

$$u_{1/3} = \frac{u_0 + 2/(\gamma-1)c_0 + \alpha(u_1 - 2/(\gamma-1)c_1)}{(1+\alpha)} \tag{6.14}$$

Thus in a program, one first form α and then use (6.14) to find $u_{1/3}$. $c_{1/3}$ and $c_{2/3}$ are then easily evaluated from (6.13). Then the density is computed as

$$\left(\frac{\rho_{1/3}}{\rho_0}\right)^{1-\gamma} = \left(\frac{c_0}{c_{1/3}}\right)^2$$
$$\left(\frac{\rho_{2/3}}{\rho_1}\right)^{1-\gamma} = \left(\frac{c_1}{c_{2/3}}\right)^2$$

We then have complete information about the states $\mathbf{u}_{1/3}$ and $\mathbf{u}_{2/3}$. We can proceed to the evaluation of the numerical viscosity integral,

$$\int_{\mathbf{u}_j}^{\mathbf{u}_{j+1}} |A(\mathbf{u})|\, d\mathbf{u} = \int_{\mathbf{u}_0}^{\mathbf{u}_{1/3}} |A(\mathbf{u})|\, d\mathbf{u} + \int_{\mathbf{u}_{1/3}}^{\mathbf{u}_{2/3}} |A(\mathbf{u})|\, d\mathbf{u} + \int_{\mathbf{u}_{2/3}}^{\mathbf{u}_1} |A(\mathbf{u})|\, d\mathbf{u} \qquad (6.15)$$

For the Euler equations of gas dynamics, because of the genuine non linearity, the eigenvalues $u - c$ and $u + c$ can change sign in at most one point on their paths. The eigenvalue u is a Riemann invariant and is constant on path 2.

We start with the smallest eigenvalue $\lambda_1 = u - c$ for the first path. The sign of λ_1 is either constant or changes at the single sonic point where

$$u_{s1} - c_{s1} = 0$$

Thus if $(u_0 - c_0)(u_{1/3} - c_{1/3}) < 0$ we include the sonic point in the path and obtain for the first third part of the integral

$$\int_{\mathbf{u}_0}^{\mathbf{u}_{1/3}} |A(\mathbf{u})|\, d\mathbf{u} = \text{sign}(u_0 - c_0)(2\mathbf{f}(\mathbf{u}_{s1}) - \mathbf{f}(\mathbf{u}_j) - \mathbf{f}(\mathbf{u}_{1/3}))$$

Here the sonic state $\mathbf{u}_{s1}$ is found from the Riemann invariants and the sonic condition

$$\begin{aligned} &u_0 + \frac{2}{\gamma-1}c_0 = u_{s1} + \frac{2}{\gamma-1}c_{s1} \\ &p_0\rho_0^{-\gamma} = p_{s1}\rho_{s1}^{-\gamma} \\ &u_{s1} - c_{s1} = 0 \end{aligned} \quad .$$

This system is easy to solve for the sonic state. If λ_1 does not change sign between the 0 and the 1/3 state then

$$\int_{\mathbf{u}_0}^{\mathbf{u}_{1/3}} |A(\mathbf{u})|\, d\mathbf{u} = \text{sign}(u_0 - c_0)(\mathbf{f}(\mathbf{u}_{1/3}) - \mathbf{f}(\mathbf{u}_j))$$

For the second part, the eigenvalue $\lambda_2 = u$ is constant, and the integral becomes

$$\int_{\mathbf{u}_{1/3}}^{\mathbf{u}_{2/3}} |A(\mathbf{u})|\, d\mathbf{u} = \text{sign}(u_{1/3})(\mathbf{f}(\mathbf{u}_{2/3}) - \mathbf{f}(\mathbf{u}_{1/3}))$$

The third part of the integral is similar to the first part, if $(u_{2/3} + c_{2/3})(u_1 + c_1) < 0$ then the sonic point

$$u_{s2} + c_{s2} = 0$$

is included and the integral from 2/3 to 1 becomes

$$\int_{\mathbf{u}_{2/3}}^{\mathbf{u}_1} |A(\mathbf{u})|\, d\mathbf{u} = \mathrm{sign}(u_{2/3} + c_{2/3})(2\mathbf{f}(\mathbf{u}_{s2}) - \mathbf{f}(\mathbf{u}_{2/3}) - \mathbf{f}(\mathbf{u}_{j+1}))$$

The sonic state is found in the same way as the state $s1$ on path 1. If λ_3 does not change sign we obtain

$$\int_{\mathbf{u}_{2/3}}^{\mathbf{u}_1} |A(\mathbf{u})|\, d\mathbf{u} = \mathrm{sign}(u_{2/3} + c_{2/3})(\mathbf{f}(\mathbf{u}_{j+1}) - \mathbf{f}(\mathbf{u}_{2/3}))$$

Summing the three integrals according to (6.15) yields the final result for the integral from $\mathbf{u}_j$ to $\mathbf{u}_{j+1}$. From this integral we obtain the numerical flux as

$$\mathbf{h}_{j+1/2}^n = \frac{1}{2}(\mathbf{f}_{j+1}^n + \mathbf{f}_j^n) - \frac{1}{2}\int_{\mathbf{u}_j}^{\mathbf{u}_{j+1}} |A(\mathbf{u})|\, d\mathbf{u}$$

Remark: Osher originally proposed to order the integration path with the largest eigenvalue first. In practice it has turned out that the method described above, with the smallest eigenvalue first, works much better. The method starting with the smallest eigenvalue is sometimes called the P-version of the scheme, while Osher's original ordering is called the O-version.

6.4. Flux vector splitting

We here give some methods which are somewhat simpler than the methods in the previous section. They are all based on flux vector splitting, the idea of which we can be understand from Engquist-Osher scheme for a scalar problem. The numerical flux for the scalar E-O scheme can be written

$$h^n_{j+1/2} = \frac{1}{2}(f_{j+1} + f_j) - \frac{1}{2}\int_{u_j}^{u_{j+1}} |f'(s)|\, ds = f^+(u_j) + f^-(u_{j+1})$$

with

$$f^+(u) = f(0) + \frac{1}{2}\int_0^u (f'(s) + |f'(s)|)\, ds \qquad f^-(u) = f(u) - f^+(u).$$

Thus the flux function is split in two parts corresponding to positive and negative wave speeds respectively.

$$f(u) = f^+(u) + f^-(u)$$

The approximation of the flux derivative becomes

$$D_-h^n_{j+1/2} = D_+f^-(u_j) + D_-f^+(u_j)$$

such that the derivative is approximated in a stable, upwind way. The Osher scheme for systems can similarly be written as a splitting of the flux function into one part corresponding to positive wave speeds and one part corresponding to negative wave speeds.

In this section we use the flux splitting technique to obtain other, simpler methods than the Osher scheme. The methods are all based on the idea that we split the flux vector

$$\mathbf{f}(\mathbf{u}) = \mathbf{f}^+(\mathbf{u}) + \mathbf{f}^-(\mathbf{u})$$

where we try to achieve that the matrices

$$\begin{aligned} A^+ &= \partial\mathbf{f}^+/\partial\mathbf{u} \\ A^- &= \partial\mathbf{f}^-/\partial\mathbf{u} \end{aligned}$$

are such that A^+ has positive eigenvalues and A^- has negative eigenvalues. The numerical flux

$$\mathbf{h}_{j+1/2} = \mathbf{f}^+(\mathbf{u}_j) + \mathbf{f}^-(\mathbf{u}_{j+1})$$

then defines a method of upwind type.

6.4.1. Steger-Warming splitting. For the first method given here, we need an additional property of the problem to be approximated. It is not hard to prove that for the Euler equations

$$\mathbf{f}(\mathbf{u}) = A(\mathbf{u})\mathbf{u} \tag{6.16}$$

where A is the Jacobian matrix $\partial\mathbf{f}/\partial\mathbf{u}$ (show the homogeniety $\mathbf{f}(\alpha\mathbf{u}) = \alpha\mathbf{f}(\mathbf{u})$ then differentiate with respect to α). If (6.16) holds we define a flux splitting as

$$\mathbf{f}(\mathbf{u}) = A^+\mathbf{u} + A^-\mathbf{u}$$

where

$$A^+ = R \begin{pmatrix} \lambda_1^+ & 0 & \dots & 0 \\ 0 & \lambda_2^+ & \dots & 0 \\ 0 & \dots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_m^+ \end{pmatrix} R^{-1}.$$

and similarly for A^-. As usual λ_i are the eigenvalues and R the matrix of eigenvectors of the Jacobian matrix A. For a scalar we define $\lambda^+ = \max(0, \lambda)$ and $\lambda^- = \min(0, \lambda)$. The flux splitting above is named after Steger and Warming.

6.4.2. van Leer splitting. Another example, which does not require the property (6.16), is the so called van Leer flux vector splitting. The method is, however, closely linked to the structure of the Euler equations. In the van Leer splitting, we use the (signed) Mach number

$$M = \frac{u}{c}$$

to determine the number of positive and negative eigenvalues. If $M > 1$ then the flow is supersonic, and all eigenvalues are positive. We then define

$$\mathbf{f}^+ = \mathbf{f} \quad \mathbf{f}^- = \mathbf{0}.$$

Similarly we define

$$\mathbf{f}^+ = \mathbf{0} \quad \mathbf{f}^- = \mathbf{f}.$$

in the case $M < -1$.

If $|M| < 1$ then there are both positive and negative eigenvalues. We define a splitting using the identity

$$M = ((M+1)^2 - (M-1)^2)/4.$$

The first component of the flux vector is

$$f_1 = \rho u = \rho c M = \rho c((M+1)^2 - (M-1)^2)/4$$

and the definition

$$f_1^+ = \rho c(M+1)^2/4 \qquad f_1^- = -\rho c(M-1)^2/4$$

gives the property $f_1^+ = f_1$ for $M \to 1$, $f_1^- = f_1$ for $M \to -1$.

The other components are treated similarly, the algebra becomes more complicated and we do not give the derivation here. The result is the following formulas

$$\begin{aligned} \mathbf{f}^+ &= \frac{\rho(u+c)^2}{4c} \begin{pmatrix} 1 & \\ (2c + (\gamma-1)u) & /\gamma \\ (2c + (\gamma-1)u)^2 & /(2(\gamma^2-1)) \end{pmatrix} \\ \mathbf{f}^+ &= \frac{\rho(u-c)^2}{4c} \begin{pmatrix} -1 & \\ (2c - (\gamma-1)u) & /\gamma \\ -(2c - (\gamma-1)u)^2 & /(2(\gamma^2-1)) \end{pmatrix} \end{aligned} .$$

It is an easy exercise to verify that $\mathbf{f} = \mathbf{f}^+ + \mathbf{f}^-$. The common factors of flux vectors can be written

$$\begin{aligned} \frac{\rho(u+c)^2}{4c} &= \frac{\rho c}{4}(M+1)^2 \\ \frac{\rho(u-c)^2}{4c} &= \frac{\rho c}{4}(M-1)^2 \end{aligned}$$

so that we have the desired property $\mathbf{f}^+ \to \mathbf{0}$ when $M \to -1$ and $\mathbf{f}^- \to \mathbf{0}$ when $M \to 1$. The van Leer flux splitting is less expensive than the Osher scheme, but gives slightly worse accuracy, especially for resolution of contact discontinuities.

6.4.3. Pressure splitting. We next describe the method of pressure splitting. It is based on the observation that

$$\mathbf{f}(\mathbf{u}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(p+e) \end{pmatrix} = u \begin{pmatrix} \rho \\ \rho u \\ (e+p) \end{pmatrix} + \begin{pmatrix} 0 \\ p \\ 0 \end{pmatrix} = u\mathbf{f}_c + \mathbf{f}_p$$

It turns out that the eigenvalues of the first term, $u\mathbf{f}_c$ are $u, u,$andγu. The eigenvalues of the second term are $0, 0$ and $-(\gamma - 1)u$. Thus it is natural to try a splitting according to the sign of u.

$$\begin{aligned} \mathbf{f}^+ &= u^+\mathbf{f}_c + \frac{1-\text{sign}(u)}{2}\mathbf{f}_p \\ \mathbf{f}^- &= u^-\mathbf{f}_c + \frac{1+\text{sign}(u)}{2}\mathbf{f}_p \end{aligned}.$$

However, when the discontinuity in the switch $(1 + \text{sign}(u))/2$ is differenced the method becomes unstable. It is important that the switch of sign is continuous, as in the first term where u^+ and u^- are continuous at $u = 0$. To overcome this difficulty, we replace the sign function by a smoother version. In many applications the function

$$g(M) = \begin{cases} M(3-M^2)/2 & |M| < 1 \\ \text{sign}(M) & |M| > 1 \end{cases}$$

is used instead of the sign function. The signed Mach number $M = u/c$ has of course the same sign as u. The total pressure splitting method then becomes

$$\begin{aligned} \mathbf{f}^+ &= u^+\mathbf{f}_c + \frac{1-g(M)}{2}\mathbf{f}_p \\ \mathbf{f}^- &= u^-\mathbf{f}_c + \frac{1+g(M)}{2}\mathbf{f}_p \end{aligned}.$$

Usually it is necessary to add an entropy fix i.e., increase the amount of artificial dissipation, for the $u\mathbf{f}_c$ terms when u is near zero, in the same way as this is done for the upwind method when the wave speed changes sign. The advantage of the pressure splitting method is that it requires a relatively few number of arithmetic operations.

6.4.4. Lax-Friedrichs splitting. Finally we show how the Lax-Friedrich scheme can be viewed as a flux splitting method. The numerical flux for a system is

$$\mathbf{h}_{j+1/2} = \frac{1}{2}(\mathbf{f}_{j+1} + \mathbf{f}_j) - \frac{1}{2\lambda}(\mathbf{u}_{j+1} - \mathbf{u}_j)$$

which we split in two parts by defining

$$\mathbf{f}^+(\mathbf{u}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \frac{1}{\lambda}\mathbf{u})$$
$$\mathbf{f}^-(\mathbf{u}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) - \frac{1}{\lambda}\mathbf{u})$$

where now $\lambda = \Delta t/\Delta x$.

Since the CFL condition $\max |a^k|\lambda \leq 1$ is used, with a^k eigenvalues to the jacobian matrix, we see that the matrices

$$\frac{\partial \mathbf{f}^+}{\partial \mathbf{u}} \qquad \frac{\partial \mathbf{f}^-}{\partial \mathbf{u}}$$

have positive and negative eigenvalues respectively. Thus the definition is reasonable. We can generalize this and define

$$\mathbf{f}^+(\mathbf{u}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + k\mathbf{u})$$
$$\mathbf{f}^-(\mathbf{u}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) - k\mathbf{u})$$

with $k > 0$. If $k = \max |a_j^k|$, the largest eigenvalue of the jacobian matrix at $\mathbf{u}_j$, the scheme is called Rusanov's method.

All the methods described in this section are inferior in shock resolution to the methods in the previous section, (Godunov, Roe and Osher). For example some of the methods in this section does not permit a steady shock solution spread over a fixed number of grid points, instead all steady shocks will be smeared out over a large number of grid points. The better schemes does admit such steady shock profiles. The schemes in this section are however somewhat simpler to implement and uses fewer arithmetic operations.

6.5. Interpretation as approximate Riemann solver

The schemes described in this chapter can be viewed as approximations to the Godunov scheme. We can construct methods in the same way as we defined the Godunov scheme in section 6.3, but with $\mathbf{w}(x/t)$, the solution of the Riemann problem, replaced by an approximate solution to the same Riemann problem. Let

$$\mathbf{w}((x - x_{j-1/2})/(t - t_n), \mathbf{u}_j, \mathbf{u}_{j-1})$$

be an approximate solution of the Riemann problem between the states $\mathbf{u}_j$ and $\mathbf{u}_{j-1}$. Assume the same set up as in the description of the Godunov scheme in section 6.3. We thus define the cell average u_j^{n+1} on the new time level as

$$\begin{aligned}\mathbf{u}_j^{n+1} =& \frac{1}{\Delta x}(\int_{x_{j-1/2}}^{x_j} \mathbf{w}((x - x_{j-1/2})/(t_{n+1} - t_n), \mathbf{u}_{j-1}, \mathbf{u}_j)\, dx \\ &+ \int_{x_j}^{x_{j+1/2}} \mathbf{w}((x - x_{j+1/2})/(t_{n+1} - t_n), \mathbf{u}_j, \mathbf{u}_{j+1})\, dx)\end{aligned} \qquad (6.17)$$

First we give a necessary condition that such an approximate solver has to satisfy.

Theorem 6.5. *If the approximate solution of the Riemann problem between $\mathbf{u}_L$ and $\mathbf{u}_R$ with jump at $x = 0$, $\mathbf{w}(x/t, \mathbf{u}_R, \mathbf{u}_L)$, is consistent with the conservation law in the sense that*

$$\int_{-\Delta x/2}^{\Delta x/2} \mathbf{w}(x/t, \mathbf{u}_R, \mathbf{u}_L)\, dx = \frac{\Delta x}{2}(\mathbf{u}_L + \mathbf{u}_R) - \Delta t(\mathbf{f}_R - \mathbf{f}_L) \qquad (6.18)$$

then (6.17) defines a scheme on conservative form, consistent with the conservation law.

We do not give the proof of this theorem, but we derive a formula for the numerical flux associated with a given approximate Riemann solver.

Start by integrating around the square $[x_j, x_{j+1/2}] \times [t_n, t_{n+1}]$ and set this integral equal to zero. We obtain

$$\begin{aligned}&\int_{x_j}^{x_{j+1/2}} \mathbf{u}_j^n\, dx - \int_{t_n}^{t_{n+1}} \mathbf{h}_{j+1/2}\, dt + \int_{x_{j+1/2}}^{x_j} \mathbf{w}((x - x_{j+1/2})/(t_{n+1} - t_n), \mathbf{u}_{j+1}, \mathbf{u}_j) - \\ &\int_{t_{n+1}}^{t_n} \mathbf{f}(\mathbf{u}_j^n)\, dt = \mathbf{0} \Leftrightarrow \\ &\mathbf{h}_{j+1/2} = \frac{\mathbf{u}_j^n \Delta x}{2\Delta t} + \mathbf{f}_j - \frac{1}{\Delta t}\int_{x_j}^{x_{j+1/2}} \mathbf{w}((x - x_{j+1/2})/(t_{n+1} - t_n), \mathbf{u}_{j+1}, \mathbf{u}_j)\end{aligned} \qquad (6.19)$$

where the numerical flux $h_{j+1/2}$ is now unknown. We only use the approximate Riemann solution on the time level t_{n+1}, but not in between the time levels. The formula (6.18) guarantees that if we integrate around the square $[x_{j+1/2}, x_{j+1}] \times [t_n, t_{n+1}]$ instead, we obtain the same numerical flux. Integrate

$$\int_{x_{j+1/2}}^{x_{j+1}} \mathbf{u}_{j+1}^n\,dx - \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{u}_{j+1}^n)\,dt + \int_{x_{j+1}}^{x_{j+1/2}} \mathbf{w}((x-x_{j+1/2})/(t_{n+1}-t_n), \mathbf{u}_{j+1}, \mathbf{u}_j) -$$

$$\int_{t_{n+1}}^{t_n} \mathbf{h}'_{j+1/2}\,dt = \mathbf{0} \Leftrightarrow$$

$$\mathbf{h}'_{j+1/2} = \frac{\mathbf{u}_{j+1}^n \Delta x}{2\Delta t} + \mathbf{f}_{j+1} + \frac{1}{\Delta t}\int_{x_{j+1/2}}^{x_{j+1}} \mathbf{w}((x-x_{j+1/2})/(t_{n+1}-t_n), \mathbf{u}_{j+1}, \mathbf{u}_j) \tag{6.20}$$

and combine (6.19) and (6.20), it is easy to see that (6.18) gives $h_{j+1/2} = h'_{j+1/2}$.

Note that we can *not*, as we did for the Godunov scheme, obtain the numerical flux as $\mathbf{f}(\mathbf{w}(0, \mathbf{u}_j, \mathbf{u}_{j-1}))$. This is because the solution between the time levels is not an exact solution of the continuous problem, and thus the contour integral (6.10) is not equal to zero. The numerical flux is uniquely determined from $\mathbf{w}$ by the formula (6.19).

It is not hard to show that the upwind scheme can be obtained in this way, if we define $\mathbf{w}(x/t, \mathbf{u}_R, \mathbf{u}_L)$ as the solution to the linearized Riemann problem

$$\mathbf{u}_t + A(\mathbf{u}_R, \mathbf{u}_L)\mathbf{u}_x = \mathbf{0}$$

$$\mathbf{u}(0,x) = \begin{cases} \mathbf{u}_L & x < 0 \\ \mathbf{u}_R & x \geq 0 \end{cases}$$

where $A(\mathbf{u}, \mathbf{v})$ is e.g., the Roe matrix (excercise).

Next we derive a simplified scheme for the Euler equations of gas dynamics. Assume an approximate solution of the following form

$$\mathbf{w}(x/t, \mathbf{u}_L, \mathbf{u}_R) = \begin{cases} \mathbf{u}_L & x < b_1 t \\ \mathbf{u}_m & b_1 t \leq x \leq b_2 t \\ \mathbf{u}_R & b_2 t < x \end{cases}$$

we thus assume that there are two waves moving with speeds b_1 and b_2, and we require $b_1 < b_2$. The intermediate state $\mathbf{u}_m$ is determined from the consistency condition (6.19), with the result

$$\mathbf{u}_m = \frac{b_2 \mathbf{u}_L - b_1 \mathbf{u}_R}{b_2 - b_1} - \frac{\mathbf{f}_R - \mathbf{f}_L}{b_2 - b_1}$$

By evaluating the integrals (6.18) we obtain the numerical flux for this method

$$h_{j+1/2} = \frac{b^+_{j+1/2}\mathbf{f}_j - b^-_{j+1/2}\mathbf{f}_{j+1}}{b^+_{j+1/2} - b^-_{j+1/2}} + \frac{b^+_{j+1/2} b^-_{j+1/2}}{b^+_{j+1/2} - b^-_{j+1/2}}(\mathbf{u}_{j+1} - \mathbf{u}_j)$$

where $b^+_{j+1/2} = \max(b_{2,j+1/2}, 0)$ and $b^-_{j+1/2} = \min(b_{1,j+1/2}, 0)$. The wave speeds are parameters we can tune to obtain a method with desired properties. For the Euler equations, we can use the largest and smallest eigenvalues

$$b_{1,j+1/2} = u_{j+1/2} - c_{j+1/2} \qquad b_{2,j+1/2} = u_{j+1/2} + c_{j+1/2}$$

where $u_{j+1/2}$ and $c_{j+1/2}$ are the velocity and the speed of sound evaluated at an intermediate point. One reasonable choice is to use the average procedure in Roe's method i.e.,

$$u_{j+1/2} = w_{1,j+1/2}u_j + w_{2,j+1/2}u_{j+1}$$
$$h_{j+1/2} = w_{1,j+1/2}h_j + w_{2,j+1/2}h_{j+1}$$
$$c_{j+1/2}^2 = (\gamma - 1)(h_{j+1/2} - \frac{1}{2}u_{j+1/2})$$

where h is the entalphy, $h = (p + e)/\rho$, and the weights are

$$w_{1,j+1/2} = \frac{\sqrt{\rho_j}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}} \qquad w_{2,j+1/2} = \frac{\sqrt{\rho_{j+1}}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}}$$

(c.f. section 6.3). This scheme is sometimes called the HLL scheme from the initials of its inventors (Harten, Lax, van Leer).

Note that the wave speeds $b_1 = -\lambda$ and $b_2 = \lambda$ gives the Lax-Friedrichs scheme.

6.6. Generalization to second and higher order of accuracy

The same ideas as were used in chapter 3 are here used for systems. Assume that a first order method is given. We obtain a second order method by using the numerical flux

$$\mathbf{h}_{j+1/2} = \mathbf{h}(\mathbf{u}_{j+1} - \mathbf{s}_{j+1}/2, \mathbf{u}_j + \mathbf{s}_j/2)$$

where $\mathbf{s}$ are slopes of a piecewise linear reconstruction and where $\mathbf{h}_{j+1/2}^n$ is the flux of the first order method. We can use inner flux limiters as described in chapter 3, adapted to systems analogously, but here we only describe piecewise linear interpolation.

One additional difficulty is to determine a good coordinate system for the interpolation. The following strategies are in use

(a) Do interpolation componentwise in the conserved variables ($\rho\ m\ e$).

(b) Do interpolation componentwise in the variables ($\rho\ u\ p$).

(c) For each $\mathbf{u}_j$ define the characteristic coordinate system spanned by the left eigenvectors of the jacobian A evaluated at $\mathbf{u}_j$, $\mathbf{l}^k(\mathbf{u}_j)$, $k = 1, \ldots, m$. All the interpolation or limiting for the cell j is then made in this coordinate system. e.g., define the characteristic variables

$$c_i^k = \mathbf{l}^k(\mathbf{u}_j)^T\mathbf{u}_{j+i}, \quad i = -1, 0, 1 \quad k = 1, \ldots, m$$

and then the componentwise slopes

$$s_j^k = B(c_1^k - c_0^k, c_0^k - c_{-1}^k)$$

where $B(x, y)$ is a limiter function described in chapter 3. Finally the slopes in the original coordinates are obtained as

$$\mathbf{s}_j = \sum_{k=1}^{m} s_j^k \mathbf{r}^k(\mathbf{u}_j)$$

with $\mathbf{r}^k$, the right eigenvectors to A.

(d) Use the Roe matrix and associated quantities. For a description of the matrix, see Section 6.3. The quantities (cf. Section 6.3) $\alpha^k_{j+1/2}$ are used to represent $\Delta_+ u_j$ in characteristic coordinates. The slope limiting can be defined as

$$s^k_j = B(\alpha^k_{j+1/2}, \alpha^k_{j-1/2})$$

with $B(x,y)$ a limiter function. In the physical coordinates the slopes are added to $\mathbf{u}$ at interfaces $j+1/2$, and we thus take

$$\mathbf{u}_j + \mathbf{s}_j/2 = \mathbf{u}_j + \frac{1}{2}\sum_{k=1}^{m} s^k_j \mathbf{r}^k_{j+1/2}.$$

At the $j-1/2$ interface we use the eigenvectors $\mathbf{r}^k_{j-1/2}$ to the matrix $A_{j-1/2}$ instead.

Note that in (c) the coordinate system is kept fixed, at a cell j when the slopes in j are computed. While in (d) the limiting is done on quantities belonging to different coordinate systems (e.g., $\alpha_{j+3/2}$ and $\alpha_{j+1/2}$).

The scheme with outer limiter (3.30) can be generalized to systems in a similar way. We can use the Roe matrix decomposition or a fixed chararacteristic coordinate system.

(d) can also be viewed as a general way to generalize schemes for scalar problems to systems. All occurencies of $\Delta_+ u_j$ in the scalar method are replaced by $\alpha^k_{j+1/2}$ for all components k. Finally the numerical flux function, $\mathbf{h}_{j+1/2}$ is evaluated by using the coordinate system spanned by $\mathbf{r}^k_{j+1/2}$. This method is usually applied to the Lax-Wendroff type TVD schemes described in Chapter 3. We give an example to clarify this.

Example 6.1 The method with the following numerical flux is a second order ENO scheme based on the Lax-Wendroff scheme, derived for a scalar conservation law.

$$h^n_{j+1/2} = \frac{1}{2}(f(u^n_{j+1}) + f(u^n_j)) + \frac{1}{2}a^+_{j+1/2}(1 - \lambda a_{j-1/2})/(1 + \lambda(a_{j+1/2} - a_{j-1/2}))s_j$$
$$- \frac{1}{2}a^-_{j+1/2}(1 - \lambda a_{j+3/2})/(1 + \lambda(a_{j+3/2} - a_{j+1/2}))s_{j+1}$$

where

$$s_j = minmod(\Delta_+ u^n_j, \Delta_- u^n_j).$$

$a_{j+1/2}$ is as usual the local wave speed (cf. Chapter 2), and $\lambda = \frac{\Delta t}{\Delta x}$. The derivation is omitted since we just want to show how to generalize this method to systems using (d) above. For systems the wave speeds are the eigenvalues of the jacobian. We thus replace $a_{j+1/2}$ with $a^k_{j+1/2}$ the eigenvalue of the Roe matrix. The coefficients $\alpha^k_{j+1/2}$ are used instead of $\Delta_+ u_j$, since by definition

$$R\alpha = \Delta_+ \mathbf{u}_j$$

or written on vector form

$$\sum_{k=1}^{m} \alpha^k_{j+1/2}\mathbf{r}^k_{j+1/2} = \Delta_+ \mathbf{u}_j.$$

The matrix R have the eigenvectors of the Roe matrix, $\mathbf{r}^k_{j+1/2}$ as columns. The numerical flux for a system becomes

$$\begin{aligned}\mathbf{h}^n_{j+1/2} =& \frac{1}{2}(\mathbf{f}_j + \mathbf{f}_{j+1}) + \\ & \frac{1}{2}\sum_{k=1}^{m}(a^{+,k}_{j+1/2}(1-\lambda a^k_{j-1/2})/(1+\lambda(a^k_{j+1/2}-a^k_{j-1/2}))s^k_j - \\ & \qquad \frac{1}{2}a^{-,k}_{j+1/2}(1-\lambda a^k_{j+3/2})/(1+\lambda(a^k_{j+3/2}-a^k_{j+1/2}))s^k_{j+1})\mathbf{r}^k_{j+1/2}\end{aligned}$$

with the slope limited as

$$s^k_j = \text{minmod}(\alpha^k_{j+1/2}, \alpha^k_{j-1/2}).$$

Here we use the notation $a^{+,k}_{j+1/2} = \max(0, a^k_{j+1/2})$ and similarly for $a^{-,k}_{j+1/2}$.

It has been observed that the ENO scheme can give a solution with spurious oscillations if the interpolation is not made in the characteristic variables.

6.7. Some test problems

In this section we have collected some test problems for the one dimensional Euler equations. We will always let $\mathbf{u}$ denote the vector with components $(\rho \;\; m \;\; e)$, the density, momentum and total energy.

First we give some Riemann problems, which have become standard to use in tests of numerical methods. The first problem is

$$\mathbf{u}_1(0,x) = \begin{cases}\begin{pmatrix}0.445\\0.311\\8.928\end{pmatrix} & x < 0\\ \begin{pmatrix}0.5\\0\\1.4275\end{pmatrix} & x \geq 0\end{cases} \tag{6.21}$$

The solution of this problem at a time >0 is given in Fig. 6.3.

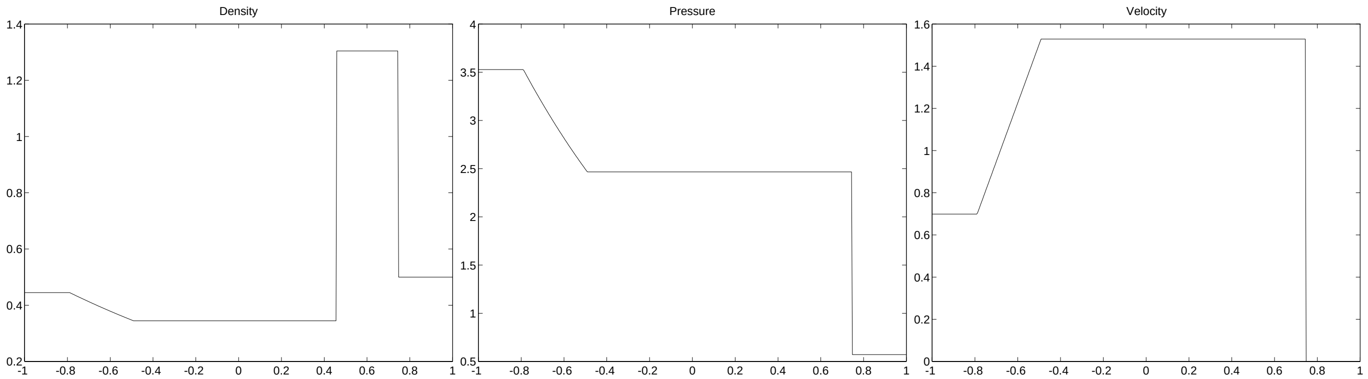


Fig. 6.3. Solution of Riemann problem (6.21).

A 1-rarefaction wave is travelling to the left. Travelling to the right, a 3-shock is followed by a contact discontinuity. Note that the pressure does not change across the contact discontinuity. The intermediate states are

$$\mathbf{u}_1 = \begin{pmatrix} 0.345 \\ 0.528 \\ 6.571 \end{pmatrix} \qquad \mathbf{u}_2 = \begin{pmatrix} 1.304 \\ 1.995 \\ 7.693 \end{pmatrix}$$

and the wave speeds

$s_1 = 2.480$ for the shock

$s_2 = u_c = 0.5290$ for the contact discontinuity

$u_L - c_L = -2.6326 \;\; u_1 - c_1 = -1.6365$ across the rarefaction wave

Problem number two is the following

$$\mathbf{u}_2(0,x) = \begin{cases} \begin{pmatrix} 1 \\ 0 \\ 2.5 \end{pmatrix} & x < 0 \\ \begin{pmatrix} 0.125 \\ 0 \\ 0.25 \end{pmatrix} & x \geq 0 \end{cases} \tag{6.22}$$

for which the solution is displayed in Fig. 6.4.

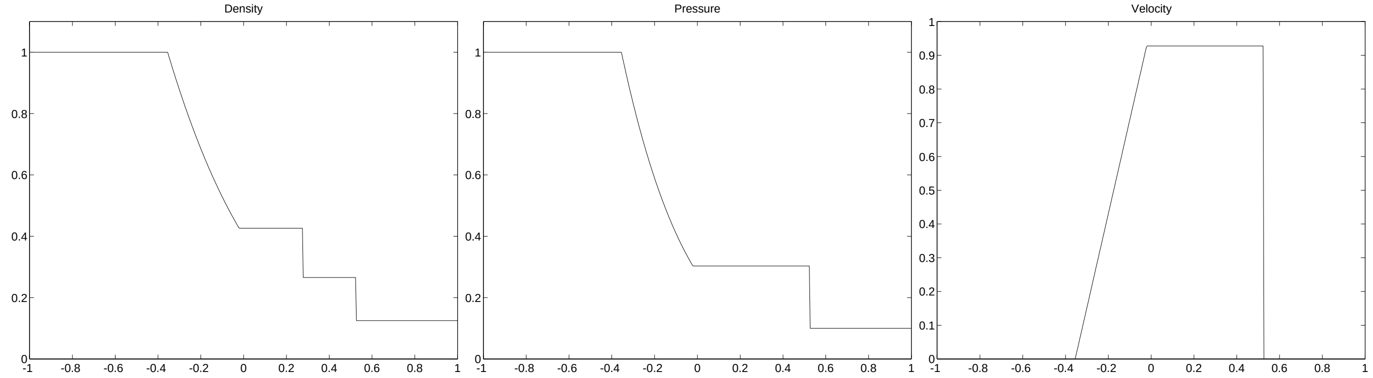


Fig. 6.4. Solution of Riemann problem (6.22).

It has similar structure to the previous problem, with a 1-rarefaction followed by a contact discontinuity and a 3-shock. The intermediate states are

$$\mathbf{u}_1 = \begin{pmatrix} 0.42632 \\ 0.39539 \\ 0.94118 \end{pmatrix} \qquad \mathbf{u}_2 = \begin{pmatrix} 0.26557 \\ 0.24631 \\ 0.87204 \end{pmatrix}$$

and the wave speeds

$$\begin{aligned} s_1 &= 1.75222 \text{ for the shock} \\ s_2 &= u_c = 0.92745 \text{ for the contact discontinuity} \\ u_L - c_L &= -1.183216 \;\; u_1 - c_1 = -0.07027 \text{ across the rarefaction wave} \end{aligned}$$

The second problem contains a rarefaction wave which is close to being transonic, and is a good test for the entropy condition.

These two problems can usually be solved without difficulties, although the quality of the solution of course differs from different methods. A more difficult problem is the following

$$\mathbf{u}_3(0,x) = \begin{cases} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} & x < 0 \\ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} & x \geq 0 \end{cases} . \tag{6.23}$$

The solution is displayed in Fig. 6.5.

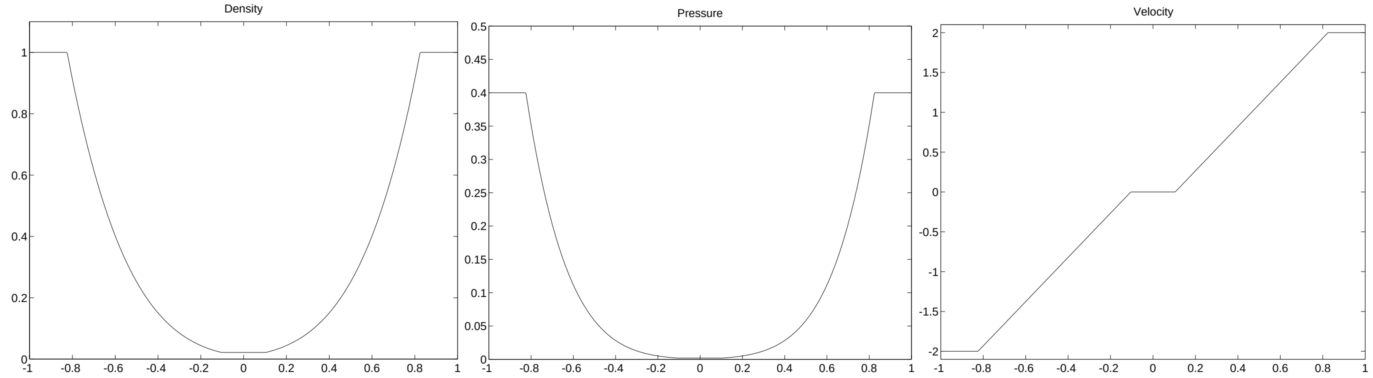


Fig. 6.5. Solution of Riemann problem (6.23).

The intermediate state is

$$\begin{pmatrix} 0.02185 \\ 0 \\ 0.004735 \end{pmatrix}$$

The state of low density and pressure will sometimes lead to difficulties with negative pressure. Of the schemes described in this chapter only Godunov and the P-version of the Osher scheme can solve this problem, without crashing because of negative pressure.

Another common test problem is the so called blast wave problem. This problem is defined on $0 < x < 1$ with solid walls at $x = 0$ and $x = 1$. The initial data is

$$\mathbf{u}_4(0,x) = \begin{cases} (\,1 \;\; 0 \;\; 2500\,)^T & x < 0.1 \\ (\,1 \;\; 0 \;\; 0.025\,)^T & 0.1 < x < 0.9 \\ (\,1 \;\; 0 \;\; 250\,)^T & 0.9 < x \end{cases} \tag{6.24}$$

At the walls the boundary conditions for the velocity

$$u(t,1) = 0 \quad u(t,0) = 0$$

are imposed. A shock wave and a rarefaction wave are formed at $t = 0.038$ they have interacted to produce the solution in Fig. 6.6

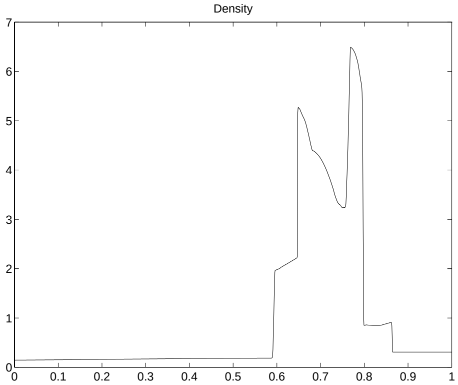


Fig. 6.6. Solution of the blast wave problem (6.24) at $t = 0.038$.

after one reflection at the boundary. The very large differences in pressure and the more complex structure of the solution makes this a more challangeing problem than a single Riemann problem.

The fifth problem is

$$\mathbf{u}_5(0,x) = \begin{cases} \begin{pmatrix} 3.857143 \\ 10.141852 \\ 39.166666 \end{pmatrix} & x < -4 \\ \begin{pmatrix} 1+\epsilon \sin 5x \\ 0 \\ 2.5 \end{pmatrix} & x \geq -4 \end{cases} \tag{6.25}$$

here one shock wave interacts with a sine wave of small amplitude. If $\epsilon = 0$ this is a shock wave moving to the right. Usually one takes $\epsilon = 0.2$. The solution for this ϵ is

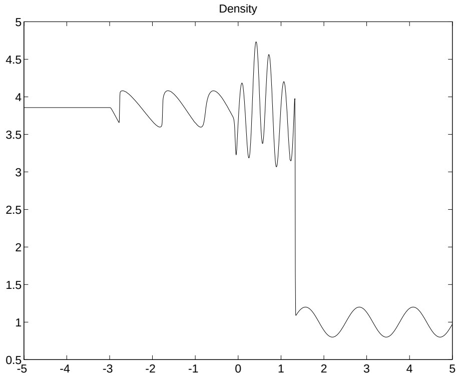


Fig. 6.6. Solution of the oscillatory problem (6.25).

The difficulty lies in resolving the oscillations that are formed behind the shock wave in the density. This is a good test for resolution, usually the higher order accurate ENO methods perform much better than second order TVD methods for this problem.

The solutions in the two last problems can not be found analytically, but in one space dimension it is possible to obtain a converged solution by putting in a very large amount of grid points. We do this to find the "exact" solution for comparison.

Exercises

1. Solve the linear Riemann problem

$$\mathbf{u}_t + A\mathbf{u}_x = \mathbf{0}$$
$$\mathbf{u}(0,x) = \begin{cases} \mathbf{u}_L & x < 0 \\ \mathbf{u}_R & x \geq 0 \end{cases}$$

   where A is a constant matrix with a basis of eigenvectors.

2. We define an approximate Riemann solver, $\mathbf{w}$ as the solution of the Riemann problem for a linear equation with the Roe matrix as coefficient matrix. The techniques in section 6.4 are used to define a difference method from this approximate Riemann solver. Show that the method thus obtained is the same as Roe's method, described in section 6.2.

3. Give an example which shows that for the Osher scheme applied to the Euler equations

$$\mathbf{h}_{j+1/2} \neq \mathbf{f}_j$$

   even if all wave speeds in $\mathbf{u}_{j+1}$ and $\mathbf{u}_j$ are positive.