3
A Short Course in Difference Methods

Computation will cure what ails you.

Clifford Truesdell, The Computer, Ruin of Science and Threat to Mankind, 1980/1982

Although front tracking can be thought of as a numerical method, and has indeed been shown to be excellent for one-dimensional conservation laws, it is not part of the standard repertoire of numerical methods for conservation laws. Traditionally, difference methods have been central to the development of the theory of conservation laws, and the study of such methods is very important in applications.

This chapter is intended to give a brief introduction to difference methods for conservation laws. The emphasis throughout will be on methods and general results rather than on particular examples. Although difference methods and the concepts we discuss can be formulated for systems, we will exclusively concentrate on scalar equations. This is partly because we want to keep this chapter introductory, and partly due to the lack of general results for difference methods applied to systems of conservation laws.

3.1 Conservative Methods

We are interested in numerical methods for the scalar conservation law in one dimension. (We will study multidimensional problems in Chapter 4.)

Thus we consider

$$u_t + f(u)_x = 0, \qquad u|_{t=0} = u_0. \tag{3.1}$$

A difference method is created by replacing the derivatives by finite differences, e.g.,

$$\frac{\Delta u}{\Delta t} + \frac{\Delta f(u)}{\Delta x} = 0. \tag{3.2}$$

Here Δt and Δx are small positive numbers. We shall use the notation

$$U_j^n = u\left(j\Delta x, n\Delta t\right) \quad \text{and} \quad U^n = \left(U_{-K}^n, \dots, U_j^n, \dots, U_K^n\right),$$

where u now is our numerical approximation to the solution of (3.1). Normally, since we are interested in the initial value problem (3.1), we know the initial approximation

$$U_j^0, \quad -K \le j \le K,$$

and we want to use (3.2) to calculate U^n for $n \in \mathbb{N}$. We will not say much about boundary conditions in this book. Often one assumes that the initial data is periodic, i.e.,

$$U_{-K+j}^0 = U_{K+j}^0, \quad \text{for } 0 \le j \le 2K,$$

which gives $U_{-K+j}^n = U_{K+j}^n$. Another commonly used device is to assume that $\partial_x f(u) = 0$ at the boundary of the computational domain. For a numerical scheme this means that

$$f\left(U_{-K-j}^n\right) = f\left(U_{-K}^n\right) \quad \text{and} \quad f\left(U_{K+j}^n\right) = f\left(U_K^n\right) \quad \text{for } j > 0.$$

For nonlinear equations, explicit methods are most common. These can be written

$$U^{n+1} = G\left(U^n, \dots, U^{n-l}\right) \tag{3.3}$$

for some function G.

◇ **Example 3.1 (A nonconservative method).**

If $f(u) = u^2/2$, then we can define an explicit method

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x} U_j^n \left(U_{j+1}^n - U_j^n\right). \tag{3.4}$$

If U^0 is given by

$$U_j^0 = \begin{cases} 0 & \text{for } j \le 0, \\ 1 & \text{for } j > 0, \end{cases}$$

then $U^n = U^0$ for all n. So the method produces a nicely converging sequence, but the limit is not a solution to the original problem. The difference method (3.4) is based on a nonconservative formulation. Henceforth, we will not discuss nonconservative schemes. ◇

We call a difference method *conservative* if it can be written in the form

$$\begin{aligned} U_j^{n+1} &= G(U_{j-1-p}^n, \dots, U_{j+q}^n) \\ &= U_j^n - \lambda\left(F\left(U_{j-p}^n, \dots, U_{j+q}^n\right) - F\left(U_{j-1-p}^n, \dots, U_{j-1+q}^n\right)\right), \end{aligned} \tag{3.5}$$

where

$$\lambda = \frac{\Delta t}{\Delta x}.$$

The function F is referred to as the *numerical flux*. For brevity, we shall often use the notation

$$\begin{aligned} G(U;j) &= G\left(U_{j-1-p}, \dots, U_{j+q}\right), \\ F(U;j) &= F\left(U_{j-p}, \dots, U_{j+q}\right), \end{aligned}$$

so that (3.5) reads

$$U_j^{n+1} = G(U^n;j) = U_j^n - \lambda\left(F\left(U^n;j\right) - F\left(U^n;j-1\right)\right).$$

Conservative methods have the property that U is conserved, since

$$\sum_{j=-K}^{K} U_j^{n+1}\Delta x = \sum_{j=-K}^{K} U_j^n \Delta x - \Delta t\left(F\left(U^n;K\right) - F\left(U^n;-K-1\right)\right).$$

If we set U_j^0 equal to the average of u_0 over the jth grid cell, i.e.,

$$U_j^0 = \frac{1}{\Delta x}\int_{j\Delta x}^{(j+1)\Delta x} u_0(x)\,dx,$$

and for the moment assume that $F\left(U^n;K\right) = F\left(U^n;-K-1\right)$, then

$$\int U^n(x)\,dx = \int u_0(x)\,dx. \tag{3.6}$$

A conservative method is said to be *consistent* if

$$F(u, \dots, u) = f(u). \tag{3.7}$$

In addition we demand that F be Lipschitz continuous in all its variables.

◇ **Example 3.2 (Some conservative methods).**

The simplest conservative method is the *upwind scheme*

$$F(U;j) = f\left(U_j\right). \tag{3.8}$$

Another common method is the *Lax–Friedrichs scheme*, usually written

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{1}{2}\lambda\left(f\left(U_{j+1}^n\right) - f\left(U_{j-1}^n\right)\right). \tag{3.9}$$

In conservation form, this reads

$$F(U^n;j) = \frac{1}{2\lambda}\left(U_j^n - U_{j+1}^n\right) + \frac{1}{2}\left(f\left(U_j^n\right) + f\left(U_{j+1}^n\right)\right).$$

Also, two-step methods are used. One is the *Richtmyer two-step Lax–Wendroff scheme*:

$$F(U;j) = f\left(\frac{1}{2}\left(U^n_{j+1} + U^n_j\right) - \lambda\left(f\left(U^n_{j+1}\right) - f\left(U^n_j\right)\right)\right). \tag{3.10}$$

Another two-step method is the *MacCormack scheme*:

$$F(U;j) = \frac{1}{2}\Big(f\left(U^n_j - \lambda\left(f\left(U^n_{j+1}\right) - f\left(U^n_j\right)\right)\right) + f\left(U^n_j\right)\Big). \tag{3.11}$$

The *Godunov scheme* is a generalization of the upwind method. Let $\tilde{u}_j$ be the solution of the Riemann problem with initial data

$$\tilde{u}_j(x,0) = \begin{cases} U^n_j & \text{for } x \le 0, \\ U^n_{j+1} & \text{for } x > 0. \end{cases}$$

The numerical flux is given by

$$F(U;j) = f\left(\tilde{u}(0,\Delta t)\right). \tag{3.12}$$

To avoid that waves from neighboring grid cells start to interact before the next time step, we cannot take too long time steps Δt. Since the maximum speed is bounded by $\max|f'(u)|$, we need to enforce the requirement that

$$\lambda\,|f'(u)| < 1. \tag{3.13}$$

The condition (3.13) is called the *Courant–Friedrichs–Lewy (CFL) condition*. If all characteristic speeds are nonnegative (nonpositive), Godunov's method reduces to the upwind (downwind) method.

The Lax–Friedrichs and Godunov schemes are both of first order in the sense that the local truncation error is of order one. (We shall return to this concept below.) However, both the Lax–Wendroff and MacCormack methods are of second order. In general, higher-order methods are good for smooth solutions, but also produce solutions that oscillate in the vicinity of discontinuities. On the other hand, lower order methods have "enough diffusion" to prevent oscillations. Therefore, one often uses *hybrid methods*. These methods usually consist of a linear combination of a lower- and a higher-order method. The numerical flux is then given by

$$F(U;j) = \theta(U;j)F_L(U;j) + \left(1 - \theta(U;j)\right)F_H(U;j), \tag{3.14}$$

where F_L denotes a lower-order numerical flux, and F_H a higher-order numerical flux. The function $\theta(U;j)$ is close to zero, where U is smooth and close to one near discontinuities. Needless to say, choosing appropriate θ's is a discipline in its own right. We have implemented a method (called *fluxlim* in Figure 3.1) that is a combination of the (second-order) MacCormack method and the (first-order) Lax–Friedrichs scheme, and this scheme is compared with the "pure" methods in this figure. We, somewhat arbitrarily used

$$\theta(U;j) = 1 - \frac{1}{1 + |\Delta_{j,\Delta x}U|},$$

where $\Delta_{j,\Delta x}U$ is an approximation to the second derivative of U with respect to x,

$$\Delta_{j,\Delta x}U = \frac{U_{j+1} - 2U_j + U_{j-1}}{\Delta x^2}.$$

Another approach is to try to generalize Godunov's method by replacing the piecewise constant data U^n by a smoother function. The simplest such replacement is by a piecewise linear function. To obtain a proper generalization one should then solve a "Riemann problem" with linear initial data to the left and right. While this is difficult to do exactly, one can use approximations instead. One such approximation leads to the following method:

$$F\left(U^n;j\right) = \frac{1}{2}\left(g_j + g_{j+1}\right) - \frac{1}{2\lambda}\Delta U^n_j.$$

Here $\Delta U^n_j = U^n_{j+1} - U^n_j$, and

$$g_j = f(u_j^{n+1/2}) + \frac{1}{2\lambda}u'_j,$$

where

$$u'_j = \operatorname{MinMod}\left(\Delta U^n_{j-1}, \Delta U^n_j\right),$$
$$u_j^{n+1/2} = U^n_j - \frac{\lambda}{2}f'\left(U^n_j\right)u'_j,$$

and

$$\operatorname{MinMod}(a,b) := \frac{1}{2}\left(\operatorname{sign}(a) + \operatorname{sign}(b)\right)\min\left(|a|, |b|\right).$$

This method is labeled *slopelim* in the figures. Now we show how these methods perform on two test examples. In both examples the flux function is given by

$$f(u) = \frac{u^2}{u^2 + (1-u)^2}. \tag{3.15}$$

The example is motivated by applications in oil recovery, where one often encounters flux functions that have a shape similar to that of f; that is, $f' \ge 0$ and $f''(u) = 0$ at a single point u. The model is called the *Buckley–Leverett* equation. The first example uses initial data

$$u_0(x) = \begin{cases} 1 & \text{for } x \le 0, \\ 0 & \text{for } x > 0. \end{cases} \tag{3.16}$$

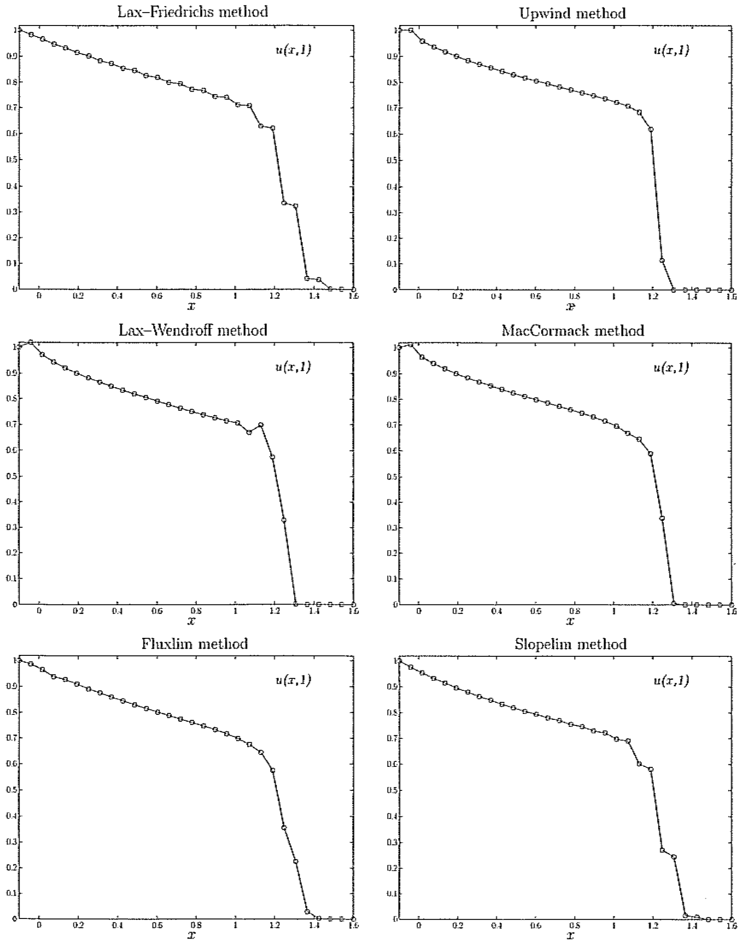

Figure 3.1. Computed solutions at time $t = 1$ for flux function (3.15) and initial data (3.16).

In Figure 3.1 we show the computed solution at time $t = 1$ for all methods, using 30 grid points in the interval $[-0.1, 1.6]$, and $\Delta x = 1.7/29$, $\Delta t = 0.5\Delta x$. The second example uses initial data

$$u_0(x) = \begin{cases} 1 & \text{for } x \in [0,1], \\ 0 & \text{otherwise,} \end{cases} \tag{3.17}$$

and 30 grid points in the interval $[-0.1, 2.6]$, $\Delta x = 2.7/29$, $\Delta t = 0.5\Delta x$. In Figure 3.2 we also show a reference solution computed by the upwind method using 500 grid points. The most notable feature of the plots in Figure 3.2 is the solutions computed by the second-order methods. We shall show that if a sequence of solutions produced by a consistent, conservative method converges, then the limit is a weak solution. The exact solution to both these problems can be calculated by the method of characteristics. ◊

The *local truncation error* of a numerical method $L_{\Delta t}$ is defined (formally) as

$$L_{\Delta t}(x) = \frac{1}{\Delta t}\left(S(\Delta t)u - S_N(\Delta t)u\right)(x), \tag{3.18}$$

where $S(t)$ is the solution operator associated with (3.1); that is, $u = S(t)u_0$ denotes the solution at time t, and $S_N(t)$ is the formal solution operator associated with the numerical method, i.e.,

$$S_N(\Delta t)u(x) = u(x) - \lambda\left(F(u;j) - F(u;j-1)\right).$$

To make matters more concrete, assume that we are studying the upwind method. Then

$$S_N(\Delta t)u(x) = u(x) - \frac{\Delta t}{\Delta x}\left(f(u(x)) - f(u(x-\Delta x))\right).$$

We say that the method is of kth order if for all smooth solutions $u(x,t)$,

$$|L_{\Delta t}(x)| = \mathcal{O}\left(\Delta t^k\right)$$

as $\Delta t \to 0$. That a method is of high order, $k \geq 2$, usually implies that it is "good" for computing smooth solutions.

◊ **Example 3.3 (Local truncation error).**

We verify that the upwind method is of first order:

$$\begin{aligned} L_{\Delta t}(x) &= \frac{1}{\Delta t}\Big(u(x, t+\Delta t) - u(x) + \frac{\Delta t}{\Delta x}\left(f(u(x)) - f(u(x-\Delta x))\right)\Big) \\ &= \frac{1}{\Delta t}\Big(u + \Delta t\, u_t + \frac{(\Delta t)^2}{2}u_{tt} + \cdots - u \\ &\qquad - \lambda\Big(f'(u)\Big(-u_x\Delta x + \frac{(\Delta x)^2}{2}u_{xx} + \cdots\Big) \\ &\qquad + f''(u)\frac{1}{2}\left(-u_x\Delta x + \cdots\right)^2\Big)\Big) \\ &= \frac{1}{\Delta t}\Big(\Delta t\left(u_t + f(u)_x\right) + \frac{(\Delta t)^2}{2}u_{tt} \\ &\qquad - \frac{\Delta t\Delta x}{2}\left(u_{xx}f'(u) + f''(u)u_x^2\right) + \cdots\Big) \end{aligned}$$

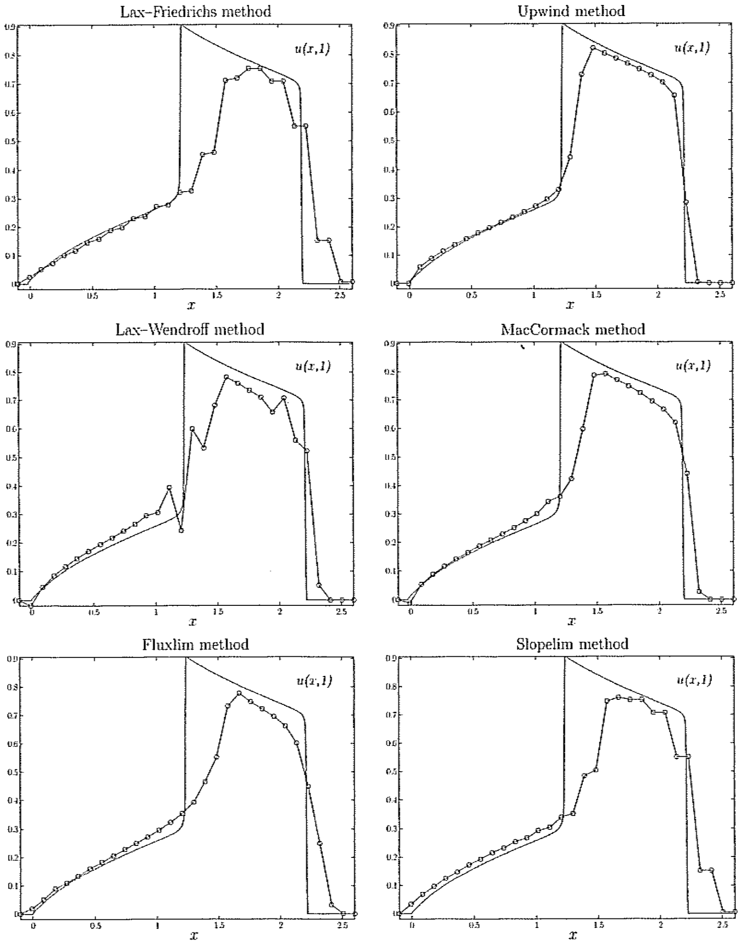

Figure 3.2. Computed solutions at time $t = 1$ for flux function (3.15) and initial data (3.17).

$$= u_t + f(u)_x + \frac{1}{2}\left(\Delta t\, u_{tt} - \Delta x \left(f'(u)u_x\right)_x\right) + \mathcal{O}\left((\Delta t)^2\right)$$

$$= u_t + f(u)_x + \frac{\Delta x}{2}\left(\lambda\, u_{tt} - \left(f'(u)u_x\right)_x\right) + \mathcal{O}\left((\Delta t)^2\right).$$

Assuming that u is a smooth solution of (3.1), we find that

$$u_{tt} = \left((f'(u))^2 u_x\right)_x,$$

and inserting this into the previous equation we obtain

$$L_{\Delta t} = \frac{\Delta t}{2\lambda}\frac{\partial}{\partial x}\left(f'(u)\left(\lambda f'(u) - 1\right)u_x\right) + \mathcal{O}\left((\Delta t)^2\right). \tag{3.19}$$

Hence, the upwind method is of first order. The above computations were purely formal, assuming sufficient smoothness for the Taylor expansion to be valid. This means that Godunov's scheme is also of first order. Similarly, computations based on the Lax–Friedrichs scheme yield

$$L_{\Delta t} = \frac{\Delta t}{2\lambda^2}\frac{\partial}{\partial x}\left(\left((\lambda f'(u))^2 - 1\right)u_x\right) + \mathcal{O}\left(\Delta t^2\right). \tag{3.20}$$

Consequently, the Lax–Friedrichs scheme is also of first order. From the above computations it also emerges that the Lax–Friedrichs scheme is *second-order* accurate on the equation

$$u_t + f(u)_x = \frac{\Delta t}{2\lambda^2}\left(\left(1 - (\lambda f'(u))^2\right)u_x\right)_x. \tag{3.21}$$

This is called the *model equation* for the Lax–Friedrichs scheme. In order for this to be well posed we must have that the coefficient of u_{xx} on the right-hand side is nonnegative. Hence

$$|\lambda f'(u)| \leq 1. \tag{3.22}$$

This is a stability restriction on λ, and is the Courant–Friedrichs–Lewy (CFL) condition that we encountered in (3.13). The model equation for the upwind method is

$$u_t + f(u)_x = \frac{\Delta t}{2\lambda}\left(f'(u)\left(1 - \lambda f'(u)\right)u_x\right)_x. \tag{3.23}$$

In order for this equation to be well posed, we must have $f'(u) \geq 0$ and $\lambda f'(u) < 1$. ◇

From the above examples, we see that first-order methods have model equations with a diffusive term. Similarly, one finds that second-order methods have model equations with a dispersive right-hand side. Therefore, the oscillations observed in the computations were to be expected.

From now on we let the function $u_{\Delta t}$ be defined by

$$u_{\Delta t}(x,t) = U_j^n, \qquad (x,t) \in [j\Delta x, (j+1)\Delta x) \times [n\Delta t, (n+1)\Delta t). \tag{3.24}$$

Observe that

$$\int_{\mathbb{R}} u_{\Delta t}(x,t)\,dx = \Delta x \sum_j U_j^n, \quad \text{for } n\Delta t \leq t < (n+1)\Delta t.$$

We briefly mentioned in Example 3.2 the fact that if $u_{\Delta t}$ converges, then the limit is a weak solution. Precisely, we have the well-known Lax–Wendroff theorem.

Theorem 3.4 (Lax–Wendroff theorem). *Let $u_{\Delta t}$ be computed from a conservative and consistent method. Assume that* T.V.$_x\left(u_{\Delta t}\right)$ *is uniformly* *bounded in Δt. Consider a subsequence $u_{\Delta t_k}$ such that $\Delta t_k \to 0$, and assume that $u_{\Delta t_k}$ converges in L^1_{loc} as $\Delta t_k \to 0$. Then the limit is a weak solution to* (3.1).

Proof. The proof uses summation by parts. Let $\varphi(x,t)$ be a test function. By the definition of U_j^{n+1},

$$\begin{aligned}\sum_{n=0}^{N}\sum_{j=-\infty}^{\infty} &\varphi(x_j,t_n)\left(U_j^{n+1}-U_j^n\right)\\ &= -\frac{\Delta t}{\Delta x}\sum_{n=0}^{N}\sum_{j=-\infty}^{\infty}\varphi(x_j,t_n)\left(F(U^n;j)-F(U^n;j-1)\right),\end{aligned}$$

where $x_j = j\Delta x$ and $t_n = n\Delta t$, and we choose $T = N\Delta t$ such that $\varphi = 0$ for $t \geq T$. After a summation by parts we get

$$\begin{aligned}-\sum_{j=-\infty}^{\infty}\varphi(x_j,0)U_j^0 &- \sum_{j=-\infty}^{\infty}\sum_{n=1}^{N}\left(\varphi(x_j,t_n)-\varphi(x_j,t_{n-1})\right)U_j^n\\ &-\frac{\Delta t}{\Delta x}\sum_{n=0}^{N}\sum_{j=-\infty}^{\infty}\left(\varphi(x_{j+1},t_n)-\varphi(x_j,t_n)\right)F(U^n;j) = 0.\end{aligned}$$

Rearranging, we find that

$$\begin{aligned}\Delta t\Delta x\sum_{n=1}^{N}\sum_{j=-\infty}^{\infty}&\left[\left(\frac{\varphi(x_j,t_n)-\varphi(x_j,t_{n-1})}{\Delta t}\right)U_j^n\right.\\ &\left.+\left(\frac{\varphi(x_{j+1},t_n)-\varphi(x_j,t_n)}{\Delta x}\right)F(U^n;j)\right]\\ &= -\Delta x\sum_{j=-\infty}^{\infty}\varphi(x_j,0)\,U_j^0. \qquad (3.25)\end{aligned}$$

This almost looks like a Riemann sum for the weak formulation of (3.1), were it not for F. To conclude that the limit is a weak solution we must show that

$$\Delta t\Delta x\sum_{n=1}^{N}\sum_{j=-\infty}^{\infty}\left|F(U^n;j)-f(U_j^n)\right| \qquad (3.26)$$

tends to zero as $\Delta t \to 0$. Using consistency, we find that (3.26) equals

$$\Delta t\Delta x\sum_{n=1}^{N}\sum_{j=-\infty}^{\infty}\left|F\left(U_{j-p}^n,\dots,U_{j+q}^n\right)-F\left(U_j^n,\dots,U_j^n\right)\right|,$$

which by the Lipschitz continuity of F is less than

$$\begin{aligned}\Delta t\Delta xM&\sum_{n=1}^{N}\sum_{j=-\infty}^{\infty}\sum_{k=-p}^{q}\left|U_{j+k}^n-U_j^n\right|\\ &\leq\frac{1}{2}(q(q-1)+p(p-1))\Delta t\,\Delta x\,M\sum_{n=1}^{N}\sum_{j=-\infty}^{\infty}\left|U_{j+1}^n-U_j^n\right|\\ &\leq (q^2+p^2)\Delta x\,M\,\mathrm{T.V.}\,(u_{\Delta t})\;T,\end{aligned}$$

where M is larger than the Lipschitz constant of F. Therefore, (3.26) is small for small Δx, and the limit is a weak solution. □

We proved in Theorem 2.14 that the solution of a scalar conservation law in one dimension possesses several properties. The corresponding properties for conservative and consistent numerical schemes read as follows:

Definition 3.5. *Let $u_{\Delta t}$ be computed from a conservative and consistent method.*

- *A method is said to be total variation stable*[1] *if the total variation of U^n is uniformly bounded, independently of Δx and Δt.*
- *We say that a numerical method is total variation diminishing (TVD) if* $\mathrm{T.V.}\left(U^{n+1}\right) \leq \mathrm{T.V.}\left(U^n\right)$ *for all $n \in \mathbb{N}_0$.*
- *A method is called monotonicity preserving if the initial data is monotone implies that so is U^n for all $n \in \mathbb{N}$.*
- *A numerical method is called L^1-contractive if it is L^1-contractive [sic!], i.e., $\left\|u_{\Delta t}(t)-v_{\Delta t}(t)\right\|_1 \leq \left\|u_{\Delta t}(0)-v_{\Delta t}(0)\right\|_1$ for all $t \geq 0$. Here $v_{\Delta t}$ is another solution with initial data v_0. Alternatively, we can of course write this as*
  $$\sum_j\left|U_j^{n+1}-V_j^{n+1}\right| \leq \sum_j\left|U_j^n-V_j^n\right|, \quad n \in \mathbb{N}_0.$$
- *A method is said to be monotone if for initial data U^0 and V^0, we have*
  $$U_j^0 \leq V_j^0, \quad j \in \mathbb{Z} \quad \Rightarrow \quad U_j^n \leq V_j^n, \quad j \in \mathbb{Z},\ n \in \mathbb{N}.$$

The above notions are strongly interrelated, as the next theorem shows.

Theorem 3.6. *For conservative and consistent methods the following hold:*

(i) *Any monotone method is L^1-contractive, assuming $u_{\Delta t}(0)-v_{\Delta t}(0) \in L^1(\mathbb{R})$.*

(ii) *Any L^1-contractive method is TVD, assuming that* $\mathrm{T.V.}\,(u_0)$ *is finite.*

(iii) *Any TVD method is monotonicity preserving.*

[1] This definition is slightly different from the standard definition of T.V. stable methods.

Proof. (i) We apply the Crandall–Tartar lemma, Lemma 2.12, with $\Omega = \mathbb{R}$, and D equal to the set of all functions in L^1 that are piecewise constant on the grid $\Delta x\mathbb{Z}$, and finally we let $T(U^0) = U^n$. Since the method is conservative (cf. (3.6)), we have that

$$\sum_j U_j^n = \sum_j U_j^0, \text{ or } \int T(U^0) = \int U^n = \int U^0.$$

Lemma 2.12 immediately implies that

$$\begin{aligned}\|u_{\Delta t} - v_{\Delta t}\|_1 &= \Delta x \sum_j \left|U_j^n - V_j^n\right| \le \Delta x \sum_j \left|U_j^0 - V_j^0\right| \\ &= \|u_{\Delta t}(0) - v_{\Delta t}(0)\|_1.\end{aligned}$$

(ii) Assume now that the method is L^1-contractive, i.e.,

$$\sum_j \left|U_j^{n+1} - V_j^{n+1}\right| \le \sum_j \left|U_j^n - V_j^n\right|.$$

Let V^n be the numerical solution with initial data

$$V_i^0 = U_{i+1}^0.$$

Then by the translation invariance induced by (3.5), $V_i^n = U_{i+1}^n$ for all n. Furthermore,

$$\begin{aligned}\text{T.V.}\left(U^{n+1}\right) &= \sum_{j=-\infty}^{\infty} \left|U_{j+1}^{n+1} - U_j^{n+1}\right| = \sum_j \left|U_j^{n+1} - V_j^{n+1}\right| \\ &\le \sum_j \left|U_j^n - V_j^n\right| = \text{T.V.}\left(U^n\right).\end{aligned}$$

(iii) Consider now a TVD method, and assume that we have monotone initial data. Since $\text{T.V.}\left(U^0\right)$ is finite, the limits

$$U_L = \lim_{j \to -\infty} U_j^0 \text{ and } U_R = \lim_{j \to \infty} U_j^0$$

exist. Then $\text{T.V.}\left(U^0\right) = |U_R - U_L|$. If U^1 were not monotone, then $\text{T.V.}\left(U^1\right) > |U_R - U_L| = \text{T.V.}\left(U^0\right)$, which is a contradiction. □

We can summarize the above theorem as follows:

monotone ⇒ L^1-contractive ⇒ TVD ⇒ monotonicity preserving.

Monotonicity is relatively easy to check for explicit methods, e.g., by calculating the partial derivatives $\partial G / \partial U^i$ in (3.3).

◇ **Example 3.7 (Lax–Friedrichs scheme).**

Recall from Example 3.2 that the Lax–Friedrichs scheme is given by

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{1}{2}\lambda\left(f\left(U_{j+1}^n\right) - f\left(U_{j-1}^n\right)\right).$$

Computing partial derivatives we obtain

$$\frac{\partial U_j^{n+1}}{\partial U_k^n} = \begin{cases} (1 - \lambda f'(U_k^n))/2 & \text{for } k = j+1, \\ (1 + \lambda f'(U_k^n))/2 & \text{for } k = j-1, \\ 0 & \text{otherwise,} \end{cases}$$

and hence we see that the Lax–Friedrichs scheme is monotone as long as the CFL condition

$$\lambda \left|f'(u)\right| < 1$$

is fulfilled. ◇

Theorem 3.8. *Let $u_0 \in L^1(\mathbb{R})$ have bounded variation. Assume that $u_{\Delta t}$ is computed with a method that is conservative, consistent, total variation stable, and uniformly bounded; that is,*

$$\text{T.V.}\left(u_{\Delta t}\right) \le M \text{ and } \left\|u_{\Delta t}\right\|_\infty \le M,$$

where M is independent of Δx and Δt.

Let $T > 0$. Then $\{u_{\Delta t}(t)\}$ has a subsequence that converges for all $t \in [0,T]$ to a weak solution $u(t)$ in $L^1_{\text{loc}}(\mathbb{R})$. Furthermore, the limit is in $C\left([0,T]; L^1_{\text{loc}}(\mathbb{R})\right)$.

Proof. We intend to apply Theorem A.8. It remains to show that

$$\int_a^b \left|u_{\Delta t}(x,t) - u_{\Delta t}(x,s)\right| \, dx \le C\left|t-s\right| + o(1), \text{ as } \Delta t \to 0, \quad s,t \in [0,T].$$

Consistency of the scheme implies, for any fixed Δt,

$$\begin{aligned}\left|U_j^{n+1} - U_j^n\right| &= \lambda \left|F(U_j^n; j) - F(U_j^n; j-1)\right| \\ &= \lambda \left|F(U_{j-p}^n, \dots, U_{j+q}^n) - F(U_{j-p-1}^n, \dots, U_{j+q-1}^n)\right| \\ &\le \lambda L \left(\left|U_{j-p}^n - U_{j-p-1}^n\right| + \dots + \left|U_{j+q}^n - U_{j+q-1}^n\right|\right),\end{aligned}$$

from which we conclude that

$$\begin{aligned}\left\|u_{\Delta t}(\,\cdot\,, t_{n+1}) - u_{\Delta t}(\,\cdot\,, t_n)\right\|_1 &= \sum_{j=-\infty}^{\infty} \left|U_j^{n+1} - U_j^n\right| \Delta x \\ &\le L(p+q+1)\text{T.V.}\left(U^n\right) \Delta t \\ &\le L(p+q+1) M \Delta t,\end{aligned}$$

where L is the Lipschitz constant of F. More generally,

$$\left\|u_{\Delta t}(\,\cdot\,, t_m) - u_{\Delta t}(\,\cdot\,, t_n)\right\|_1 \le L(p+q+1)M\left|n-m\right|\Delta t.$$

Now let $\tau_1, \tau_2 \in [0,T]$, and choose $\tilde{t}_1, \tilde{t}_2 \in \{n\Delta t \mid 0 \le n \le T/\Delta t\}$ such that

$$0 \le \tau_j - \tilde{t}_j < \Delta t \text{ for } j = 1, 2.$$

By construction $u_{\Delta t}(\tau_j) = u_{\Delta t}(\tilde{t}_j)$, and hence

$$\begin{aligned}
&\left\|u_{\Delta t}(\,\cdot\,,\tau_1) - u_{\Delta t}(\,\cdot\,,\tau_2)\right\|_1 \\
&\quad\le \left\|u_{\Delta t}(\,\cdot\,,\tau_1) - u_{\Delta t}(\,\cdot\,,\tilde{t}_1)\right\|_1 + \left\|u_{\Delta t}(\,\cdot\,,\tilde{t}_1) - u_{\Delta t}(\,\cdot\,,\tilde{t}_2)\right\|_1 \\
&\qquad + \left\|u_{\Delta t}(\,\cdot\,,\tilde{t}_2) - u_{\Delta t}(\,\cdot\,,\tau_2)\right\|_1 \\
&\quad\le (p+q+1)L\,M\left|\tilde{t}_1 - \tilde{t}_2\right| \le (p+q+1)L\,M\left|\tau_1 - \tau_2\right| + \mathcal{O}\left(\Delta t\right).
\end{aligned}$$

Observe that this estimate is uniform in $\tau_1, \tau_2 \in [0,T]$. We conclude that

$$u_{\Delta t} \to u \text{ in } C([0,T]; L^1([a,b]))$$

for a sequence $\Delta t \to 0$. The Lax–Wendroff theorem then says that this limit is a weak solution. □

At this point it is convenient to introduce the concept of *entropy pairs* or *entropy/entropy flux pairs.*[2] Recall that a pair of functions $(\eta(u), q(u))$ with η convex is called an entropy pair if

$$q'(u) = f'(u)\eta'(u). \tag{3.27}$$

The reason for introducing this concept is that the entropy condition can now be reformulated using (η, q). To see this, assume that u is a solution of the viscous conservation law

$$u_t + f(u)_x = \varepsilon u_{xx}. \tag{3.28}$$

Assume, or consult Appendix B, that this equation has a unique twice-differentiable solution. Hence, multiplying by $\eta'(u)$ yields (cf. (2.10))

$$\eta(u)_t + q(u)_x = \varepsilon\eta'(u)u_{xx} = \varepsilon\left(\eta'(u)u_x\right)_x - \varepsilon\eta''(u)\left(u_x\right)^2.$$

If η' is bounded, and $\eta'' > 0$, then the first term on the right of the above equation tends to zero as a distribution as $\varepsilon \to 0$, while the second term is nonpositive. Consequently, if the solution of (3.1) is to be the limit of the solutions of (3.28) as $\varepsilon \to 0$, the solution of (3.1) must satisfy (cf. (2.12))

$$\eta(u)_t + q(u)_x \le 0 \tag{3.29}$$

as a distribution. Choosing $\eta(u) = |u-k|$ we recover the Kružkov entropy condition; see (2.17). We have demonstrated that that if a function satisfies (2.46) for all k, then it satisfies (3.29) for all convex η and vice versa; see Remark 2.1. Hence, the Kružkov entropy condition is equivalent to demanding (3.29) for all convex η.

The analogue of an entropy pair for difference schemes reads as follows. Write

$$a \vee b = \max(a,b) \quad \text{and} \quad a \wedge b = \min(a,b),$$

[2] We have already encountered an entropy pair with $\eta(u) = |u-k|$ and $q(u) = \operatorname{sign}(u-k)\left(f(u) - f(k)\right)$ when we introduced the Kružkov entropy condition in Chapter 2.

and observe the trivial identity

$$|a-b| = a \vee b - a \wedge b.$$

Then we define the *numerical entropy flux* Q by

$$Q(U;j) = F(U \vee k; j) - F(U \wedge k; j),$$

or explicitly,

$$\begin{aligned}
&Q\left(U_{j-p}, \dots, U_{j+p'}\right) \\
&\qquad = F\left(U_{j-p} \vee k, \dots, U_{j+p'} \vee k\right) - F\left(U_{j-p} \wedge k, \dots, U_{j+p'} \wedge k\right).
\end{aligned}$$

We have that Q is consistent with the usual entropy flux, i.e.,

$$Q\left(u, \dots, u\right) = \operatorname{sign}\left(u-k\right)\left(f(u) - f(k)\right).$$

Returning to monotone difference schemes, we have the following result.

Theorem 3.9. *Under the assumptions of Theorem 3.8, the approximate solutions computed by a conservative, consistent, and monotone difference method converge to the entropy solution as* $\Delta t \to 0$.

Proof. Theorem 3.8 allows us to conclude that $u_{\Delta t}$ has a subsequence that converges in $C([0,T]; L^1([a,b]))$ to a weak solution. It remains to show that the limit satisfies a discrete Kružkov form. By a direct calculation we find that

$$\left|U_j^n - k\right| - \lambda\left(Q\left(U^n; j\right) - Q\left(U^n; j-1\right)\right) = G(U^n \vee k; j) - G(U^n \wedge k; j).$$

Using that $U_j^{n+1} = G(U^n; j)$ and that $k = G(k; j)$, the monotonicity of the scheme implies that

$$\begin{aligned}
G(U^n \vee k; j) &\ge G(U^n; j) \vee G(k; j) = G(U^n; j) \vee k, \\
-G(U^n \wedge k; j) &\ge -G(U^n; j) \wedge G(k; j) = -G(U^n; j) \wedge k.
\end{aligned}$$

Therefore,

$$\left|U_j^{n+1} - k\right| - \left|U_j^n - k\right| + \lambda\left(Q\left(U^n; j\right) - Q\left(U^n; j-1\right)\right) \le 0. \tag{3.30}$$

Applying the technique used in proving the Lax–Wendroff theorem to (3.30) gives that the limit u satisfies

$$\iint \left(|u-k|\,\varphi_t + \operatorname{sign}\left(u-k\right)\left(f(u) - f(k)\right)\varphi_x\right)\,dx\,dt \ge 0.$$

□

Note that we can also use the above theorem to conclude the existence of weak entropy solutions to scalar conservation laws.

Now we shall examine the local truncation error of a general conservative, consistent, and monotone method. Since this can be written

$$\begin{aligned}
U_j^{n+1} &= G\left(U^n; j\right) = G\left(U_{j-p-1}^n, \dots, U_{j+q}^n\right) \\
&= U_j^n - \lambda\left(F\left(U_{j+q}^n, \dots, U_{j-p}^n\right) - F\left(U_{j-p-1}^n, \dots, U_{j+q-1}^n\right)\right),
\end{aligned}$$

we write

$$G = G(u_1, \dots, u_{p+q+1}) \quad \text{and} \quad F = F(u_1, \dots, u_{p+q}).$$

We assume that F, and hence G, is three times continuously differentiable with respect to all arguments, and write the derivatives with respect to the ith argument as

$$\partial_i G(u_1, \dots, u_{p+q+1}) \quad \text{and} \quad \partial_i F(u_1, \dots, u_{p+q}).$$

We set $\partial_i F = 0$ if $i = 0$ or $i = p + q + 1$. Throughout this calculation, we assume that the jth slot of G contains U_j^n, so that $G(u_1, \dots, u_{p+q+1}) = u_j - \lambda(\dots)$. By consistency we have that

$$G(u, \dots, u) = u \quad \text{and} \quad F(u, \dots, u) = f(u).$$

Using this we find that

$$\sum_{i=1}^{p+q} \partial_i F(u, \dots, u) = f'(u), \tag{3.31}$$

$$\partial_i G = \delta_{i,j} - \lambda \left(\partial_{i-1} F - \partial_i F\right), \tag{3.32}$$

and

$$\partial_{i,k}^2 G = -\lambda \left(\partial_{i-1,k-1}^2 F - \partial_{i,k}^2 F\right). \tag{3.33}$$

Therefore,

$$\sum_{i=1}^{p+q+1} \partial_i G(u, \dots, u) = \sum_{i=1}^{p+q+1} \delta_{i,j} = 1. \tag{3.34}$$

Furthermore,

$$\begin{aligned}
\sum_{i=1}^{p+q+1} (i-j)\partial_i G(u, \dots, u) &= \sum_{i=1}^{p+q+1} (i-j)\delta_{i,j} \\
&\quad - \lambda(i-j)\left(\partial_{i-1} F(u, \dots, u) - \partial_i F(u, \dots, u)\right) \\
&= -\lambda \sum_{i=1}^{p+q} ((i+1) - i))\partial_i F(u, \dots, u) \\
&= -\lambda f'(u).
\end{aligned} \tag{3.35}$$

We also find that

$$\begin{aligned}
&\sum_{i,k=1}^{p+q+1} (i-k)^2 \partial_{i,k}^2 G(u, \dots, u) \\
&\qquad = -\lambda \sum_{i,k=1}^{p+q+1} (i-k)^2 \left(\partial_{i-1,k-1}^2 F(u, \dots, u) - \partial_{i,k}^2 F(u, \dots, u)\right)
\end{aligned}$$

$$\begin{aligned}
&= -\lambda \sum_{i,k=1}^{p+q} \left(((i+1) - (k+1))^2 - (i-k)^2\right) \partial_{i,k}^2 F(u, \dots, u) \\
&= 0.
\end{aligned} \tag{3.36}$$

Having established this, we now let $u = u(x,t)$ be a smooth solution of the conservation law (3.1). We are interested in applying G to $u(x,t)$, i.e., to calculate

$$G(u(x - p\Delta x, t) \dots, u(x,t), \dots, u(x + q\Delta x, t)).$$

Set $u_i = u(x + (i-j)\Delta x, t)$ for $i = 1, \dots, p + q + 1$. Then we find that

$$\begin{aligned}
&G(u_1, \dots, u_{p+q+1}) \\
&\quad = G(u_j, \dots, u_j) + \sum_{i=1}^{p+q+1} \partial_i G(u_j, \dots, u_j)\,(u_i - u_j) \\
&\qquad + \frac{1}{2} \sum_{i,k=1}^{p+q+1} \partial_{i,k}^2 G(u_j, \dots, u_j)\,(u_i - u_j)\,(u_k - u_j) + \mathcal{O}\left(\Delta x^3\right) \\
&\quad = u(x,t) + u_x(x,t)\Delta x \sum_{i=1}^{p+q+1} (i-j)\partial_i G(u_j, \dots, u_j) \\
&\qquad + \frac{1}{2} u_{xx}(x,t)\Delta x^2 \sum_{i=1}^{p+q+1} (i-j)^2 \partial_i G(u_j, \dots, u_j) \\
&\qquad + \frac{1}{2} u_x^2(x,t)\Delta x^2 \sum_{i,k=1}^{p+q+1} (i-j)(k-j)\partial_{i,k}^2 G(u_j, \dots, u_j) + \mathcal{O}\left(\Delta x^3\right) \\
&\quad = u(x,t) + u_x(x,t)\Delta x \sum_{i=1}^{p+q+1} (i-j)\partial_i G(u_j, \dots, u_j) \\
&\qquad + \frac{1}{2}\Delta x^2 \sum_{i=1}^{p+q+1} (i-j)^2 \left[\partial_i G(u_j, \dots, u_j) u_x(x,t)\right]_x \\
&\qquad - \frac{1}{2}\Delta x^2 u_x^2(x,t) \sum_{i,k}^{p+q+1} \left((i-j)^2 - (i-j)(k-j)\right) \partial_{i,k}^2 G(u_j, \dots, u_j) \\
&\qquad + \mathcal{O}\left(\Delta x^3\right).
\end{aligned}$$

Next we observe, since $\partial_{i,k}^2 G = \partial_{k,i}^2 G$ and using (3.36), that

$$
\begin{aligned}
0 = \sum_{i,k}(i-k)^2\partial^2_{i,k}G &= \sum_{i,k}((i-j)-(k-j))^2\partial^2_{i,k}G \\
&= \sum_{i,k}((i-j)^2 - 2(i-j)(k-j))\partial^2_{i,k}G + \sum_{i,k}(k-j)^2\partial^2_{k,i}G \\
&= 2\sum_{i,k}((i-j)^2 - (i-j)(k-j))\partial^2_{i,k}G.
\end{aligned}
$$

Consequently, the penultimate term in the Taylor expansion of G above is zero, and we have that

$$
\begin{aligned}
G(u(x-p\Delta x,t),\dots,u(x+q\Delta x,t)) &= u(x,t) - \Delta t f(u(x,t))_x \\
&+ \frac{\Delta x^2}{2}\sum_i (i-j)^2\left[\partial_i G(u(x,t),\dots,u(x,t))u_x\right]_x + \mathcal{O}\left(\Delta x^3\right). \qquad (3.37)
\end{aligned}
$$

Since u is a smooth solution of (3.1), we have already established that

$$
u(x,t+\Delta t) = u(x,t) - \Delta t f(u)_x + \frac{\Delta t^2}{2}\left[\partial\left(f'(u)\right)^2 u_x\right]_x + \mathcal{O}\left(\Delta t^3\right).
$$

Hence, we compute the local truncation error as

$$
\begin{aligned}
L_{\Delta t} &= \frac{\Delta t}{2\lambda^2}\left[\left(\sum_{i=1}^{p+q+1}(i-j)^2\partial_i G(u,\dots,u) - \lambda^2(f'(u))^2\right)u_x\right]_x \\
&=: \frac{\Delta t}{2\lambda^2}\left[\beta(u)u_x\right]_x + \mathcal{O}\left(\Delta t^2\right). \qquad (3.38)
\end{aligned}
$$

Thus if $\beta > 0$, then the method is of first order. What we have done so far is valid for any conservative and consistent method where the numerical flux function is three times continuously differentiable. Next, we utilize that $\partial_i G \geq 0$, so that $\sqrt{\partial_i G}$ is well-defined. This means that

$$
\begin{aligned}
-\lambda f'(u) &= \sum_{i=1}^{p+q+1}(i-j)\partial_i G(u,\dots,u) \\
&= \sum_{i=1}^{p+q+1}(i-j)\sqrt{\partial_i G(u,\dots,u)}\sqrt{\partial_i G(u,\dots,u)}.
\end{aligned}
$$

Using the Cauchy–Schwarz inequality and (3.34) we find that

$$
\begin{aligned}
\lambda^2\left(f'(u)\right)^2 &\leq \sum_{i=1}^{p+q+1}(i-j)^2\partial_i G(u,\dots,u)\sum_{i=1}^{p+q+1}\partial_i G(u,\dots,u) \\
&= \sum_{i=1}^{p+q+1}(i-j)^2\partial_i G(u,\dots,u).
\end{aligned}
$$

Thus, $\beta(u) \geq 0$. Furthermore, the inequality is strict if more than one term in the right-hand sum is different from zero. If $\partial_i G(u,\dots,u) = 0$ except for $i = k$ for some k, then $G(u_1,\dots,u_{p+q+1}) = u_k$ by (3.34). Hence the scheme is a linear translation, and by consistency $f(u) = cu$, where $c = (j-k)\lambda$. Therefore, monotone methods for nonlinear conservation laws are at most first-order accurate. This is indeed their main drawback. To recapitulate, we have proved the following theorem:

Theorem 3.10. *Assume that the numerical flux F is three times continuously differentiable, and that the corresponding scheme is monotone. Then the method is at most first-order accurate.*

3.2 Error Estimates

> Let others bring order to chaos.
> I would bring chaos to order instead.
>
> *Kurt Vonnegut, Breakfast of Champions (1973)*

The concept of local error estimates is based on formal computations, and indicates how the method performs in regions where the solution is smooth. Since the convergence of the methods discussed was in L^1, it is reasonable to ask how far the approximated solution is from the true solution in this space.

In this section we will consider functions u that are maps $t \mapsto u(t)$ from $[0,\infty\rangle$ to $L^1_{\text{loc}}\cap BV(\mathbb{R})$ such that the one-sided limits $u(t\pm)$ exist in L^1_{loc}, and for definiteness we assume that this map is right continuous. Furthermore, we assume that

$$
\left\|u(t)\right\|_\infty \leq \left\|u(0)\right\|_\infty, \quad \text{T.V.}\left(u(t)\right) \leq \text{T.V.}\left(u(0)\right).
$$

We denote this class of functions by $\mathcal{K}$. From Theorem 2.14 we know that solutions of scalar conservation laws are in the class $\mathcal{K}$.

It is convenient to introduce *moduli of continuity in time* (see Appendix A)

$$
\begin{aligned}
\nu_t(u,\sigma) &= \sup_{|\tau|\leq\sigma}\left\|u(t+\tau)-u(t)\right\|_1, \quad \sigma > 0, \\
\nu(u,\sigma) &= \sup_{0\leq t\leq T}\nu_t(u,\sigma).
\end{aligned}
$$

From Theorem 2.14 we have that

$$
\nu(u,\sigma) \leq |\sigma|\left\|f\right\|_{\text{Lip}}\text{T.V.}\left(u_0\right) \qquad (3.39)
$$

for weak solutions of conservation laws.

Now let $u(x,t)$ be any function in $\mathcal{K}$, not necessarily a solution of (3.1). In order to measure how far u is from being a solution of (3.1) we insert u

in the Kružkov form (cf. (2.19))

$$\Lambda_T(u,\phi,k) = \int_0^T \int \left(|u-k|\,\phi_t + q(u,k)\phi_x\right) dx\,ds \tag{3.40}$$
$$- \int |u(x,T)-k|\,\phi(x,T)\,dx + \int |u_0(x)-k|\,\phi(x,0)\,dx.$$

If u is a solution, then $\Lambda_T \geq 0$ for all constants k and all nonnegative test functions ϕ. We shall now use the special test function

$$\Omega(x,x',s,s') = \omega_{\varepsilon_0}(s-s')\omega_\varepsilon(x-x'),$$

where

$$\omega_\varepsilon(x) = \frac{1}{\varepsilon}\omega\left(\frac{x}{\varepsilon}\right)$$

and $\omega(x)$ is an even C^∞ function satisfying

$$0 \leq \omega \leq 1, \quad \omega(x) = 0 \quad \text{for } |x| > 1, \quad \int \omega(x)\,dx = 1.$$

Let $v(x',s')$ be the unique weak solution of (3.1), and define

$$\Lambda_{\varepsilon,\varepsilon_0}(u,v) = \int_0^T \int \Lambda_T\left(u, \Omega(\,\cdot\,,x',\,\cdot\,,s'), v(x',s')\right) dx'ds'.$$

The comparison result reads as follows.

Theorem 3.11 (Kuznetsov's lemma). *Let $u(\,\cdot\,,t)$ be a function in $\mathcal{K}$, and v be a solution of* (3.1). *If $0 < \varepsilon_0 < T$ and $\varepsilon > 0$, then*

$$\left\|u(\,\cdot\,,T-) - v(\,\cdot\,,T)\right\|_1 \leq \|u_0 - v_0\|_1 + \text{T.V.}\,(v_0)\,(2\varepsilon + \varepsilon_0\|f\|_{\text{Lip}})$$
$$+ \nu(u,\varepsilon_0) - \Lambda_{\varepsilon,\varepsilon_0}(u,v), \tag{3.41}$$

where $u_0 = u(\,\cdot\,,0)$ and $v_0 = v(\,\cdot\,,0)$.

Proof. We use special properties of the test function Ω, namely that

$$\Omega(x,x',s,s') = \Omega(x',x,s,s') = \Omega(x,x',s',s) = \Omega(x',x,s',s) \tag{3.42}$$

and

$$\Omega_x = -\Omega_{x'}, \quad \text{and} \quad \Omega_s = -\Omega_{s'}. \tag{3.43}$$

Using (3.42) and (3.43), we find that

$$\Lambda_{\varepsilon,\varepsilon_0}(u,v) = -\Lambda_{\varepsilon,\varepsilon_0}(v,u) - \int_0^T \iint \Omega(x,x',s,T)\big(\,|u(x,T) - v(x',s)|$$
$$+ |v(x',T) - u(x,s)|\,\big)\,dx\,dx'\,ds$$
$$+ \int_0^T \iint \Omega(x,x',s,0)\big(\,|v_0(x') - u(x,s)|$$
$$+ |u_0(x) - v(x',s)|\,\big)\,dx\,dx'\,ds$$

$$:= -\Lambda_{\varepsilon,\varepsilon_0}(v,u) - A + B.$$

Since v is a weak solution, $\Lambda_{\varepsilon,\varepsilon_0}(v,u) \geq 0$, and hence

$$A \leq B - \Lambda_{\varepsilon,\varepsilon_0}(u,v).$$

Therefore, we would like to obtain a lower bound on A and an upper bound on B, the lower bound on A involving $\|u(T) - v(T)\|_1$ and the upper bound on B involving $\|u_0 - v_0\|_1$. We start with the lower bound on A.

Let ρ_ε be defined by

$$\rho_\varepsilon(u,v) = \iint \omega_\varepsilon(x-x')\,|u(x) - v(x')|\,dx\,dx'. \tag{3.44}$$

Then

$$A = \int_0^T \omega_{\varepsilon_0}(T-s)\left(\rho_\varepsilon(u(T),v(s)) + \rho_\varepsilon(u(s),v(T))\right) ds.$$

Now by a use of the triangle inequality,

$$\|u(x,T) - v(x',s)\| + |u(x,s) - v(x',T)|$$
$$\geq |u(x,T) - v(x,T)| + |u(x,T) - v(x,T)|$$
$$- |v(x,T) - v(x',T)| - |u(x,T) - u(x,s)|$$
$$- |v(x',T) - v(x',s)| - |v(x,T) - v(x',T)|\,.$$

Hence

$$\rho_\varepsilon(u(T),v(s)) + \rho_\varepsilon(u(s),v(T)) \geq 2\|u(T) - v(T)\|_1 - 2\rho_\varepsilon(v(T),v(T))$$
$$- \|u(T) - u(s)\|_1 - \|v(T) - v(s)\|_1.$$

Regarding the upper estimate on B, we similarly have that

$$B = \int_0^T \omega_{\varepsilon_0}(s)\left[\rho_\varepsilon(u_0,v(s)) + \rho_\varepsilon(u(s),v_0)\right] ds,$$

and we also obtain

$$\rho_\varepsilon(u_0,v(s)) + \rho_\varepsilon(u(s),v_0) \leq 2\|u_0 - v_0\|_1 + 2\rho_\varepsilon(v_0,v_0)$$
$$+ \|u_0 - u(s)\|_1 + \|v_0 - v(s)\|_1.$$

Since v is a solution, it satisfies the TVD property, and hence

$$\rho_\varepsilon(v(T),v(T)) = \int \int_{-\varepsilon}^{\varepsilon} \omega_\varepsilon(z)\,|v(x+z,T) - v(x,T)|\,dz\,dx$$
$$\leq \int_{-\varepsilon}^{\varepsilon} \omega_\varepsilon(z) \sup_{|z|\leq\varepsilon}\left(\int |v(x+z,T) - v(x,T)|\,dx\right) dz$$
$$= |\varepsilon| \int_{-\varepsilon}^{\varepsilon} \omega_\varepsilon(z)\text{T.V.}\,(v(T))\,dz \leq |\varepsilon|\,\text{T.V.}\,(v_0)\,,$$

using (A.10). By the properties of ω,

$$\int_0^T \omega_\varepsilon(T-s)\,ds = \int_0^T \omega_\varepsilon(s)\,ds = \frac{1}{2}.$$

Applying (3.39) we obtain (recall that $\varepsilon_0 < T$)

$$\begin{aligned}\int_0^T \omega_{\varepsilon_0}(T-s)\|v(T)-v(s)\|_1\,ds &\le \int_0^T \omega_{\varepsilon_0}(T-s)\,(T-s)\|f\|_{\text{Lip}}\text{T.V.}\,(v_0)\,ds\\ &\le \frac{1}{2}\varepsilon_0\|f\|_{\text{Lip}}\text{T.V.}\,(v_0)\end{aligned}$$

and

$$\int_0^T \omega_{\varepsilon_0}(s)\|v_0-v(s)\|_1\,ds \le \frac{1}{2}\varepsilon_0\|f\|_{\text{Lip}}\text{T.V.}\,(v_0)\,.$$

Similarly,

$$\int_0^T \omega_{\varepsilon_0}(T-s)\|u(T)-u(s)\|_1\,ds \le \frac{1}{2}\nu\,(u,\varepsilon_0)$$

and

$$\int_0^T \omega_{\varepsilon_0}(s)\|u_0-u(s)\|_1\,ds \le \frac{1}{2}\nu\,(u,\varepsilon_0)\,.$$

If we collect all the above bounds, we should obtain the statement of the theorem. □

Observe that in the special case where u is a solution of the conservation law (3.1), we know that $\Lambda_{\varepsilon,\varepsilon_0}(u,v) \ge 0$, and hence we obtain, when $\varepsilon, \varepsilon_0 \to 0$, the familiar stability result

$$\|u(\,\cdot\,,T)-v(\,\cdot\,,T)\|_1 \le \|u_0-v_0\|_1.$$

We shall now show in three cases how Kuznetsov's lemma can be used to give estimates on how fast a method converges to the entropy solution of (3.1).

◇ **Example 3.12 (The smoothing method).**

While not a proper numerical method, the smoothing method provides an example of how the result of Kuznetsov may be used. The smoothing method is a (semi)numerical method approximating the solution of (3.1) as follows: Let $\omega_\delta(x)$ be a standard mollifier with support in $[-\delta,\delta]$, and let $t_n = n\Delta t$. Set $u^0 = u_0 * \omega_\delta$. For $0 \le t < \Delta t$ define u^1 to be the solution of (3.1) with initial data u^0. If Δt is small enough, u^1 remains differentiable for $t < \Delta t$. In the interval $[(n-1)\Delta t, n\Delta t)$, we define u^n to be the solution of (3.1), with $u^n\,(x,(n-1)\Delta t) = u^{n-1}(\,\cdot\,,t_n-) * \omega_\delta$.

The advantage of doing this is that u^n will remain differentiable in x for all times, and the solution in the strips $[t_n, t_{n+1})$ can be found by, e.g., the method of characteristics. To show that u^n is differentiable, we calculate

$$\begin{aligned}|u^n_x(x,t_{n-1})| &= \left|\int u^{n-1}_x(y,t_{n-1})\omega_\delta(x-y)\,dy\right|\\ &\le \frac{1}{\delta}\text{T.V.}\,(u^{n-1}(t_{n-1})) \le \frac{\text{T.V.}\,(u_0)}{\delta}.\end{aligned}$$

Let $\mu(t) = \max_x |u_x(x,t)|$. Using that u is a classical solution of (3.1), we find by differentiating (3.1) with respect to x that

$$u_{xt} + f'(u)u_{xx} + f''(u)u_x^2 = 0.$$

Write

$$\mu(t) = u_x(x_0(t),t),$$

where $x_0(t)$ is the location of the maximum of $|u_x|$. Then

$$\begin{aligned}\mu'(t) &= u_{xx}(x_0(t),t)x_0'(t) + u_{xt}(x_0(t),t)\\ &\le u_{xt}(x_0(t),t) = -f''(u)\big(u_x(x_0(t),t)\big)^2\\ &\le c\mu(t)^2,\end{aligned}$$

since $u_{xx} = 0$ at an extremum of u_x. Thus

$$\mu'(t) \le c\mu^2(t), \tag{3.45}$$

where $c = \|f''\|_\infty$. The idea is now that (3.45) has a blowup at some finite time, and we choose Δt less than this time. We shall be needing a precise relation between the Δt and δ and must therefore investigate (3.45) further. Solving (3.45) we obtain

$$\mu(t) \le \frac{\mu\,(t_n)}{1-c\mu\,(t_n)\,(t-t_n)} \le \frac{\text{T.V.}\,(u_0)}{\delta - c\text{T.V.}\,(u_0)\,\Delta t}.$$

So if

$$\Delta t < \frac{\delta}{c\text{T.V.}\,(u_0)}, \tag{3.46}$$

the method is well-defined. Choosing $\Delta t = \delta/(2c\text{T.V.}\,(u_0))$ will do.

Since u is an exact solution in the strips $[t_n, t_{n+1})$, we have

$$\begin{aligned}&\int_{t_n}^{t_{n+1}}\int \left(|u-k|\,\phi_t + q(u,k)\phi_x\right)dx\,dt\\ &+\int \Big(|u(x,t_n+)-k|\,\phi(x,t_n) - |u(x,t_{n+1}-)-k|\,\phi(x,t_{n+1})\Big)\,dx \ge 0.\end{aligned}$$

Summing these inequalities, and setting $k = v(y,s)$ where v is an exact solution of (3.1), we obtain

$$\Lambda_T(u, \Omega, v(y,s)) \geq -\sum_{n=0}^{N-1} \int \Omega\,(x,y,t_n,s) \Big(|u(x,t_n+) - v(y,s)| \\ - |u(x,t_n-) - v(y,s)| \Big)\, dx,$$

where we use the test function $\Omega(x,y,t,s) = \omega_{\varepsilon_0}(t-s)\omega_\varepsilon(x-y)$. Integrating this over y and s, and letting ε_0 tend to zero, we get

$$\liminf_{\varepsilon_0 \to 0} \Lambda_{\varepsilon,\varepsilon_0}(u,v) \geq -\sum_{n=0}^{N-1} \left(\rho_\varepsilon(u(t_n+), v(t_n)) - \rho_\varepsilon(u(t_n-), v(t_n))\right).$$

Using this in Kuznetsov's lemma, and letting $\varepsilon_0 \to 0$, we obtain

$$\|u(T) - v(T)\|_1 \leq \left\|u_0 - u^0\right\|_1 + 2\varepsilon\,\mathrm{T.V.}\,(u_0) \\ + \sum_{n=0}^{N-1} \left(\rho_\varepsilon(u(t_n+), v(t_n)) - \rho_\varepsilon(u(t_n-), v(t_n))\right)), \tag{3.47}$$

where we have used that $\lim_{\varepsilon_0 \to 0} \nu_t\,(u, \varepsilon_0) = 0$, which holds because u is a solution of the conservation law in each strip $[t_n, t_{n+1}\rangle$.

To obtain a more explicit bound on the difference of u and v, we investigate $\rho_\varepsilon(\omega_\delta * u, v) - \rho_\varepsilon(u,v)$, where ρ_ε is defined by (3.44),

$$\begin{aligned}\rho_\varepsilon(u * \omega_\delta, v) - \rho_\varepsilon(u,v) &\leq \iiint_{|z|\leq 1} \omega_\varepsilon(x-y)\omega(z)\Big(|u(x+\delta z) - v(y)| \\ &\qquad - |u(x) - v(y)| \Big)\, dx\,dy\,dz \\ &= \frac{1}{2} \iiint_{|z|\leq 1} (\omega_\varepsilon(x-y) - \omega_\varepsilon(x+\delta z - y))\,\omega(z) \\ &\qquad \times (|u(x+\delta z) - v(y)| - |u(x) - v(y)|)\, dx\,dy\,dz,\end{aligned}$$

which follows after writing $\iiint = \frac{1}{2}\iiint + \frac{1}{2}\iiint$ and making the substitution $x \to x - \delta z$, $z \to -z$ in one of these integrals. Therefore,

$$\begin{aligned}\rho_\varepsilon(u * \omega_\delta, v) - \rho_\varepsilon(u,v) &\leq \frac{1}{2} \iiint_{|z|\leq 1} |\omega_\varepsilon(y+\delta z) - \omega_\varepsilon(y)| \\ &\qquad \times \omega(z)\,|u(x+\delta z) - u(x)|\, dx\,dy\,dz \\ &\leq \frac{1}{2}\mathrm{T.V.}\,(\omega_\varepsilon)\;\mathrm{T.V.}\,(u)\;\delta^2 \\ &\leq \mathrm{T.V.}\,(u)\,\frac{\delta^2}{\varepsilon},\end{aligned}$$

by the triangle inequality and a further substitution $y \mapsto x - y$. Since $N = T/\Delta t$, the last term in (3.47) is less than

$$N\,\mathrm{T.V.}\,(u_0)\,\frac{\delta^2}{\varepsilon} \leq (\mathrm{T.V.}\,(u_0))^2\; 2cT\frac{\delta}{\varepsilon},$$

using (3.46). Furthermore, we have that

$$\left\|u^0 - u_0\right\|_1 \leq \delta\mathrm{T.V.}\,(u_0)\,.$$

Letting $K = \mathrm{T.V.}\,(u_0)\,c$, we find that

$$\|u(T) - v(T)\|_1 \leq 2\mathrm{T.V.}\,(u_0)\left[\delta + \varepsilon + \frac{KT\delta}{\varepsilon}\right],$$

using (3.47). Minimizing with respect to ε, we find that

$$\|u(T) - v(T)\|_1 \leq 2\mathrm{T.V.}\,(u_0)\left(\delta + 2\sqrt{KT\delta}\,\right). \tag{3.48}$$

So, we have shown that the smoothing method is of order $\frac{1}{2}$ in the smoothing coefficient δ. ◇

◇ **Example 3.13 (The method of vanishing viscosity).**

Another (semi)numerical method for (3.1) is the method of vanishing viscosity. Here we approximate the solution of (3.1) by the solution of

$$u_t + f\,(u)_x = \delta u_{xx}, \quad \delta > 0, \tag{3.49}$$

using the same initial data. Let u^δ denote the solution of (3.49). Due to the dissipative term on the right-hand side, the solution of (3.49) remains a classical (twice differentiable) solution for all $t > 0$. Furthermore, the solution operator for (3.49) is TVD. Hence a numerical method for (3.49) will (presumably) not experience the same difficulties as a numerical method for (3.1). If (η, q) is a convex entropy pair, we have, using the differentiability of the solution, that

$$\eta(u)_t + q(u)_x = \delta\eta'(u)u_{xx} = \delta\left(\eta(u)_{xx} - \eta''(u)u_x^2\right).$$

Multiplying by a nonnegative test function φ and integrating by parts, we get

$$\iint \left(\eta(u)\varphi_t + q(u)\varphi_x\right)\, dx\,dt \geq \delta \iint \eta(u)_x\varphi_x\, dx\,dt,$$

where we have used the convexity of η. Applying this with $\eta = |u^\delta - u|$ and $q = F(u^\delta, u)$ we can bound $\lim_{\varepsilon_0 \to 0} \Lambda_{\varepsilon,\varepsilon_0}(u^\delta, u)$ as follows:

$$\begin{aligned}
-\lim_{\varepsilon_0 \to 0} &\Lambda_{\varepsilon,\varepsilon_0}(u^\delta, u) \\
&\le \delta \int_0^T \iint \left| \frac{\partial \omega_\varepsilon(x-y)}{\partial x} \right| \frac{\partial \left| u^\delta(x,t) - u(y,t) \right|}{\partial x} \, dx \, dy \, dt \\
&\le \delta \int_0^T \iint \left| \frac{\partial \omega_\varepsilon(x-y)}{\partial x} \right| \left| \frac{\partial u^\delta(x,t)}{\partial x} \right| dx \, dy \, dt \\
&\le 2\text{T.V.}\left(u^\delta\right) T \frac{\delta}{\varepsilon} \\
&\le 2T\,\text{T.V.}\left(u_0\right) \frac{\delta}{\varepsilon}.
\end{aligned}$$

Now letting $\varepsilon_0 \to 0$ in (3.41) we obtain

$$\left\| u^\delta(T) - u(T) \right\|_1 \le \min_{\varepsilon} \left(2\varepsilon + \frac{2T\delta}{\varepsilon} \right) \text{T.V.}\left(u_0\right) = 2\text{T.V.}\left(u_0\right) \sqrt{T\delta}.$$

So the method of vanishing viscosity also has order $\frac{1}{2}$. ◇

◇ **Example 3.14 (Monotone schemes).**

We will here show that monotone schemes converge in L^1 to the solution of (3.1) at a rate of $(\Delta t)^{1/2}$. In particular, this applies to the Lax–Friedrichs scheme.

Let $u_{\Delta t}$ be defined by (3.24), where U_j^n is defined by (3.5), that is,

$$U_j^{n+1} = U_j^n - \lambda \Big(F\left(U_{j-p}^n, \dots, U_{j+p'}^n\right) - F\left(U_{j-1-p}^n, \dots, U_{j-1+p'}^n\right) \Big), \tag{3.50}$$

for a scheme that is assumed to be monotone; cf. Definition 3.5. In the following we use the notation

$$\eta_j^n = \left| U_j^n - k \right|, \quad q_j^n = f\left(U_j^n \vee k\right) - f\left(U_j^n \wedge k\right).$$

We find that

$$\begin{aligned}
-\Lambda_T&(u_{\Delta t}, \phi, k) \\
&= -\sum_j \sum_{n=0}^{N-1} \int_{x_j}^{x_{j+1}} \int_{t_n}^{t_{n+1}} \left(\eta_j^n \phi_t(x,s) + q_j^n \phi_x(x,s) \right) ds \, dx \\
&\quad - \sum_j \int_{x_j}^{x_{j+1}} \eta_j^0 \phi(x,0) \, dx + \sum_j \int_{x_j}^{x_{j+1}} \eta_j^N \phi(x,T) \, dx
\end{aligned}$$

$$\begin{aligned}
&= -\sum_j \Bigg[\sum_{n=0}^{N-1} \int_{x_j}^{x_{j+1}} \eta_j^n \left(\phi(x, t_{n+1}) - \phi(x, t_n) \right) dx \\
&\qquad + \int_{x_j}^{x_{j+1}} \eta_j^0 \phi(x,0) \, dx - \int_{x_j}^{x_{j+1}} \eta_j^N \phi(x,T) \, dx \\
&\qquad + \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} q_j^n \left(\phi(x_{j+1}, s) - \phi(x_j, s) \right) ds \Bigg] \\
&= \sum_j \sum_{n=0}^{N-1} \Bigg(\left(\eta_j^{n+1} - \eta_j^n \right) \int_{x_j}^{x_{j+1}} \phi(x, t_{n+1}) \, dx \\
&\qquad + \left(q_j^n - q_{j-1}^n \right) \int_{t_n}^{t_{n+1}} \phi(x_j, s) \, ds \Bigg)
\end{aligned}$$

by a summation by parts. Recall that we define the numerical entropy flux by

$$Q_j^n = Q(U^n; j) = F(U^n \vee k; j) - F(U^n \wedge k; j).$$

Monotonicity of the scheme implies, cf. (3.30), that

$$\eta_j^{n+1} - \eta_j^n + \lambda(Q_j^n - Q_{j-1}^n) \le 0.$$

For a nonnegative test function ϕ we obtain

$$\begin{aligned}
-\Lambda_T&(u_{\Delta t}, \phi, k) \\
&\le \sum_j \sum_{n=0}^{N-1} \Bigg(-\lambda(Q_j^n - Q_{j-1}^n) \int_{x_j}^{x_{j+1}} \phi(x, t_{n+1}) \, dx \\
&\qquad + (q_j^n - q_{j-1}^n) \int_{t_n}^{t_{n+1}} \phi(x_j, s) \, ds \Bigg) \\
&= \sum_j \sum_{n=0}^{N-1} \Bigg[\lambda(Q_j^n - q_j^n) \left(\int_{x_{j+1}}^{x_{j+2}} \phi(x, t_{n+1}) \, dx - \int_{x_j}^{x_{j+1}} \phi(x, t_{n+1}) \, dx \right) \\
&\qquad + (q_j^n - q_{j-1}^n) \left(\int_{t_n}^{t_{n+1}} \phi(x_j, s) \, ds - \lambda \int_{x_j}^{x_{j+1}} \phi(x, t_{n+1}) \, dx \right) \Bigg].
\end{aligned}$$

We also have that

$$\left| Q_j^n - q_j^n \right| \le 2 \|f\|_{\text{Lip}} \sum_{m=-p}^{p'} \left| U_{j+m}^n - U_j^n \right|,$$

and

$$\left| q_j^n - q_{j-1}^n \right| \le 2 \|f\|_{\text{Lip}} \left| U_j^n - U_{j-1}^n \right|,$$

which implies that

$$
\begin{aligned}
-\Lambda_T(u_{\Delta t},\phi,k)
&\le 2\|f\|_{\mathrm{Lip}} \sum_j \sum_{n=0}^{N-1} \Bigg[\Big(\sum_{m=-p}^{p'} \big|U^n_{j+m}-U^n_j\big| \Big) \\
&\qquad \times \int_{x_j}^{x_{j+1}} \left|\phi(x+\Delta x,t_{n+1})-\phi(x,t_{n+1})\right| dx \\
&\qquad + \left|U^n_j - U^n_{j-1}\right| \\
&\qquad \times \left| \int_{t_n}^{t_{n+1}} \phi(x_j,s)\,ds - \lambda \int_{x_j}^{x_{j+1}} \phi(x,t_{n+1})\,dx \right| \Bigg].
\end{aligned}
$$

Next, we subtract $\phi(x_j,t_n)$ from the integrand in each of the latter two integrals. Since $\Delta t=\lambda\Delta x$, the extra terms cancel, and we obtain

$$
\begin{aligned}
-\Lambda_T(u_{\Delta t},\phi,k) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.51)& \\
\le 2\|f\|_{\mathrm{Lip}} \sum_j \sum_{n=0}^{N-1} \Bigg[\Big(\sum_{m=-p}^{p'} \big|U^n_{j+m}-U^n_j\big| \Big)& \\
\times \int_{x_j}^{x_{j+1}} \left|\phi(x+\Delta x,t_{n+1})-\phi(x,t_{n+1})\right| dx& \\
+ \left|U^n_j - U^n_{j-1}\right| \left(\int_{t_n}^{t_{n+1}} \left|\phi(x_j,s)-\phi(x_j,t_n)\right| ds \right.& \\
\left. + \lambda \int_{x_j}^{x_{j+1}} \left|\phi(x,t_{n+1})-\phi(x_j,t_n)\right| dx \right) \Bigg].&
\end{aligned}
$$

Let $v=v(x,t)$ denote the unique weak solution of (3.1), and let $k=v(x',s')$. Choose the test function as $\phi(x,s)=\omega_\varepsilon(x-x')\omega_{\varepsilon_0}(s-s')$, and observe that

$$
\begin{aligned}
\int_0^T \int_{\mathbb{R}} \left|\omega_\varepsilon(x+\Delta x-x')-\omega_\varepsilon(x-x')\right| \omega_{\varepsilon_0}(t_n-s')\;dx'\,ds'& \\
\le \Delta x\,\mathrm{T.V.}\,(\omega_\varepsilon) \le 2\frac{\Delta x}{\varepsilon}.&
\end{aligned}
$$

Similarly,

$$
\int_0^T \int_{\mathbb{R}} \omega_\varepsilon(x_j-x')\left|\omega_{\varepsilon_0}(s-s')-\omega_{\varepsilon_0}(t_n-s')\right| dx'\,ds' \le 2\frac{\Delta t}{\varepsilon_0},
$$

whenever $|s-t_n|\le\Delta t$, and

$$
\int_0^T \int_{\mathbb{R}} \Big| \omega_\varepsilon(x-x')\omega_{\varepsilon_0}(t_{n+1}-s')
$$

$$
- \omega_\varepsilon(x_j-x')\omega_{\varepsilon_0}(t_n-s')\Big|\,dx'\,ds' \le 2\Big(\frac{\Delta t}{\varepsilon_0}+\frac{\Delta x}{\varepsilon}\Big).
$$

Integrating (3.51) over (x',s') with $0\le s'\le T$ we obtain

$$
\begin{aligned}
&-\Lambda_{\varepsilon,\varepsilon_0}(u_{\Delta t},v) \\
&\quad\le 4\|f\|_{\mathrm{Lip}} \sum_{n=0}^{N-1} \Bigg[\sum_j \sum_{m=-p}^{p'} \left|U^n_{j+m}-U^n_j\right| \frac{\Delta x}{\varepsilon}\Delta x \\
&\qquad\qquad + \sum_j \left|U^n_j-U^n_{j-1}\right| \left(\frac{\Delta t}{\varepsilon_0}\Delta t + \lambda\big(\frac{\Delta x}{\varepsilon}+\frac{\Delta t}{\varepsilon_0}\big)\Delta x \right) \Bigg] \\
&\quad\le 4\|f\|_{\mathrm{Lip}}\mathrm{T.V.}\,(u_{\Delta t}(0)) \\
&\qquad \times \sum_{n=0}^{N-1} \left[\frac{1}{2}(p+p'+1)^2\frac{(\Delta x)^2}{\varepsilon} + \frac{(\Delta t)^2}{\varepsilon_0} + \lambda\big(\frac{(\Delta x)^2}{\varepsilon}+\frac{\Delta x\Delta t}{\varepsilon_0}\big) \right] \\
&\quad\le K\,T\,\mathrm{T.V.}\,(u_{\Delta t}(0)) \left(\frac{1}{\varepsilon}+\frac{1}{\varepsilon_0}\right)\Delta t
\end{aligned}
$$

for some constant K, by using the estimate

$$
\sum_j \sum_{m=-p}^{p'} \left|U^n_{j+m}-U^n_j\right| \le \frac{1}{2}(p+p'+1)^2\mathrm{T.V.}\,(U^n).
$$

Recalling Kuznetsov's lemma

$$
\begin{aligned}
\|u_{\Delta t}(T)-v(T)\|_1 \le \|u_{\Delta t}(0)-v_0\|_1 + \mathrm{T.V.}\,(v_0)\,(2\varepsilon+\varepsilon_0\|f\|_{\mathrm{Lip}})& \\
+ \frac{1}{2}\left(\nu_T(u_{\Delta t},\varepsilon_0)+\nu_0(u_{\Delta t},\varepsilon_0)\right) - \Lambda_{\varepsilon,\varepsilon_0}.&
\end{aligned}
$$

We have that

$$
\mathrm{T.V.}\,(u_{\Delta t}(\,\cdot\,,t)) \le \mathrm{T.V.}\,(u_{\Delta t}(\,\cdot\,,0))
$$

and

$$
\nu_t\,(u_{\Delta t},\varepsilon) \le K_1\,(\Delta t+\varepsilon)\,\mathrm{T.V.}\,(u_{\Delta t}(\,\cdot\,,0)).
$$

Choose the initial approximation such that

$$
\|u_{\Delta t}(0)-v_0\|_1 \le \Delta x\,\mathrm{T.V.}\,(v_0).
$$

This implies

$$
\begin{aligned}
&\|u_{\Delta t}(T)-v(T)\|_1 \\
&\quad\le \mathrm{T.V.}\,(v_0)\,(\Delta x+2\varepsilon+\varepsilon_0\|f\|_{\mathrm{Lip}}) \\
&\qquad + \mathrm{T.V.}\,(u_{\Delta t}(\,\cdot\,,0)) \left(K_1(\Delta t+\varepsilon_0) + KT\Delta t\left(\frac{1}{\varepsilon_0}+\frac{1}{\varepsilon}\right)\right) \\
&\quad\le K_2\mathrm{T.V.}\,(v_0)\left[\Delta t+\varepsilon+\frac{T\Delta t}{\varepsilon}+\varepsilon_0+\frac{T\Delta t}{\varepsilon_0}\right].
\end{aligned}
$$

Minimizing with respect to ε_0 and ε, we obtain the final bound

$$\left\|u_{\Delta t}(T) - v(T)\right\|_1 \leq K_4 \mathrm{T.V.}\,(v_0)\left(\Delta t + 4\sqrt{T\Delta t}\right). \tag{3.52}$$

Thus, as promised, we have shown that monotone schemes are of order $(\Delta t)^{1/2}$. ◊

If one uses Kuznetsov's lemma to estimate the error of a scheme, one must estimate the modulus of continuity $\tilde{\nu}_t\,(u, \varepsilon_0)$ and the term $\Lambda_{\varepsilon,\varepsilon_0}(u, v)$. In other words, one must obtain regularity estimates on the *approximation* u. Therefore, this approach gives a posteriori error estimates, and perhaps the proper use for this approach should be in adaptive methods, in which it would provide error control and govern mesh refinement. However, despite this weakness, Kuznetsov's theory is still actively used.

3.3 A Priori Error Estimates

We shall now describe an application of a variation of Kuznetsov's approach, in which we obtain an error estimate for the method of vanishing viscosity, without using the regularity properties of the viscous approximation. Of course, this application only motivates the approach, since regularity of the solutions of parabolic equations is not difficult to obtain elsewhere. Nevertheless, it is interesting in its own right, since many difference methods have (3.53) as their model equation. We first state the result.

Theorem 3.15. *Let $v(x,t)$ be a solution of* (3.1) *with initial value v_0, and let u solve the equation*

$$u_t + f(u)_x = \left(\delta(u)u_x\right)_x, \qquad u(x,0) = u_0(x), \tag{3.53}$$

in the classical sense, with $\delta(u) > 0$. Then

$$\left\|u(T) - v(T)\right\|_1 \leq 2\left\|u_0 - v_0\right\|_1 + 4\mathrm{T.V.}\,(v_0)\sqrt{8T\|\delta\|_v},$$

where

$$\|\delta\|_v = \sup_{\substack{t\in[0,T]\\ x\in\mathbb{R}}} \bar{\delta}\left(v(x-,t), v(x+,t)\right)$$

and

$$\bar{\delta}(a,b) = \frac{1}{b-a}\int_a^b \delta(c)\,dc.$$

This result is not surprising, and in some sense is weaker than the corresponding result found by using Kuznetsov's lemma. The new element here is that the proof does *not* rely on any smoothness properties of the function

u, and is therefore also considerably more complicated than the proof using Kuznetsov's lemma.

Proof. The proof consists in choosing new Λ's, and using a special form of the test function φ. Let ω^∞ be defined as

$$\omega^\infty(x) = \begin{cases} \frac{1}{2} & \text{for } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We will consider a family of smooth functions ω such that $\omega \to \omega^\infty$. To keep the notation simple we will not add another parameter to the functions ω, but rather write $\omega \to \omega^\infty$ when we approach the limit. Let

$$\varphi(x,y,t,s) = \omega_\varepsilon(x-y)\omega_{\varepsilon_0}(t-s)$$

with $\omega_\alpha(x) = (1/\alpha)\,\omega(x/\alpha)$ as usual. In this notation

$$\omega_\varepsilon^\infty(x) = \begin{cases} 1/(2\varepsilon) & \text{for } |x| \leq \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

In the following we will use the entropy pair

$$\eta(u,k) = |u-k| \quad \text{and} \quad q(u,k) = \operatorname{sign}(u-k)\left(f(u) - f(k)\right),$$

and except where explicitly stated, we always let $u = u(y,s)$ and $v = v(x,t)$. Let $\eta_\sigma(u,k)$ and $q_\sigma(u,k)$ be smooth approximations to η and q such that

$$\eta_\sigma(u) \to \eta(u) \quad \text{as } \sigma \to 0, \qquad q_\sigma(u,k) = \int \eta_\sigma'(z-k)(f(z) - f(k))\,dz.$$

For a test function φ define

$$\Lambda_T^\sigma(u,k) = \int_0^T \int \eta_\sigma'(u-k)\left(u_s + f(u)_y - \left(\delta(u)u_y\right)_y\right)\varphi\,dy\,ds$$

(which is clearly zero because of (3.53)) and

$$\Lambda_{\varepsilon,\varepsilon_0}^\sigma(u,v) = \int_0^T \int \Lambda_T^\sigma(u, v(x,t))\,dx\,dt.$$

Note that since u satisfies (3.53), $\Lambda_{\varepsilon,\varepsilon_0}^\sigma = 0$ for every v. We now split $\Lambda_{\varepsilon,\varepsilon_0}^\sigma$ into two parts. Writing (cf. (2.10))

$$\begin{aligned}
(u_s + &f(u)_x - (\delta(u)u_y)_y)\eta_\sigma'(u-k) \\
&= \eta(u-k)_s + ((f(u) - f(k))'\eta_\sigma'(u-k)u_y - (\delta(u)u_y)_y\eta_\sigma'(u-k) \\
&= \eta_\sigma(u-k)_s + q_\sigma(u,k)_u u_y - (\delta(u)u_y)_y\eta_\sigma'(u-k) \\
&= \eta_\sigma(u-k)_s + q_\sigma(u,k)_y - (\delta(u)\eta_\sigma(u-k)_y)_y + \eta_\sigma''(u-k)\delta(u)(u_y)^2 \\
&= \eta_\sigma(u-k)_s + (q_\sigma(u,k) - \delta(u)\eta_\sigma(u-k)_y)_y + \eta''(u-k)\delta(u)(u_y)^2,
\end{aligned}$$

we may introduce

$$\Lambda_1^\sigma(u,v) = \int_0^T \int \int_0^T \int \eta_\sigma''(u-v)\delta(u)\,(u_y)^2\,\varphi\,dy\,ds\,dx\,dt,$$

$$\begin{aligned}\Lambda_2^\sigma(u,v)& \\ &= \int_0^T \int \int_0^T \int \Big(\eta_\sigma(u-v)_s + \big(q_\sigma(u,v) - \delta(u)\eta_\sigma(u-v)_y\big)_y\Big)\,\varphi\,dy\,ds\,dx\,dt,\end{aligned}$$

such that $\Lambda_{\varepsilon,\varepsilon_0}^\sigma = \Lambda_1^\sigma + \Lambda_2^\sigma$. Note that if $\delta(u) > 0$, we always have $\Lambda_1^\sigma \geq 0$, and hence $\Lambda_2^\sigma \leq 0$. Then we have that

$$\Lambda_2 := \limsup_{\sigma\to 0} \Lambda_2^\sigma \leq 0.$$

To estimate Λ_2, we integrate by parts:

$$\begin{aligned}\Lambda_2(u,v)& \\ &= \int_0^T \int \int_0^T \int \left(-\eta(u-v)\varphi_s - q(u,v)\varphi_y + V(u,v)\varphi_{yy}\right)\,dy\,ds\,dx\,dt \\ &\quad + \int_0^T \iint \eta(u(T)-v)\varphi|_{s=T}\,dy\,dx\,dt - \int_0^T \iint \eta(u_0-v)\varphi|_{s=0}\,dy\,dx\,dt \\ &= \int_0^T \int \int_0^T \int \left(\eta(u-v)\varphi_t + F(u,v)\varphi_x - V(u,v)\varphi_{xy}\right)\,dy\,ds\,dx\,dt \\ &\quad + \int_0^T \iint \eta(u(T)-v)\varphi|_{s=T}\,dy\,dx\,dt - \int_0^T \iint \eta(u_0-v)\varphi|_{s=0}\,dy\,dx\,dt,\end{aligned}$$

where

$$V(u,v) = \int_u^v \delta(s)\eta'(s-v)\,ds.$$

Now define (the "dual of Λ_2")

$$\begin{aligned}\Lambda_2^* := &- \int_0^T \int \int_0^T \int \left(\eta(u-v)\varphi_t + q(u,v)\varphi_x - V(u,v)\varphi_{xy}\right)\,dy\,ds\,dx\,dt \\ &- \int_0^T \iint \eta(u-v(T))\varphi\,\Big|_{t=0}^{t=T}\,dx\,dy\,ds.\end{aligned}$$

Then we can write

$$\begin{aligned}\Lambda_2 = &-\Lambda_2^* \\ &\underbrace{+ \int_0^T \iint \left(\eta(u(T)-v)\varphi\right)|_{s=T}\,dy\,dx\,dt}_{\Phi_1} \\ &\underbrace{- \int_0^T \iint \left(\eta(u_0-v)\varphi\right)|_{s=0}\,dy\,dx\,dt}_{\Phi_2} \\ &\underbrace{+ \int_0^T \iint \left(\eta(u-v(T))\varphi\right)|_{t=T}\,dx\,dy\,ds}_{\Phi_3} \\ &\underbrace{- \int_0^T \iint \left(\eta(u_0-v_0)\varphi\right)|_{t=0}\,dx\,dy\,ds}_{\Phi_4} \\ =:& -\Lambda_2^* + \Phi.\end{aligned}$$

We will need later that

$$\Phi = \Lambda_2^* + \Lambda_2 \leq \Lambda_2^*. \tag{3.54}$$

Let

$$\Omega_{\varepsilon_0}(t) = \int_0^t \omega_{\varepsilon_0}(s)\,ds$$

and

$$e(t) = \left\|u(t) - v(t)\right\|_1 = \int \eta(u(x,t) - v(x,t))\,dx.$$

To continue estimating, we need the following proposition.

Proposition 3.16.

$$\begin{aligned}\Phi \geq \Omega_{\varepsilon_0}(T)e(T) - \Omega_{\varepsilon_0}(T)e(0) + \int_0^T \omega_{\varepsilon_0}(T-t)e(t)\,dt - \int_0^T \omega_{\varepsilon_0}(t)e(t)\,dt& \\ - 4\Omega_{\varepsilon_0}(T)\left(\varepsilon_0\|f\|_{\mathrm{Lip}} + \varepsilon\right)\mathrm{T.V.}\left(v_0\right).&\end{aligned}$$

Proof (of Proposition 3.16). We start by estimating Φ_1. First note that

$$\begin{aligned}\eta(u(y,T) - v(x,t)) &= |u(y,T) - v(x,t)| \\ &\geq |u(y,T) - v(y,T)| \\ &\qquad - |v(y,T) - v(y,t)| - |v(y,t) - v(x,t)| \\ &= \eta(u(y,T) - v(y,T)) \\ &\qquad - |v(y,T) - v(y,t)| - |v(y,t) - v(x,t)|.\end{aligned}$$

Thus

$$\begin{aligned}\Phi_1 \geq &\int_0^T \iint \eta(u(y,T) - v(y,T))\varphi|_{s=T}\,dy\,dx\,dt \\ &- \int_0^T \iint |v(y,T) - v(y,t)|\,\varphi|_{s=T}\,dy\,dx\,dt \\ &- \int_0^T \iint |v(y,t) - v(x,t)|\,\varphi|_{s=T}\,dy\,dx\,dt \\ \geq &\ \Omega_{\varepsilon_0}(T)e(T) - \Omega_{\varepsilon_0}(T)\left(\varepsilon_0\|f\|_{\mathrm{Lip}} + \varepsilon\right)\mathrm{T.V.}\left(v_0\right).\end{aligned}$$

Here we have used that v is an exact solution. The estimate for Φ_2 is similar, yielding

$$\Phi_2 \geq -\Omega_{\varepsilon_0}(T)e(0) - \Omega_{\varepsilon_0}(T)\,(\varepsilon_0\|f\|_{\mathrm{Lip}} + \varepsilon)\,\mathrm{T.V.}\,(v_0)\,.$$

To estimate Φ_3 we proceed in the same manner:

$$\eta(u(y,s) - v(x,T)) \geq \eta(u(y,s) - v(y,s)) - |v(y,s) - v(x,s)| \\ - |v(x,s) - v(x,T)|\,.$$

This gives

$$\Phi_3 \geq \int_0^T \omega_{\varepsilon_0}(T-t)e(t)\,dt - \Omega_{\varepsilon_0}(T)\,(\varepsilon_0\|f\|_{\mathrm{Lip}} + \varepsilon)\,\mathrm{T.V.}\,(v_0)\,,$$

while by the same reasoning, the estimate for Φ_4 reads

$$\Phi_4 \geq -\int_0^T \omega_{\varepsilon_0}(t)e(t)\,dt - \Omega_{\varepsilon_0}(T)\,(\|f\|_{\mathrm{Lip}}\varepsilon_0 + \varepsilon)\,\mathrm{T.V.}\,(v_0)\,.$$

The proof of Proposition 3.16 is complete. □

To proceed further, we shall need the following Gronwall-type lemma:

Lemma 3.17. *Let θ be a nonnegative function that satisfies*

$$\Omega_{\varepsilon_0}^{\infty}(\tau)\theta(\tau) + \int_0^{\tau} \omega_{\varepsilon_0}^{\infty}(\tau - t)\theta(t)\,dt \leq C\,\Omega_{\varepsilon_0}^{\infty}(\tau) + \int_0^{\tau} \omega_{\varepsilon_0}^{\infty}(t)\theta(t)\,dt, \tag{3.55}$$

for all $\tau \in [0,T]$ and some constant C. Then

$$\theta(\tau) \leq 2C.$$

Proof (of Lemma 3.17). If $\tau \leq \varepsilon_0$, then for $t \in [0,\tau]$, $\omega_{\varepsilon_0}^{\infty}(t) = \omega_{\varepsilon_0}^{\infty}(\tau - t) = 1/(2\varepsilon_0)$. In this case (3.55) immediately simplifies to $\theta(t) \leq C$.

For $\tau > \varepsilon_0$, we can write (3.55) as

$$\theta(\tau) \leq C + \frac{1}{\Omega_{\varepsilon_0}^{\infty}(\tau)} \int_0^{\varepsilon_0} \left(\omega_{\varepsilon_0}^{\infty}(t) - \omega_{\varepsilon_0}^{\infty}(\tau - t)\right)\theta(t)\,dt.$$

For $t \in [0,\varepsilon_0]$ we have $\theta(t) \leq C$, and this implies

$$\theta(\tau) \leq C\left(1 + \frac{1}{\Omega_{\varepsilon_0}^{\infty}(\tau)} \int_0^{\varepsilon_0} \left(\omega_{\varepsilon_0}^{\infty}(t) - \omega_{\varepsilon_0}^{\infty}(\tau - t)\right)\,dt\right) \leq 2C.$$

This concludes the proof of the lemma. □

Now we can continue the estimate of $e(T)$.

Proposition 3.18. *We have that*

$$e(T) \leq 2e(0) + 8\,(\varepsilon + \varepsilon_0\|f\|_{\mathrm{Lip}})\,\mathrm{T.V.}\,(v_0) + 2\lim_{\omega\to\omega^\infty}\,\sup_{t\in[0,T]}\frac{\Lambda_2^*(u,v)}{\Omega_{\varepsilon_0^\infty}(t)}.$$

Proof (of Proposition 3.18). Starting with the inequality (3.54), using the estimate for Φ from Proposition 3.16, we have, after passing to the limit $\omega \to \omega^\infty$, that

$$\Omega_{\varepsilon_0}^{\infty}(T)e(T) + \int_0^T \omega_{\varepsilon_0}^{\infty}(T-t)e(t)\,dt \leq \Omega_{\varepsilon_0}^{\infty}(t)e(0) + \int_0^T \omega_{\varepsilon_0}^{\infty}(t)e(t)\,dt \\ + 4\Omega_{\varepsilon_0}^{\infty}(t)\,(\varepsilon + \varepsilon_0\|f\|_{\mathrm{Lip}})\,\mathrm{T.V.}\,(v_0) \\ + \Omega_{\varepsilon_0}^{\infty}(T)\lim_{\omega\to\omega^\infty}\,\sup_{t\in[0,T]}\frac{\Lambda_2^*(u,v)}{\Omega_{\varepsilon_0}^{\infty}(t)}.$$

We apply Lemma 3.17 with

$$C = 4\,(\varepsilon + \varepsilon_0\|f\|_{\mathrm{Lip}})\,\mathrm{T.V.}\,(v_0) + \lim_{\omega\to\omega^\infty}\,\sup_{t\in[0,T]}\frac{\Lambda_2^*(u,v)}{\Omega_{\varepsilon_0}^{\infty}(t)} + e(0)$$

to complete the proof. □

To finish the proof of the theorem, it remains only to estimate

$$\lim_{\omega\to\omega^\infty}\,\sup_{t\in[0,T]}\frac{\Lambda_2^*(u,v)}{\Omega(t)}.$$

We will use the following inequality:

$$\left|\frac{V(u,v^+) - V(u,v^-)}{v^+ - v^-}\right| \leq \frac{1}{v^+ - v^-}\int_{v^-}^{v^+} \delta(s)\,ds. \tag{3.56}$$

Since v is an entropy solution to (3.1), we have that

$$\Lambda_2^* \leq -\int_0^T \int \int_0^T \int V(u,v)\varphi_{xy}\,dy\,ds\,dx\,dt. \tag{3.57}$$

Since v is of bounded variation, it suffices to study the case where v is differentiable except on a countable number of curves $x = x(t)$. We shall bound Λ_2^* in the case that we have one such curve; the generalization to more than one is straightforward. Integrating (3.57) by parts, we obtain

$$\Lambda_2^* \leq \int_0^T \int \Psi(y,s)\,dy\,ds, \tag{3.58}$$

where Ψ is given by

$$\Psi(y,s) = \int_0^T \Big(\int_{-\infty}^{x(t)} V(u,v)_v\,v_x\varphi_y\,dx \\ + \frac{[\![V]\!]}{[\![v]\!]}[\![v]\!]\varphi_y|_{x=x(t)} + \int_{x(t)}^{\infty} V(u,v)_v\,v_x\varphi_y\,dx\Big)\,dt.$$

As before, $[\![a]\!]$ denotes the jump in a, i.e., $[\![a]\!] = a(x(t)+,t) - a(x(t)-,t)$. Using (3.56), we obtain

$$\begin{aligned}|\Psi(y,s)| \leq &\|\delta\|_v \int_0^T \Big(\int_{-\infty}^{x(t)} |v_x|\,|\varphi_y|\,dx \\ &+ |[\![v]\!]|\,\big|\varphi_y|_{x=x(t)}\big| + \int_{x(t)}^{\infty} |v_x|\,|\varphi_y|\,dx\Big)\,dt.\end{aligned} \tag{3.59}$$

Let D be given by

$$D(x,t) = \int_0^T \int |\varphi_y|\,dy\,ds.$$

A simple calculation shows that

$$D(x,t) = \frac{1}{\varepsilon}\int_0^T \omega_{\varepsilon_0}(t-s)\,ds \int |\omega'(y)|\,dy \leq \frac{1}{\varepsilon}\int_0^T \omega_{\varepsilon_0}(t-s)\,ds.$$

Consequently,

$$\begin{aligned}\int_0^T \sup_x D(x,t)\,dt &\leq \frac{1}{\varepsilon}\int_0^T\int_0^T \omega_{\varepsilon_0}(t-s)\,ds\,dt \\ &= \frac{2}{\varepsilon}\int_0^T (T-t)\omega_{\varepsilon_0}(t)\,dt \\ &\leq \frac{2T\Omega(T)}{\varepsilon}.\end{aligned}$$

Inserting this in (3.59), and the result in (3.58), we find that

$$\Lambda_2^*(u,v,T) \leq \frac{2}{\varepsilon} T\,\mathrm{T.V.}\,(v_0)\,\|\delta\|_v \Omega(T).$$

Summing up, we have now shown that

$$e(T) \leq 2e(0) + 8\,(\varepsilon + \varepsilon_0\|f\|_{\mathrm{Lip}})\,\mathrm{T.V.}\,(v_0) + \frac{4}{\varepsilon} T\,\mathrm{T.V.}\,(v_0)\,\|\delta\|_v.$$

We can set ε_0 to zero, and minimize over ε, obtaining

$$\|u(T) - v(T)\|_1 \leq 2\|u_0 - v_0\|_1 + 4\mathrm{T.V.}\,(v_0)\,\sqrt{8T\|\delta\|_v}.$$

The theorem is proved. □

The main idea behind this approach to getting a priori error estimates, is to choose the "Kuznetsov-type" form $\Lambda_{\varepsilon,\varepsilon_0}$ such that

$$\Lambda_{\varepsilon,\varepsilon_0}(u,v) = 0$$

for every function v, and then writing $\Lambda_{\varepsilon,\varepsilon_0}$ as a sum of a nonnegative and a nonpositive part. Given a numerical scheme, the task is then to prove a discrete analogue of the previous theorem.

3.4 Measure-Valued Solutions

> You try so hard, but you don't understand ...
>
> *Bob Dylan, Ballad of a Thin Man (1965)*

Monotone methods are at most first-order accurate. Consequently, one must work harder to show that higher-order methods converge to the entropy solution. While this is possible in one space dimension, i.e., in the above setting, it is much more difficult in several space dimensions. One useful tool to aid the analysis of higher-order methods is the concept of *measure-valued solutions*. This is a rather complicated concept, which requires a background from analysis beyond this book. Therefore, the presentation in this section is brief, and is intended to give the reader a first flavor, and an idea of what this method can accomplish.

Consider the case where a numerical scheme gives a sequence U_j^n that is uniformly bounded in $L^\infty(\mathbb{R}\times[0,\infty))$, and with the L^1-norm Lipschitz continuous in time, but such that there is no bound on the total variation. We can still infer the existence of a weak limit

$$u_{\Delta t} \overset{*}{\rightharpoonup} u,$$

but the problem is to show that

$$f(u_{\Delta t}) \overset{*}{\rightharpoonup} f(u).$$

Here, we have introduced the concept of weak-$*$ L^∞ convergence. A sequence $\{u_n\}$ that is bounded in L^∞ is said to converge weakly-$*$ to u if for all $v \in L^1$,

$$\int u_n v\,dx \to \int uv\,dx, \quad \text{as } n \to \infty.$$

Since $u_{\Delta t}$ is bounded, $f(u_{\Delta t})$ is also bounded and converges weakly, and thus

$$f(u_{\Delta t}) \overset{*}{\rightharpoonup} \bar{f},$$

but $\bar{f}$ is in general not equal to $f(u)$. We provide a simple example of the problem.

◇ **Example 3.19.**

Let $u_n = \sin(nx)$ and $f(u) = u^2$. Then

$$\left|\int \sin(nx)\varphi(x)\,dx\right| \leq \frac{1}{n}\left|\int \cos(nx)\varphi'(x)\,dx\right| \leq \frac{C}{n} \to 0 \text{ as } n \to \infty.$$

On the other hand, $f(u_n) = \sin^2(nx) = (1-\cos(2nx))/2$, and hence a similar estimate shows that

$$\left|\int (f(u_n) - \frac{1}{2})\varphi(x)\,dx\right| \leq \frac{C}{n} \to 0 \text{ as } n \to \infty.$$

3.5 Notes

The Lax–Friedrichs scheme was introduced by Lax in 1954; see [94]. Godunov discussed what has later become the Godunov scheme in 1959 as a method to study gas dynamics; see [60]. The Lax–Wendroff theorem, Theorem 3.4, was first proved in [97]. Theorem 3.8 was proved by Oleĭnik in her fundamental paper [110]; see also [130]. Several of the fundamental results concerning monotone schemes are due to Crandall and Majda [39], [38]. Theorem 3.10 is due to Harten, Hyman, and Lax; see [62].

The error analysis is based on the fundamental analysis by Kuznetsov, [89], where one also can find a short discussion of the examples we have analyzed, namely the smoothing method, the method of vanishing viscosity, as well as monotone schemes. Our presentation of the a priori estimates follows the approach due to Cockburn and Gremaud; see [31] and [32], where also applications to numerical methods are given.

The concept of measure-valued solutions is due to DiPerna, and the key result, Corollary 3.21, can be found in [45], while Lemma 3.25 is to be found in [44]. The proof of Lemma 3.25 and Remark 3.26 are due to H. Hanche-Olsen. Our presentation of the uniqueness of measure-valued solutions, Theorem 3.24, is taken mainly from Szepessy, [134]. Theorem 3.27 is due to Coquel and LeFloch, [35]; see also [36], where several extensions are discussed.

Exercises

3.1 Show that the Lax–Wendroff and the MacCormack methods are of second order.

3.2 The Engquist–Osher (or generalized upwind) method, see [46], is a conservative difference scheme with a numerical flux defined as follows:

$$F(U;j) = f^{\mathrm{EO}}\left(U_j, U_{j+1}\right), \quad \text{where}$$

$$f^{\mathrm{EO}}(u,v) = \int_0^u \max(f'(s),0)\,ds + \int_0^v \min(f'(s),0)\,ds + f(0).$$

a. Show that this method is consistent and monotone.

b. Find the order of the scheme.

c. Show that the Engquist–Osher flux f^{EO} can be written

$$f^{\mathrm{EO}}(u,v) = \frac{1}{2}\left(f(u) + f(v) - \int_u^v |f'(s)|\,ds\right).$$

d. If $f(u) = u^2/2$, show that the numerical flux can be written

$$f^{\mathrm{EO}}(u,v) = \frac{1}{2}\left(\max(u,0)^2 + \min(v,0)^2\right).$$

Generalize this simple expression to the case where $f''(u) \neq 0$ and $\lim_{|u|\to\infty} |f(u)| = \infty$.

3.3 Why does the method

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

not give a viable difference scheme?

3.4 We study a nonconservative method for Burgers' equation. Assume that $U_j^0 \in [0,1]$ for all j. Then the characteristic speed is nonnegative, and we define

$$U_j^{n+1} = U_j^n - \lambda U_j^{n+1}\left(U_j^n - U_{j-1}^n\right), \quad n \geq 0, \tag{3.89}$$

where $\lambda = \Delta t/\Delta x$.

a. Show that this yields a monotone method, provided that a CFL condition holds.

b. Show that this method is consistent and determine the truncation error.

3.5 Assume that $f'(u) > 0$ and that $f''(u) \geq 2c > 0$ for all u in the range of u_0. We use the upwind method to generate approximate solutions to

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x); \tag{3.90}$$

i.e., we set

$$U_j^{n+1} = U_j^n - \lambda\left(f(U_j^n) - f(U_{j-1}^n)\right).$$

Set

$$V_j^n = \frac{U_j^n - U_{j-1}^n}{\Delta x}.$$

a. Show that

$$\begin{aligned} V_j^{n+1} &= \left(1 - \lambda f'(U_{j-1}^n)\right)V_j^n + \lambda f'(U_{j-1}^n)V_{j-1}^n \\ &\quad - \frac{\Delta t}{2}\left(f''(\eta_{j-1/2})\left(V_j^n\right)^2 + f''(\eta_{j-3/2})\left(V_{j-1}^n\right)^2\right), \end{aligned}$$

where $\eta_{j-1/2}$ is between U_j^n and U_{j-1}^n.

b. Next, assume inductively that

$$V_j^n \leq \frac{1}{(n+2)c\Delta t}, \quad \text{for all } j,$$

and set $\hat{V}^n = \max(\max_j V_j^n, 0)$. Then show that

$$\hat{V}^{n+1} \leq \hat{V}^n - c\Delta t\left(\hat{V}^n\right)^2.$$

c. Use this to show that

$$\hat{V}^n \leq \frac{\hat{V}^0}{1+\hat{V}^0 cn\Delta t}.$$

d. Show that this implies that

$$U_i^n - U_j^n \leq \Delta x(i-j)\frac{\hat{V}^0}{1+\hat{V}^0 cn\Delta t},$$

for $i \geq j$.

e. Let u be the entropy solution of (3.90), and assume that $0 \leq \max_x u_0'(x) = M < \infty$, show that for almost every x, y, and t we have that

$$\frac{u(x,t)-u(y,t)}{x-y} \leq \frac{M}{1+cMt}. \tag{3.91}$$

This is the Oleĭnik entropy condition for convex scalar conservation laws.

3.6 Assume that f is as in the previous exercise, and that u_0 is periodic with period p.

a. Use uniqueness of the entropy solution to (3.90) to show that the entropy solution $u(x,t)$ is also periodic in x with period p.

b. Then use the Oleĭnik entropy condition (3.91) to deduce that

$$\sup_x u(x,t) - \inf_x u(x,t) \leq \frac{Mp}{1+cMt}.$$

Thus $\lim_{t\to\infty} u(x,t) = \bar{u}$ for some constant $\bar{u}$.

c. Use conservation to show that

$$\bar{u} = \frac{1}{p}\int_0^p u_0(x)\,dx.$$

3.7 Assume that $g(x)$ is a continuously differentiable function with period 2π. Then we have that the Fourier representation

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}\left(a_k\cos(kx) + b_k\sin(kx)\right)$$

holds pointwise, where

$$a_0 = \frac{1}{\pi}\int_0^{2\pi} g(x)\,dx \quad \text{and} \quad \begin{cases} a_k = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} g(x)\cos(kx)\,dx, \\ b_k = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} g(x)\sin(kx)\,dx, \end{cases}$$

for $k \geq 1$.

a. Use this to show that

$$g(nx) \overset{*}{\rightharpoonup} \frac{a_0}{2}.$$

b. Find a regular measure ν such that for any continuously differentiable h,

$$h(\sin(nx)) \overset{*}{\rightharpoonup} \int h(\lambda)\,d\nu(\lambda).$$

Thus we have found an explicit form of the Young measure associated with the sequence $\{\sin(nx)\}$.

3.8 We shall consider a scalar conservation law with a "fractal" function as the initial data. Define the set of piecewise linear functions

$$\mathcal{D} = \{\phi(x) = Ax + B \mid x \in [a,b], A, B \in \mathbb{R}\},$$

and the map

$$F(\phi) \quad = \quad \begin{cases} 2D(x-a)+\phi(a) & \text{for } x \in [a, a+L/3], \\ -D(x-a)+\phi(a) & \text{for } x \in [a+L/3, a+2L/3], \\ 2D(x-b)+\phi(b) & \text{for } x \in [a+2L/3, b], \end{cases}$$

$\phi \in \mathcal{D}$, where $L = b-a$ and $D = (\phi(b)-\phi(a))/L$. For a nonnegative integer k introduce $\chi_{j,k}$ as the characteristic function of the interval $I_{j,k} = [j/3^k, (j+1)/3^k]$, $j = 0,\ldots,3^{k+1}-1$. We define functions $\{v_k\}$ recursively as follows. Let

$$v_0(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ x & \text{for } 0 \leq x \leq 1, \\ 1 & \text{for } 1 \leq x \leq 2, \\ 3-x & \text{for } 2 \leq x \leq 3, \\ 0 & \text{for } 3 \leq x. \end{cases}$$

Assume that $v_{j,k}$ is linear on $I_{j,k}$ and let

$$v_k = \sum_{j=-3^k}^{3^k-1} v_{j,k}\chi_{j,k}, \tag{3.92}$$

and define the next function v_{k+1} by

$$v_{k+1} = \sum_{j=0}^{3^{k+1}-1} F(v_{j,k})\chi_{j,k} = \sum_{j=0}^{3^{k+2}-1} v_{j,k+1}\chi_{j,k+1}. \tag{3.93}$$

In the left part of Figure 3.3 we show the effect of the map F, and on the right we show $v_5(x)$ (which is piecewise linear on $3^6 = 729$ segments).

a. Show that the sequence $\{v_k\}_{k>1}$ is a Cauchy sequence in the supremum norm, and hence we can define a continuous function v by setting

$$v(x) = \lim_{k\to\infty} v_k(x).$$

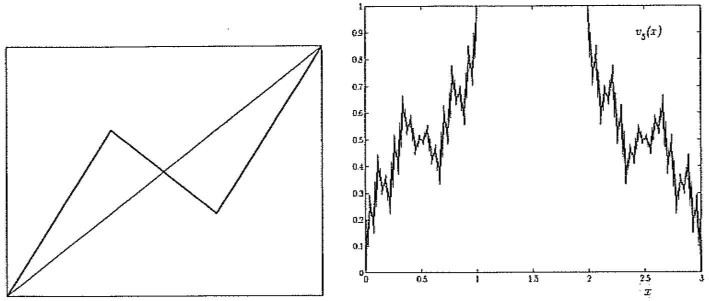

Figure 3.3. Left: the construction of $F(\phi)$ from ϕ. Right: $v_5(x)$.

b. Show that v is not of bounded variation, and determine the total variation of v_k.

c. Show that

$$v(j/3^k) = v_k(j/3^k),$$

for any integers $j = 0, \dots, 3^{k+1}$, $k \in \mathbb{N}$.

d. Assume that f is a C^1 function on $[0,1]$ with $0 \le f'(u) \le 1$. We are interested in solving the conservation law

$$u_t + f(u)_x = 0, \quad u_0(x) = v(x).$$

To this end we shall use the upwind scheme defined by (3.8), with $\Delta t = \Delta x = 1/3^k$, and

$$U_j^0 = v(j\Delta x).$$

Show that $u_{\Delta t}(x,t)$ converges to an entropy solution of the conservation law above.

4
Multidimensional Scalar Conservation Laws

> Just send me the theorems,
> then I shall find the proofs.[1]
>
> *Chrysippus told Cleanthes, 3th century* B.C.

Our analysis has so far been confined to scalar conservation laws in one dimension. Clearly, the multidimensional case is considerably more important. Luckily enough, the analysis in one dimension can be carried over to higher dimensions by essentially treating each dimension separately. This technique is called *dimensional splitting.* The final results are very much the natural generalizations one would expect.

The same splitting techniques of dividing complicated differential equations into several simpler parts, can in fact be used to handle other problems. These methods are generally denoted *operator splitting methods* or *fractional steps methods.*

4.1 Dimensional Splitting Methods

We will in this section show how one can analyze scalar multidimensional conservation laws by dimensional splitting, which amounts to solving one

[1] Lucky guy! Paraphrased from Diogenes Laertius, *Lives of Eminent Philosophers*, c. A.D. 200.

Stability for some non-strictly hyperbolic systems of conservation laws (these are really only "quasi-systems") has been proved by Winther and Tveito [142] and Klingenberg and Risebro [84].

We end this chapter with a suitable quotation:

> This is really easy:
>
> $|\text{what you have}| \le |\text{what you want}|$
> $\qquad + |\text{what you have} - \text{what you want}|$
>
> *Rinaldo Colombo, private communication*

Exercises

7.1 Show that the solution of the Cauchy problem obtained by the front-tracking construction of Chapter 6 is an entropy solution in the sense of conditions **A**–**C** on pages 267–268.

7.2 The proof of Theorem 7.8 is detailed only in the genuinely nonlinear case. Do the necessary estimates in the case of a linearly degenerate wave family.

Appendix A
Total Variation, Compactness, etc.

> I hate T.V. I hate it as much as peanuts.
> But I can't stop eating peanuts.
>
> *Orson Welles, The New York Herald Tribune (1956)*

A key concept in the theory of conservation laws is the notion of *total variation*, T.V. (u), of a function u of one variable. We define

$$\text{T.V.}\,(u) := \sup \sum_i |u(x_i) - u(x_{i-1})|\,. \tag{A.1}$$

The supremum in (A.1) is taken over all finite partitions $\{x_i\}$ such that $x_{i-1} < x_i$. The set of all functions with finite total variation on I we denote by $BV(I)$. Clearly, functions in $BV(I)$ are bounded. We shall omit explicit mention of the interval I if (we think that) this is not important, or if it is clear which interval we are referring to.

For any finite partition $\{x_i\}$ we can write

$$\begin{aligned}\sum_i |u(x_{i+1}) - u(x_i)| &= \sum_i \max\left(u(x_{i+1}) - u(x_i), 0\right)\\ &\quad - \sum_i \min\left(u(x_{i+1}) - u(x_i), 0\right)\\ &=: p + n.\end{aligned}$$

Then the total variation of u can be written

$$\text{T.V.}\,(u) = P + N := \sup p + \sup n. \tag{A.2}$$

We call P the positive, and N the negative variation, of u. If for the moment we consider the finite interval $I = [a, x]$, and partitions with $a = x_1 < \cdots < x_n = x$, we have that

$$p_a^x - n_a^x = u(x) - u(a),$$

where we write p_a^x and n_a^x to indicate which interval we are considering. Hence

$$p_a^x \leq N_a^x + u(x) - u(a).$$

Taking the supremum on the left-hand side we obtain

$$P_a^x - N_a^x \leq u(x) - u(a).$$

Similarly, we have that $N_a^x - P_a^x \leq u(a) - u(x)$, and consequently

$$u(x) = P_a^x - N_a^x + u(a). \tag{A.3}$$

In other words, any function $u(x)$ in BV can be written as a difference between two increasing functions,[1]

$$u(x) = u_+(x) - u_-(x), \tag{A.4}$$

where $u_+(x) = u(a) + P_a^x$ and $u_-(x) = N_a^x$. Let ξ_j denote the points where u is discontinuous. Then we have that

$$\sum_j |u(\xi_j+) - u(\xi_j-)| \leq \text{T.V.}\,(u) < \infty,$$

and hence we see that there can be at most a countable set of points where $u(\xi+) \neq u(\xi-)$.

Equation (A.3) has the very useful consequence that if a function u in BV is also differentiable, then

$$\int |u'(x)|\, dx = \text{T.V.}\,(u)\,. \tag{A.5}$$

This equation holds, since

$$\int |u'(x)|\, dx = \int \Big(\frac{d}{dx} P_a^x + \frac{d}{dx} N_a^x \Big)\, dx = P + N = \text{T.V.}\,(u)\,.$$

We can also relate the total variation with the shifted L^1-norm. Define

$$\lambda(u, \varepsilon) = \int |u(x+\varepsilon) - u(x)|\, dx. \tag{A.6}$$

If $\lambda(u,\varepsilon)$ is a (nonnegative) continuous function in ε with $\lambda(u, 0) = 0$, we say that it is a *modulus of continuity* for u. More generally, we will use the name modulus of continuity for any continuous function $\lambda(u,\varepsilon)$ vanishing at $\varepsilon = 0$[2] such that $\lambda(u,\varepsilon) \geq \|u(\,\cdot\, + \varepsilon) - u\|_p$, where $\|\cdot\|_p$ is the L^p-norm. We

[1] This decomposition is often called the Jordan decomposition of u.

[2] This is *not* an exponent, but a footnote! Clearly, $\lambda(u,\varepsilon)$ is a modulus of continuity if and only if $\lambda(u,\varepsilon) = o\,(1)$ as $\varepsilon \to 0$.

will need a convenient characterization of total variation (in one variable), which is described in the following lemma.

Lemma A.1. *Let u be a function in L^1. If $\lambda(u,\varepsilon)/\,|\varepsilon|$ is bounded as a function of ε, then u is in BV and*

$$\text{T.V.}\,(u) = \lim_{\varepsilon\to 0} \frac{\lambda(u,\varepsilon)}{|\varepsilon|}. \tag{A.7}$$

Conversely, if u is in BV, then $\lambda(u,\varepsilon)/\,|\varepsilon|$ is bounded, and thus (A.7) *holds. In particular, we shall frequently use*

$$\lambda(u,\varepsilon) \leq |\varepsilon|\,\text{T.V.}\,(u) \tag{A.8}$$

if u is in BV.

Proof. Assume first that u is a smooth function. Let $\{x_i\}$ be a partition of the interval in question. Then

$$|u\,(x_i) - u\,(x_{i-1})| = \left| \int_{x_{i-1}}^{x_i} u'(x)\, dx \right| \leq \lim_{\varepsilon\to 0} \int_{x_{i-1}}^{x_i} \left| \frac{u(x+\varepsilon) - u(x)}{\varepsilon} \right| dx.$$

Summing this over i we get

$$\text{T.V.}\,(u) \leq \liminf_{\varepsilon\to 0} \frac{\lambda(u,\varepsilon)}{|\varepsilon|}$$

for differentiable functions $u(x)$. Let u be an arbitrary bounded function in L^1, and u_k be a sequence of smooth functions such that $u_k(x) \to u(x)$ for almost all x, and $\|u_k - u\|_1 \to 0$. The triangle inequality shows that

$$|\lambda\,(u_k,\varepsilon) - \lambda\,(u,\varepsilon)| \leq 2\|u_k - u\|_1 \to 0.$$

Let $\{x_i\}$ be a partition of the interval. We can now choose u_k such that $u_k\,(x_i) = u\,(x_i)$ for all i. Then

$$\sum |u\,(x_i) - u\,(x_{i-1})| \leq \liminf_{\varepsilon\to 0} \frac{\lambda(u_k,\varepsilon)}{|\varepsilon|}.$$

Therefore,

$$\text{T.V.}\,(u) \leq \liminf_{\varepsilon\to 0} \frac{\lambda(u,\varepsilon)}{|\varepsilon|}.$$

Furthermore, we have

$$\begin{aligned}\int |u(x+\varepsilon)-u(x)|\,dx &= \sum_j \int_{(j-1)\varepsilon}^{j\varepsilon} |u(x+\varepsilon)-u(x)|\,dx\\ &= \sum_j \int_0^{\varepsilon} |u(x+j\varepsilon)-u(x+(j-1)\varepsilon)|\,dx\\ &= \int_0^{\varepsilon} \sum_j |u(x+j\varepsilon)-u(x+(j-1)\varepsilon)|\,dx\\ &\le \int_0^{\varepsilon} \mathrm{T.V.}\,(u)\\ &= |\varepsilon|\,\mathrm{T.V.}\,(u)\,.\end{aligned}$$

Thus we have proved the inequalities

$$\frac{\lambda(u,\varepsilon)}{|\varepsilon|} \le \mathrm{T.V.}\,(u) \le \liminf_{\varepsilon\to 0} \frac{\lambda(u,\varepsilon)}{|\varepsilon|} \le \limsup_{\varepsilon\to 0} \frac{\lambda(u,\varepsilon)}{|\varepsilon|} \le \mathrm{T.V.}\,(u)\,, \quad \text{(A.9)}$$

which imply the lemma. □

Observe that we trivially have

$$\tilde\lambda(u,\varepsilon) := \sup_{|\sigma|\le|\varepsilon|} \lambda(u,\sigma) \le |\varepsilon|\,\mathrm{T.V.}\,(u)\,. \quad \text{(A.10)}$$

For functions in L^p care has to be taken as to which points are used in the supremum, since these functions in general are not defined pointwise. The right choice here is to consider only points x_i that are points of *approximate continuity*[3] of u. Lemma A.1 remains valid.

We include a useful characterization of total variation.

Theorem A.2. *Let u be a function in $L^1(I)$ where I is an interval. Assume $u \in BV(I)$. Then*

$$\mathrm{T.V.}\,(u) = \sup_{\phi\in C_0^1(I),\,|\phi|\le 1} \int_I u(x)\phi_x(x)\,dx. \quad \text{(A.11)}$$

Conversely, if the right-hand side of (A.11) *is finite for an integrable function u, then $u \in BV(I)$ and* (A.11) *holds.*

[3] A function u is said to be approximately continuous at x if there exists a measurable set A such that $\lim_{r\to 0} \|[x-r,x+r]\cap A\|/\|[x-r,x+r]\| = 1$ (here $|B|$ denotes the measure of the set B), and u is continuous at x relative to A. (Every Lebesgue point is a point of approximate continuity.) The supremum (A.1) is then called the essential variation of the function. However, in the theory of conservation laws it is customary to use the name total variation in this case, too, and we will follow this custom here.

Proof. Assume that u has finite total variation on I. Let ω be a nonnegative function bounded by unity with support in $[-1,1]$ and unit integral. Define

$$\omega_\varepsilon(x) = \frac{1}{\varepsilon}\omega\Big(\frac{x}{\varepsilon}\Big),$$

and

$$u^\varepsilon = \omega_\varepsilon * u. \quad \text{(A.12)}$$

Consider points $x_1 < x_2 < \cdots < x_n$ in I. Then

$$\begin{aligned}&\sum_i |u^\varepsilon(x_i)-u^\varepsilon(x_{i-1})|\\ &\qquad\le \int_{-\varepsilon}^{\varepsilon} \omega_\varepsilon(x) \sum_i |u(x_i-x)-u(x_{i-1}-x)|\,dx\\ &\qquad\le \mathrm{T.V.}\,(u)\,. \end{aligned}\quad \text{(A.13)}$$

Using (A.5) and (A.13) we obtain

$$\begin{aligned}\int |(u^\varepsilon)'(x)|\,dx &= \mathrm{T.V.}\,(u^\varepsilon)\\ &= \sup \sum_i |u^\varepsilon(x_i)-u^\varepsilon(x_{i-1})|\\ &\le \mathrm{T.V.}\,(u)\,.\end{aligned}$$

Let $\phi \in C_0^1$ with $|\phi| \le 1$. Then

$$\begin{aligned}\int u^\varepsilon(x)\phi'(x)\,dx &= -\int (u^\varepsilon)'(x)\phi(x)\,dx\\ &\le \int |(u^\varepsilon)'(x)|\,dx\\ &\le \mathrm{T.V.}\,(u)\,,\end{aligned}$$

which proves the first part of the theorem.

Now let u be such that

$$\|Du\| := \sup_{\substack{\phi\in C_0^1\\ |\phi|\le 1}} \int u(x)\phi_x(x)\,dx < \infty.$$

First we infer that

$$\begin{aligned}-\int (u^\varepsilon)'(x)\phi(x)\,dx &= \int u^\varepsilon(x)\phi'(x)\,dx\\ &= -\int (\omega_\varepsilon * u)(x)\phi'(x)\,dx\\ &= -\int u(x)(\omega_\varepsilon * \phi)'(x)\,dx\\ &\le \|Du\|\,.\end{aligned}$$

Using that (see Exercise A.1)

$$\|f\|_1 = \sup_{\substack{\phi \in C_0^1, \\ |\phi| \le 1}} \int f(x)\phi(x)\, dx,$$

we conclude that

$$\int |(u^\varepsilon)'(x)|\, dx \le \|Du\|. \tag{A.14}$$

Next we show that $u \in L^\infty$. Choose a sequence $u_j \in BV \cap C^\infty$ such that (see, e.g., [47, p. 172])

$$u_j \to u \text{ a.e.}, \quad \|u_j - u\|_1 \to 0, \quad j \to \infty, \tag{A.15}$$

and

$$\int \left|u_j'(x)\right|\, dx \to \|Du\|, \quad j \to \infty. \tag{A.16}$$

For any y, z we have

$$u_j(z) = u_j(y) + \int_y^z u_j'(x)\, dx.$$

Averaging over some bounded interval $J \subseteq I$ we obtain

$$|u_j| \le \frac{1}{|J|} \int_J |u_j(y)|\, dy + \int_I \left|u_j'(x)\right|\, dx, \tag{A.17}$$

which shows that the u_j are uniformly bounded, and hence $u \in L^\infty$. Thus

$$u^\varepsilon(x) \to u(x)$$

as $\varepsilon \to 0$ at each point of approximate continuity of u. Using points of approximate continuity $x_1 < x_2 < \cdots < x_n$ we conclude that

$$\begin{aligned} \sum_i |u(x_i) - u(x_{i-1})| &= \lim_{\varepsilon \to 0} \sum_i |u^\varepsilon(x_i) - u^\varepsilon(x_{i-1})| \\ &\le \limsup_{\varepsilon \to 0} \int |(u^\varepsilon)'(x)|\, dx \\ &\le \|Du\|. \end{aligned} \tag{A.18}$$

□

For a function u of two variables (x, y) the total variation is defined by

$$\mathrm{T.V.}_{x,y}(u) = \int \mathrm{T.V.}_x(u)(y)\, dy + \int \mathrm{T.V.}_y(u)(x)\, dx. \tag{A.19}$$

The extension to functions of n variables is obvious.

Total variation is used to obtain compactness. The appropriate compactness statement is Kolmogorov's compactness theorem. We say that a subset M of a complete metric space X is *(strongly) compact* if any infinite subset of it contains a (strongly) convergent sequence. A set is *relatively compact* if its closure is compact. A subset of a metric space is called *totally bounded* if it is contained in a finite union of balls of radius ε for any $\varepsilon > 0$ (we call this finite union an ε-net). Our starting theorem is the following result.

Theorem A.3. *A subset M of a complete metric space X is relatively compact if and only if it is totally bounded.*

Proof. Consider first the case where M is relatively compact. Assume that there exists an ε_0 for which there is no finite ε_0-net. For any element $u_1 \in M$ there exists an element $u_2 \in M$ such that $\|u_1 - u_2\| \ge \varepsilon_0$. Since the set $\{u_1, u_2\}$ is not an ε_0-net, there has to be an $u_2 \in M$ such that $\|u_1 - u_3\| \ge \varepsilon_0$ and $\|u_2 - u_3\| \ge \varepsilon_0$. Continuing inductively construct a sequence $\{u_j\}$ such that

$$\|u_j - u_k\| \ge \varepsilon_0, \quad j \ne k,$$

which clearly cannot have a convergent subsequence, which yields a contradiction. Hence we conclude that there has to exist an ε-net for every ε.

Assume now that we can find a finite ε-net for M for every $\varepsilon > 0$, and let M_1 be an arbitrary infinite subset of M. Construct an ε-net for M_1 with $\varepsilon = \frac{1}{2}$, say $\{u_1^{(1)}, \dots, u_{N_1}^{(1)}\}$. Now let $M_1^{(j)}$ be the set of those $u \in M_1$ such that $\|u - u_j^{(1)}\| \le \frac{1}{4}$. At least one of $M_1^{(1)}, \dots, M_1^{(N_1)}$ has to be infinite, since M_1 is infinite. Denote (one of) this by M_2 and the corresponding element u_2. On this set we construct an ε-net with $\varepsilon = \frac{1}{4}$. Continuing inductively we construct a nested sequence of subsets $M_{k+1} \subset M_k$ for $k \in \mathbb{N}$ such that M_k has an ε-net with $\varepsilon = 1/2^k$, say $\left\{u_1^{(k)}, \dots, u_{N_k}^{(k)}\right\}$. For arbitrary elements u, v of M_k we have $\|u - v\| \le \|u - u_k\| + \|u_k - v\| \le 1/2^{k-1}$. The sequence $\{u_k\}$ with $u_k \in M_k$ is convergent, since

$$\|u_{k+m} - u_k\| \le \frac{1}{2^{k-1}},$$

proving that M_1 contains a convergent sequence. □

A result that simplifies our argument is the following.

Lemma A.4. *Let M be a subset of a metric space X. Assume that for each $\varepsilon > 0$, there is a totally bounded set A such that* $\operatorname{dist}(f, A) < \varepsilon$ *for each $f \in M$. Then M is totally bounded.*

Proof. Let A be such that $\operatorname{dist}(f, A) < \varepsilon$ for each $f \in M$. Since A is totally bounded, there exist points $x_1, \dots, x_n$ in X such that $A \subseteq \cup_{j=1}^n \mathcal{B}_\varepsilon(x_j)$, where

$$\mathcal{B}_\varepsilon(y) = \{z \in X \mid \|z - y\| \le \varepsilon\}.$$

For any $f \in M$ there exists by assumption some $a \in A$ such that $\|a - f\| < \varepsilon$. Furthermore, $\|a - x_j\| < \varepsilon$ for some j. Thus $\|f - x_j\| < 2\varepsilon$, which proves

$$M \subseteq \bigcup_{j=1}^{n} \mathcal{B}_{2\varepsilon}(x_j).$$

Hence M is totally bounded. □

We can state and prove Kolomogorov's compactness theorem.

Theorem A.5 (Kolmogorov's compactness theorem). *Let M be a subset of $L^p(\Omega)$, $p \in [1, \infty)$, for some open set $\Omega \subseteq \mathbb{R}^n$. Then M is relatively compact if and only if the following three conditions are fulfilled:*

(i) *M is bounded in $L^p(\Omega)$, i.e.,*

$$\sup_{u \in M} \|u\|_p < \infty.$$

(ii) *We have*

$$\|u(\,\cdot\, + \varepsilon) - u\|_p \le \lambda(|\varepsilon|)$$

for a modulus of continuity λ that is independent of $u \in M$ (we let u equal zero outside Ω).

(iii)

$$\lim_{\alpha \to \infty} \int_{\{x \in \Omega \mid |x| \ge \alpha\}} |u(x)|^p \, dx = 0 \textit{ uniformly for } u \in M.$$

Remark A.6. In the case Ω is bounded, condition **(iii)** is clearly superfluous.

Proof. We start by proving that conditions **(i)**–**(iii)** are sufficient to show that M is relatively compact. Let φ be a nonnegative and continuous function such that $\varphi \le 1$, $\varphi(x) = 1$ on $|x| \le 1$, and $\varphi(x) = 0$ whenever $|x| \ge 2$. Write $\varphi_r(x) = \varphi(x/r)$. From condition **(iii)** we see that $\|\varphi_r u - u\| \to 0$ as $r \to \infty$. Using Lemma A.4 we see that it suffices to show that $M_r = \{\varphi_r u \mid u \in M\}$ is totally bounded. Furthermore, we see that M_r satisfies **(i)** and **(ii)**. In other words, we need to prove only that **(i)** and **(ii)** together with the existence of some R so that $u = 0$ whenever $u \in M$ and $|x| \ge R$ imply that M is totally bounded. Let ω_ε be a mollifier, that is,

$$\omega \in C_0^\infty, \quad 0 \le \omega \le 1, \quad \int \omega \, dx = 1, \quad \omega_\varepsilon(x) = \frac{1}{\varepsilon^n} \omega\Big(\frac{x}{\varepsilon}\Big).$$

Then

$$\|u * \omega_\varepsilon - u\|_p^p = \int |u * \omega_\varepsilon(x) - u(x)|^p \, dx$$

$$\begin{aligned}
&= \int \left| \int_{\mathcal{B}_\varepsilon} \big(u(x-y) - u(x)\big)\omega_\varepsilon(y) \, dy \right|^p dx \\
&\le \int \int_{\mathcal{B}_\varepsilon} |u(x-y) - u(x)|^p \, dy \, \|\omega_\varepsilon\|_q^p \, dx \\
&= \varepsilon^{np/q-p} \|\omega\|_q^p \int_{\mathcal{B}_\varepsilon} \int |u(x-y) - u(x)|^p \, dx \, dy \\
&\le \varepsilon^{np/q-p} \|\omega\|_q^p \int_{\mathcal{B}_\varepsilon} \max_{|z| \le \varepsilon} \lambda(|z|) \, dy \\
&= \varepsilon^{n+np/q-p} \|\omega\|_q^p \, |\mathcal{B}_1| \max_{|z| \le \varepsilon} \lambda(|z|),
\end{aligned}$$

where $1/p + 1/q = 1$ and

$$\mathcal{B}_\varepsilon = \mathcal{B}_\varepsilon(0) = \{z \in \mathbb{R}^n \mid \|z\| \le \varepsilon\}.$$

Thus

$$\|u * \omega_\varepsilon - u\|_p \le \varepsilon^{n-1} \|\omega\|_q \, |\mathcal{B}_1|^{1/p} \max_{|z| \le \varepsilon} \lambda(|z|), \tag{A.20}$$

which together with **(ii)** proves uniform convergence as $\varepsilon \to 0$ for $u \in M$. Using Lemma A.4 we see that it suffices to show that $N_\varepsilon = \{u * \omega_\varepsilon \mid u \in M\}$ is totally bounded for any $\varepsilon > 0$.

Hölder's inequality yields

$$|u * \omega_\varepsilon(x)| \le \|u\|_p \|\omega_\varepsilon\|_q,$$

so by **(i)**, functions in N_ε are uniformly bounded. Another application of Hölder's inequality implies

$$\begin{aligned}
|u * \omega_\varepsilon(x) - u * \omega_\varepsilon(y)| &= \left| \int \big(u(x-z) - u(y-z)\big)\omega_\varepsilon(z) \, dz \right| \\
&\le \|u(\,\cdot\, + x - y) - u\|_p \, \|\omega_\varepsilon\|_q,
\end{aligned}$$

which together with **(ii)** proves that N_ε is equicontinuous. The Arzela–Ascoli theorem implies that N_ε is relatively compact, and hence totally bounded in $C(\mathcal{B}_{R+r})$. Since the natural embedding of $C(\mathcal{B}_{R+r})$ into $L^p(\mathbb{R}^n)$ is bounded, it follows that N_ε totally bounded in $L^p(\mathbb{R}^n)$ as well. Thus we have proved that conditions **(i)**–**(iii)** imply that M is relatively compact.

To prove the converse, we assume that M is relatively compact. Condition **(i)** is clear. Now let $\varepsilon > 0$. Since M is relatively compact, we can find functions $u_1, \dots, u_m$ in $L^p(\mathbb{R}^n)$ such that

$$M \subseteq \bigcup_{j=1}^{m} \mathcal{B}_\varepsilon(u_j).$$

Furthermore, since $C_0(\mathbb{R}^n)$ is dense in $L^p(\mathbb{R}^n)$, we may as well assume that $u_j \in C_0(\mathbb{R}^n)$. Clearly, $\|u_j(\,\cdot\, + y) - u_j\|_p \to 0$ as $y \to 0$, and so there is some $\delta > 0$ such that $\|u_j(\cdot + y) - u_j\|_p \le \varepsilon$ whenever $|y| < \delta$. If $u \in M$ and $|y| < \delta$, then pick some j such that $\|u - u_j\|_p < \varepsilon$, and obtain

$$\begin{aligned}\|u(\cdot + z) - u\|_p &\le \|u(\cdot + z) - u_j(\cdot + z)\|_p \\ &\quad + \|u_j(\cdot + z) - u_j\|_p + \|u_j - u\|_p \\ &= 2\|u_j - u\|_p + \|u_j(\cdot + z) - u_j\|_p \\ &\le 3\varepsilon,\end{aligned}$$

proving **(ii)**.

When r is large enough, $\chi_{B_r} u_j = u_j$ for all j, and then, with the same choice of j as above, we obtain

$$\|\chi_{B_r} u - u\|_p \le \|\chi_{B_r}(u - u_j)\|_p + \|u - u_j\|_p \le 2\|u - u_j\|_p \le 2\varepsilon,$$

which proves **(iii)**. □

Helly's theorem is a simple corollary of Kolmogorov's compactness theorem.

Corollary A.7 (Helly's theorem). *Let $\{h^\delta\}$ be a sequence of functions defined on an interval $[a, b]$, and assume that this sequence satisfies*

$$\mathrm{T.V.}\left(h^\delta\right) < M, \qquad \textit{and} \qquad \left\|h^\delta\right\|_\infty < M,$$

where M is some constant independent of δ. Then there exists a subsequence h^{δ_n} that converges almost everywhere to some function h of bounded variation.

Proof. It suffices to apply (A.8) (for $p = 1$) together with the boundedness of the total variation to show that condition **(ii)** in Kolmogorov's compactness theorem is satisfied. □

We remark that one can prove that the convergence in Helly's theorem is at every point, not only almost everywhere; see Exercise A.2.

The application of Kolmogorov's theorem in the context of conservation laws relies on the following result.

Theorem A.8. *Let $u_\eta \colon \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ be a family of functions such that for each positive T,*

$$|u_\eta(x, t)| \le C_T, \quad (x, t) \in \mathbb{R}^n \times [0, T]$$

for a constant C_T independent of η. Assume in addition for all compact $B \subset \mathbb{R}^n$ and for $t \in [0, T]$ that

$$\sup_{|\xi| \le |\rho|} \int_B |u_\eta(x + \xi, t) - u_\eta(x, t)|\, dx \le \nu_{B,T}(|\rho|),$$

for a modulus of continuity ν. Furthermore, assume for s and t in $[0, T]$ that

$$\int_B |u_\eta(x, t) - u_\eta(x, s)|\, dx \le \omega_{B,T}(|t - s|) \text{ as } \eta \to 0,$$

for some modulus of continuity ω_T. Then there exists a sequence $\eta_j \to 0$ such that for each $t \in [0, T]$ the function $\{u_{\eta_j}(t)\}$ converges to a function $u(t)$ in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$. The convergence is in $C([0, T]; L^1_{\mathrm{loc}}(\mathbb{R}^n))$.

Proof. Kolmogorov's theorem implies that for each fixed $t \in [0, T]$ and for any sequence $\eta_j \to 0$ there exists a subsequence (still denoted by η_j) $\eta_j \to 0$ such that $\{u_{\eta_j}(t)\}$ converges to a function $u(t)$ in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$.

Consider now a dense countable subset E of the interval $[0, T]$. By possibly taking a further subsequence (which we still denote by $\{u_{\eta_j}\}$) we find that

$$\int_B \left|u_{\eta_j}(x, t) - u(x, t)\right| dx \to 0 \text{ as } \eta_j \to 0, \text{ for } t \in E.$$

Now let $\varepsilon > 0$ be given. Then there exists a positive δ such that $\omega_{B,T}(\tilde\delta) \le \varepsilon$ for all $\tilde\delta \le \delta$. Fix $t \in [0, T]$. We can find a $t_k \in E$ with $|t_k - t| \le \delta$. Thus

$$\int_B |u_{\tilde\eta}(x, t) - u_{\tilde\eta}(x, t_k)|\, dx \le \omega_{B,T}(|t - t_k|) \le \varepsilon \text{ for } \tilde\eta \le \eta$$

and

$$\int_B \left|u_{\eta_{j_1}}(x, t_k) - u_{\eta_{j_2}}(x, t_k)\right| dx \le \varepsilon \text{ for } \eta_{j_1}, \eta_{j_2} \le \eta \text{ and } t_k \in E.$$

The triangle inequality yields

$$\begin{aligned}&\int_B \left|u_{\eta_{j_1}}(x, t) - u_{\eta_{j_2}}(x, t)\right| dx \\ &\quad\le \int_B \left|u_{\eta_{j_1}}(x, t) - u_{\eta_{j_1}}(x, t_k)\right| dx + \int_B \left|u_{\eta_{j_1}}(x, t_k) - u_{\eta_{j_2}}(x, t_k)\right| dx \\ &\qquad + \int_B \left|u_{\eta_{j_2}}(x, t_k) - u_{\eta_{j_2}}(x, t)\right| dx \\ &\quad\le 3\varepsilon,\end{aligned}$$

proving that for each $t \in [0, T]$ we have that $u_\eta(t) \to u(t)$ in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$. The bounded convergence theorem then shows that

$$\sup_{t \in [0,T]} \int_B |u_\eta(x, t) - u(x, t)|\, dx\, dt \to 0 \text{ as } \eta \to 0,$$

thereby proving the theorem. □

but with different flux functions. Let u and v be the weak solutions of

$$u_t + f(u)_x = 0, \quad v_t + g(v)_x = 0 \tag{2.62}$$

with initial data

$$u(x,0) = v(x,0) = \begin{cases} u_l & \text{for } x < 0, \\ u_r & \text{for } x > 0. \end{cases}$$

We assume that both f and g are continuous and piecewise linear with a finite number of breakpoints. The solutions u and v of (2.62) will be piecewise constant functions of x/t that are equal outside a finite interval in x/t. We need to estimate the difference in L^1 between the two solutions.

Lemma 2.11. *The following inequality holds*:

$$\frac{d}{dt}\|u-v\|_1 \le \sup_u |f'(u) - g'(u)|\, |u_l - u_r|\,, \tag{2.63}$$

where the supremum is over all u between u_l and u_r.

Proof. Assume that $u_l \le u_r$; the case $u_l \ge u_r$ is similar. Consider first the case where f and g both are convex. Without loss of generality we may assume that f and g have common breakpoints $u_l = w_1 < w_2 < \cdots < w_n = u_r$, and let the speeds be denoted by

$$f'|_{\langle w_j, w_{j+1}\rangle} = s_j \text{ and } g'|_{\langle w_j, w_{j+1}\rangle} = \tilde{s}_j.$$

Then

$$\int_{u_l}^{u_r} |f'(u) - g'(u)|\, du = \sum_{j=1}^{n-1} |s_j - \tilde{s}_j|\,(w_{j+1} - w_j).$$

Let σ_j be an ordering, that is, $\sigma_j < \sigma_{j+1}$, of all the speeds $\{s_j, \tilde{s}_j\}$. Then we may write

$$u(x,t)|_{x\in\langle\sigma_j t, \sigma_{j+1} t\rangle} = u_{j+1},$$
$$v(x,t)|_{x\in\langle\sigma_j t, \sigma_{j+1} t\rangle} = v_{j+1},$$

where both u_{j+1} and v_{j+1} are from the set of all possible breakpoints, namely $\{w_1, w_2, \dots, w_n\}$, and $u_j \le u_{j+1}$ and $v_j \le v_{j+1}$. Thus

$$\|u(\,\cdot\,,t) - v(\,\cdot\,,t)\|_1 = t\sum_{j=1}^{m} |u_{j+1} - v_{j+1}|\,(\sigma_{j+1} - \sigma_j).$$

We easily see that

$$\begin{aligned}\frac{d}{dt}\|u(\,\cdot\,,t) - v(\,\cdot\,,t)\|_1 &= \int_{u_l}^{u_r} |f'(u) - g'(u)|\, du \\ &\le \sup_u |f'(u) - g'(u)|\, |u_l - u_r|\,.\end{aligned}$$

The case where f and g are not necessarily convex is more involved. We will show that

$$\int_{u_l}^{u_r} |f'_{\smile}(u) - g'_{\smile}(u)|\, du \le \int_{u_l}^{u_r} |f'(u) - g'(u)|\, du \tag{2.64}$$

when the convex envelopes are taken on the interval $[u_l, u_r]$. To this end we use the following general lemma:

Lemma 2.12 (Crandall–Tartar). *Let D be a subset of $L^1(\Omega)$, where Ω is some measure space. Assume that if ϕ and ψ are in D, then also $\phi\vee\psi = \max\{\phi,\psi\}$ is in D. Assume furthermore that there is a map $T\colon D \to L^1(\Omega)$ such that*

$$\int_\Omega T(\phi) = \int_\Omega \phi, \quad \phi \in D.$$

Then the following statements, valid for all $\phi, \psi \in D$, are equivalent:

(i) *If $\phi \le \psi$, then $T(\phi) \le T(\psi)$.*

(ii) *$\int_\Omega (T(\phi) - T(\psi))^+ \le \int_\Omega (\phi - \psi)^+$, where $\phi^+ = \phi \vee 0$.*

(iii) *$\int_\Omega |T(\phi) - T(\psi)| \le \int_\Omega |\phi - \psi|$.*

Proof of Lemma 2.12. For completeness we include a proof of the lemma. Assume (i). Then $T(\phi\vee\psi) - T(\phi) \ge 0$, which trivially implies $T(\phi) - T(\psi) \le T(\phi\vee\psi) - T(\psi)$, and thus $(T(\phi) - T(\psi))^+ \le T(\phi\vee\psi) - T(\psi)$. Furthermore,

$$\int_\Omega (T(\phi) - T(\psi))^+ \le \int_\Omega (T(\phi\vee\psi) - T(\psi)) = \int_\Omega (\phi\vee\psi - \psi) = \int_\Omega (\phi - \psi)^+,$$

proving (ii). Assume now (ii). Then

$$\begin{aligned}\int_\Omega |T(\phi) - T(\psi)| &= \int_\Omega (T(\phi) - T(\psi))^+ + \int_\Omega (T(\psi) - T(\phi))^+ \\ &\le \int_\Omega (\phi - \psi)^+ + \int_\Omega (\psi - \phi)^+ \\ &= \int_\Omega |\phi - \psi|\,,\end{aligned}$$

which is (iii). It remains to prove that (iii) implies (i). Let $\phi \le \psi$. For real numbers we have $x^+ = (|x| + x)/2$. This implies

$$\begin{aligned}\int_\Omega (T(\phi) - T(\psi))^+ &= \frac{1}{2}\int_\Omega |T(\phi) - T(\psi)| + \frac{1}{2}\int_\Omega (T(\phi) - T(\psi)) \\ &\le \int_\Omega |\phi - \psi| + \int_\Omega (\phi - \psi) = 0.\end{aligned}$$

□

To apply this lemma in our context, we let D be the set of all piecewise constant functions on $[u_l, u_r]$. For any piecewise linear and continuous flux