双曲型偏微分方程数值方法

汤华中

北京大学数学科学学院

July 7, 2019

Contents

1 引言

力学中的很多方程是双曲的, 所以对双曲方程(组)的研究具有重要的科学意义.

模型双曲方程为波方程或波动方程. 一维波动方程可写为

$$u_{tt} - c^2 u_{xx} = 0,\ u = u(x,t), \tag{1.1}$$

这里c是正实数.

d'Alembert公式

$$u(x,t) = \frac{f(x-ct) + f(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\ ds, \tag{1.2}$$

给出(1.1)的满足下列初始条件的解

$$u(x,0) = f(x),\quad u_t(x,0) = g(x).$$

由(1.2)知, 在点(x_0,t_0)处的解依赖于:

- $f(x)$在点$x_0 + ct_0$ 和$x_0 - ct_0$的值,
- $g(x)$在区间$x_0 - ct_0 \le x \le x_0 + ct_0$内的值.

区间$[x_0-ct_0, x_0+ct_0]$为u在(x_0, t_0)处的依赖区间. 由(x_0, t_0), $(x_0-ct_0, 0)$, $(x_0 + ct_0, 0)$构成的三角形内区域为u在(x_0, t_0)处的依赖区域. u在(x_0, t_0)处的影响区域: $\{(x, t)|t \ge t_0, x \in [x_0-c(t-t_0), x_0+c(t - t_0)]$.

利用线性自变量变换, 除了低阶项(对方程的定性理解不重要)外, 任何如下形式的方程($B^2 - AC > 0$)

$$A\frac{\partial^2 u}{\partial x^2} + 2B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + (\text{低阶导数项}) = 0,$$

都可以变成(1.1). 这个定义类似于平面双曲线的定义.

双曲型PDE的其它例子有:

- 对流方程 /输运方程

$$u_t + a u_x = 0,\ a \in \mathbb{R},\ u = u(x,t).$$

- Burgers 方程

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0, \quad u = u(x,t).$$

- 电磁波方程

$$\left(v_{ph}^2 \nabla^2 - \frac{\partial^2}{\partial t^2}\right) \mathbf{E} = \mathbf{0}$$

$$\left(v_{ph}^2 \nabla^2 - \frac{\partial^2}{\partial t^2}\right) \mathbf{B} = \mathbf{0}$$

其中$v_{ph} = \dfrac{1}{\sqrt{\mu\varepsilon}}$是在磁导率为$\mu$和介电常数为$\varepsilon$的介质中的光速(即相速度).

- Maxwell 方程组

$$\nabla \cdot \mathbf{E} = 0 \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t},$$
$$\nabla \cdot \mathbf{B} = 0 \qquad \nabla \times \mathbf{B} = \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}.$$

它们分别为高斯定律, 法拉第感应定律, 高斯磁定律, 安培环路定律.

- Euler方程组

$$\frac{\partial \boldsymbol{U}}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \boldsymbol{F_j}(\boldsymbol{U}) = 0,$$

其中$\boldsymbol{U} = (\rho, \rho u_1, \cdots, \rho u_d, E)^T$, $p = p(\rho, e)$,

$$\boldsymbol{F}_j = (\rho, \rho u_1 u_j, \cdots, \rho u_{j-1} u_j, \rho u_{j-1} u_j + p, \rho u_{j+1} u_j, \cdots, \rho u_d u_j, (E+p) u_j)^T.$$

- Dirac方程

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = \left(\beta m c^2 + c \left(\sum_{n=1}^{3} \alpha_n p_n \right) \right) \psi(x,t).$$

$\hbar$等于Planck常数除以2π, p_1, p_2, p_3为动量分量, c为光速, m为电子的静止质量. 4×4 矩阵α_k, β都是Hermitian, 满足

$$\alpha_i^2 = \beta^2 = I_4, \ \alpha_i \alpha_j + \alpha_j \alpha_i = 0, \alpha_i \beta + \beta \alpha_i = 0 \alpha_i \beta + \beta \alpha_i = 0.$$

- 弹塑性流体力学方程组 .....

守恒是物理世界的基本原理: 物质可四处运动并重新分布, 但它不会出现或消失. 科学与工程中很多问题(如流体、气体动力学、交通流、相对论流体等)的数学描述是守恒律方程, 时间依赖的PDE, 通常是双曲的, 非线性的, 但具有一个简单形式. 在过去的四十多年中, 我们看到了与非线性守恒律的数值方法有关的大量活动. 这种活动已经影响到应用科学的各个分支, 从航空到石油勘探、图像处理等. Because there exist many special difficulties associated with solving these equations (e.g. shock wave formation and discontinuous solutions etc.) and numerical methods based on simple finite-difference or continuous finite element approximations may behave well for smooth solutions but can give disastrous results when discontinuities or shock waves are present, the study of numerical methods for hyperbolic conservations laws is an important, interesting, fascinating, and challenging field of research.

一维标量守恒律方程具有如下散度形式

$$u_t + f(u)_x = 0. \tag{1.3}$$

这里u 是守恒量, f是通量. 这类方程通常描述输运现象等. 在固定空间区间$[a,b]$上对(1.3)积分, 则有

$$\begin{aligned}\frac{d}{dt}\int_a^b u\,dx &= f\big(u(a,t)\big) - f\big(u(b,t)\big)\\ &= [\text{左端}a\text{处净入流}] - [\text{右端}b\text{处净出流}].\end{aligned}$$

在区间$[a,b]$内u的总量的改变仅是由区间边界点处u的流动引起. 如果$f(u)$ 关于u可微, u可微, 则(1.3)可写为拟线性形式

$$u_t + a(u)u_x = 0, \quad a(u) = f'(u), \tag{1.4}$$

对于光滑解, (1.4)等价于(1.3). 但是, 如果u在某点x_0处不连续, (1.4)中左端的第二项一般不能定义, 因为它是间断函数$a(u)$和

在x_0处含一个Dirac质量的广义导数u_x的乘积, 因此, (1.4)仅在连续函数类中有意义.

On the other hand, working with the equation in the divergence form (1.3) allows us to consider discontinuous solutions as well, interpreted in distributional sense. More precisely, a locally integrable function $u = u(x,t)$ is a *weak solution* of (1.3) provided that

$$\iint \left(u\phi_t + f(u)\phi_x\right)\, dxdt = 0,$$

for any differentiable function with compact support $\phi \in C_0^1$. Unfortunately, the weak solutions may in general be not unique. The correct or physically relevant solution has then to be properly characterized by some principle such as the *entropy condition*, and numerical approximation has to respect such characterization otherwise it would converge to a nonphysical weak solution. The problem of lack of uniqueness for weak solutions is intrinsic

in the theory of hyperbolic conservation laws. The above facts give rise to the difficulty in studying numerical methods for quasilinear hyperbolic conservation laws.

Numerical methods for hyperbolic conservation laws have been rapidly developed in recent decades. Since the TVD (total variation diminishing) concept [25] is presented, various high resolution schemes such as TVD, TVB (total variation bounded), MmB (Maxima minima Bound preserving) [95], ENO (essentially non-oscillatory) [29, 67, 68], WENO (Weighted ENO) [50, 37, 66] schemes, and RKDG (Runge-Kutta discontinuous Galerkin) methods [8] etc. have been developed and applied successfully to many practical problems in fluid dynamics etc. The study of the convergence and stability as well as the error estimation of conservative numerical methods are also important topics and tasks, while the convergence of numerical methods for hyperbolic conservation laws depends on corresponding discrete entropy

condition and some kinds of nonlinear stability such as the total variation stability. However, there exists some relationship between the entropy condition and nonlinear stability of numerical methods. In the 1D scalar case many important issues are well understood now. In practice, numerical schemes may first be designed for the 1D scalar equation and then are extended to the 1D system or multidimensional equation or system, since the difficulties encountered in the 1D system and multidimensional case are already met in the 1D scalar equation. On the other hand, only when they are proved to perform well in this setting, it is safe to extend them to the 1D system or multidimensional case.

2 线性方程的有限差分格式

用有限差商近似微商的有限差分格式简单、高效、灵活、但限于结构网格.

2.1 标量情形

假设a 是常数, 考虑1D对流方程

$$u_t + au_x = 0, \quad x \in \Omega \subseteq \mathbb{R}, \quad t > 0, \tag{2.1}$$

满足$t = 0$时初始条件$u(x, 0) = u_0(x)$的精确解为$u(x, t) = u_0(x - at)$, 它沿着特征线$x - at$ =constant保持常数, 见图1.

给定(x, t) 上半平面的均匀网格剖分$\{(x_j, t_n)|x_j = jh, t = n\tau, j \in \mathbb{Z}, n \in \mathbb{Z}^+ \cup \{0\}\}$, 这里$h, \tau$ 分别为空间和时间方向的网

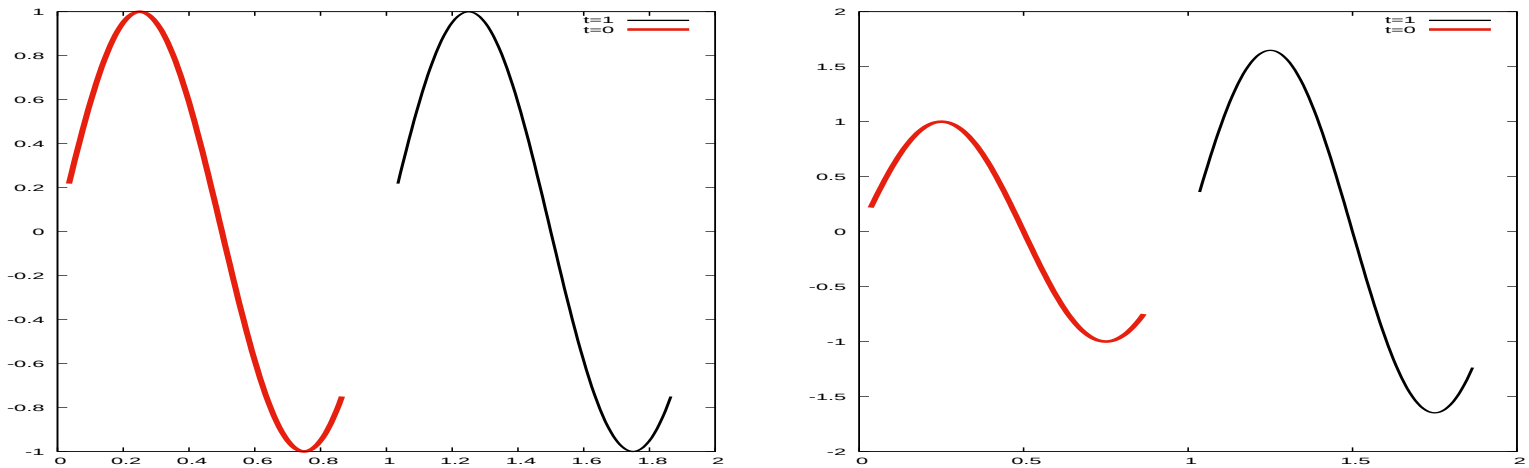


Figure 1: 初值问题$u_t + u_x = bt$, $u(x,0) = \sin(2\pi x)$在$t = 1$时的解$u(x,t) = u_0(x-at)e^{bt}$. 左: $b = 0$; 右: $b = 0.5$.

格步长. 可用差商逐点逼近偏导数u_x为

$$u_x(x_j) = \begin{cases} \frac{1}{2h}\big(u(x_{j+1}) - u(x_{j-1})\big) + \mathcal{O}(h^2), \\ \frac{1}{h}\big(u(x_j) - u(x_{j-1})\big) + \mathcal{O}(h), \\ \frac{1}{h}\big(u(x_{j+1}) - u(x_j)\big) + \mathcal{O}(h), \\ \frac{1}{12h}\big(u(x_{j+2}) - 8u(x_{j+1}) + 8u(x_{j-1}) - u(x_{j-2})\big) + \mathcal{O}(h^4), \\ \big(\frac{25}{12h}u(x_j) - \frac{4}{h}u(x_{j-1}) + \frac{3}{h}u(x_{j-2}) - \frac{4}{3h}u(x_{j-3}) + \frac{1}{4h}u(x_{j-4})\big) + \mathcal{O}(h^4), \end{cases}$$

等等, 这里为了简单起见已经略去了自变量t. 时间偏导数u_t可以类似地近似, Taylor级数展开或多项式拟合等通常用于高阶导数的逼近.

Using u_j^n to approximate the point value $u(x_j, t_n)$, Eq. (2.1) may be approximated at the grid node (x_j, t_n) by the following FDS or difference equations.

- The backward differencing scheme

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_j^n - u_{j-1}^n}{h} = 0, \tag{2.2}$$

- The forward differencing scheme

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_{j+1}^n - u_j^n}{h} = 0, \tag{2.3}$$

- The central differencing scheme

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0, \tag{2.4}$$

- The leap frog scheme

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\tau} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0, \tag{2.5}$$

- The Crank-Nicolson scheme

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_{j+1}^n - u_{j-1}^n}{4h} + a\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{4h} = 0, \quad (2.6)$$

- The Lax-Friedrichs (LF) scheme

$$\frac{u_j^{n+1} - (u_{j+1}^n + u_{j-1}^n)/2}{\tau} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0, \quad (2.7)$$

- The Lax-Wendroff (LW) scheme

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} - \tau a^2\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2h^2} = 0, \quad (2.8)$$

- The Beam-Warming (BW) scheme

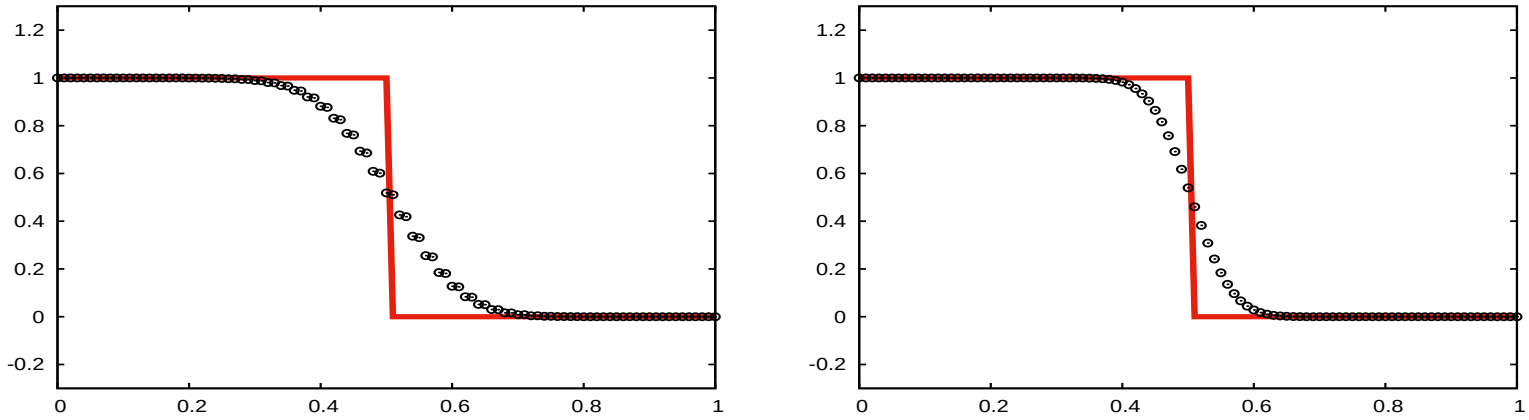
$$\frac{u_j^{n+1}-u_j^n}{\tau}+a\frac{3u_j^n-4u_{j-1}^n+u_{j-2}^n}{2h}-\tau a^2\frac{u_j^n-2u_{j-1}^n+u_{j-2}^n}{2h^2}=0. \tag{2.9}$$

The LF scheme (2.7) is regarded as a modification of (2.4) by replacing u_j^n with $(u_{j+1}^n+u_{j-1}^n)/2$, while the LW scheme (2.8) is derived by replacing the time derivatives in the Taylor series expansion of the solution in time with spatial derivatives via the original equation (2.1), and then approximating spatial derivatives via central differences. The BW scheme (2.9) can also be derived by replacing the time derivatives in the Taylor series expansion of $u(x,t_{n+1})$ in time with spatial derivatives via (2.1), and then approximating spatial derivatives via second-order backward differences.

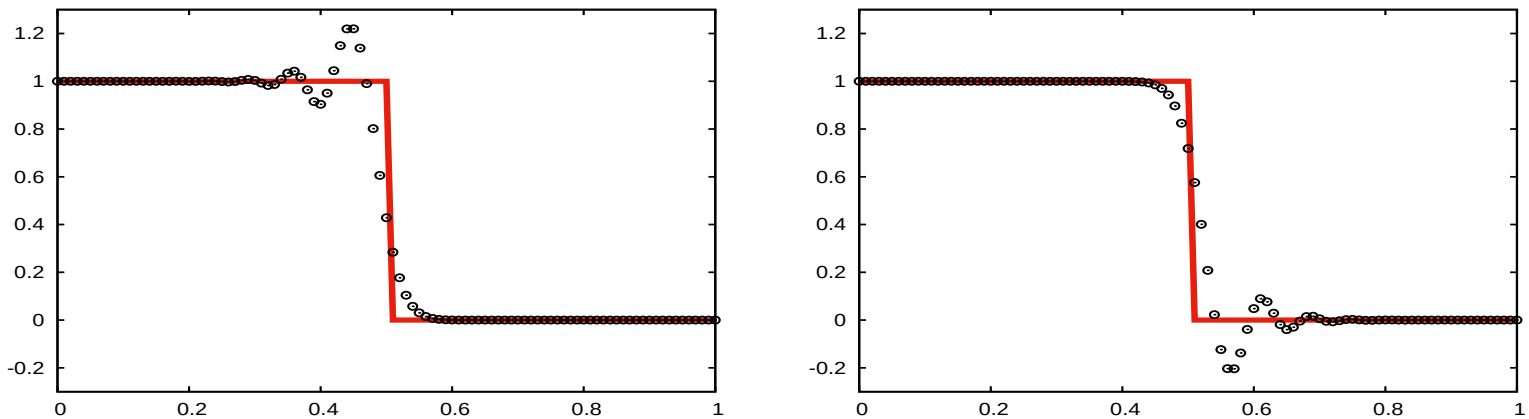
Example 2.1 *Consider the initial value problem of* (2.1) *with initial date*

$$u_0(x) = \begin{cases} 1, & x < 0, \\ 0, & x > 0, \end{cases}$$

and take $a = 1$ *and* $\tau/h = 0.5$ *for different numerical schemes: (a) Lax-Friedrichs, (b) upwind scheme, (c) Lax-Wendroff and (d) Beam-Warming. Figs. 2-3 plot the resulting numerical and exact solution* $u(x,t) = u_0(x - at)$ *at time* $t = 0.5$ *for the cases of* $h = 0.01$ *and* $h = 0.0025$ *for the computational domain* $[0, 1]$. *They shows their different behaviours in resolving the discontinuity and the resolution can be improved as the mesh is refined.*

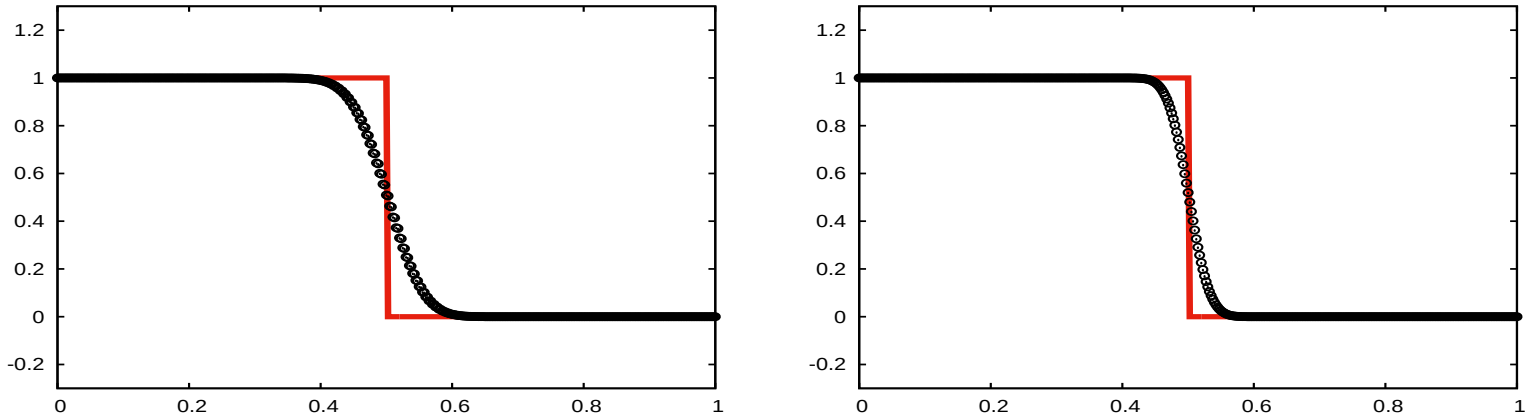


(a) Left: LF scheme (2.7); right: Upwind scheme (2.2).

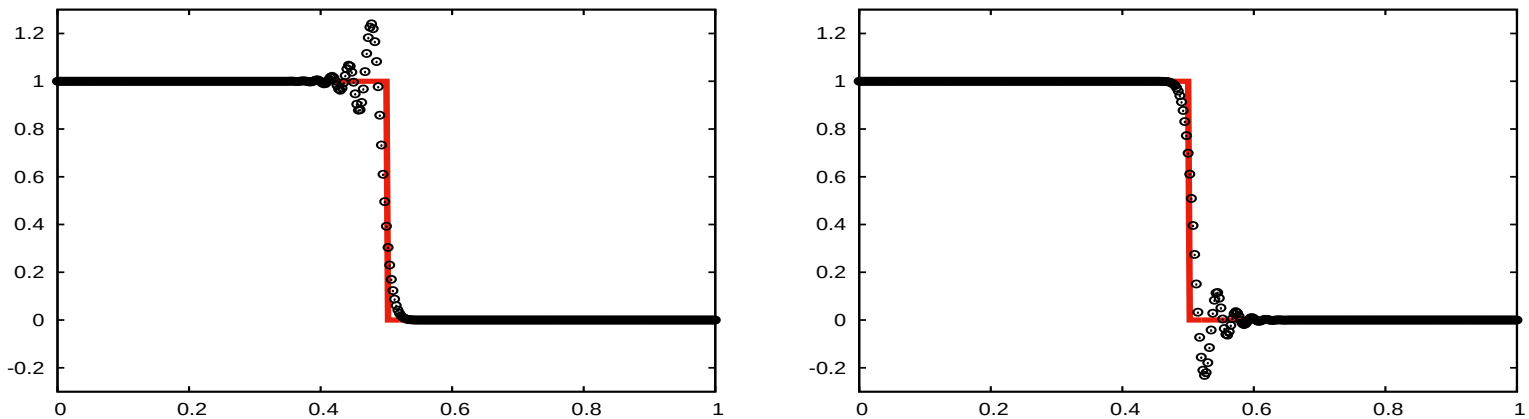


(b) Left: LW scheme (2.8); right: BW scheme (2.9).

Figure 2: Example 2.1: The solutions at $t = 0.5$ with $h = 0.01$.



(a) Left: LF scheme (2.7); right: Upwind scheme (2.2).



(b) Left: LW scheme (2.8); right: BW scheme (2.9).

Figure 3: Example 2.1: The solutions at $t = 0.5$ with $h = 0.0025$.

Several basic concepts are introduced for numerical methods below.

Definition 2.1 (Convergence) *A FDS* $\mathcal{L}_{h,\tau}u_j^n = f_j^n$ *approximating a PDE* $\mathcal{L}u(x,t) = f$ *is point-wisely* **convergent** *if for any* x *and* t*, as* $(x_j, t_n) = (jh, n\tau)$ *tends to* (x,t)*, the global error* $e(x_j, t_n; h, \tau) := u_j^n - u(x_j, t_n)$ *converges to zero for* $\max\{h, \tau\} \to 0$.

Definition 2.2 (Consistency) *A FDS* $\mathcal{L}_{h,\tau}u_j^n = f_j^n$ *is* **consistent** *with a given PDE* $\mathcal{L}u = f$*, if for any smooth function* $w(x,t)$*, the local truncation error*

$$\mathcal{L}w(x,t) - \mathcal{L}_{h,\tau}w(x,t) \to 0 \text{ as } \tau, h \to 0.$$

Definition 2.3 (Stability) *A finite difference scheme is* **stable** *if the errors made at one time step do not cause the errors to increase as the computations are continued.*

In general, the consistency is most easily checked, by using the Taylor expansions after expanding a solution of the FDS as a smooth function. For example, the truncation error of the schemes (2.2)-(2.4) are

$$\left(\frac{ah}{2}u_{xx}-\frac{\tau}{2}u_{tt}+\cdots\right)_j^n,\ \left(\frac{-ah}{2}u_{xx}-\frac{\tau}{2}u_{tt}+\cdots\right)_j^n,\ \left(-\frac{\tau}{2}u_{tt}-\frac{ah^2}{6}u_{xx}+\cdots\right)_j^n,$$

respectively, while that of the LF scheme (2.7) is

$$\left(-\frac{\tau}{2}u_{tt}+\frac{h^2}{2\tau}u_{xx}-\frac{ah^2}{6}u_{xxx}+\cdots\right)_j^n.$$

A number of well-established methods establish stability or lack thereof, for example, the well-known *Courant-Friedrichs-Lewy (CFL) condition* [14] states that the ratio of the time step-size to the spatial step-size should exceed the speed at which waves propagate in the case of hyperbolic PDE. Unfortunately, the proof of convergence is more difficult than the stability.

Theorem 2.1 (Lax equivalence theorem [43]) *A consistent FDS with a linear PDE for which the initial-value problem is well posed is convergent if and only if it is stable.*

It is a fundamental theorem in the analysis of the FDS for the linear PDEs.

One simple but not general method for analyzing the stability of a numerical scheme is the *energy method*, where the sum of all squares of the numerical solution, will be calculated and its tendency in time will be evaluated. It should be noted that for this stability method, periodic boundary conditions are usually required.

Example 2.2 *The scheme* (2.2) *may be rewritten as follow*

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

where $\nu = a\tau/h$ *is so-called* Courant number. *Assume that* $\sum_{j=-\infty}^{\infty} (u_j^0)^2 h < \infty$ *and the periodic boundary conditions are specified. Squaring on both sides and summing over the index* j *results in*

$$\sum_{j=-\infty}^{\infty} (u_j^{n+1})^2 h = \sum_{j=-\infty}^{\infty} \Big((1-\nu)^2 (u_j^n)^2 + 2(1-\nu)\nu u_j^n u_{j-1}^n + \nu^2 (u_{j-1}^n)^2 \Big) h. \tag{2.10}$$

Using *Schwartz's inequality* *gives*

$$\sum_{j=-\infty}^{\infty} u_j^n u_{j-1}^n h \le \left(\sum_{j=-\infty}^{\infty} (u_j^n)^2 \right)^{1/2} h \cdot \left(\sum_{j=-\infty}^{\infty} (u_{j-1}^n)^2 \right)^{1/2} h.$$

With shifting by one index in a periodic domain, one has

$$\sum_{j=-\infty}^{\infty} (u_j^n)^2 h = \sum_{j=-\infty}^{\infty} (u_{j-1}^n)^2 h.$$

Substituting above two equations into (2.10) *gives*

$$\sum_{j=-\infty}^{\infty}(u_j^{n+1})^2 h \le \sum_{j=-\infty}^{\infty}\left((1-\nu)^2+2(1-\nu)\nu+\nu^2\right)(u_j^n)^2 h = \sum_{j=-\infty}^{\infty}(u_j^n)^2 h,$$

under the condition $\nu(1-\nu)\ge 0$, *so that the total energy of the numerical solution does not increase as* n *increases. Thus the scheme* (2.2) *is stable if* $0\le\nu\le 1$.

Similarly, (2.3) *is stable if* $-1\le\nu\le 0$. *Both schemes* (2.2) *and* (2.3) *belong to the* upwind schemes, *which attempt to approximate hyperbolic equations* (2.1) *by using differencing biased in the direction determined by the sign of the characteristic speed* a. *The unified form of* (2.2) *and* (2.3) *is as follows*

$$u_j^{n+1} = u_j^n - \frac{\tau a}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{\tau|a|}{2h}(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \quad (2.11)$$

or

$$u_j^{n+1} = u_j^n - \frac{\tau a}{h} \begin{cases} u_{j+1}^n - u_j^n, & a < 0, \\ u_j^n - u_{j-1}^n, & a > 0. \end{cases}$$

It may be derived by tracing the solution $u(x_j, t_{n+1})$ *along the characteristic line* $x = x_j - a(t_{n+1} - t)$ *back to* $t = t_n$

$$u(x_j, t_{n+1}) = u\big(x_j - a(t_{n+1} - t_n), t_n\big) = u(x_j - a\tau, t_n),$$

and then approximating $u\big(x_j - a\tau, t_n\big)$ *via the interpolation under the CFL condition [14]*

$$\left|\frac{a\tau}{h}\right| \le 1.$$

In fact, if a *is positive, then* $x_{j-1} \le x_j - a\tau \le x_j$ *and* $u\big(x_j - a\tau, t_n\big)$ *may be linearly interpolated by using the point values* u_j^n *and* u_{j-1}^n*. Similarly, if* a *is negative, then* $x_j \le x_j - a\tau \le x_{j+1}$*,*

the point values u_j^n *and* u_{j+1}^n *are used to interpolate the value of* $u\big(x_j - a\tau, t_n\big)$.

Historically, the origin of upwind methods can be traced back to the work [17] of Courant, Isaacson, and Rees who proposed the CIR method. Moreover, the BW scheme (2.9) does also belong to upwind schemes. On a temporal-spatial grid, a domain of influence and a domain of dependence for a certain numerical scheme may be drawn for each point. Another description of the CFL condition *[14] is that the discrete solution must not be independent of data that determine the solution of the associated PDE, and thus the physical domain of dependence as defined by the PDE must be inside the numerical domain of dependence.*

One drawback of the energy method is that a new strategy for each scheme has to be found how to calculate the energy of the numerical solution. A more generic method for testing

the stability of numerical schemes is the *von Neumann method* (also known as *Fourier method*), which is however also straightforward only for periodic boundary conditions. Furthermore it is only applied to linear schemes with constant coefficients. The *von Neumann method* is a procedure used to check the linear stability of the FDS as applied to linear PDEs. It is based on the Fourier decomposition of the numerical error. Each periodic function $\varepsilon(x,t)$ may be represented by an infinite Fourier series

$$\varepsilon(x,t) = \sum_{k=-\infty}^{\infty} \xi_k(t) e^{ikx},$$

where $i = \sqrt{-1}$. A key property is that each Fourier mode $\xi_k(t)e^{ikx}$ in the above equation is an eigenfunction of the (spatial) derivative operator

$$\frac{d}{dx} e^{ikx} = ike^{ikx}. \tag{2.12}$$

The property (2.12) may be used in the discrete space for studying the stability of the FDS. More precisely, the solution u_j^n of the linear FDS may be represented by a finite Fourier series

$$u_j^n = \sum_{k=-N}^{N} \xi_k^n e^{ikjh}. \tag{2.13}$$

Each Fourier mode is also an eigenfunction of the linear (spatial) difference operator, for example, the forward difference operator acting on "vector" e^{ikjh} results in a product of a scalar and the "vector"

$$\frac{1}{h}(e^{ikh} - 1)e^{ikjh}.$$

After one time step, the solution of the linear FDS for a specific discrete Fourier mode $u_j^n = \xi_k^n e^{ikjh}$ will be of the form

$$\xi_k^{n+1} = G_k \xi_k^n,$$

with the complex so-called *amplification factor* G_k. For linear schemes with constant coefficients, the factor G_k will be independent on the time step so that

$$\xi_k^n = (G_k)^n \xi_k^0.$$

For the stability of a linear FDS, it will be required that each Fourier component of (2.13) is bounded, i.e.

$$|\xi_k^n| \leq |G_k|^n |\xi_k^0| \leq c.$$

The *von Neumann stability criterion* is therefore formulated as

$$|G_k| \leq 1 + c\tau,$$

with c being independent of k, τ and h. The above inequality guarantees that a consistent scheme converges for $\max\{h, \tau\} \to 0$

due to Theorem 2.1. For problems with bounded solutions, it may be limited to

$$|G_k| \leq 1.$$

Example 2.3 *As an illustration, the linear convection equation* (2.1) *with* $a > 0$ *and the first-order accurate upwind scheme* (2.2) *will be tested again. Expressing the solution at each point as a specific Fourier component* $u_j^n = \xi_k^n e^{ikx_j}$ *and substituting it into* (2.2) *give its amplification factor*

$$G_k = \xi_k^{n+1}/\xi_k^n = 1 - \nu(1 - e^{-ikh}) = (1 - \nu) + \nu e^{-ikh}.$$

The square of the modulus of G_k *is equal to*

$$\begin{aligned} |G_k|^2 =& (1 - \nu(1 - \cos(kh)))^2 + \nu^2 \sin^2(kh) \\ =& 1 - 2\nu(1 - \nu)(1 - \cos(kh)). \end{aligned}$$

Thus, the condition $|G_k| \leq 1$ *is fulfilled if*

$$\nu(1-\nu) \geq 0.$$

For $a \geq 0$*, it is equivalent to* $\nu \leq 1$*.*

Similarly, one may know by using the von Neumann method that the BW method (2.9) *is stable for* $0 \leq \nu \leq 2$*.*

Besides the above approaches for the stability, a heuristic method is sufficiently with the aid of the *modified equation* of the FDS [93], which is derived by first expanding each term of a difference scheme in a Taylor series and then eliminating time derivatives higher than first order by the algebraic manipulations described herein. Generally, the modified equation of the rth-order accurate FDS is

$$u_t + a u_x = \sum_{s=r}^{\infty} \nu_s h^s \frac{\partial^s u}{\partial x^s}, \tag{2.14}$$

where ν_s is the coefficient of the truncation error of the FDS.

Example 2.4 *The modified equation of the upwind scheme* (2.2) *is*

$$u_t + au_x = \frac{1}{2}ah\left(1-\nu\right)u_{xx} + \cdots .$$

The terms at the right-hand side gives the the local truncation error of (2.2)*, and the leading term of the local truncation error is the second order diffusion term. Dropping the higher-order terms at the right-hand side obtains a familiar convection-diffusion equation. The heuristic stability condition of* (2.2) *is that the coefficient of the leading term of the local truncation error is positive, equivalently,* $ah(1-\nu) > 0$*. It is almost the same as the condition obtained by using the von Neumann method.*

The above example discusses the special case of $r = 2$. For the general case of $r > 2$ in (2.14), the error between the solutions

of (2.1) and (2.14) satisfies

$$\varepsilon_t + a\varepsilon_x = \sum_{s=r}^{\infty} \nu_s h^s \frac{\partial^s \varepsilon}{\partial x^s}. \tag{2.15}$$

Assuming that the error at $t = 0$ is

$$\varepsilon(x, 0) = Ae^{ikx},$$

where k is the wave number and A is the initial amplitude. Eq. (2.15) has the solution of the form

$$\varepsilon(x, t) = Ae^{i(kx-\omega t)} = Ae^{\alpha(k)t} e^{ik[x-(a-\beta(k))t]}, \quad t \geq 0,$$

with

$$\alpha(k) = \sum \nu_{2s} h^{2s} (-1)^s k^{2s}, \quad \beta(k) = \sum \nu_{2s+1} h^{2s+1} (-1)^s k^{2s}.$$

Corresponding *dispersion relation* is

$$\omega(k) = ak - \beta(k)k + i\alpha(k),$$

which determines how time oscillations $e^{-i\omega t}$ are linked to spatial oscillations e^{ikx} of wave number k. In other words, the dispersion relation is the function for which the plane waves $e^{ik\cdot x}e^{-i\omega(k)t}$ solve the equation. From the dispersion relation $\omega = \omega(k)$, the *phase speed* and the *group velocity* of a wave may be calculated by $\omega(k)/k$ and $d\omega(k)/dk$, respectively. The former is the rate at which the phase of any one frequency component of the wave travels, while the latter is the velocity at which the overall shape of the waves' amplitudes known as the modulation or envelope of the wave propagates through space.

Example 2.5 *The modified equation of the LW scheme* (2.8) *is*

$$u_t + au_x = -\frac{1}{6}ah^2\left(1-\nu^2\right)u_{xxx} + \cdots .$$

Dropping the higher-order terms at the right-hand side obtains a dispersive equation with the dispersion relation $\omega(k) = ak - \frac{1}{6}ah^2(1-\nu^2)k^3$. *The* group velocity $\omega'(k) = a\left(1 - \frac{1}{2}h^2(1-\nu^2)k^2\right)$ *is less than* a *for all wave numbers* k *if* $|\nu| < 1$. *The readers are referred to [81] for the group velocity in FDS. The oscillation behind of the main hump, see Figs. 2-3*

Example 2.6 *The modified equation of the BW scheme (2.9) is*

$$u_t + au_x = \frac{1}{6}ah^2(2-\nu)(1-\nu)u_{xxx} + \cdots .$$

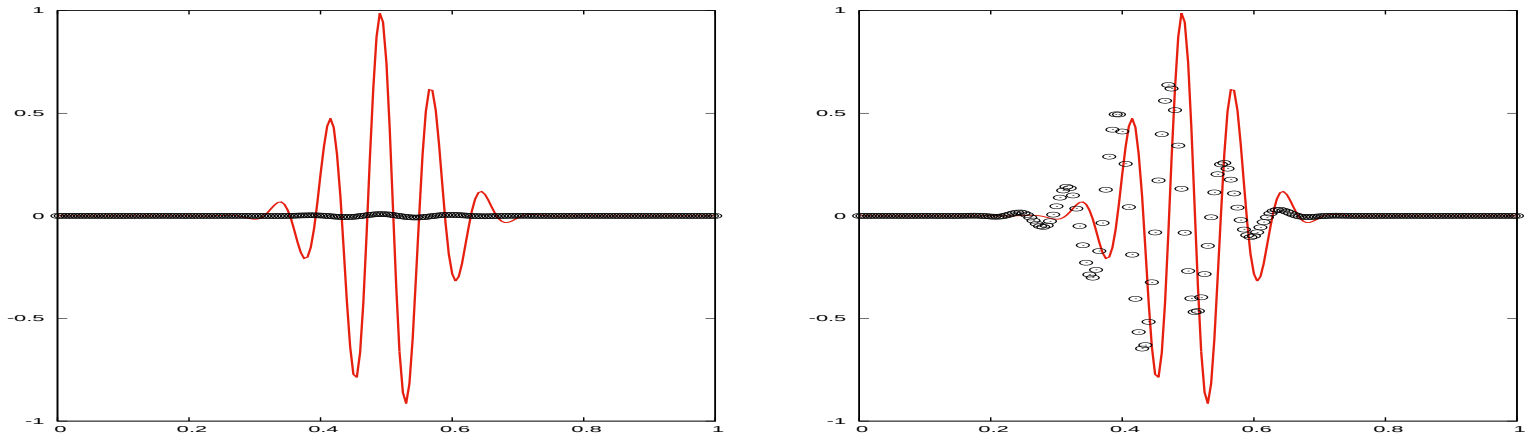
Dropping the higher-order terms at the RHS again obtains a dispersive equation with the dispersion relation $\omega(k) = ak + \frac{1}{6}ah^2(2-\nu)(1-\nu)k^3$. *The group velocity* $\omega'(k) = a\left(1 + \frac{1}{2}h^2(2-\nu)(1-\nu)k^2\right)$. *Thus, if* $0 < \nu < 1$, *the group velocity is greater than* a *for all wave numbers and the oscillations move ahead of the main hump,*

see Figs. 2-3; if $1 < \nu < 2$, *then the group velocity is less than a and the oscillations fall behind.*

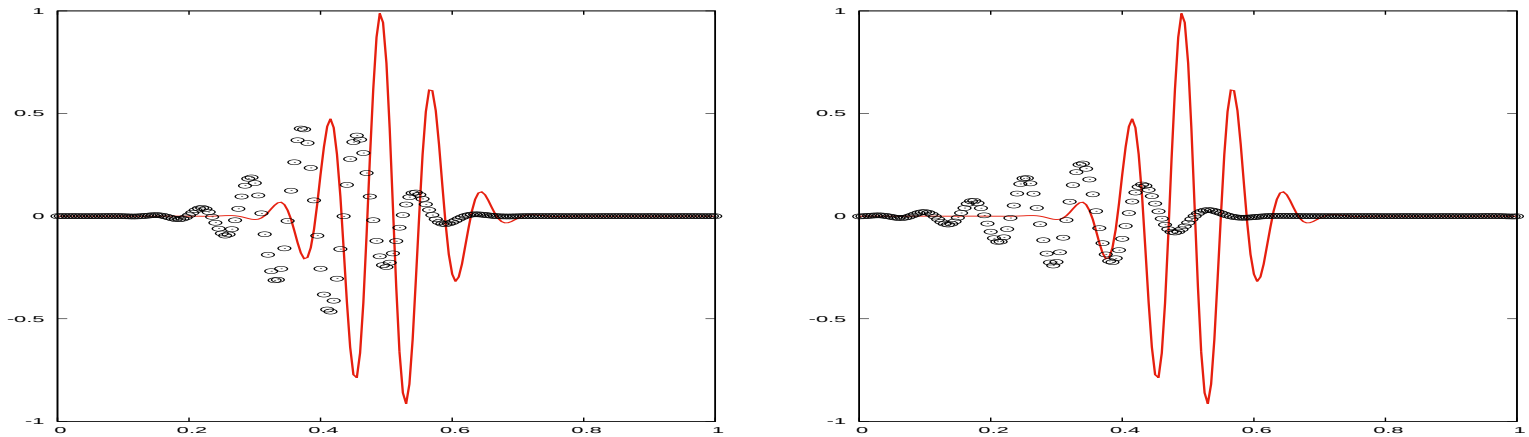
Exercise 2.1 (Propagation of a wave packet) *Solve the initial value problem of* (2.1) *with* $a = 1$ *and the initial data consisting of a sine wave modulated by a Gaussian centered at* $x = 0.5$

$$u(x, 0) = e^{-100(x-0.5)^2} \sin(kx),$$

with k *chosen so that there are 16 grid points per wavelength:* $kh = 2\pi/16 \approx 0.4$. *The computational domain and the spatial step size may be taken as* $[0, 1]$ *and* $h = 1/200$, *respectively. The output time* $t = 2, 4, 8$, *the time step size* $\tau = 0.8h$, *and* $k = 80$.



(a) The solutions at $t = 2$. Left: Upwind scheme; right: LW scheme.



(b) The solutions at $t = 4, 8$ by LW scheme.

Figure 4: Propagation of a wave packet in Exercise 2.1.

Exercise 2.2 (Highly discontinuous data) *Numerically solve the initial value problem of* (2.1) *with* $a = 1$ *and the highly discontinuous initial data*

$$u(x,0) = \begin{cases} -\xi \sin(1.5\pi\xi^2), & \text{if } -1 \le \xi < -1/3, \\ |\sin(2\pi\xi)|, & \text{if } |\xi| < 1/3, \\ 2\xi - 1 - \sin(3\pi\xi)/6, & \text{if } 1/3 < \xi \le 1, \end{cases}$$

which is periodic in x *with period 2, where* $\xi = x - 0.3$ *if* $-0.7 \le x \le 1$, *but* $\xi = x - 0.3 + 2$ *if* $-1 \le x < -0.7$.

Fig. 4 shows propagation of a wave packet in Exercise 2.1 obtained by using the first-order accurate upwind scheme and the second-order accurate LW scheme respectively, while Fig. 5 shows the solutions for Exercise 2.2. Those exercises may validate that high-order accurate schemes gives superior performance for

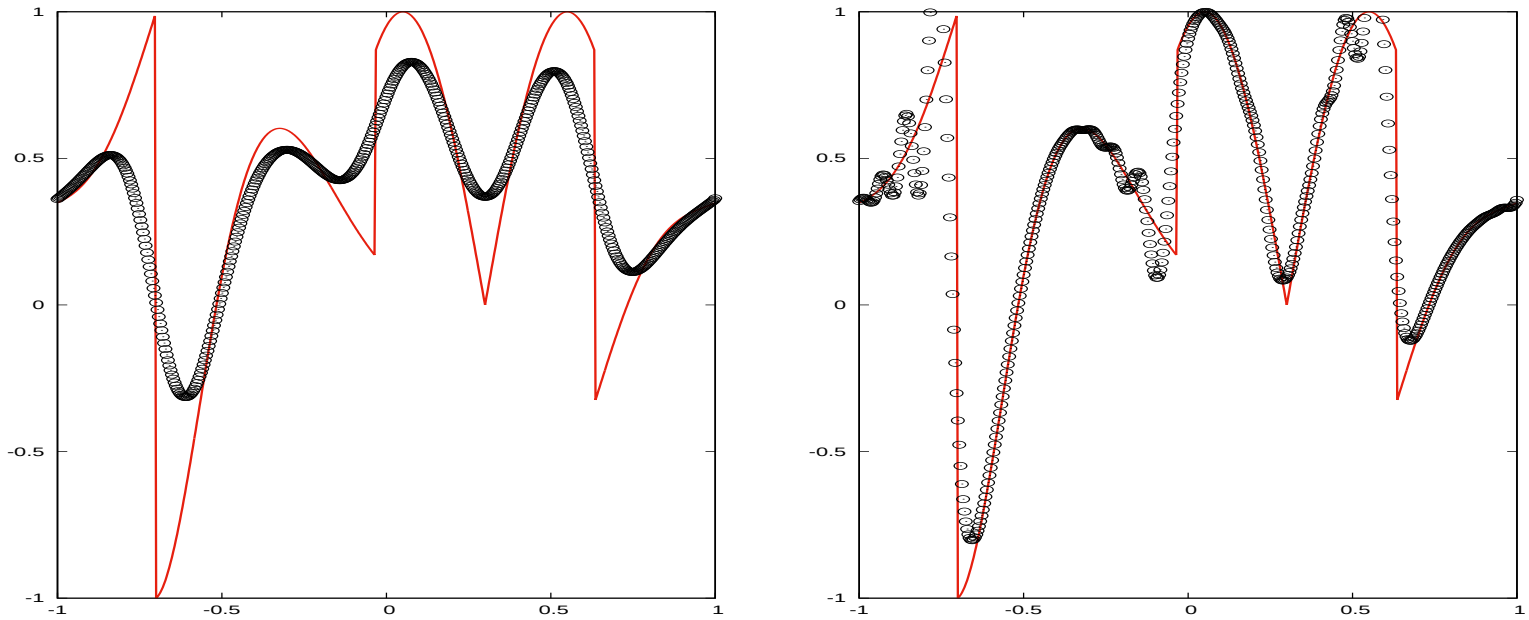


Figure 5: The solutions at $t = 8$ for the highly discontinuous initial data in Exercise 2.2 obtained by using the first-order accurate upwind scheme (left) and the second-order accurate LW scheme (right) with 501 points and $\tau = 0.8h$.

equal resolution (same mesh), but the second-order accurate LW scheme suffers from the numerical oscillations moving behind of the discontinuities.

2.2 方程组情形

Consider the hyperbolic system

$$\boldsymbol{U}_t + \boldsymbol{A}\boldsymbol{U}_x = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{2.16}$$

with the initial data $\boldsymbol{U}(x,0) = \boldsymbol{U}_0(x)$ at time $t = 0$, where $\boldsymbol{U} \in \mathbb{R}^m$, and the constant matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ are really diagonalizable, that is to say, there exists an invertible matrix $\boldsymbol{R}$ such that

$$\boldsymbol{L}\boldsymbol{A}\boldsymbol{R} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m), \quad \boldsymbol{L}\boldsymbol{R} = \boldsymbol{I},$$

here $\boldsymbol{R} = (\boldsymbol{R}^{(1)}, \boldsymbol{R}^{(2)}, \cdots, \boldsymbol{R}^{(m)})$, $\lambda_i \in \mathbb{R}$ is the ith eigenvalue of $\boldsymbol{A}$, and $\boldsymbol{R}^{(i)}$ denotes corresponding right eigenvector of $\boldsymbol{A}$. The hyperbolic system (2.16) may be symmetrized as follows

$$\left(\boldsymbol{L}^T \boldsymbol{L}\right) \boldsymbol{U}_t + \left(\boldsymbol{L}^T \boldsymbol{\Lambda} \boldsymbol{L}\right) \boldsymbol{U}_x = 0, \tag{2.17}$$

which the coefficient matrix $\boldsymbol{L}^T \boldsymbol{L}$ is real symmetric and positive definite, while the matrix $\boldsymbol{L}^T \boldsymbol{\Lambda} \boldsymbol{L}$ is real symmetric too. Symmetric hyperbolic system is important because corresponding initial-value problem is well-posed. Roughly speaking, to say that a system is well-posed is to say that the growth of its solution is bounded in a well-defined way, specifically, growth in an L^2 norm cannot be faster than exponential. Assume that the problem is defined over an unbounded spatial domain $\mathbb{R}$ without boundary, and take the inner product of $\boldsymbol{U}^T$ (the transpose of $\boldsymbol{U}$) with

(2.17) to get

$$\boldsymbol{U}^T\left(\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{U}_t+\boldsymbol{U}^T\left(\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{L}\right)\boldsymbol{U}_x=0. \tag{2.18}$$

Due to the symmetry of $\boldsymbol{L}^T\boldsymbol{L}$ and $\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{L}$, (2.18) can be rewritten as follows

$$\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{U}\right)_t+\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{L}\boldsymbol{U}\right)_x=0.$$

Now integrating it over $\mathbb{R}$ gets

$$\begin{aligned}0=&\frac{d}{dt}\int_{\mathbb{R}}\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{U}\right)\ dx+\int_{\mathbb{R}}\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{L}\boldsymbol{U}\right)_x\ dx\\ =&\frac{d}{dt}\int_{\mathbb{R}}\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{U}\right)\ dx+\frac{1}{2}\Big(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{L}\boldsymbol{U}\Big)_{-\infty}^{+\infty}.\end{aligned}$$

The above second term vanishes because there is no boundary. The quantity $E(t) := \int_{\mathbb{R}}\frac{1}{2}\left(\boldsymbol{U}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{U}\right)$ can be interpreted as the

total energy of system (2.16) or (2.17). Note that the positivity of $\boldsymbol{L}^T\boldsymbol{L}$ implies that $E(t)$ is a positive definite function of the state $\boldsymbol{U}$. The above deduction shows that the energy $E(t)$ of the system (2.16) or (2.17) is conserved in time.

If there exist a differentiable convex function $\eta(\boldsymbol{U})$ and a scalar function $q(\boldsymbol{U})$ for the system (2.16) satisfying $q'(\boldsymbol{U}) = \eta'(\boldsymbol{U})\boldsymbol{A}$, then (2.16) can also be symmetrized by the change of variables $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{V})$ with the aid of the function pair $(\eta(\boldsymbol{U}), q(\boldsymbol{U}))$. More precisely, if define the change of variables by $\boldsymbol{V}^T = \eta'(\boldsymbol{U})$, then (2.16) becomes

$$\boldsymbol{U}'(\boldsymbol{V})\boldsymbol{V}_t + \boldsymbol{A}\boldsymbol{U}'(\boldsymbol{V})\boldsymbol{V}_x = 0.$$

Define the conjugate functions

$$\eta^*(\boldsymbol{V}) := \boldsymbol{V}^T\boldsymbol{U}(\boldsymbol{V}) - \eta(\boldsymbol{U}(\boldsymbol{V})), \quad q^*(\boldsymbol{V}) := \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{U}(\boldsymbol{V}) - q(\boldsymbol{U}(\boldsymbol{V})),$$

and get their gradient vectors $(\eta^*)'(\boldsymbol{V}) = (\boldsymbol{U}(\boldsymbol{V}))^T$ and $(q^*)'(\boldsymbol{V}) = (\boldsymbol{A}\boldsymbol{U}(\boldsymbol{V}))^T$, which imply $(\eta^*)''(\boldsymbol{V}) = \boldsymbol{U}'(\boldsymbol{V})$ and $(q^*)''(\boldsymbol{V}) = \boldsymbol{A}\boldsymbol{U}'(\boldsymbol{V})$ are symmetric. On the other hand, $\boldsymbol{U}'(\boldsymbol{V})$ is positive-definite because the differentiation of $\boldsymbol{V}^T = \eta'(\boldsymbol{U})$ and the convexity of η with respect to $\boldsymbol{V}$ gives $\boldsymbol{U}'(\boldsymbol{V}) = \big(\eta''(\boldsymbol{U}(\boldsymbol{V}))\big)^{-1} > 0$.

Some schemes may be extended to the system (2.16) componentwisely. For example, the LF scheme becomes

$$\frac{\boldsymbol{U}_j^{n+1} - (\boldsymbol{U}_{j+1}^n + \boldsymbol{U}_{j-1}^n)/2}{\tau} + \boldsymbol{A}\frac{\boldsymbol{U}_{j+1}^n - \boldsymbol{U}_{j-1}^n}{2h} = 0.$$

The following mainly discusses the characteristic decomposition and the upwind-type scheme of (2.16). Multiplying (2.16) with $\boldsymbol{L}$ gives

$$\frac{\partial(\boldsymbol{L}\boldsymbol{U})}{\partial t} + \boldsymbol{L}\boldsymbol{A}\boldsymbol{R}\frac{\partial(\boldsymbol{L}\boldsymbol{U})}{\partial x} = 0.$$

Define the *characteristic variables* $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{U}$, then

$$\frac{\partial \boldsymbol{W}}{\partial t} + \boldsymbol{\Lambda}\frac{\partial \boldsymbol{W}}{\partial x} = 0, \quad \text{or} \quad \frac{\partial w_i}{\partial t} + \lambda_i \frac{\partial w_i}{\partial x} = 0, \ i = 1, 2, \cdots, m.$$

They are m decoupled scalar convection equations, thus the discussion on the convection equation in Section 2.1 is completely available for them now.

The solution of the initial value problem of (2.16) is

$$\boldsymbol{U}(x,t) = \sum_{i=1}^{m} w_i^{(0)}(x - \lambda_i t)\boldsymbol{R}^{(i)}, \quad \boldsymbol{W}^{(0)}(x) = \boldsymbol{L}\boldsymbol{U}_0(x).$$

Exercise 2.3 *Write the solution of the initial value problem of* (2.16) *with Riemann data*

$$\boldsymbol{U}(x,0) = \boldsymbol{U}_0(x) = \begin{cases} \boldsymbol{U}_L, & x < 0, \\ \boldsymbol{U}_R, & x > 0. \end{cases}$$

If assuming that

$$\boldsymbol{U}_L = \sum_{i=1}^{m} \alpha_i \boldsymbol{R}^{(i)}, \quad \boldsymbol{U}_R = \sum_{i=1}^{m} \beta_i \boldsymbol{R}^{(i)},$$

then the solution at the point (x, t) *is*

$$\boldsymbol{U}(x,t) = \sum_{i=1}^{i_0} \beta_i \boldsymbol{R}^{(i)} + \sum_{i=i_0}^{m} \alpha_i \boldsymbol{R}^{(i)},$$

where the integer $i_0 \in [1, m]$ *ensures that*

$$x - \lambda_{i_0} t > 0, \quad x - \lambda_{i_0+1} t < 0.$$

The upwind scheme (2.11) for the equation of the characteristic variable w_i is

$$\frac{(w_i)_j^{n+1} - (w_i)_j^n}{\tau} + \lambda_i \frac{(w_i)_{j+1}^n - (w_i)_{j-1}^n}{2h} - |\lambda_i| \frac{(w_i)_{j+1}^n - 2(w_i)_j^n + (w_i)_{j-1}^n}{2h} = 0,$$

and thus the first-order accurate upwind scheme for (2.16) can be obtained by multiplying the matrix $\boldsymbol{R}$ from the left

$$\frac{\boldsymbol{U}_j^{n+1} - \boldsymbol{U}_j^n}{\tau} + \boldsymbol{A} \frac{\boldsymbol{U}_{j+1}^n - \boldsymbol{U}_{j-1}^n}{2h} - |\boldsymbol{A}| \frac{\boldsymbol{U}_{j+1}^n - 2\boldsymbol{U}_j^n + \boldsymbol{U}_{j-1}^n}{2h} = 0,$$

where $|\boldsymbol{A}| = \boldsymbol{R}|\boldsymbol{\Lambda}|\boldsymbol{L}$ and $|\boldsymbol{\Lambda}| = \mathrm{diag}\{|\lambda_1|, \cdots, |\lambda_m|\}$.

The LW scheme of (2.16) may be written into

$$\frac{\boldsymbol{U}_j^{n+1} - \boldsymbol{U}_j^n}{\tau} + \boldsymbol{A} \frac{\boldsymbol{U}_{j+1}^n - \boldsymbol{U}_{j-1}^n}{2h} - \tau \boldsymbol{A}^2 \frac{\boldsymbol{U}_{j+1}^n - 2\boldsymbol{U}_j^n + \boldsymbol{U}_{j-1}^n}{2h^2} = 0,$$

where $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{R}\boldsymbol{\Lambda}^2\boldsymbol{L}$. To avoid the matrix multiplication $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A}$ or the Jacobian evaluation in the quasilinear system

(see §5), a two-step procedure may be used. What follows is Richtmyer's two-step LW scheme

$$\begin{aligned}
\boldsymbol{U}_{j+\frac{1}{2}}^{n+1/2} =& \frac{1}{2}\left(\boldsymbol{U}_{j+1}^{n} + \boldsymbol{U}_{j}^{n}\right) - \frac{\tau}{2h}\boldsymbol{A}\left(\boldsymbol{U}_{j+1}^{n} - \boldsymbol{U}_{j}^{n}\right), \\
\boldsymbol{U}_{j}^{n+1} =& \boldsymbol{U}_{j}^{n} - \frac{\tau}{h}\boldsymbol{A}\left(\boldsymbol{U}_{j+\frac{1}{2}}^{n+1/2} - \boldsymbol{U}_{j-\frac{1}{2}}^{n+1/2}\right).
\end{aligned}$$

Another method of this same type is proposed by MacCormack [51]. The MacCormack scheme uses first forward differencing and then backward differencing to approximate the spatial derivative in (2.16) and is of the form

$$\begin{aligned}
\overline{\boldsymbol{U}}_{j}^{n+1} =& \boldsymbol{U}_{j}^{n} - \frac{\tau}{h}\boldsymbol{A}\left(\boldsymbol{U}_{j+1}^{n} - \boldsymbol{U}_{j}^{n}\right), \\
\boldsymbol{U}_{j}^{n+1} =& \frac{1}{2}\left(\boldsymbol{U}_{j}^{n} + \overline{\boldsymbol{U}}_{j}^{n+1}\right) - \frac{\tau}{2h}\boldsymbol{A}\left(\overline{\boldsymbol{U}}_{j}^{n+1} - \overline{\boldsymbol{U}}_{j-1}^{n+1}\right).
\end{aligned}$$

The order of differencing can be reversed as follows

$$\begin{aligned}\overline{\boldsymbol{U}}_j^{n+1} =& \boldsymbol{U}_j^n - \frac{\tau}{h}\boldsymbol{A}\left(\boldsymbol{U}_j^n - \boldsymbol{U}_{j-1}^n\right),\\ \boldsymbol{U}_j^{n+1} =& \frac{1}{2}\left(\boldsymbol{U}_j^n + \overline{\boldsymbol{U}}_j^{n+1}\right) - \frac{\tau}{2h}\boldsymbol{A}\left(\overline{\boldsymbol{U}}_{j+1}^{n+1} - \overline{\boldsymbol{U}}_j^{n+1}\right).\end{aligned}$$

The MacCormack method is implemented component-wisely and thus well suited for nonlinear equations. For linear equation (2.1) with constant coefficient, it is equivalent to the LW scheme.

3 Conservative schemes for conservation laws

This section begins to introduce the conservative FDS for the quasilinear hyperbolic conservation law (1.3).

3.1 Definition and properties of conservative schemes

For the sake of convenience, our attention is still paid to the case of the uniform mesh: $x_j = jh$, $j \in \mathbb{Z}$, where h is the spatial step-size.

Definition 3.1 *If the right-hand side of the explicit scheme*

$$u_j^{n+1} = H(u_{j-l}^n, \cdots, u_{j+l}^n), \tag{3.1}$$

may be written as follows

$$H(u_{j-l}^n, \cdots, u_{j+l}^n) = u_j^n - \lambda\big(\hat{f}_{j+\frac{1}{2}}^n - \hat{f}_{j-\frac{1}{2}}^n\big), \tag{3.2}$$

where $\lambda := \tau/h$, h, τ *denote the stepsizes,* $\hat{f}_{j+\frac{1}{2}} = \hat{f}(u_{j-l+1}, \cdots, u_{j+l})$ *is numerical flux function,* u_j^n *is an approximation of the point*

value $u(x_j, t_n)$, *then* (3.1)-(3.2) *is called as a* **conservative finite difference scheme**. *If* $\hat{f}$ *satisfies*

$$\hat{f}(w, \cdots, w) = f(w),$$

then (3.1)-(3.2) *is consistent with* (1.3).

Remark 3.1 *The above definition may be extended to the implicit scheme and multi time level scheme. Below are two examples*

$$u_j^{n+1} = u_j^n - \lambda\big(\hat{f}_{j+\frac{1}{2}}^{n+1} - \hat{f}_{j-\frac{1}{2}}^{n+1}\big),$$

and

$$u_j^{n+1} = u_j^{n-1} - 2\lambda\big(\hat{f}_{j+\frac{1}{2}}^{n} - \hat{f}_{j-\frac{1}{2}}^{n}\big).$$

Remark 3.2 *If* u_j^n *in Definition 3.1 is an approximation of the*

cell average value of $u(x,t_n)$ *over the cell* $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$, *i.e.,*

$$u_j^n \approx \frac{1}{h}\int_{x_j-h/2}^{x_j+h/2} u(x,t_n)\,dx,$$

then corresponding scheme is called as **conservative finite volume scheme**, *which is built on the integral form of* (1.3) *over the finite control volume.*

Remark 3.3 *Assuming that as* $j \to \pm\infty$, $\hat{f}_{j+\frac{1}{2}} \to 0$, *then the two time level conservative scheme* (3.1)-(3.2) *implies the total mass conservation*

$$\sum_{j\in\mathbb{Z}} u_j^{n+1} h = \sum_{j\in\mathbb{Z}} u_j^n h.$$

Example 3.1 *The LF scheme [42] for (1.3)*

$$u_j^{n+1} = \frac{u_{j+1}^n + u_{j-1}^n}{2} - \frac{\lambda}{2}(f(u_{j+1}^n) - f(u_{j-1}^n)), \tag{3.3}$$

may be rewritten into the conservative form (3.1)-(3.2) with numerical flux $\hat{f}_{j+1/2} = \frac{1}{2}(f_j + f_{j+1}) - \frac{1}{2\lambda}(u_{j+1} - u_j)$, *where* $f_j = f(u_j)$. ∎

Example 3.2 *The usual upwind scheme in a conservative form for (1.3) is*

$$\begin{aligned} u_j^{n+1} =& u_j^n - \frac{\lambda}{2}(f(u_{j+1}^n) - f(u_{j-1}^n)) + \frac{\lambda|a(u^n_{j+\frac{1}{2}})|}{2}(u_{j+1}^n - u_j^n) \\ & - \frac{\lambda|a(u^n_{j-\frac{1}{2}})|}{2}(u_j^n - u_{j-1}^n), \end{aligned} \tag{3.4}$$

where $u^n_{j+\frac{1}{2}} = \frac{1}{2}(u^n_j + u^n_{j+1})$ *and* $a(u) = f'(u)$. *Its numerical flux is* $\hat{f}_{j+1/2} = \frac{1}{2}(f_j + f_{j+1}) - \frac{1}{2}|a(u_{j+\frac{1}{2}})|(u_{j+1} - u_j)$. ∎

Example 3.3 *The LW scheme [44] is derived as follows. Replacing the time derivatives in the Taylor series expansion of the solution in time with spatial derivatives via the original PDE (1.3) gives*

$$\begin{aligned} u(x_j, t_{n+1}) =& u(x_j, t_n) + \tau(u_t)^n_j + \frac{1}{2}\tau^2(u_{tt})^n_j + \mathcal{O}(\tau^3) \\ =& u(x_j, t_n) - \tau(f_x)^n_j + \frac{1}{2}\tau^2\partial_x(af_x) + \mathcal{O}(\tau^3) \\ =& u(x_j, t_n) - \tau(f_x)^n_j + \frac{1}{2}\tau^2\partial_x(a^2u_x) + \mathcal{O}(\tau^3), \end{aligned}$$

and then approximating spatial derivatives via central differences and neglecting the higher-order terms and replacing $u(x_j, t_n)$ *with*

u_j^n *lead to*

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(f(u_{j+1}^n) - f(u_{j-1}^n)) + \frac{\lambda^2 a^2(u_{j+\frac{1}{2}}^n)}{2}(u_{j+1}^n - u_j^n) - \frac{\lambda^2 a^2(u_{j-\frac{1}{2}}^n)}{2}(u_j^n - u_{j-1}^n). \tag{3.5}$$

It is a second-order accurate three-point scheme in time and space in the sense of truncation error, and its numerical flux is $\hat{f}_{j+1/2} = \frac{1}{2}(f_j + f_{j+1}) - \frac{\lambda}{2}a^2(u_{j+\frac{1}{2}})(u_{j+1} - u_j)$. ∎

Remark 3.4 *The term* $a(u_{j+1/2})$ *in* (3.4) *and* (3.5) *may be replaced with*

$$a_{j+1/2} = \begin{cases} \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j}, & u_{j+1} \neq u_j, \\ a(u_j), & u_{j+1} = u_j. \end{cases} \tag{3.6}$$

Example 3.4 *Richtmyer's two-step LW scheme [60] for (1.3) is*

$$\begin{cases} u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}(u_j^n + u_{j+1}^n) - \frac{\lambda}{2}\left(f(u_{j+1}^n) - f(u_j^n)\right), \\ u_j^{n+1} = u_j^n - \lambda\left(f(u_{j+\frac{1}{2}}^{n+\frac{1}{2}}) - f(u_{j-\frac{1}{2}}^{n+\frac{1}{2}})\right), \end{cases} \tag{3.7}$$

which may be constructed from

$$\begin{aligned} u(x_j, t_{n+1}) =& u(x_j, t_n) + \tau(u_t)_j^n + \frac{1}{2}\tau^2(u_{tt})_j^n + \mathcal{O}(\tau^3) \\ =& u(x_j, t_n) + \tau(\overline{u}_t)_j^n + \mathcal{O}(\tau^3), \quad \overline{u} := u + \frac{\tau}{2}u_t. \end{aligned}$$

∎

Example 3.5 *MacCormack's scheme [51] for (1.3) becomes*

$$\begin{cases} \overline{u}_j^* = u_j^n - \lambda\left(f(u_{j+1}^n) - f(u_j^n)\right), \\ u_j^{n+1} = \frac{1}{2}(u_j^n + \overline{u}_j^*) - \frac{\lambda}{2}\left(f(\overline{u}_j^*) - f(\overline{u}_{j-1}^*)\right), \end{cases} \tag{3.8}$$

or

$$\begin{cases} \overline{u}_j^* = u_j^n - \lambda \left(f(u_j^n) - f(u_{j-1}^n) \right), \\ u_j^{n+1} = \frac{1}{2}(u_j^n + \overline{u}_j^*) - \frac{\lambda}{2} \left(f(\overline{u}_{j+1}^*) - f(\overline{u}_j^*) \right), \end{cases} \tag{3.9}$$

which may be derived from

$$\begin{aligned} u(x_j, t_{n+1}) =& u(x_j, t_n) + \tau (u_t)_j^n + \frac{1}{2}\tau^2 (u_{tt})_j^n + \mathcal{O}(\tau^3) \\ =& \frac{1}{2} u(x_j, t_n) + \frac{1}{2}(\overline{u} + \tau \overline{u}_t)_j^n + \mathcal{O}(\tau^3), \quad \overline{u} := u + \tau u_t. \end{aligned}$$

∎

Example 3.6 *The Godunov scheme [22] may be considered as a conservative finite volume method which needs solve local Riemann problems at each inter-cell boundary exactly. Define* $I_j :=$

$(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$, *calculate the cell-average value of the initial data*

$$\bar{u}_j^0 = \frac{1}{h}\int_{I_j} u(x,0)\,dx, \tag{3.10}$$

and assume that the time step-size satisfies a more severe CFL condition

$$\frac{\tau}{h}\max_u\{|f'(u)|\} \le \frac{1}{2}. \tag{3.11}$$

After that, we perform the following steps.

`Step (1).` *For $n \ge 0$, reconstruct a (discontinuous) piecewise constant function approximating the initial data at $t = t_n$*

$$u_h(x,t_n) := \bar{u}_j^n, \quad x \in I_j,\ t \in [t_n, t_{n+1}), \tag{3.12}$$

which naturally form local Riemann problems at each inter-cell

boundary

$$\begin{cases} \text{Eq. (1.3)}, \qquad t \in [t_n, t_{n+1}), \\ u(x,t_n) = \begin{cases} u_h(x_{j+\frac{1}{2}} - 0, t_n), & x < x_{j+\frac{1}{2}}, \\ u_h(x_{j+\frac{1}{2}} + 0, t_n), & x > x_{j+\frac{1}{2}}. \end{cases} \end{cases} \tag{3.13}$$

Fig. 6 shows the profile of the piecewise constant function in (3.12) and the wave structure of the local Riemann problems (3.13).

The exact solution of the Riemann problem (3.13) is denoted by $\omega(x,t)$ which is of the form

$$\omega(x,t) = \omega\left(\frac{x - x_{j+\frac{1}{2}}}{t - t_n}, u_h(x_{j+\frac{1}{2}} - 0, t_n), u_h(x_{j+\frac{1}{2}} + 0, t_n)\right).$$

Step (2). *Calculate the cell-average value at* $t = t_{n+1}$ *by*

$$\bar{u}_j^{n+1} = \frac{1}{h} \int_{I_j} \omega(x, t_{n+1}) \, dx. \tag{3.14}$$

Under the CFL condition (3.11)*, one has*

$$\begin{aligned} \bar{u}_j^{n+1} =& \frac{1}{h} \int_{x_j - \frac{1}{2}h}^{x_j} \omega \left(\frac{x - x_{j-\frac{1}{2}}}{\tau}, \bar{u}_{j-1}^n, \bar{u}_j^n \right) \, dx \\ &+ \frac{1}{h} \int_{x_j}^{x_j + \frac{1}{2}h} \omega \left(\frac{x - x_{j+\frac{1}{2}}}{\tau}, \bar{u}_j^n, \bar{u}_{j+1}^n \right) \, dx. \end{aligned} \tag{3.15}$$

The CFL condition (3.11) *ensures that the primary waves emanating from two neighboring nodes* $x_{j-\frac{1}{2}}$ *and* $x_{j+\frac{1}{2}}$ *do not interact each other within a time interval* $[t_n, t_{n+1}]$.

The above Godunov method is time-consuming in solving exactly the Riemann problem and the evaluation of integrals. In

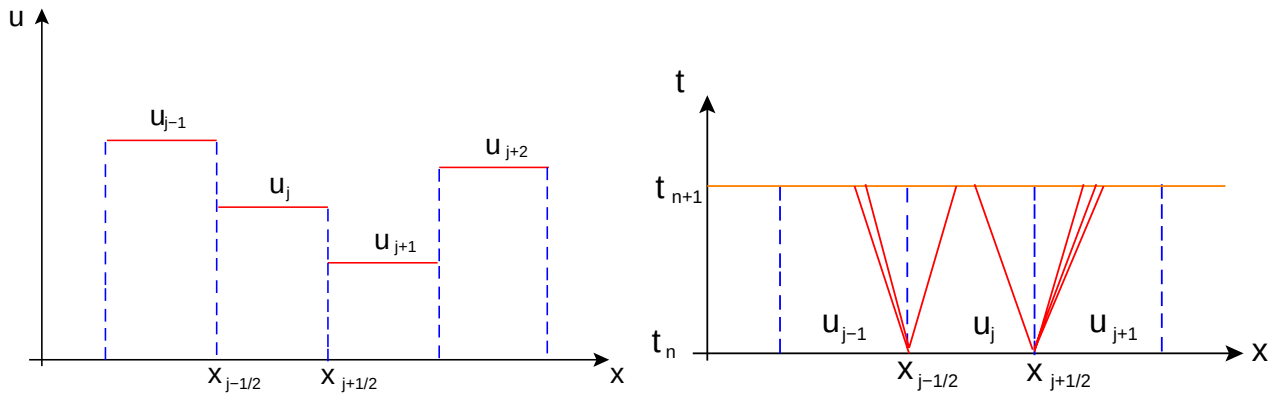


Figure 6: Piecewise constant function (3.12) and wave structure of local Riemann problem (3.13).

practice, the Godunov method may be implemented as follows. Because $\omega(x,t)$ *is the exact solution of the Riemann problem, integrating the conservation law* (1.3) *over the control volume* $I_j \times [t_n, t_{n+1}]$ *gives*

$$\int_{I_j} \omega(x, t_{n+1})\ dx - \int_{I_j} \omega(x, t_n)\ dx + \int_{t_n}^{t_{n+1}} f\left(\omega(x_{j+\frac{1}{2}}, t)\right)\ dt - \int_{t_n}^{t_{n+1}} f\left(\omega(x_{j-\frac{1}{2}}, t)\right)\ dt = 0. \tag{3.16}$$

Dividing it by τh *gives*

$$\bar{u}_j^{n+1} - \bar{u}_j^n + \lambda\left[f\left(\omega(0; \bar{u}_j^n, \bar{u}_{j+1}^n)\right) - f\left(\omega(0; \bar{u}_{j-1}^n, \bar{u}_j^n)\right)\right] = 0. \tag{3.17}$$

Here has noted the fact that $\omega(x_{j+1/2}, t) = \omega(0; \bar{u}_j^n, \bar{u}_{j+1}^n)$. *In the*

present case, the CFL condition (3.11) *may be relaxed as*

$$\frac{\tau}{h}\max\{|f'(u)|\} \leq 1.$$

The above Godunov scheme may be considered as an extension of the upwind scheme (2.11). ∎

Exercise 3.1 *Are two versions of the Godunov scheme equivalent to each other?*

Example 3.7 *This example lists numerical fluxes of several upwind type schemes in the conservative form for the quasilinear equation* (1.3).

a). *Roe's scheme [61]*

$$\hat{f}(u_j, u_{j+1}) = \frac{f(u_j) + f(u_{j+1})}{2} - \frac{1}{2}|\hat{a}_{j+1/2}|(u_{j+1} - u_j), \quad (3.18)$$

where $\hat{a}_{j+1/2}$ *is defined by* $\hat{a}_{j+1/2}(u_{j+1}-u_j)=f(u_{j+1})-f(u_j)$, *similar to* $a_{j+1/2}$ *in Remark 3.4.*

b). *Huang's scheme [34]*

$$\hat{f}(u_j,u_{j+1})=\frac{f(u_j)+f(u_{j+1})}{2}-\frac{1}{2}sign\left(a\left(\frac{u_j+u_{j+1}}{2}\right)\right)(f(u_{j+1})-f(u_j)).$$

c). *The scheme of Engquist and Osher [19, 20]*

$$\hat{f}(u_j,u_{j+1})=\frac{1}{2}(f(u_j)+f(u_{j+1}))-\frac{1}{2}\int_{u_j}^{u_{j+1}}|a(u)|\,du=f^+(u_j)+f^-(u_{j+1}),\tag{3.19}$$

where

$$f^+(u)=\int_0^u\max\{f'(v),0\}dv+f(0),f^-(u)=\int_0^u\min\{f'(v),0\}dv.$$

When those upwind type schemes are applied to the linear convection equation (2.1), they reduce to (2.2) if $a>0$, *otherwise (2.3).* ∎

If the grid function $\{u_j^n\}$ is expanded to the entire upper half space, define

$$\delta = \max\{h, \tau\},$$

and the step function

$$u_\delta(x,t) = u_j^n, \quad t \in [t_n, t_{n+1}), \ x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}),$$

then the following well-known theorem holds for the conservative scheme.

Theorem 3.1 (Lax-Wendroff theorem [44]) *Assuming that the conservative scheme (3.1)-(3.2) is consistent with (1.3). If the solution of (3.1)-(3.2), $u_\delta(x,t)$, satisfying the initial data*

$$u_\delta(x,0) = u_j^0 = \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\, dx, \tag{3.20}$$

is almost everywhere bounded and converges to $u(x,t)$ *as* $\delta \to 0$, *then* $u(x,t)$ *is a weak solution of the initial value problem of* (1.3).

Proof: Rewrite (3.1)–(3.2) as follows

$$\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{\hat{f}^n_{j+\frac{1}{2}} - \hat{f}^n_{j-\frac{1}{2}}}{h} = 0,$$

or in the step function $u_\delta(x,t)$

$$\frac{u_\delta(x,t+\tau) - u_\delta(x,t)}{\tau} + \frac{\hat{f}_\delta(x+\frac{h}{2},t) - \hat{f}_\delta(x-\frac{h}{2},t)}{h} = 0,$$

where $\hat{f}_\delta(x+\frac{h}{2},t) = \hat{f}\Big(u_\delta\big(x-(l-1)h,t\big),\cdots,u_\delta(x+lh,t)\Big)$.

Multiply it by any test function $\varphi(x,t) \in C_0^\infty(\mathbb{R}\times[0,T))$, and integrate with respect to x and t, then

$$\iint\limits_{t>0} \varphi(x,t)\frac{u_\delta(x,t+\tau)-u_\delta(x,t)}{\tau}+\frac{\hat{f}_\delta(x+\frac{h}{2},t)-\hat{f}_\delta(x-\frac{h}{2},t)}{h}dxdt=0.$$

Using the independent variable substitutions such as $t+\tau \to t$ and $x \pm h/2 \to x$ gives

$$\iint\limits_{t\geq\tau} \frac{\varphi(x,t-\tau)-\varphi(x,t)}{\tau}u_\delta(x,t)\ dxdt$$

$$+\iint\limits_{t>0} \frac{\varphi(x-\frac{h}{2},t)-\varphi(x+\frac{h}{2},t)}{h}\hat{f}(x,t)\ dxdt$$

$$-\frac{1}{\tau}\int_0^\tau\int_{\mathbb{R}} \varphi(x,t)u_\delta(x,t)\ dxdt=0. \qquad (3.21)$$

Because $u_\delta(x,t)$ is almost everywhere bounded and converges to $u(x,t)$ as $\delta \to 0$, the value of $\hat{f}_\delta(x,t)$ tends to $\hat{f}(u,u,\cdots,u) = f(u)$ by using the consistency, while the limit of the last integral term in (3.21) is

$$\lim_{\tau\to 0} \frac{\int_0^\tau \int_{\mathbb{R}} \varphi u_\delta dx dt}{\tau} = \lim_{\tau\to 0} \frac{\int_{\mathbb{R}} \varphi(x,\tau) u_\delta(x,\tau)\ dx}{1} = \int_{\mathbb{R}} \varphi(x,0) u_\delta(x,0)\ dx.$$

Moreover, one has

$$\lim_{\tau\to 0} \frac{\varphi(x,t-\tau)-\varphi(x,t)}{\tau} = \lim_{\tau\to 0} \frac{\frac{d}{d\tau}\left[\varphi(x,t-\tau)-\varphi(x,t)\right]}{1} = -\frac{\partial \varphi}{\partial t}(x,t),$$

$$\lim_{h\to 0} \frac{\varphi(x-\frac{h}{2},t)-\varphi(x+\frac{h}{2})}{h} = \lim_{h\to 0} \frac{\frac{d}{dh}\left[\varphi(x-\frac{h}{2},t)-\varphi(x+\frac{h}{2},t)\right]}{1} = -\frac{\partial \varphi}{\partial x}(x,t).$$

Thus, as $\delta \to 0$, (3.21) tends to

$$\iint_{t\geq 0} \left[\frac{\partial \varphi}{\partial t} u + \frac{\partial \varphi}{\partial x} f(u)\right]\ dx dt + \int_\infty^\infty \varphi(x,0) u_0(x)\ dx = 0,$$

which means that $u(x,t)$ is a weak solution of the initial value problem of (1.3). ∎

Remark 3.5 *The LW theorem tells us that even though the classical solution of the initial value problem of* (1.3) *does not exist, the conservative scheme is still meaningful and may be used to give an approximate weak solution of the initial value problem of* (1.3). *However, the LW theorem cannot tell us when the solution of conservative scheme does converge, and whether the limiting function* $u(x,t)$ *is admissible. To answer those questions, one still need some nonlinear stabilities of the numerical scheme and the discrete entropy inequality or condition.*

The *discrete entropy condition* is introduced in the following.

Definition 3.2 *If for any strictly convex entropy* $\eta(u)$ *and corresponding entropy flux* $q(u) = \int^u \eta'(z)f'(z)\,dz$ *of* (1.3) *there*

exists a $2l$*-variable function* $\hat{q}(u_1, \cdots, u_{2l})$ *satisfying*

$$\hat{q}(u, \cdots, u) = q(u), \tag{3.22}$$

such that the solution of the conservative scheme (3.1)-(3.2) *satisfies the inequality*

$$\eta_j^{n+1} \leq \eta_j^n - \lambda \left(\hat{q}_{j+\frac{1}{2}}^n - \hat{q}_{j-\frac{1}{2}}^n \right), \ \lambda = \frac{\tau}{h}, \tag{3.23}$$

where $\eta_j^n = \eta(u_j^n)$ *and* $\hat{q}_{j+\frac{1}{2}}^n = \hat{q}(u_{j-l+1}^n, \cdots, u_{j+l}^n)$*, then we say that the scheme* (3.1)-(3.2) *satisfies the entropy condition* (3.23) *and call* $\hat{q}_{j+\frac{1}{2}}$ *as* **numerical entropy flux**.

Theorem 3.2 *Assuming that the conservative numerical scheme* (3.1)-(3.2) *is consistent with conservation laws* (1.3)*, and satisfies the discrete entropy condition* (3.23)*. If as* $\delta \to 0$*, the solution of* (3.1)*,* $u_\delta(x, t)$*, satisfying the initial data* (3.20)*, is almost*

everywhere bounded and converges to $u(x,t)$, then $u(x,t)$ is the unique entropy solution of the initial value problem of (1.3).

Proof: It is similar to the proof of the LW theorem, thus is omitted here. ∎

Remark 3.6 *The condition (3.23) is necessary to ensure that the solution of the conservative numerical scheme converges to the admissible solution. It still implies that under the assumption that the initial function $u_0(x)$ satisfies*

$$u_0(x) = u_*, \quad |x| \geq M,$$

the solution of the conservative numerical scheme satisfying (3.23) is bounded in L^2-norm, i.e.

$$\sum_j |u_j^n - u_*|^2 h \leq c \sum_j \eta(u_0(x_j))h \quad n = 1, 2, \cdots,$$

where $\eta(u_*) = 0$ *and the constant* c *does not depend on* n*.*

Remark 3.7 *The above concepts and theoretical results for conservative numerical scheme may be extended to the multidimensional scalar case.*

The scalar equation (1.3) with $f''(u) > 0$ holds the following conclusion [92].

Lemma 3.3 *The piece-wisely smooth solution* u *of the scalar conservation law* (1.3) *with* $f''(u) > 0$ *satisfies*

$$\frac{\partial}{\partial t}\left(\frac{1}{2}u^2\right) + \frac{\partial}{\partial x}\int_0^u \xi f'(\xi)\, d\xi \leq 0,$$

in the weak sense, then for any entropy pair $\{\eta(u), q(u)\}$ u *satisfies the inequality*

$$\eta_t + q_x \leq 0.$$

3.2 Monotone schemes

Consider the initial value problem of the scalar conservation law (1.3) with initial data

$$u(x,0) = u_0(x). \tag{3.24}$$

Definition 3.3 *If the function* $H(u^n; j)$ *in the FDS*

$$u_j^{n+1} = H(u_{j-l}^n, \cdots, u_{j+l}^n) =: H(u^n; j), \tag{3.25}$$

is non-decreasing with respect to its every independent variable u_ℓ^n, *i.e. if for all* u_ℓ *and* v_ℓ *such that* $u_\ell \le v_\ell$ *one has* $H(..., u_\ell, ...) \le H(..., v_\ell, ...)$, $\ell = j-l, \cdots, j+l$, *then* (3.25) *is called as* **monotone scheme**.

Definition 3.4 *If the FDS* (3.25) *satisfies that*

$$v_j^{n+1} = H(v^n; j) \ge u_j^{n+1} = H(u^n; j), \quad \forall j \in \mathbb{Z},$$

under the assumption that $v_j^n \geq u_j^n$ *for all* $j \in \mathbb{Z}$, *then* (3.25) *is called as* **monotone scheme**.

Lemma 3.4 *The above definitions are equivalent to each other.*

Proof: (1) If assuming that $v_j^n \geq u_j^n$ for all $j \in \mathbb{Z}$, and the function $H(u^n; j)$ is non-decreasing with respect to its every independent variable u_ℓ^n, then we have

$$\begin{aligned} v_j^{n+1} - u_j^{n+1} =& H(v_{j-l}^n, v_{j-l+1}^n, \cdots, v_{j+l}^n) - H(u_{j-l}^n, u_{j-l+1}^n, \cdots, u_{j+l}^n) \\ =& H(v_{j-l}^n, v_{j-l+1}^n, \cdots, v_{j+l}^n) - H(u_{j-l}^n, v_{j-l+1}^n, \cdots, v_{j+l}^n) \\ & + H(u_{j-l}^n, v_{j-l+1}^n, \cdots, v_{j+l}^n) - \cdots - H(u_{j-l}^n, u_{j-l+1}^n, \cdots, u_{j+l}^n) \geq 0. \end{aligned}$$

(2) Conversely, the proof may be completed by choosing some special initial data $\{u^n\}$. For example, for each $p = -l, -l+1, \cdots, l$, taking $v_j^n = u_j^n$ for all j except for that $v_{j+p}^n = u_{j+p}^n + \delta$ with arbitrary $\delta > 0$ gives two initial data $\{v_j^n\}$ and $\{u_j^n\}$ satisfying $v_j^n \geq u_j^n$ for all $j \in \mathbb{Z}$. Due to Definition 3.4, the

inequality $v_j^{n+1} = H(v^n; j) \geq u_j^{n+1} = H(u^n; j)$ holds. It implies that the function $H(u^n; j)$ is non-decreasing with respect to its independent variable u_{j+p}^n, $-l \leq p \leq l$. ∎

Remark 3.8 *Monotone scheme may be non-conservative, and extended to the multi-time level and implicit scheme.*

Lemma 3.5 *The three-point conservative scheme*

$$u_j^{n+1} = u_j^n - \lambda\big(\hat{f}(u_j^n, u_{j+1}^n) - \hat{f}(u_{j-1}^n, u_j^n)\big) =: H(u^n; j), \quad (3.26)$$

is monotone, if $\hat{f}_1 \geq 0$, $\hat{f}_2 \leq 0$, *and* $1-\lambda(\hat{f}_1(u_j, u_{j+1})-\hat{f}_2(u_{j-1}, u_j) \geq 0$, *where* $\hat{f}_1 := \frac{\partial \hat{f}}{\partial u}(u, v)$ *and* $\hat{f}_2 := \frac{\partial \hat{f}}{\partial v}(u, v)$.

Remark 3.9 *The LF, Engquist-Osher, and Godunov schemes for* (1.3) *are monotone under the CFL condition, while the LW and upwind schemes are not monotone.*

Theorem 3.6 *[30] If the* $(2l+1)$*-point conservative monotone scheme*

$$u_j^{n+1} = u_j^n - \lambda\big(\hat{f}(u_{j-l+1}^n, \cdots, u_{j+l}^n) - \hat{f}(u_{j-l}^n, \cdots, u_{j+l-1}^n)\big) =: H(u^n; j), \tag{3.27}$$

is consistent with the conservation law (1.3)*, then its modified equation is of the form*

$$u_t + f(u)_x = \tau[\beta(u,\lambda)u_x]_x, \tag{3.28}$$

which implies that the truncation error of (3.27) *is first order, where*

$$\beta(u,\lambda) = \frac{1}{2\lambda^2}\left[\sum_{k=-l}^{l} k^2 H_k(u, \cdots, u) - \lambda^2 a^2(u)\right] \geq 0. \tag{3.29}$$

Theorem 3.7 *[30] If the conservative monotone scheme* (3.27) *is consistent with the conservation law* (1.3)*, then it satisfies the discrete entropy condition* (3.23)*.*

Proof: For scalar equation (1.3), any convex function $\eta(u)$ is convex entropy, and corresponding entropy flux is

$$q(u) = \int_0^u \eta'(z) f'(z)\, dz.$$

Introduce notations $z \vee u := \max\{z, u\}$ and $z \wedge u := \min\{z, u\}$, which imply $z \vee u - z \wedge u = |u - z|$. The convex entropy $\eta(U)$ may also be expressed as

$$\eta(u) = \frac{1}{2} \int_{\mathbb{R}} \eta'' \left[|u - z| + \mathrm{sign}(z)(u - z) \right] \, dz + \eta'(0)u + \eta(0),$$

where $\eta'' \geq 0$.

Thanks to the definition of H, one has

$$H(z \vee u_{j-l}, \cdots, z \vee u_{j+l}) = z \vee u_j^n - \lambda \Delta_+ \hat{f}(z \vee u_{j-l}, \cdots, z \vee u_{j+l-1}),$$
$$H(z \wedge u_{j-l}, \cdots, z \wedge u_{j+l}) = z \wedge u_j^n - \lambda \Delta_+ \hat{f}(z \wedge u_{j-l}, \cdots, z \wedge u_{j+l-1}).$$

Subtracting those two equations gives

$$H(z \vee u_{j-l}, \cdots, z \vee u_{j+l}) - H(z \wedge u_{j-l}, \cdots, z \wedge u_{j+l}) \\ = |u_j^n - z| - \lambda \Delta_+ \Big[\hat{f}(z \vee u_{j-l}, \cdots, z \vee u_{j+l-1}) \\ - \hat{f}(z \wedge u_{j-l}, \cdots, z \wedge u_{j+l-1}) \Big]. \tag{3.30}$$

Owning to the fact that $z \vee u \geq z$ and $z \vee u \geq u$ and the monotonicity of the function H gives

$$H(z \vee u_{j-l}, \cdots, z \vee u_{j+l}) \geq H(z, \cdots, z) \vee H(u_{j-l}, \cdots, u_{j+l}) = z \vee u_j^{n+1}, \\ H(z \wedge u_{j-l}^n, \cdots, z \wedge u_{j+l}^n) \leq H(z, \cdots, z) \wedge H(u_{j-l}, \cdots, u_{j+l}) = z \wedge u_j^{n+1}.$$

Subtracting those two inequalities gives

$$H(z \vee u_{j-l}^n, \cdots, x \vee u_{j+l}^n) - H(z \wedge u_{j-l}^n, \cdots, z \wedge u_{j+l}^n) \\ \geq z \vee u_j^{n+1} - z \wedge u_j^{n+1} = |u_j^{n+1} - z|. \tag{3.31}$$

Using (3.27) and (3.30)-(3.31) gives

$$\begin{aligned}\eta(u_j^{n+1}) \le & \frac{1}{2}\int_{\mathbb{R}} \eta''(z)\Big[H(z\vee u_{j-l}^n \cdots z\vee u_{j+l}^n) - H(z\wedge u_{j-l}^n \cdots, z\wedge u_{j+l}^n) \\ & + \mathrm{sign}(z)(u_j^{n+1}-z)\Big]\, dz + \eta'(0)u_j^{n+1} + \eta(0) \\ = & \eta(u_j^n) - \frac{\tau}{h}\Delta_+ \hat{q}(u_{j-l}, \cdots, u_{j+l-1}),\end{aligned}$$

where the numerical flux is defined by

$$\begin{aligned}\hat{q}(u_{j-l+1}, \cdots, u_{j+l}) = & \frac{1}{2}\int_{-\infty}^{+\infty} \eta''(z)\Big[\hat{f}(z\vee u_{j-l+1}, \cdots, z\vee u_{j+l}) \\ & - \hat{f}(z\wedge u_{j-l+1}, \cdots, z\wedge u_{j+l}) \\ & + \mathrm{sign}(z)(\hat{f}(u_{j-l+1}, \cdots, u_{j+l}) - f(z))\Big]\, dz \\ & + \eta'(0)\big(\hat{f}(u_{j-l+1}, \cdots, u_{j+l}) - f(0)\big),\end{aligned}$$

which satisfies the consistent condition

$$\begin{aligned}\hat{q}(u,\cdots,u) =& \frac{1}{2}\int_{-\infty}^{+\infty}\eta''(z)\Big[f(z\vee u)-f(z\wedge u)\\ &+\operatorname{sign}(z)(f(u)-f(z))\Big]\,dz+\eta'(0)(f(u)-f(0))\\ =&\int_0^u \eta'(z)f'(z)\,dz = q(z).\end{aligned}$$

The proof is completed. ∎

Combining Theorem 3.2 with Theorem 3.7 gives the following conclusion.

Theorem 3.8 *Assuming the conservative monotone scheme (3.27) is consistent with (1.3). If as $\delta \to 0$, the solution of (3.27), $u_\delta(x,t)$, satisfying the initial data (3.20), is almost everywhere bounded and converges to $u(x,t)$, then $u(x,t)$ is a admissible solution of the initial value problem of (1.3).*

The proof of convergence of the conservative monotone scheme requires the nonlinear stabilities, which imply some compactness.

Theorem 3.9 *[18] Assume that the grid functions* $\{u_j\}$ *and* $\{v_j\}$ *are bounded in* $L^1 \cap L^\infty \cap BV$*, then the conservative monotone scheme* (3.27) *satisfies the following properties:*

$$\text{if } u_j \le v_j, \ \forall j \in \mathbb{Z}, \ \text{then } H(u;j) \le H(v;j), \ \forall j \in \mathbb{Z},$$

$$\min_j\{u_j\} \le H(u;j) \le \max_j\{u_j\}, \ \forall j \in \mathbb{Z},$$

$$||H(u;j) - H(v;j)||_1 \le ||u - v||_1,$$

$$||H(u;j)||_1 \le ||u||_1,$$

$$||H(u;j) - H(v;j)||_{TV} \le ||u - v||_{TV}.$$

where $||u||_1 := h\sum_{j\in\mathbb{Z}} |u_j|$ *and* $||u||_{TV} := \sum_{j\in\mathbb{Z}} |u_{j+1} - u_j|$.

Lemma 3.10 (Pre-compactness [18]) *Assume that* $\{u_\delta(x,t)\}$ *is a set of functions defined in* $\mathbb{R}\times\mathbb{R}^+$, *where* δ *is a positive number tending to 0. If* $\{u_\delta\}$ *satisfies*

$$\begin{aligned}
&\|u_\delta(\cdot,t)\|_1 \le C,\\
&\sup_{|h|\neq 0}\left(\frac{1}{|h|}\|u_\delta(\cdot+h,t)-u_\delta(\cdot,t)\|_1\right)\le C,\\
&\|u_\delta(\cdot,t+\tau)-u_\delta(\cdot,t)\|_1 \le C(|\tau|+\delta),
\end{aligned}$$

where C *is a constant independent on* δ *and* $t\in[0,T]$, *then* $\{u_\delta(x,t)\}$ *contains a subsequence of uniform convergence in* L^1_{loc}*-norm for* $t\in[0,T]$, *where* T *is an arbitrarily given positive number and* $L^1_{loc}(\mathbb{R})$ *is the set of locally integrable function defined in* $\mathbb{R}$, *that is,* $f\in L^1_{loc}(\mathbb{R})$ *if it is integrable in each compact set of* $\mathbb{R}$.

Using the above pre-compactness lemma and the nonlinear stabilities may further prove convergence of the conservative monotone scheme (3.27) for (1.3).

Theorem 3.11 (Convergence of monotone scheme) *If the conservative monotone scheme (3.27) is consistent with (1.3), the numerical flux $\hat{f}$ satisfies Lipschitz condition, $u_\delta(x,t)$ is the solution of the scheme (3.27) satisfying (3.20), where $u_0(x) \in BV(\mathbb{R})$ has compact support, and τ/h keeps constant, then as $\delta \to 0$, in the sense of $L^1_{loc}(\mathbb{R})$-norm, for $t \in [0,T]$, $u_\delta(x,t)$ uniformly converges to the unique admissible solution of (1.3), that is,*

$$\lim_{\delta\to 0} \sup_{0\le t\le T} \int_{\mathbb{R}} |u_\delta(x,t) - u(x,t)|\, dx = 0,$$

where $u(x,t)$ is the admissible solution of (1.3) and T is arbitrarily given positive constant.

The above initial data may be relaxed.

Theorem 3.12 (Convergence of monotone scheme) *If the conservative monotone scheme (3.27) is consistent with (1.3), the numerical flux $\hat{f}$ satisfies Lipschitz condition, $u_\delta(x,t)$ is the solution of the scheme (3.27) satisfying initial data (3.20), where $u_0(x) \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, and τ/h keeps constant, then as $\delta \to 0$, in the sense of $L^1_{loc}(\mathbb{R})$-norm, for $t \in [0,T]$, $u_\delta(x,t)$ uniformly converges to the unique admissible solution of (1.3), where T is arbitrarily given positive constant.*

3.3 Nonlinear stability

Section 3.2 has shown that the conservative monotone scheme (3.27) for scalar conservation law (1.3) has many nonlinear stabilities, satisfies discrete entropy condition, and its solution converges to the unique solution of (1.3), but the $(2l+1)$-point

conservative monotone scheme is only first-order accurate in the sense of truncation error, see Theorem 3.6. Moreover, it is revealed in [74] that not all monotone schemes are non-oscillatory in the sense that no new extreme point is produced. Fig. 7 shows the solution of 1D inviscid Burgers' equation, i.e. (1.3) with the flux

$$f(u) = \frac{1}{2}u^2, \tag{3.32}$$

obtained by using the LF scheme. In view of this, the "non-oscillatory" mentioned in the following generally refers to the "essentially non-oscillatory", unless otherwise specified.

The relationship between several nonlinear stabilities is discussed here under the assumption that the mesh is uniform and initial data $\{u_j^0\}$ satisfy

$$TV(u^0) < \infty, \ ||u^0||_\infty := \max_{j\in\mathbb{Z}}\{|u_j^0|\} < \infty, \ ||u^0||_1 < \infty.$$

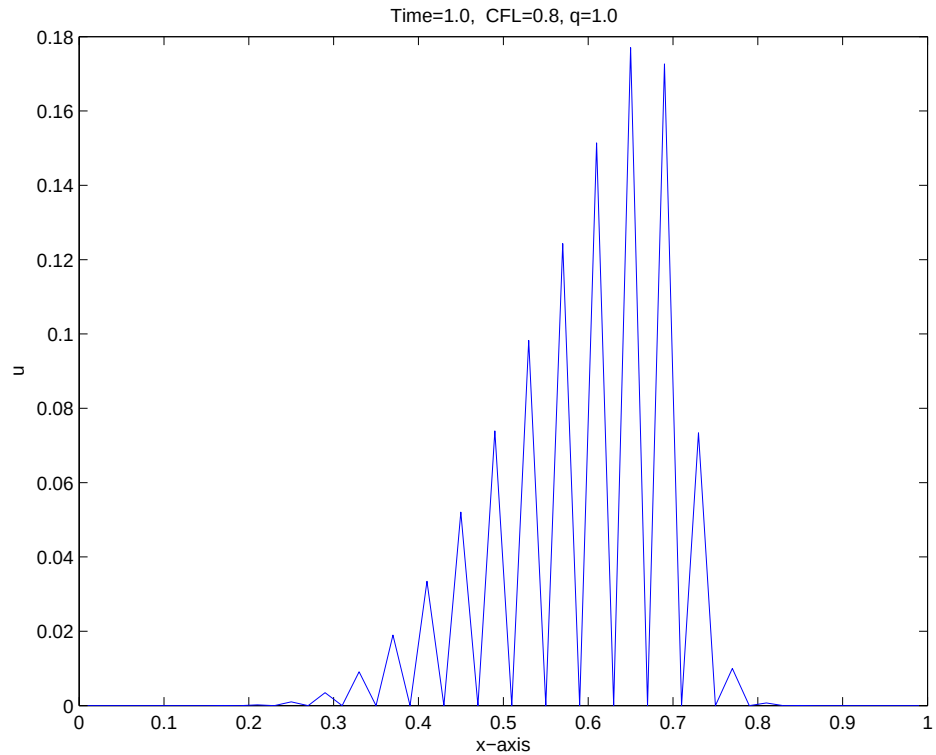


Figure 7: Solution of Burgers' equation obtained by LF scheme.

Definition 3.5 (TVD scheme) *Assume that* $TV(u^0) < \infty$. *If the scheme*

$$u_j^{n+1} = H(u_{j-l}^n, \cdots, u_{j+l}^n) =: H(u^n; j), \tag{3.33}$$

satisfies

$$TV(u^{n+1}) \le TV(u^n), \quad \forall n \ge 0,$$

then (3.33) *is called as* **TVD** *scheme [25].*

Theorem 3.13 *[25] Assume that the scheme* (3.33) *may be written into the incremental form*

$$u_j^{n+1} = u_j^n + C_{j+\frac{1}{2}}^n \Delta_x^+ u_j^n - D_{j-\frac{1}{2}}^n \Delta_x^- u_j^n, \tag{3.34}$$

where $\Delta_x^+ u_j^n := u_{j+1}^n - u_j^n$ *and* $\Delta_x^- u_j^n := u_j^n - u_{j-1}^n$. *If the incremental coefficients in* (3.34) *satisfy*

$$D_{j-\frac{1}{2}}^n \ge 0,\ C_{j+\frac{1}{2}}^n \ge 0,\ 1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n \ge 0, \quad \forall j, n,$$

then (3.33) *is TVD.*

Proof: Translating (3.34) from j to $j+1$, and then subtracting (3.34) from it leads to

$$\begin{aligned}\Delta_x^+ u_j^{n+1} = \Delta_x^+ u_j^n + C_{j+\frac{3}{2}}^n \Delta_x^+ u_{j+1}^n - D_{j+\frac{1}{2}}^n \Delta_x^+ u_j^n \\ - C_{j+\frac{1}{2}}^n \Delta_x^+ u_j^n + D_{j-\frac{1}{2}}^n \Delta_x^- u_j^n.\end{aligned}$$

Taking the absolute value, and using the triangle inequality and the known conditions gives

$$|\Delta_x^+ u_j^{n+1}| \le (1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n)|\Delta_x^+ u_j^n| + C_{j+\frac{3}{2}}^n |\Delta_x^+ u_{j+1}^n| + D_{j-\frac{1}{2}}^n |\Delta_x^- u_j^n|.$$

Summing it with respect to j, and translating the last two terms in j results in $TV(u^{n+1}) \le TV(u^n)$. ∎

Theorem 3.14 *If the incremental coefficients in (3.34) satisfy*

$$D^n_{j-\frac{1}{2}} \geq 0,\ C^n_{j+\frac{1}{2}} \geq 0,\ 1 - C^n_{j+\frac{1}{2}} - D^n_{j-\frac{1}{2}} \geq 0, \quad \forall j, n,$$

then (3.33) satisfies the local extremum principle

$$\min\{u^n_j, u^n_{j\pm1}\} \leq u^{n+1}_j \leq \max\{u^n_j, u^n_{j\pm1}\}.$$

Remark 3.10 *It is possible that the local extremum principle of the monotone scheme (e.g. Lax-Friedrichs scheme) in Theorem 3.9 cannot be proved by using Theorem 3.14, but can be directly completed by using Definition 3.4.*

Lemma 3.15 *Let $\mathcal{Q}$ is a finite difference operator in the following form*

$$(\mathcal{Q} \cdot u)_j = u_j + C_{j+\frac{1}{2}} \Delta u_j - D_{j-\frac{1}{2}} \Delta u_{j-1},$$

then, (1) if for all j, the inequalities

$$C_{j+\frac{1}{2}} \geq 0, \quad D_{j+\frac{1}{2}} \geq 0, \quad C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} \leq 1,$$

hold, then $\mathcal{Q}$ is TVD; (2) if for all j, the coefficients satisfy

$$-\infty < C \leq C_{j+\frac{1}{2}}, \quad D_{j+\frac{1}{2}} \leq 0,$$

then $\mathcal{Q}$ is TVI (total variation increasing).

Definition 3.6 (Monotonicity-preserving scheme) *If the initial data $\{u_j^0\}$ is monotone with respect to j, and the solution $\{u_j^n\}$ of the scheme (3.33) has the same monotonicity in j as that of $\{u_j^0\}$, $\forall n \in \mathbb{N}$, then the scheme (3.33) is called as* **monotonicity-preserving**.

Lemma 3.16 *The TVD scheme with finite points is monotonicity-preserving.*

Proof: Assume that the initial data satisfy $u_j^0 \geq u_{j+1}^0$ for all j and $TV(u^0) = |u_{-\infty}^0 - u_{+\infty}^0| < \infty$. Since the dependence domain is finite, $u_j^n \to u_{\pm\infty}^0$ for any finite n as $j \to \pm\infty$. Thus $TV(u^n) \geq TV(u^0)$. The fact that the scheme is TVD implies $TV(u^n) = TV(u^0)$. Hence the data $\{u_j^n\}$ are non-oscillatory and satisfy $u_j^n \geq u_{j+1}^n$ for all j, otherwise $TV(u^n)$ should be bigger than $TV(u^0)$. ∎

Theorem 3.17 (Godunov) *Any linear monotonicity-preserving scheme is at most first-order accurate.*

Proof: Combining Lemma 3.20 and the accuracy of the monotone scheme may complete the proof. ∎

Definition 3.7 (L^1-contraction scheme) *If for two grid functions u^n, v^n, and $u^n - v^n$ has compact support, the grid functions*

$u_j^{n+1} = H(u^n; j)$ *and* $v_j^{n+1} = H(v^n; j)$ *satisfy*

$$||u^{n+1} - v^{n+1}||_1 \leq ||u^n - v^n||_1,$$

then the numerical method $u_j^{n+1} = H(u^n; j)$ *is called as* L^1**-contraction**.

Lemma 3.18 *The* L^1*-contraction scheme is TVD.*

Proof: For any grid function $\{u_j^n\}$ satisfying $TV(u^n) < \infty$ and $\| u^n - v^n \|_1 < \infty$, where the grid function $\{v_j^n\}$ is defined by (only translation) $v_j^n := u_{j-1}^n$ for all j. Since the scheme has the translation invariance, $v_j^{n+1} = H(v^n; j)$. Thanks to the L^1-contraction, one has

$$TV(u^{n+1}) = \frac{1}{h} \| u^{n+1} - v^{n+1} \|_1 \leq \frac{1}{h} \| u^n - v^n \|_1 = TV(u^n).$$

∎

Lemma 3.19 *[18] The conservative monotone scheme is L^1-contractive.*

Proof: Let $u \vee v := \max\{u, v\}$ and $u \wedge v := \min\{u, v\}$. Assume that two grid functions $\{u_j^n\}$ and $\{v_j^n\}$ satisfy $||u^n - v^n||_1 < \infty$. The monotonicity of the scheme implies

$$|u_j^{n+1} - v_j^{n+1}| = |H(u^n; j) - H(v^n; j)| \leq H(u^n \vee v^n; j) - H(u^n \wedge v^n; j).$$

If summing it with respect to j and using the conservativeness of the scheme and $\hat{f}_{j+\frac{1}{2}} \to 0$ as $j \to \pm\infty$, then

$$\begin{aligned}\sum_j |u_j^{n+1} - v_j^{n+1}| &\leq \sum_j H(u \vee v; j) - \sum_j H(u \wedge v; j) \\ &= \sum_j (u_j^n \vee v_j^n - u_j^n \wedge v_j^n) = \sum_j |u_j^n - v_j^n|.\end{aligned}$$

Thus, $\sum_j |u_j^{n+1} - v_j^{n+1}|h \le \sum_j |u_j^n - v_j^n|h$. ∎

The second-order accurate LW scheme is not monotonicity-preserving, so that the overshoot and undershoot may be produced near the shock wave or discontinuity, see Fig. 8(b). Thus, it is important to use the monotonicity-preserving or nonlinear stable scheme to resolve the shock wave or other discontinuity.

In general, there is the following relation schema

$$\{\text{conservative monotone schemes}\} \subset \{\ell_1\text{-contraction schemes}\}$$
$$\subset \{\text{TVD schemes }\} \subset \{\text{monotonicity-preserving schemes}\}.$$

However, some converse relations may be derived occasionally.

Lemma 3.20 *The linear monotonicity-preserving is monotone.*

Proof: Consider two grid data $\{u_j^n\}$ and $\{v_j^n\}$, where $u_j^n = v_j^n$

as $j \neq j_0$ and $u_{j_0}^n < v_{j_0}^n$. To prove the conclusion, it has to prove $u_j^{n+1} \leq v_j^{n+1}$.

Define a monotone non-decreasing grid function

$$w_j^n = \begin{cases} u_{j_0}^n, & j < j_0, \\ v_{j_0}^n, & j \geq j_0. \end{cases}$$

Thus, the solution of the monotonicity-preserving scheme, $w_j^{n+1} = H(w^n; j,$ is monotone non-decreasing in j. It is worth noting that, for all j, $w_j^n = w_{j-1}^n + (v_j^n - u_j^n)$. Because the scheme is linear, $w_j^{n+1} = w_{j-1}^{n+1} + (v_j^{n+1} - u_j^{n+1})$, which implies that $v_j^{n+1} = u_j^{n+1} + (w_j^{n+1} - w_{j-1}^{n+1}) \geq u_j^{n+1}$. ∎

Lemma 3.21 ([71]) *A three-point monotonicity-preserving scheme is TVD.*

Proof: From incremental form $u_j^{n+1} = u_j^n + C_{j+\frac{1}{2}}^n \Delta_x^+ u_j^n - D_{j-\frac{1}{2}}^n \Delta_x^+ u_{j-1}^n$, one has

$$\Delta_x^+ u_j^{n+1} = \Delta_x^+ u_j^n \left(1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n\right) + C_{j+\frac{3}{2}}^n \Delta_x^+ u_{j+1}^n + D_{j-\frac{1}{2}}^n \Delta_x^+ u_{j-1}^n.$$

If taking $u_{j-1}^n = u_j^n = u_{j+1}^n$, that is, $\Delta_x^+ u_{j-1}^n = \Delta_x^+ u_j^n = 0$, then

$$\Delta_x^+ u_j^{n+1} = C_{j+\frac{3}{2}}^n \Delta_x^+ u_{j+1}^n.$$

Thanks to the monotonicity-preserving, the sign of $\Delta_x^+ u_j^{n+1}$ has to be the same as that of $\Delta_x^+ u_{j+1}^n$, thus $C_{j+\frac{3}{2}}^n$ is non-negative for arbitrary u_{j+1} and u_{j+2}. Similarly, if letting $\Delta_x^+ u_{j-1}^n = \Delta_x^+ u_{j+1}^n = 0$ or $\Delta_x^+ u_j^n = \Delta_x^+ u_{j+1}^n = 0$, then $D_{j-\frac{1}{2}}^n \geq 0$, $1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n \geq 0$. ∎

A more general two-time-level conservative scheme is given as follows [26]

$$u_j^{n+1} = u_j^n - \lambda \left(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}} \right), \tag{3.35}$$

where $\hat{f}_{j+\frac{1}{2}} = \hat{f}\left(u_{j-r+1}^n, \cdots, u_{j+r}^n; u_{j-l+1}^{n+1}, \cdots, u_{j+l}^{n+1}\right)$ is the numerical flux, being Lipschitz continuous and satisfying the consistency $\hat{f}(u, \cdots, u; u, \cdots, u) = f(u)$.

Lemma 3.22 *Assume that (3.35) may be written as follows*

$$\begin{cases} \mathcal{L} \cdot u^{n+1} = \mathcal{R} \cdot u^n, \\ (\mathcal{L} \cdot u)_j = u_j + \eta\lambda(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}), \\ (\mathcal{R} \cdot u) = u_j - (1-\eta)\lambda(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}), \end{cases} \tag{3.36}$$

where $\eta \in [0,1]$. *If* $\mathcal{R}$ *is a TVD operator, i.e.* $TV(\mathcal{R}\cdot v) \le TV(v)$, $\mathcal{L}$ *is a TVI operator, i.e.* $TV(\mathcal{L}\cdot u) \ge TV(u)$, *then the scheme* $u_j^{n+1} = (\mathcal{L}^{-1}\mathcal{R}\cdot u)_j^n$ *is TVD.*

Theorem 3.23 *If*

$$D_{j-\frac{1}{2}}(t) \ge 0,\ C_{j+\frac{1}{2}}(t) \ge 0, \qquad \forall j \in \mathbb{Z},\ t \ge 0,$$

then the semi-discrete scheme in incremental form

$$\frac{du_j(t)}{dt} = C_{j+\frac{1}{2}}(t)\Delta_x^+ u_j(t) - D_{j-\frac{1}{2}}(t)\Delta_x^- u_j(t), \tag{3.37}$$

is TVD.

Proof: Translating (3.37) from j to $j+1$, and subtracting (3.37)

from it results in

$$\begin{aligned}\frac{d\Delta_x^+ u_j(t)}{dt} =& C_{j+\frac{3}{2}}(t)\Delta_x^+ u_{j+1}(t) + D_{j-\frac{1}{2}}(t)\Delta_x^- u_j(t) \\ & - (C_{j+\frac{1}{2}}(t) + D_{j+\frac{1}{2}}(t))\Delta_x^+ u_j(t). \end{aligned} \tag{3.38}$$

Multiplying it by $s_j = \text{sign}\{\Delta_x^+ u_j(t)\}$ both sides gives

$$\begin{aligned} s_j \frac{d\Delta_x^+ u_j(t)}{dt} =& -\big(C_{j+\frac{1}{2}}(t) + D_{j+\frac{1}{2}}(t)\big)|\Delta_x^+ u_j(t)| \\ & + s_j\big(C_{j+\frac{3}{2}}(t)\Delta_x^+ u_{j+1}(t) + D_{j-\frac{1}{2}}(t)\Delta_x^- u_j(t)\big). \end{aligned}$$

Under the known conditions, one has

$$\begin{aligned} \frac{d|\Delta_x^+ u_j(t)|}{dt} \le& -\big(C_{j+\frac{1}{2}}(t) + D_{j+\frac{1}{2}}(t)\big)|\Delta_x^+ u_j(t)| \\ & + C_{j+\frac{3}{2}}(t)|\Delta_x^+ u_{j+1}(t)| + D_{j-\frac{1}{2}}(t)|\Delta_x^- u_j(t)|. \end{aligned}$$

Summing it in terms of j, and translating the last two terms in j gives $\frac{d}{dt}TV(u(t)) \leq 0$. ∎

Remark 3.11 *Besides two TVD conditions in Theorem 3.23, the explicit, fully-discrete scheme corresponding to (3.37) should still satisfy*

$$\tau C_{j+\frac{1}{2}} + \tau D_{j+\frac{1}{2}} \leq A, \quad A = const > 0. \tag{3.39}$$

It is because the time step-size τ should be constrained after the time derivative is discretized. The restriction (3.39) is related to the third condition in Theorem 3.13.

Theorem 3.24 *[35] The semi-discrete, multi-point difference scheme*

$$\frac{d}{dt}u_j = \sum_{q=-Q}^{Q-1} C_q(j)(u_{j-q} - u_{j-q-1}). \tag{3.40}$$

is TVD if the conditions

$$(1). \quad C_{-1}(j-1) \geq C_{-2}(j-2) \geq \cdots \geq C_{-Q}(j-Q) \geq 0,$$
$$(2). -C_0(j) \geq -C_1(j+1) \geq \cdots \geq -C_{Q-1}(j+Q-1) \geq 0,$$

hold.

Proof: Translating (3.40) from j to $j+1$ and subtracting (3.40)

from it gives

$$\begin{aligned}\frac{d}{dt}\Delta_x^+ u_j = &-C_{Q-1}(j)(u_{j-Q+1}-u_{j-Q})\\ &+[C_{Q-1}(j+1)-C_{Q-2}(j)](u_{j-Q+2}-u_{j-Q+1})\\ &+\cdots+[C_2(j+1)-C_1(j)](u_{j-1}-u_{j-2})\\ &+[C_1(j+1)-C_0(j)](u_j-u_{j-1})+[C_0(j+1)-C_{-1}(j)]\Delta u_j\\ &+[C_{-1}(j+1)-C_{-2}(j)](u_{j+2}-u_{j+1})+\cdots\\ &+[C_{-Q+1}(j+1)-C_{-Q}(j)](u_{j+Q}-u_{j+Q-1})\\ &+C_{-Q}(j+1)(u_{j+Q+1}-u_{j+Q}).\end{aligned}$$

Similar to the proof of Theorem 3.23, the proof may be completed. ∎

3.4 High resolution schemes

This section introduce several high resolution schemes, which have become a class of important numerical methods for quasi-linear hyperbolic conservation laws. There are two different ways for the calculation of the shock wave. The first is *shock-fitting* or *-tracking* methods, in which the shock wave is first accurately identified and considered as an inner boundary, and then the numerical method is used to solve the governing equations in the smooth region of the solution, while the Rankine-Hugoniot discontinuity conditions or the characteristic relations across the shock wave are used to specify the inner boundary conditions. The second way is the *shock-capturing method*, in which the unified scheme is used in the computational domain with no need to know whether there is a shock wave in the domain. In other words, the shock-capturing method does not identify the shock

wave as the inner boundary, but it automatically captures the shock wave with a unified scheme or code in the discontinuous and smooth area. Thanks to its simplicity, it has been the most widely used to numerically calculate the shock wave.

A more representative shock-capturing method is introduced by von Neumann and Richtmyer in 1950 [91] and becomes quite effective in simulating the unsteady compressible fluid flow with the shock wave, contact discontinuity and their interactions. The idea of the von Neumann and Richtmyer method is to add an artificial viscosity into the original equations governing fluid flow, see Section 3.4.1. Most of the conservative schemes introduced above such as the upwind, LF, LW schemes can be classified as the classic shock-capturing methods, while the high resolution schemes introduced below belong to the modern shock-capturing methods.

The main features of the high resolution schemes are as fol-

lows:

- at least second-order accurate in the smooth region of the solution,
- no spurious oscillation or wiggle in the solution of the numerical scheme,
- more sharp resolution of the shock wave than the first-order accurate scheme. More generally, small number of mesh points containing the wave in comparison with a 1st-order scheme with similar grid accuracy.

Why do we need to discuss high resolution scheme? To answer this question, let us see a numerical example.

Example 3.8 *Use the first-order accurate upwind scheme, second-order accurate LW scheme, and MUSCL (Monotone Upstream*

Schemes for Conservation Laws) scheme [87] which is a high resolution shock-capturing scheme, to respectively solve the Riemann problem of (2.1) with initial data

$$u(x,0) = \begin{cases} 1.4, & x < -0.5, \\ -1., & x > -0.5. \end{cases}$$

The computed solutions are displayed in Fig. 8. We see that the resolution of the discontinuity by the upwind scheme is more lower than the LW scheme, but the numerical oscillations behind the discontinuity appear in the solution computed by the LW scheme. The MUSCL scheme resolves more sharply the discontinuity without spurious oscillation or wiggle than other schemes.

To construct a high resolution scheme, the high-order accurate scheme is employed as soon as possible, but it should be modified as a high-order accurate nonlinear stable scheme so

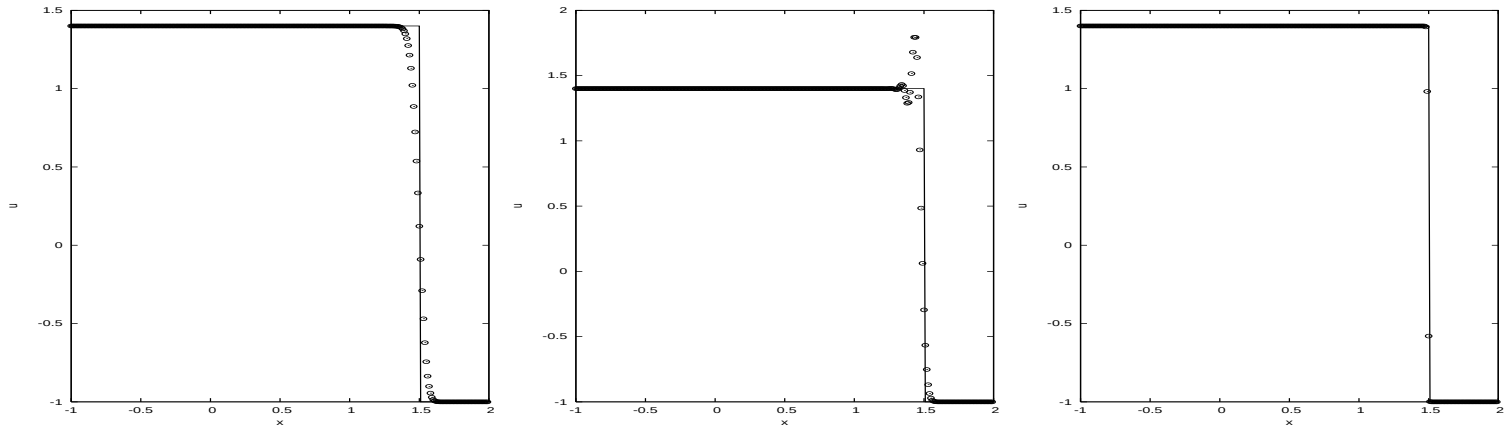


Figure 8: Example 3.8: Computed solutions at $t = 20$ for $a = 0.1$ obtained by using first-order accurate upwind scheme, second-order accurate LW scheme, and MUSCL scheme (from left to right), respectively.

that it can be very accurate and non-oscillatory in the vicinity of the discontinuity and thus capture the discontinuity with high resolution. The importance of Godunov's theorem (Theorem

3.17) in designing the high resolution shock-capturing scheme is recognized in 1970s. Here are three more representative works. In 1971, Boris developed a second-order accurate convection scheme, SHASTA (Sharp and Smooth Transport Algorithm) [3]. In 1972, van Leer did a non-oscillatory modification of the LW scheme [83], and Kolgan proposed a second-order accurate Godunov scheme in space for the Euler equations with a limiter, which made the scheme monotonicity-preserving in the scalar case [40]. The Kolgan's article [40] originally appears in Russian in the TsAGI Research Notes, and remains little known within and completely unknown outside the USSR, but is a true milestone in the history of CFD [90]. These methods are more or less different, but what they have in common is the nonlinear switch function (i.e. limiter), which plays a key role in preventing spurious oscillations near the discontinuity or in the region of a large gradient variation of the solution. Such a theme has

been persisted in developing the modern shock-capturing methods. Subsequently, 4 series papers of Boris and co-workers were published on the nonlinear limiter and transport scheme in the Journal of Computational Physics, and van Leer also published a series of other four papers entitled "Towards ultimate conservative difference scheme" in the same journal.

3.4.1 Artificial viscosity method

Intuitive approach to construct high resolution shock-capturing scheme is as follows: add an additional artificial viscosity term (e.g. in proportion to u_{xx}) to the original equation and then discretize it by a high-order accurate scheme (e.g. the LW scheme). The artificial viscosity should approach to 0 as $\max\{\tau, h\} \to 0$ in order to ensure the consistency. Moreover, the artificial viscosity can quickly vanish in the smooth region of the solution so

that the resulting scheme is high-order accurate in the smooth region of the solution. The artificial viscosity method can be traced back to the von Neumann and Richtmyer work on the shock calculation in fluid mechanics [91].

Example 3.9 *The Lagrangian method in fluid dynamics is based on the following system governing invisid compressible fluid flow*

$$\frac{D\rho}{Dt} = -\rho\nabla \boldsymbol{u}, \quad \frac{D\boldsymbol{u}}{Dt} = -\frac{1}{\rho}\nabla p, \quad \frac{De}{Dt} + p\frac{D}{Dt}(\frac{1}{\rho}) = 0,$$

where e is the specific internal energy. To solve the above system, the method of von Neumann and Richtmyer is to introduce artificial viscosity terms into the above system by replacing the pressure p with $p + q$, and then the spatial derivatives are approximated by using the central difference quotations, where q is

defined by

$$q = \begin{cases} l_0^2 \rho \left(\frac{\partial u}{\partial x}\right)^2, & \text{if } \frac{\partial u}{\partial x} < 0, \\ 0, & \text{if } \frac{\partial u}{\partial x} \geq 0, \end{cases}$$

here l_0 has dimensions of length. Typically l_0 is chosen to be some small multiple of the grid spacing h, e.g. $l_0 = a_0 h$ with the constant $a_0 \approx 2$.

Example 3.10 *Consider the modification of the LW scheme for the convection equation* (2.1)

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\nu^2}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \tau Q(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \tag{3.41}$$

where the Courant number $\nu = a\frac{\tau}{h}$ and the coefficient Q in the artificial viscosity term is assumed to be constant. The truncation

error of (3.41) *is equal to*

$$L(x,t) := \frac{1}{\tau}[u_j^{n+1} - H(u;j)] = L^{LW}(x,t) - Q[u(x+h,t) - 2u(x,t) + u(x-h,t)]$$
$$= L^{LW}(x,t) - Qh^2 u_{xx}(x,t) + \mathcal{O}(h^4) = \mathcal{O}(\tau^2), \quad as\ \tau \to 0,$$

which implies that the scheme (3.41) *is second-order accurate in the sense of truncation error. We expect that the artificial viscosity in* (3.41) *may prevent the spurious oscillations generated by the LW scheme, see Example 3.8, and the scheme* (3.41) *is monotonicity-preserving. However, unfortunately, this is not the case, because* (3.41) *is still linear and second-order accurate when* Q *is constant, thus it is not monotonicity-preserving or "non-oscillatory" due to Godunov's theorem (Theorem 3.17).*

In order to derive a high-order accurate monotonicity-preserving scheme, Q *in* (3.41) *has to rely on the solution* $\{u_j^n\}$*, so that* (3.41) *should be nonlinear even for the convection equation* (2.1) *with constant coefficient. To this end, the artificial viscosity in*

(3.41) *should be replaced with a nonlinear, conservative artificial viscosity term*

$$\tau\left(Q\big(u^n; j+\frac{1}{2}\big)(u^n_{j+1}-u^n_j) - Q\big(u^n; j-\frac{1}{2}\big)(u^n_j - u^n_{j-1})\right), \tag{3.42}$$

where the artificial viscosity coefficient $Q(u^n; j+\frac{1}{2})$ *depends on some point values of the solution such as* $u^n_{j-p}\cdots u^n_{j+q}$, *and follows the rule: it becomes small or vanishing in the smooth area while big enough near the discontinuity in order to preserve monotonicity (as well as satisfying the entropy condition).*

More generally, after adding a nonlinear artificial viscosity into a given high-order accurate conservative scheme with numerical flux $\hat{f}^H(u; j+\frac{1}{2})$ for (1.3), the numerical flux of the new

conservative scheme should be of the form

$$\hat{f}^{\text{new}}(u; j+\frac{1}{2}) = \hat{f}^{H}(u; j+\frac{1}{2}) - \tau Q\big(u; j+\frac{1}{2}\big)(u_{j+1}-u_j), \quad (3.43)$$

where the artificial viscosity $Q(u; j+\frac{1}{2})$ is only needed near the discontinuity so that it should depend on the solution and be bigger near the discontinuity than in the smooth region. The idea of the artificial viscosity method is very direct and simple, but the key task is to present a suitable formula of $Q(u; j+\frac{1}{2})$, which introduces enough dissipation, preserves the monotonicity, and does not smearing the shock wave out. Practically, it is quite hard to reach such target. In view of such reason, the different ways should be found to develop high resolution shock-capturing methods.

3.4.2 Slope limiter method

The key idea of the slope limiter method is to use some piecewise higher-order function (e.g. piecewise linear or quadratic function etc.) to approximate the solution instead of the piecewise constant function in the Godunov method. The slope limiter method may be classified as the finite volume method and can be traced back to Kolgan's article [40], which featured a Godunov-type scheme for the Euler equations with second-order spatial accuracy and a limiter. Kolgan's method is the generalized MUSCL scheme [55] to be mentioned. Other relative works may be found in [83, 84, 85, 86, 87, 15, 24].

As we know, the Godunov method is only first-order accurate. Is it possible to improve the accuracy of the Godunov scheme? Fig. 9 shows that the piecewise constant function $u_h(x,t)$ reconstructed by the cell averages $\{\bar{u}_j^n\}$ is generally not a good

approximation of the original function $u(x,t)$, where the continue solid line denotes the profile of $u(x,t)$, while the others u_h and w_h denote the piecewise constant and linear functions approximating u respectively. Hence, in order to present a high-order accurate Godunov scheme, the reconstruction may be used to give high-order accurate approximation of the solution at each time level.

Example 3.11 (An extension of Godunov method) *The Godunov method may be replaced with the three steps.*

(1). Given the approximation of the initial cell averages $\{\bar{u}_j^n\}$, *reconstruct a piecewise linear function* $\tilde{u}_h(x,t_n)$ *as*

$$\tilde{u}_h(x,t_n) := \bar{u}_j^n + s_j^n(x - x_j), \quad x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) =: I_j, \quad (3.44)$$

here s_j *is an approximation slope of* u *over the cell* I_j, *i.e.* $s_j \approx (u_x)_j$.

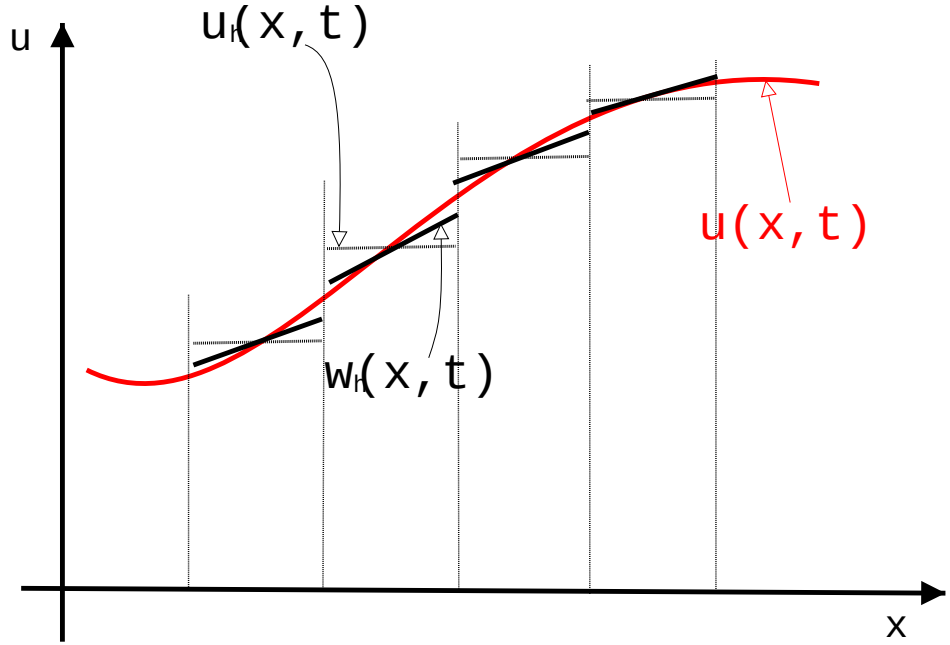


Figure 9: Piecewise constant and linear functions $u_h(x,t)$ and $w_h(x,t)$ approximating the general function $u(x,t)$.

(2). Solve the initial value problem of (1.3) *with*

$$u(x, t_n) = \tilde{u}_h(x, t_n),$$

to get the solution $u(x,t)$, $t_n \le t < t_{n+1}$.

(3). Calculate the cell averages $\{\bar{u}_j^{n+1}\}$ *at* $t = t_{n+1}$ *by*

$$\bar{u}_j^{n+1} = \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_{n+1} - 0)\, dx.$$

Obviously, if $s_j^n = 0$ *for all* (j,n), *then the above method reduces to the original Godunov method. If* $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$, *then for arbitrary* s_j^n, *the cell average of* $\tilde{u}_h(x, t_n)$ *over the cell* I_j *is equal to* $\bar{u}_j^n$, *thus the method preserves the mass conservation. The piecewise higher-order approximation may be used in the reconstruction step, see Fig.* 10 *for a schematic diagram. Two typical*

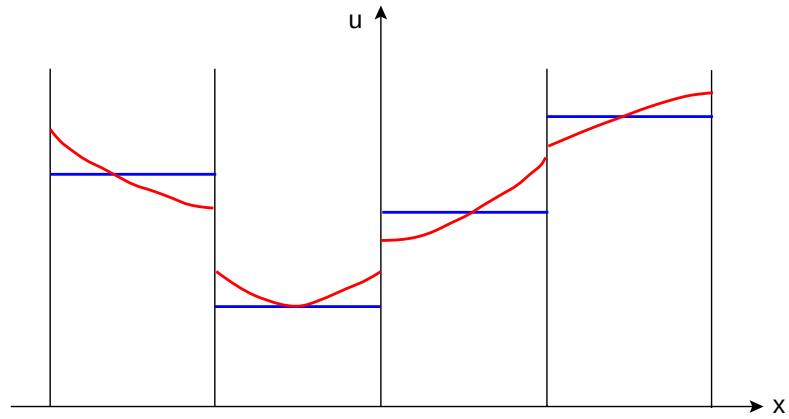


Figure 10: Schematic diagram of higher-order reconstruction.

examples are the piecewise parabolic method (PPM) of Colella and Woodward [16] and ENO method of Harten et al. [29].

In order to implement the method in Example 3.11, two key problems has to be solved: How to choose the approximate slope s_j^n? And how to solve the initial value problem in the second step? If consider a simple case of that $f = au$ with a constant a,

then the exact solution of the initial value problem in Example 3.11 is given by

$$\tilde{u}(x, t_{n+1}) = \tilde{u}_h(x - a\tau, t_n),$$

and thus the cell averages in the third step may further be calculated as follows

$$\begin{aligned}
\bar{u}_j^{n+1} =& \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}(x, t_{n+1})\, dx = \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}_h(x - a\tau, t_n)\, dx \\
=& \frac{1}{h} \int_{x_{j-\frac{1}{2}}-a\tau}^{x_{j+\frac{1}{2}}-a\tau} \tilde{u}_h(\xi, t_n)\, d\xi \overset{a\geq 0}{=} \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}-a\tau} [\bar{u}_j^n + s_j^n(\xi - x_j)]\, d\xi \\
&+ \frac{1}{h} \int_{x_{j-\frac{1}{2}}-a\tau}^{x_{j-\frac{1}{2}}} [\bar{u}_{j-1}^n + s_{j-1}^n(\xi - x_{j-1})]\, d\xi \\
=& \bar{u}_j^n - \nu(\bar{u}_j^n - \bar{u}_{j-1}^n) - \frac{h}{2}\nu(1-\nu)(s_j^n - s_{j-1}^n).
\end{aligned}$$

Here we have used the assumption that $0 < a\frac{\tau}{h} < 1$, but the above result may be extended to the other cases. Hence, for the linear convection equation (2.1), a scheme based on the piecewise linear reconstruction is of the form

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \nu(\bar{u}_{j_1}^n - \bar{u}_{j_1-1}^n) - \frac{h}{2}\nu(\mathrm{sign}(\nu) - \nu)(s_{j_1}^n - s_{j_1-1}^n), \quad (3.45)$$

with the numerical flux

$$\hat{f}_{j+\frac{1}{2}} = a\bar{u}_{j_1} + \frac{a}{2}(\mathrm{sign}(\nu) - \nu)hs_{j_1},$$

where

$$j_1 = \begin{cases} j, & \text{if } a > 0, \\ j+1, & \text{if } a < 0. \end{cases}$$

If $s_j^n = 0$ for all (j, n), then (3.45) reduces to the first-order

accurate upwind scheme. If

$$s_j^n = \frac{\Delta_x^+ \bar{u}_j^n}{h}, \tag{3.46}$$

for all (j, n), then (3.45) becomes the second-order accurate LW scheme. Thus it is possible to get a second-order accurate scheme by the initial reconstruction. However, the oscillations may be caused by such incautious choice of the approximate slope. Fig. 11 gives a diagrammatic sketch of a geometric interpretation of the better and bad choices of the approximate slope s_j^n. The bad slope may result in monotonicity-violating or a larger total variation of piecewise linear function $\tilde{u}_h(x, t_n)$ than that of $\{\bar{u}_j^n\}$. A cure in the slope limiter method is to present the approximate slope via a nonlinear limiter function instead of the incautious choice, e.g (3.46), such that the reconstructed function $\tilde{u}_h(x, t_n)$

satisfies some nonlinear stability such as

$$TV(\tilde{u}_h(\cdot, t_n)) \leq TV(u^n). \tag{3.47}$$

A simplest approximate slope satisfying (3.47) is

$$s_j = \frac{1}{h}\text{minmod}(\Delta_x^+ u_j, \Delta_x^- u_j),$$

where the function $\text{minmod}(a, b)$ is defined by

$$\text{minmod}(a, b) = \frac{\text{sign}(a) + \text{sign}(b)}{2} \min(|a|, |b|) = \begin{cases} a, & |a| \leq |b|, ab > 0, \\ b, & |b| < |a|, ab > 0, \\ 0, & ab \leq 0. \end{cases}$$

Let $r_j = \dfrac{\Delta_x^- u_j}{\Delta_x^+ u_j}$, the above approximate slope can be rewritten as

$s_j^n = \frac{1}{h}\Delta_x^+ u_j \phi(r_j^n)$ where

$$\phi(r) = \max\{0, \min(1, r)\} = \begin{cases} 0, & \text{if } r \le 0, \\ r, & \text{if } 0 \le r \le 1, \\ 1, & \text{if } r \ge 1. \end{cases}$$

Geometrically, the restriction that $\phi(r) = 0$ if $r \le 0$ is reasonable, otherwise the slope will result in a reconstructed function $\tilde{u}_h(x, t_n)$ with a larger total variation than $TV(u^n)$ so that (3.47) is violated.

Assuming that the second and third steps in Example 3.11 have been exactly solved so that they are TVD for the scalar equation (1.3), then combing them with (3.47), a second-order accurate TVD method can be gotten. Unfortunately, for a quasilinear equation (1.3), the initial value problem in the second step belongs to generalized Riemann problem (GRP) [2], which is quite

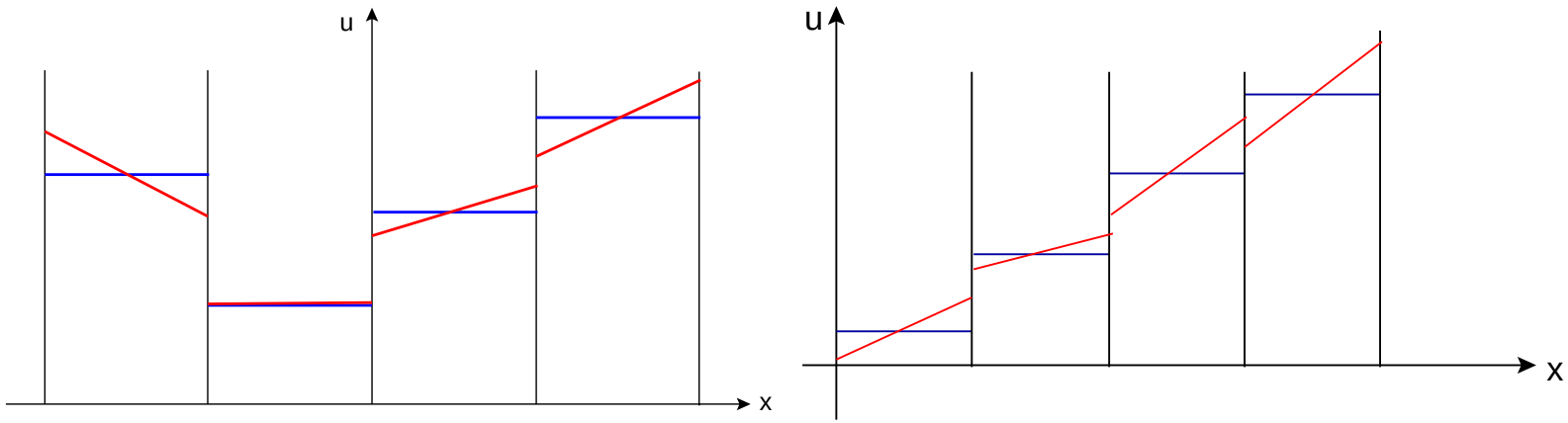


Figure 11: Piecewise linear reconstruction with a good (left) or bad (right) slope.

hard to be solved so that the above extension of the Godunov method seems generally impractical. One possible way is to use the approximate solution of the GRP instead of the exact. For example, reconstruct a piecewise function $\tilde{f}(u)$ approximating $f(u)$ by the given data $\{\bar{u}_j^n, f(\bar{u}_j^n)\}$, and then solve exactly the

initial value problem of "approximate conservation law"

$$u_t + \tilde{f}(u)_x = 0, \qquad t_n \le t < t_{n+1},$$
$$u(x, t_n) = \bar{u}_j^n + s_j^n (x - x_j), \quad x \in I_j.$$

An alternate is to use the Riemann data to replace the initial data in the above second step, so that the generalized MUSCL type method [55] is derived.

Example 3.12 (A second-order accurate Godunov method) *The extension of Godunov method in Example 3.11 may be modified as follows.*

(1). According to the given (approximate) cell averages $\{\bar{u}_j^n\}$, *reconstruct a piecewise linear function* $\tilde{u}_h(x, t_n)$

$$\tilde{u}_h(x, t_n) = \bar{u}_j^n + s_j^n (x - x_j), \quad x \in I_j, \tag{3.48}$$

where s_j is the approximate slope defined in the cell I_j, i.e. $s_j \approx (u_x)_j$.

(2). Solving exactly the local Riemann problem

$$\text{Eq. (1.3)}, \quad t_n \le t < t_{n+1},$$

$$u(x, t_n) = \begin{cases} u^L_{j+\frac{1}{2}}, & x < x_{j+\frac{1}{2}}, \\ u^R_{j+\frac{1}{2}}, & x > x_{j+\frac{1}{2}}, \end{cases}$$

gives $\tilde{u}(x, t_{n+1} - 0)$, where

$$\begin{aligned} u^L_{j+\frac{1}{2}} =& \tilde{u}_h(x_{j+\frac{1}{2}} - 0, t_n) = u^n_j + \frac{h}{2} s^n_j, \\ u^R_{j+\frac{1}{2}} =& \tilde{u}_h(x_{j+\frac{1}{2}} + 0, t_n) = u^n_{j+1} - \frac{h}{2} s^n_{j+1}. \end{aligned}$$

(3). Calculate the cell averages $\{\bar{u}_j^{n+1}\}$ *by*

$$\bar{u}_j^{n+1} = \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}(x, t_{n+1} - 0)\, dx.$$

The above procedure has been considered by Kolgan [40] for 1D Euler equations and may be further extended to a more general case. After giving any three-point scheme of first-order accuracy

$$u_j^{n+1} = u_j^n - \lambda\left(\hat{f}_{j+\frac{1}{2}}^n - \hat{f}_{j-\frac{1}{2}}^n\right), \quad \hat{f}_{j+\frac{1}{2}}^n = \hat{f}(u_j^n, u_{j+1}^n),$$

and the piecewise linear reconstruction (3.48), the generalized MUSCL method [55] may be given by

$$u_j^{n+1} = u_j^n - \lambda\left(\hat{f}(u_{j+\frac{1}{2}}^L, u_{j+\frac{1}{2}}^R) - \hat{f}(u_{j-\frac{1}{2}}^L, u_{j-\frac{1}{2}}^R)\right). \tag{3.49}$$

Here the mesh is assumed to be uniform and the notation u_j^n is uniformly used to denote the approximate cell average of solution.

Lemma 3.25 *If the numerical flux $\hat{f}_{j+\frac{1}{2}}$ is Lipschitz continuous, and $u^L_{j+\frac{1}{2}} - u_{j+\frac{1}{2}} = \mathcal{O}(h^2), u^R_{j+\frac{1}{2}} - u_{j+\frac{1}{2}} = \mathcal{O}(h^2)$, where $u_{j+\frac{1}{2}} = \frac{1}{2}(u_j + u_{j+1})$, then (3.49) is second-order accurate in space.*

Proof: It only needs prove $\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}) = \mathcal{O}(h^2)$. Thanks to the consistency and Lipschitz continuity of $\hat{f}_{j+\frac{1}{2}}$, we have

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

where L_i denotes the Lipschitz constant, $i = 1, 2$. ∎

Two sufficient TVD conditions for the semi-discrete MUSCL schemes corresponding to (3.49) will be given below, where "*the scheme with numerical flux $\hat{f}(u_j, u_{j+1})$ is TVD*" means that if it is rewritten in an incremental form

$$-\frac{1}{h}\Delta_x^-(\hat{f}(u_j,u_{j+1})) = -\frac{1}{h}\frac{\hat{f}(u_j,u_{j+1})-f(u_j)}{u_{j+1}-u_j}\Delta_x^+\bar{u}_j$$
$$-\frac{1}{h}\frac{-\hat{f}(u_{j-1},u_j)+f(u_j)}{u_j-u_{j-1}}\Delta_x^-\bar{u}_j =: C_{j+\frac{1}{2}}\Delta_x^+\bar{u}_j - D_{j-\frac{1}{2}}\Delta_x^-\bar{u}_j,$$

then the incremental coefficients satisfy $C_{j+\frac{1}{2}} \geq 0$ and $D_{j-\frac{1}{2}} \geq 0$. It is obvious that the semi-discrete E and monotone schemes satisfy the above condition.

Theorem 3.26 *Assume that the scheme with numerical flux $\hat{f}(u_j, u_{j+1})$ is TVD. The semi-discrete MUSCL scheme with the numerical*

flux $\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}})$ *is also TVD if*

$$D^1_j := \frac{u^R_{j+\frac{1}{2}} - u^L_{j+\frac{1}{2}}}{\Delta^+_x u_j} \geq 0,\ D^2_j := \frac{u^R_{j-\frac{1}{2}} - u^L_{j+\frac{1}{2}}}{\Delta^+_x u_j} \leq 0,\ D^3_j := \frac{u^R_{j-\frac{1}{2}} - u^L_{j+\frac{1}{2}}}{\Delta^-_x u_j} \leq 0. \tag{3.50}$$

Proof: Rewrite $-\Delta^+_x \hat{f}(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}})$ as follows

$$\begin{aligned}
&\Big[-\Big(\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}}) - f(u^L_{j+\frac{1}{2}})\Big) + (\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^L_{j+\frac{1}{2}}))\Big] \\
&- \Big[\Big(\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^R_{j-\frac{1}{2}})\Big) - (\hat{f}(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^R_{j-\frac{1}{2}}))\Big] \\
&= C_{j+\frac{1}{2}} \Delta^+_x u_j - D_{j-\frac{1}{2}} \Delta^-_x u_j,
\end{aligned}$$

where

$$
\begin{aligned}
C_{j+\frac{1}{2}} &= -\frac{\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}}) - f(u^L_{j+\frac{1}{2}})}{u^R_{j+\frac{1}{2}} - u^L_{j+\frac{1}{2}}} \cdot D^1_j \\
&\quad + \frac{\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^L_{j+\frac{1}{2}})}{u^R_{j-\frac{1}{2}} - u^L_{j+\frac{1}{2}}} \cdot D^2_j, \\
D_{j-\frac{1}{2}} &= \frac{\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^R_{j-\frac{1}{2}})}{u^R_{j-\frac{1}{2}} - u^L_{j+\frac{1}{2}}} \cdot D^3_j \\
&\quad - \frac{\hat{f}(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - f(u^R_{j-\frac{1}{2}})}{u^R_{j-\frac{1}{2}} - u^L_{j-\frac{1}{2}}} \cdot D^1_{j-1}.
\end{aligned}
$$

Thus, $C_{j+\frac{1}{2}} \geq 0$ and $D_{j-\frac{1}{2}} \geq 0$ for all $j \in \mathbb{Z}$. Thanks to Theorem 3.23, the semi-discrete MUSCL scheme with numerical flux

$\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}})$ is TVD. ∎

Theorem 3.27 *Assume that the numerical flux $\hat{f}(u_j, u_{j+1})$ satisfies the first two conditions in Lemma 3.5, that is, $\hat{f}_1 \geq 0$ and $\hat{f}_2 \leq 0$, where $\hat{f}_1 := \frac{\partial \hat{f}}{\partial u}(u, v)$ and $\hat{f}_2 := \frac{\partial \hat{f}}{\partial v}(u, v)$. The semi-discrete MUSCL scheme with the numerical flux $\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}})$ is TVD if*

$$\frac{u^R_{j+\frac{1}{2}} - u^R_{j-\frac{1}{2}}}{u_{j+1} - u_j} \geq 0, \quad \frac{u^L_{j+\frac{1}{2}} - u^L_{j-\frac{1}{2}}}{u_j - u_{j-1}} \geq 0. \tag{3.51}$$

Proof: Rewrite $-\Delta_x^+ \hat{f}(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}})$ as

$$
\begin{aligned}
&- \left(\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}}) - \hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}})\right) - \left(\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}}) - \hat{f}(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}})\right) \\
&= -\int_0^1 \hat{f}_2(u^L_{j+\frac{1}{2}}, u^R_{j-\frac{1}{2}} + \theta\Delta_x^+ u^R_{j-\frac{1}{2}})\, d\theta \cdot \frac{\Delta_x^+ u^R_{j-\frac{1}{2}}}{\Delta_x^+ u_j} \cdot \Delta_x^+ u_j \\
&- \int_0^1 \hat{f}_1(u^L_{j-\frac{1}{2}} + \theta\Delta_x^+ u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}})\, d\theta \cdot \frac{\Delta_x^+ u^L_{j-\frac{1}{2}}}{\Delta_x^+ u_{j-1}} \cdot \Delta_x^+ u_{j-1},
\end{aligned}
$$

Because $\hat{f}_1 \geq 0$ and $\hat{f}_2 \leq 0$, $C_{j+\frac{1}{2}} \geq 0$ and $D_{j-\frac{1}{2}} \geq 0$ so that semi-discrete MUSCL scheme with flux $\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}})$ is TVD. ∎

The following analyzes the conditions in (3.50) or (3.51) on the approximate slopes in the piecewise linear reconstruction.

Let $\bar{s}_j = hs_j$, three conditions in (3.50) become

$$1 - \frac{1}{2}\frac{\bar{s}_{j+1} + \bar{s}_j}{\Delta_x^+ u_j} \geq 0, \quad \frac{\bar{s}_j}{\Delta_x^+ u_j} \geq 0, \quad \frac{\bar{s}_j}{\Delta_x^- u_j} \geq 0, \tag{3.52}$$

which implies that it is necessary that $\bar{s}_j = 0$ if $\Delta_x^+ u_j \Delta_x^- u_j < 0$. Thus the semi-discrete MUSCL scheme with the numerical flux $\hat{f}(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}})$ in Theorem 3.26 has to reduce to first-order accuracy at the extreme point. The conditions on the second-order accuracy are

$$\begin{aligned}
u_{j+1} - \frac{\bar{s}_{j+1}}{2} &= u_{j+\frac{1}{2}} + \mathcal{O}(h^2), \\
u_j + \frac{\bar{s}_j}{2} &= u_{j+\frac{1}{2}} + \mathcal{O}(h^2), \\
\frac{1}{2}(\bar{s}_{j+1} + \bar{s}_j) &= \Delta_x^+ u_j + \mathcal{O}(h^2).
\end{aligned}$$

Since $\Delta_x^- u_j = \Delta_x^+ u_j + \mathcal{O}(h^2)$, a second-order accurate, semi-discrete TVD scheme may be derived if $\bar{s}_j = \text{minmod}(\Delta_x^+ u_j, \Delta_x^- u_j)$.

Two conditions in (3.51) become

$$1 - \frac{\bar{s}_{j+1} - \bar{s}_j}{2\Delta_x^+ u_j} \geq 0, \quad 1 + \frac{\bar{s}_{j+1} - \bar{s}_j}{2\Delta_x^+ u_j} \geq 0,$$

which implies the inequality

$$|\bar{s}_{j+1} - \bar{s}_j| \leq 2|\Delta_+ u_j|.$$

Because $|\bar{s}_{j+1} - \bar{s_j}| \leq \max\{|\bar{s}_{j+1}|, |\bar{s}_j|\}$ when $\bar{s}_{j+1}\bar{s_j} > 0$, the slopes may be limited by

$$|\bar{s}_{j+1}| \leq 2|\Delta_+ u_j|, \quad |\bar{s}_j| \leq 2|\Delta_+ u_j|. \tag{3.53}$$

In practice, $\bar{s}_j$ and $\bar{s}_{j+1}$ are always assumed to have the same

sign and $\bar{s}_j$ is usually chosen as follows

$$\bar{s}_j = \begin{cases} B(\Delta_x^+ u_j, \Delta_x^- u_j), & \Delta_x^+ u_j \Delta_x^- u_j > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.54}$$

where $B(x, y)$ is a Lipschitz continuous function, has the same sign as those of its independent variables, and satisfies $B(x, x) = x$. Obviously, if

$$|B(x, y)| \leq 2 \min(|x|, |y|), \tag{3.55}$$

then the slopes defined in (3.54) satisfy (3.53), because

$$\begin{aligned} \max\{|\bar{s}_j|, |\bar{s}_{j+1}|\} &\leq 2 \max\Big(\min\left(|\Delta_x^+ u_j|, |\Delta_x^- u_j|\right), \min\left(|\Delta_x^+ u_{j+1}|, |\Delta_x^+ u_j|\right)\Big) \\ &\leq 2|\Delta_x^+ u_j|. \end{aligned}$$

Here are several examples on the choices of $B(x, y)$.

Example 3.13 *For the minmod slope limiter [64],* $B(x,y) = sign(x)\min(|x|,|y|)$ *satisfies* (3.55).

Example 3.14 *For the van Leer slope limiter [84],* $B(x,y) = \frac{2xy}{x+y}$. *Its advantages is that* $B(x,y)$ *is a smooth function with respect to its independent variables, and satisfies* (3.55), *that is, for* $xy > 0$, *one has*

$$\begin{cases} |B| \le \frac{2|x|}{1+|\frac{y}{x}|} \le 2|x|, & \text{if } |x| \le |y|, \\ |B| \le \frac{2|y|}{1+|\frac{x}{y}|} \le 2|y|, & \text{if } |y| \le |x|. \end{cases}$$

Example 3.15 *For the Superbee slope limiter [64],*

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

Example 3.16 *For the van Albada slope limiter [82],*

$$B(x,y)=\frac{x^2y+y^2x}{x^2+y^2},\quad xy\le 0 \text{ or } xy>0.$$

Remark 3.12 *There are many various slope limiters in the literature. In practical computations, one needs carefully select a limiter function such that the resulting scheme is nonlinear stable. Some theoretical analysis at least for the scalar equation is necessary to choose the slope limiter.*

Remark 3.13 *The condition (3.47) is very harsh. It is possible to design a TVD scheme with the initial reconstruction violating (3.47). The high resolution shock-capturing methods introduced above are completely based on the TVD rule. It is also possible to use the local extreme principle, see Theorem 3.14,*

or monotonicity-preserving rule to design high resolution shock-capturing methods.

3.4.3 Piecewise parabolic method

This section introduces the piecewise parabolic method (PPM) of Colella and Woodward [16] which is a higher-order accurate extension of the Godunov method based on the piecewise parabolic polynomial reconstruction via the primitive function. Only by given the cell averages $\{\bar{u}_j(t)\}$, how to construct a polynomial approximating u that is accurate point-wisely to high order?

At a fixed t considered as a parameter, the primitive function of u is defined by

$$w(x,t) = \int_{x_{\frac{1}{2}}}^{x} u(\xi,t)\ d\xi,$$

where the lower limit of the integral $x_{\frac{1}{2}}$ may be arbitrary because we are only interested in the derivative of w with respect to x, i.e.

$$\frac{d}{dx}w(x,t) = u(x,t), \tag{3.56}$$

which admits us to get a good approximation of u and the point value of u at some point if there is a good approximation of w with respect to x.

A key observation is as follows. If the cell averages of u is given denoted by $\{\bar{u}_j\}$, then the point value of w at the mesh point $x_{j+\frac{1}{2}}$ may be calculated by

$$\begin{aligned} w_{j+\frac{1}{2}} :=& w(x_{j+\frac{1}{2}},t) = \int_{x_{\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(\xi,t)\, d\xi = \left(\int_{x_{\frac{1}{2}}}^{x_{\frac{3}{2}}} + \cdots + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \right) u(\xi,t)\, d\xi \\ =& h\sum_{i=1}^{j} \bar{u}_i(t), \end{aligned} \tag{3.57}$$

which implies

$$\frac{w_{j+\frac{1}{2}} - w_{j-\frac{1}{2}}}{h} = \bar{u}_j(t).$$

If $\{\bar{u}_j\}$ are exact, then the above point value of $w(x,t)$ is also exact.

If w is sufficiently smooth, e.g. $w(x,t) \in C^{q+1}$, then the polynomial interpolation can be used to present a polynomial approximation (as a whole) of w to arbitrary precision. Specially, in order to approximate w within the interval $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, the point values $w_{j+\frac{1}{2}-\ell}$, $w_{j+\frac{1}{2}-\ell+1}$, $\cdots$,$w_{j+\frac{1}{2}-\ell+q}$ (for some integer ℓ) can be used to interpolate a unique polynomial with degree of q (the choice of ℓ is shown below). If $p_j(x)$ denotes such polynomial, then

$$p_j(x) = w(x,t) + \mathcal{O}(h^{q+1}),\ x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]. \tag{3.58}$$

Thanks to (3.56), we get

$$p'_j(x) = u(x,t) + \mathcal{O}(h^q), \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}],$$

which may give the approximate left- and right-limitations of u at the cell interfaces $x_{j\pm\frac{1}{2}}$

$$u^R_{j-\frac{1}{2}} := p'_j(x_{j-\frac{1}{2}}), \quad u^L_{j+\frac{1}{2}} := p'_j(x_{j+\frac{1}{2}}).$$

Similarly, we can also obtain the interpolation polynomials $p_{j-1}(x)$ and $p_{j+1}(x)$ of $w(x,t)$ within $[x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}]$ and $[x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$ respectively, and define

$$u^L_{j-\frac{1}{2}} := p'_{j-1}(x_{j-\frac{1}{2}}), \qquad u^R_{j+\frac{1}{2}} := p'_{j+1}(x_{j+\frac{1}{2}}).$$

Based on them, a high-order spatial accurate scheme may be given in the form of (3.49), but it is not non-oscillatory.

Example 3.17 *This example introduces the PPM for the linear convection equation* (2.1) *with the positive constant* $a > 0$. *Assume that the mesh is uniform with the spatial and time step sizes* h *and* τ *respectively, satisfying*

$$\nu = \tau a/h < 1,$$

and the cell averages $\{\bar{u}_j\}$ *are given.*

First, calculate the point values $\{w_{j+\frac{1}{2}}\}$ *of the primitive function* $w(x,t)$ *by* (3.57), *and then use five point values* $\{(x_{j+\ell+\frac{1}{2}}, w_{j+\ell+\frac{1}{2}}),$ $\ell = 0, \pm 1, \pm 2\}$ *to interpolate a Lagrangian polynomial with degree of* 4 *approximating* $w(x,t)$, *denoted by* $p(x)$. *After that, calculate the value of the first-order derivative of* $p(x)$ *with respect to* x *at*

$x_{j+\frac{1}{2}}$, *i.e.*

$$\begin{aligned} u_{j+\frac{1}{2}} :=& p'(x_{j+\frac{1}{2}}) = \bar{u}_j + \frac{1}{2}\Delta_x^+ \bar{u}_j + \frac{1}{6}(\delta\bar{u}_j - \delta\bar{u}_{j+1}) \\ =& \frac{1}{12}(7\bar{u}_j + 7\bar{u}_{j+1} - \bar{u}_{j-1} - \bar{u}_{j+2}), \end{aligned} \tag{3.59}$$

where

$$\Delta_x^+ \bar{u}_j = \bar{u}_{j+1} - \bar{u}_j, \quad \delta\bar{u}_j = \frac{1}{2}(\Delta_x^+ \bar{u}_j + \Delta_x^- \bar{u}_j).$$

Next, using the cell average $\bar{u}_j$ *and the point values* $u_{j\pm\frac{1}{2}}$ *to give a polynomial with degree of 2 within the cell* $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ *approximating* $u(x,t)$ *by*

$$u_h(x) = a_0 + a_1(x - x_j)/h + a_2(x - x_j)^2/2h^2, \ x \in I_j, \tag{3.60}$$

where a_0, a_1, a_2 *are determined by*

$$u_h(x_{j\pm\frac{1}{2}}) = u_{j\pm\frac{1}{2}}, \quad \frac{1}{h}\int_{I_j} u_h(x)\, dx = \bar{u}_j.$$

It is not difficult to know that

$$a_1 = u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}},\ a_0 = \frac{1}{4}(6\bar{u}_j - u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}),\ a_2 = 6(u_{j+\frac{1}{2}} + u_{j-\frac{1}{2}} - 2\bar{u}_j).$$

Based on the mass conservation principle "the rate at which mass enters a system is equal to the rate at which mass leaves the system", within a time stepsize $\tau > 0$, *the "mass" over the cell* I_j *is changed by*

$$h\bar{u}_j^{n+1} - h\bar{u}_j^n = \hat{f}_{j-\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}},$$

where the mass leaving the cell interface $x_{j+\frac{1}{2}}$ *is equal to*

$$\hat{f}_{j+\frac{1}{2}} = \int_{x_{j+\frac{1}{2}}-a\tau}^{x_{j+\frac{1}{2}}} u_h(x)\, dx.$$

Substituting (3.60) *into the above equation gives the explicit formulation of the numerical flux* $\hat{f}_{j+\frac{1}{2}}$ *as follows*

$$\hat{f}_{j+\frac{1}{2}} = h\nu \left\{ (a_0 + \frac{1}{2}a_1 + \frac{1}{8}a_2) - \frac{1}{4}(a_2 + 2a_1)\nu + \frac{1}{6}a_2\nu^2 \right\}. \tag{3.61}$$

which is not numerical oscillation-free.

In practical computations, to avoid the numerical oscillation near the discontinuity, $\delta\bar{u}_j$ *in* (3.59) *is replaced with*

$$\delta_m\bar{u}_j = \begin{cases} \min\{|\delta\bar{u}_j|, 2|\Delta_x^+\bar{u}_j|, 2|\Delta_x^-\bar{u}_j|\} sign(\delta\bar{u}_j), & \Delta_x^+\bar{u}_j\Delta_x^-\bar{u}_j > 0, \\ 0, & \Delta_x^+\bar{u}_j\Delta_x^-\bar{u}_j \le 0, \end{cases}$$

so that the point value $u_{j+\frac{1}{2}}$ *in* (3.59) *is changed by*

$$u_{j+\frac{1}{2}} := \bar{u}_j + \frac{1}{2}\Delta_x^+ \bar{u}_j + \frac{1}{6}(\delta_m \bar{u}_j - \delta_m \bar{u}_{j+1}), \tag{3.62}$$

which is between $\bar{u}_j$ *and* $\bar{u}_{j+1}$.

In order that there is no overshoot or undershoot in the interpolation polynomial $u_h(x)$ *within the cell, in other words, there is no new extreme point in* $u_h(x)$, *the point values* $u_{j\pm\frac{1}{2}}$ *in* (3.62) *are still revised necessarily.*

Define $u^L_{j+\frac{1}{2}} := u_{j+\frac{1}{2}}$ *and* $u^R_{j-\frac{1}{2}} := u_{j-\frac{1}{2}}$, *which are given by* (3.62), *and do the following steps.*

- *If* $(u^L_{j+\frac{1}{2}} - \bar{u}_j)(\bar{u}_j - u^R_{j-\frac{1}{2}}) \le 0$, *which means that* $\bar{u}_j$ *is a local extreme point, then redefine* $u^L_{j+\frac{1}{2}} := \bar{u}_j, \quad u^R_{j-\frac{1}{2}} := \bar{u}_j$, *which means that the interpolation function is revised as a constant in* I_j.

- *If $\bar{u}_j$ is between $u^L_{j+\frac{1}{2}}$ and $u^R_{j-\frac{1}{2}}$, and sufficiently approaches one of them, then some values of the interpolation polynomial in (3.60), which may be rewritten as follows*

$$u_h(x) = u^R_{j-\frac{1}{2}} + \frac{x - x_{j-\frac{1}{2}}}{h}\left(\Delta u_j + u_{6,j}\left(1 - \frac{x - x_{j-\frac{1}{2}}}{h}\right)\right), \ x \in I_j, \tag{3.63}$$

is not between $u^L_{j+\frac{1}{2}}$ and $u^R_{j-\frac{1}{2}}$, where

$$\Delta u_j := u^L_{j+\frac{1}{2}} - u^R_{j-\frac{1}{2}}, \quad u_{6,j} := 6\left(\bar{u}_j - \frac{1}{2}(u^L_{j+\frac{1}{2}} + u^R_{j-\frac{1}{2}})\right).$$

The extreme point x^* of the interpolation polynomial (3.63) is the solution of the equation

$$\frac{(x - x_{j-\frac{1}{2}})}{h} = \frac{\Delta u_j + u_{6,j}}{2u_{6,j}},$$

while the value of the interpolation polynomial $u_h(x)$ *at this extreme point is*

$$u_h(x_*) = u^R_{j-\frac{1}{2}} + \frac{(\Delta u_j + u_{6,j})^2}{4u_{6,j}},$$

which implies that

$$\begin{cases} u_h(x_*) \geq \max\{u^R_{j-\frac{1}{2}}, u^L_{j+\frac{1}{2}}\}, & \text{if } u_{6,j} > 0, \\ u_h(x_*) \leq \min\{u^R_{j-\frac{1}{2}}, u^L_{j+\frac{1}{2}}\}, & \text{if } u_{6,j} < 0. \end{cases}$$

Therefore, in order that there is no overshoot or undershoot in the interpolation polynomial within the cell, the extreme point x_* *should be outside of the cell* I_j*, that is,*

$$\frac{\Delta u_j + u_{6,j}}{2u_{6,j}} \notin [0, 1],$$

which is equivalent to

$$|\Delta u_j| \geq |u_{6,j}|. \tag{3.64}$$

If the condition (3.64) is violated, the value $u^L_{j+\frac{1}{2}}$ *(or* $u^R_{j-\frac{1}{2}}$*) should be revised such that the revised polynomial is monotone within the cell* I_j*, and its derivative is reset as 0 at the cell interface* $x_{j-\frac{1}{2}}$ *(or* $x_{j+\frac{1}{2}}$*). The detailed operations are listed as follows.*

- *If* $\Delta u_j u_{6,j} > (\Delta u_j)^2$*, which implies that* $|u^L_{j+\frac{1}{2}} - u^R_{j-\frac{1}{2}}| > 3(u^L_{j+\frac{1}{2}} - \bar{u}_j) * sign(u^L_{j+\frac{1}{2}} - u^R_{j-\frac{1}{2}})$ *such that* $\bar{u}_j$ *is closer to* $u^L_{j+\frac{1}{2}}$*, then* $u^R_{j-\frac{1}{2}} := 3\bar{u}_j - 2u^L_{j+\frac{1}{2}}$*. Based on this modification of* $u^R_{j-\frac{1}{2}}$*,* $u'_h(x_{j+\frac{1}{2}}) = 0$*.*
- *If* $\Delta u_j u_{6,j} < -(\Delta u_j)^2$*, which implies that* $3(\bar{u}_j - u^R_{j-\frac{1}{2}}) *$

sign$(u^L_{j+\frac{1}{2}} - u^R_{j-\frac{1}{2}}) \leq |u^R_{j-\frac{1}{2}} - u^L_{j+\frac{1}{2}}|$ *such that* $\bar{u}_j$ *is closer to* $u^R_{j-\frac{1}{2}}$, *then* $u^L_{j+\frac{1}{2}} := 3\bar{u}_j - 2u^R_{j-\frac{1}{2}}$. *Based on this modification of* $u^R_{j-\frac{1}{2}}$, $u'_h(x_{j-\frac{1}{2}}) = 0$.

Finally, according to the sign of the characteristic speed a *of the linear convection equation* (2.1), *the revised values* $u^L_{j+\frac{1}{2}}$ *and* $u^R_{j+\frac{1}{2}}$ *may be used in* (3.49).

3.4.4 Flux limiter method

This section introduces the flux limiter method, in which the limiter function is used to limit the flux function in order to derive a non-oscillatory scheme. Using the terminology "flux limiter" means that the limiter acts on the flux function $f(u)$, while the

terminology "slope limiter" used above implies that the limiter acts on the solution u.

Assume that for the equation (1.3), a high-order accurate conservative scheme (e.g. the LW scheme) with the numerical flux $\hat{f}^H_{j+\frac{1}{2}}$ performs well in the smooth area of the solution u, and a low-order accurate conservative scheme (e.g. monotone scheme) with the numerical flux $\hat{f}^L_{j+\frac{1}{2}}$ is monotonicity-preserving or TVD etc. near the discontinuity. It is expected that mixing them may give a high resolution scheme with numerical flux

$$\hat{f}_{j+\frac{1}{2}} = \text{mix}\left\{\hat{f}^L_{j+\frac{1}{2}}, \hat{f}^H_{j+\frac{1}{2}}\right\} \to \begin{cases} \hat{f}^H_{j+\frac{1}{2}}, & \text{in smooth area,} \\ \hat{f}^L_{j+\frac{1}{2}}, & \text{near the discontinuity.} \end{cases} \tag{3.65}$$

To do that, the numerical flux $\hat{f}^H_{j+\frac{1}{2}}$ is first cast into form

$$\hat{f}^H_{j+\frac{1}{2}} = \hat{f}^L_{j+\frac{1}{2}} + (\hat{f}^H - \hat{f}^L)_{j+\frac{1}{2}}, \tag{3.66}$$

which implies that the numerical flux of the high-order accurate scheme is equal to a sum of the numerical flux of the low-order accurate scheme and a correction or anti-diffusion term. The correction term $\hat{f}^H - \hat{f}^L$ compensates excessive dissipation of the low-order accurate scheme so that a high-order accuracy is obtained. An example is that $\hat{f}^{LW} - \hat{f}^L = \frac{1}{2\lambda}(|\nu_{j+\frac{1}{2}}| - \nu^2_{j+\frac{1}{2}})\Delta^+_x u_j$ plays a anti-diffusion role because $|\nu_{j+\frac{1}{2}}| - \nu^2_{j+\frac{1}{2}} \geq 0$ when $|\nu| \leq 1$. The anti-diffusion term may be explicit or implicit.

In the flux limiter method, the correction term (3.66) is lim-

ited to derive a high resolution method with numerical flux

$$\begin{aligned}\hat{f}_{j+\frac{1}{2}} =& \hat{f}^L_{j+\frac{1}{2}} + \phi_{j+\frac{1}{2}} \left(\hat{f}^H_{j+\frac{1}{2}} - \hat{f}^L_{j+\frac{1}{2}} \right) \\ =& \hat{f}^H_{j+\frac{1}{2}} - (1-\phi_{j+\frac{1}{2}})(\hat{f}^H - \hat{f}^L)_{j+\frac{1}{2}} = \phi_{j+\frac{1}{2}} \hat{f}^H_{j+\frac{1}{2}} + (1-\phi_{j+\frac{1}{2}}) \hat{f}^L_{j+\frac{1}{2}},\end{aligned} \tag{3.67}$$

where $\phi_{j+\frac{1}{2}}$ is the so-called flux limiter depending on the solution u, and approaching to 1 when u is smooth at x_j, otherwise tending to 0. The flux limiter method may also be implemented in a multi-step form (predictor-corrector procedure)

$$\begin{aligned} u^*_j =& H^L(u^n; j), && \text{——first-order TVD scheme} \\ u^{n+1}_j =& u^*_j + M(u^*; j), && \text{——corrector} \end{aligned}$$

or

$$\begin{aligned} u^*_j =& H^H(u^n; j), && \text{——high-order scheme} \\ u^{n+1}_j =& u^*_j + F(u^*; j). && \text{——filter} \end{aligned}$$

One of the earliest high resolution method is the Flux-Corrected Transport (FCT) of Boris and Book [4], which may be considered as a flux limiter method. The FCT method is a simple and effective method by adding the most anti-diffusion possible but without increasing numerical oscillation or the total variation of the solution.

Example 3.18 (A simplest FCT) *For the linear convection equation* (2.1) *with a constant* a*, let us consider the transport scheme given in [4]*

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n) + (\frac{1}{8} + \frac{\nu^2}{2})(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \quad (3.68)$$

It is a simple two-sided differencing of the convection term plus a strong diffusion, which is the basis of FCT. The equation (3.68) *may be considered as a modification of the LW scheme with a*

diffusion

$$\frac{1}{8}(u_{j+1} - 2u_j + u_{j-1}).$$

Denote the solution of (3.68) *by* $u_j^{n+1,*}$, *which may contain numerical oscillation.*

The anti-diffusive stage in the FCT method is to remove the excessive diffusion to give the nonnegative solution

$$u_j^{n+1} = u_j^{n+1,*} - \{(u_{j+1}^{n+1,*} - u_j^{n+1,*})_{lim} - (u_j^{n+1,*} - u_{j-1}^{n+1,*})_{lim}\},$$

where the limited slopes are defined by

$$(\Delta_x^+ u_j)_{lim} = \min\{|\Delta_x^- u_j|, \text{“ } \tfrac{1}{8} \text{”}|\Delta_x^+ u_j|, |\Delta_x^+ u_{j+1}|\} sign(\Delta_x^+ u_j),$$

if $sign(\Delta_x^+ u_j) = sign(\Delta_x^- u_j) = sign(\Delta_x^+ u_{j+1})$, *otherwise the value of* $(\Delta_x^+ u_j)_{lim}$ *is 0. The quotation marks* “ $\frac{1}{8}$ ” *indicate that more exact cancellation of errors can be achieved if one expends*

a small computational effort by including at least rough approximations to the convection velocity or wavenumber-dependent corrections.

Harten and Zwas also introduced a self-adjusting hybrid scheme [32] similar to the above for shock computations. A complete review and studying the TVD property of the flux limiter method is presented by Sweby in [70].

Consider the linear convection equation (2.1) with constant coefficient and take $\hat{f}^H = \hat{f}^{LW}$ and $\hat{f}^L = \hat{f}^U$ (first-order accurate upwind scheme). If $a > 0$, then the LW scheme may be rewritten as

$$u_j^{n+1} = u_j^n - \nu(u_j^n - u_{j-1}^n) - \frac{1}{2}\nu(1-\nu)(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \tag{3.69}$$

Compared to (3.66), the LW scheme may be considered as a scheme obtained by adding a correction term into the first-order

accurate upwind scheme, and its numerical flux becomes

$$\hat{f}^{LW}_{j+\frac{1}{2}} = au_j + \frac{1}{2}a(1-\nu)(u_{j+1} - u_j).$$

Because the LW scheme is not TVD, it needs to be modified to get a high resolution flux limiter scheme satisfying some nonlinear stability (e.g. TVD property). For example, (3.69) is modified as

$$u_j^{n+1} = u_j^n - \nu\Delta_x^- u_j^n - \Delta_x^-(\varphi_j^n \frac{1}{2}\nu(1-\nu)\Delta_x^+ u_j^n), \tag{3.70}$$

which is obtained by adding a limited anti-diffusive term into the first-order accurate upwind scheme, where φ_j^n is limiter and nonnegative such that the sign of anti-diffusive term is kept.

The remaining task is to determine the choice of φ_j. Similar to the works of Roe [62], van Leer [84], and Chakravarthy and Osher [7], the limiter is taken as a function of the solution gradient

satisfying $\varphi_j = \varphi(r_j)$, where $r_j = \frac{\Delta_x^- u_j}{\Delta_x^+ u_j}$. Near the extreme point of u, where $u_x = 0$, the smoothing indicator r_j will stop, since its denominator approaches to 0 so that r_j becomes arbitrarily large, or negative even though the solution is smooth. Moreover, besides the derived scheme satisfies some nonlinear stability (e.g. TVD property), the choice of $\varphi(r)$ should ensure that the magnitude of the anti-diffusive term is the largest possible.

Lemma 3.28 *Assume that $\varphi(r)$ is bounded. The flux limiter scheme (3.70) is consistent with the convection equation (2.1), and second-order accurate in the smooth area of u and away from the extreme point of u if φ is Lipschitz continuous at $r = 1$ and $\varphi(1) = 1$.*

Lemma 3.29 *The flux limiter scheme (3.70) for (2.1) with $a >$*

0 *is TVD if*

$$\Phi \leq \min\left\{\frac{2}{1-\nu}, \frac{2}{\nu}\right\}, \tag{3.71}$$

where Φ *is the upper bound of* $|\varphi(r_j)/r_j - \varphi(r_{j-1})|$, *i.e.*

$$|\varphi(r_j)/r_j - \varphi(r_{j-1})| \leq \Phi. \tag{3.72}$$

Proof: It needs to rewrite (3.70) in an incremental form of (3.34) with the incremental coefficients satisfying conditions in Theorem 3.13. If rewrite (3.70) in an incremental form of (3.34) with $C_{j+\frac{1}{2}} = -\frac{1}{2}\nu(1-\nu)\varphi_j$ and $D_{j-\frac{1}{2}} = \nu - \frac{1}{2}\nu(1-\nu)\varphi_{j-1}$, then it is bad because $C_{j+\frac{1}{2}} < 0$ when φ_j approaches to 1. Now (3.70) is

cast into another incremental form

$$u_j^{n+1} = u_j^n - \left[\nu + \frac{1}{2}\nu(1-\nu)\frac{\varphi_j \Delta_x^+ u_j^n - \varphi_{j-1}\Delta_x^- u_j^n}{\Delta_x^- u_j^n}\right]\Delta_x^- u_j^n$$
$$=: u_j^n - D_{j-\frac{1}{2}}^n \Delta_x^- u_j^n,$$

where $C_{j+\frac{1}{2}}^n = 0$ and

$$D_{j-\frac{1}{2}}^n = \nu\left\{1 + \frac{1}{2}(1-\nu)\left[\frac{\varphi(r_j)}{r_j} - \varphi(r_{j-1})\right]\right\}. \tag{3.73}$$

Thanks to (3.72), one has

$$\nu\left(1 - \frac{1}{2}(1-\nu)\Phi\right) \le D_{j-\frac{1}{2}} \le \nu\left(1 + \frac{1}{2}(1-\nu)\Phi\right).$$

Combing this inequality with the hypothesis (3.71), the incremental coefficient in (3.73) satisfies

$$0 \leq D_{j-\frac{1}{2}} \leq 1,$$

so that the scheme (3.70) is TVD.

The hypothesis (3.71) may be strengthened as $\Phi \leq 2$. It is reasonable because under the CFL condition $0 \leq \nu \leq 1$, the inequality $2 \leq \min\{\frac{2}{1-\nu}, \frac{2}{\nu}\}$ holds. If $\varphi(r)$ is nonnegative and $\varphi(r) = 0$ when $r \leq 0$, then the inequalities (3.72) and $\Phi \leq 2$ imply

$$0 \leq \frac{\varphi(r)}{r}, \varphi(r) \leq 2, \tag{3.74}$$

which is a sufficient condition for that the scheme (3.70) for (2.1) with $a > 0$ is TVD. Fig. 12 show the TVD region with boundaries

in three straight lines $\varphi(r) = 0, 2$ and $2r$, where the straight lines $\varphi(r) = 1$ and r correspond to the LW and BW schemes, respectively, and the shadow domain is the region of the second-order accurate TVD schemes.

The condition (3.74) suggests that a possible choice of the limiter $\varphi(r)$ in the flux limiter scheme (3.70) seems

$$\varphi(r) = \begin{cases} \min\{2r, 2\}, & r > 0, \\ 0, & r \leq 0, \end{cases}$$

which corresponds to the top boundaries $\varphi = 2$ and $2r$ in Fig. 12, and ensures that the biggest anti-diffusion may be added to the first-order accurate upwind scheme. Unfortunately, with such limiter, the flux limiter scheme (3.70) is not second-order accurate in the sense of truncation error, see Lemma 3.28, although it is TVD. Hence, an ideal limiter φ in the flux limiter scheme (3.70)

should be satisfied the condition (3.74) and $\varphi(1) = 1$ in order that (3.70) becomes a high resolution scheme. A kind of limiter functions for (3.70) satisfying those requirements may be defined by

$$\varphi_\theta(r) = \max(0, \min(\theta r, 1), \min(r, \theta)), \ 1 \leq \theta \leq 2, \tag{3.75}$$

which is a monotone function with respect to r, and has the symmetry

$$\frac{\varphi_\theta(r)}{r} = \varphi_\theta\left(\frac{1}{r}\right).$$

If $\theta = 1$, then $\varphi_\theta(r)$ reduces to the minmod limiter function, i.e. the bottom boundary of the shadow domain in Fig. 12, while it becomes the Super-bee limiter corresponding to the top boundary of the shadow domain in Fig. 12 when θ is taken as 2.

Exercise 3.2 *Give an example to show that the scheme (3.70) with $\varphi(r)$ satisfying $\varphi(r) > 0$ when $r \leq 0$ is total variation increasing.*

Any 2nd-order accurate scheme depending on $\{u_{j-2}, u_{j-1}, u_j, u_{j+1}\}$ for the linear convection equation (2.1) with $a > 0$ may be regarded as a weighting average of the LW and BW schemes via the limiter function

$$\varphi(r) = (1 - \omega(r))\varphi^{\mathrm{LW}}(r) + \omega(r)\varphi^{\mathrm{BW}}(r) = 1 + \omega(r)(r - 1),\ 0 \leq \omega(r) \leq 1,$$

which may be limited within the shadow region in Fig. 12 because $\varphi^{\mathrm{LW}}(r) = 1$ and $\varphi^{\mathrm{BW}}(r) = r$. If $\omega(r) = 1/2$, the Fromm scheme is derived. When $\omega(r) \in [0, 1]$, $\varphi(r)$ is considered as the internal average (interpolation) of φ^{LW} and φ^{BW}. If using the extrapolation, the derived scheme is over-compressive, so that a sine wave is compressed as a square wave.

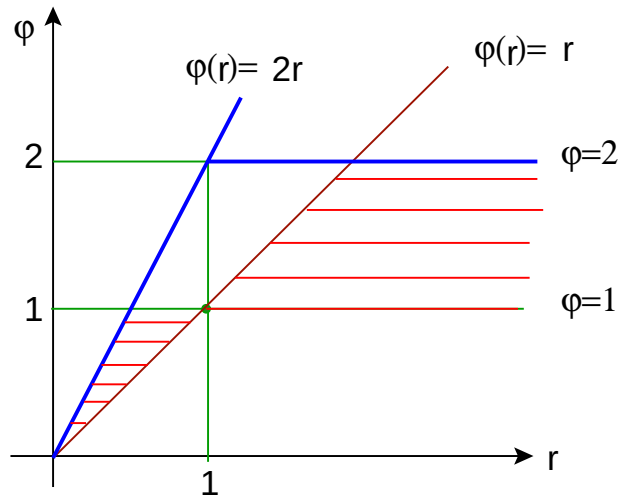


Figure 12: The shadow region is the region of second-order accurate TVD schemes. The region with boundaries $\varphi(r) = 0, 2$ and $2r$ is TVD region. $\varphi(r) = 1$: LW scheme; $\varphi(r) = r$: BW scheme.

Example 3.19 (van Leer limiter) *van Leer [84] gave by weighting the LW and WB schemes as follows*

$$u_j^{n+1} = u_j^n - \nu\Delta u_{j-\frac{1}{2}}^n - \frac{\nu}{4}(1-\nu)(\Delta u_{j+\frac{1}{2}}^n - \Delta u_{j-\frac{3}{2}}^n) + \frac{\nu}{4}(1-\nu)\left\{s(\theta_j)(\Delta u_{j+\frac{1}{2}}^n - \Delta u_{j-\frac{1}{2}}^n) - s(\theta_{j-1})(\Delta u_{j-\frac{1}{2}}^n - \Delta u_{j-\frac{3}{2}}^n)\right\},$$

where $s(\theta) = \frac{|\theta|-1}{|\theta|+1}$, *and*

$$\theta_j = \frac{\Delta_x^+ u_j}{\Delta_x^- u_j} = 1/r_j,$$

denotes the "smoothing monitor". The above scheme can be rewritten as

$$u_j^{n+1} = u_j^n - \nu\Delta_x^- u_j^n - \Delta_x^-\left\{\frac{1}{2}\left((1-s(\theta_j)) + (1+s(\theta_j))/\theta_j\right)\frac{1}{2}(1-\nu)\nu\Delta u_{j+\frac{1}{2}}^n\right\},$$

which is identical to (3.70) if setting

$$\varphi_j = \frac{1}{2}\left((1-s(\theta_j)) + (1+s(\theta_j))/\theta_j\right) = \frac{|r_j| + r_j}{1+|r_j|},$$

which may be rewritten as

$$\varphi_{VL}(r) = \frac{|r| + r}{1 + |r|} = \begin{cases} 0, & r \le 0, \\ \frac{2r}{1+r}, & r > 0. \end{cases}$$

Example 3.20 *The LW scheme for the convection equation* (2.1) *with constant coefficient can also be cast into form*

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(1+\nu)\Delta_x^- u_j^n - \frac{\nu}{2}(1-\nu)\Delta_x^+ u_j^n,$$

where $\nu = a\tau/h$. *Based on this formula, Roe's TVD-LW scheme*

[63] is obtained as follows

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(1+\nu)\Delta_x^- u_j^n - \frac{\nu}{2}(1-\nu)\Delta_x^+ u_j^n$$
$$- \frac{1}{2}|\nu|(1-|\nu|)(1-\theta_{j-\frac{1}{2}})\Delta_x^+ u_{j-1}^n$$
$$+ \frac{1}{2}|\nu|(1-|\nu|)(1-\theta_{j+\frac{1}{2}})\Delta_x^+ u_j^n. \tag{3.76}$$

Let us check the TVD condition of the scheme (3.76). Introduce two new notations

$$r_{j+\frac{1}{2}}^- := \frac{\Delta_x^- u_j^n}{\Delta_x^+ u_j^n} = \frac{1}{r_{j+\frac{1}{2}}^+}.$$

(1) If $\nu > 0$, then rewrite (3.76) as

$$u_j^{n+1} = u_j^n - D_{j-\frac{1}{2}}^n \Delta_x^- u_j^n,$$

with

$$D^n_{j-\frac{1}{2}} = \nu\left[1 - \frac{1}{2}(1-\nu)\theta_{j-\frac{1}{2}} + \frac{1}{2}(1-\nu)\theta_{j+\frac{1}{2}}/r^-_{j+\frac{1}{2}}\right].$$

The inequality $0 \le D^n_{j-\frac{1}{2}} \le 1$ *holds if*

$$\theta_{j-\frac{1}{2}} - \theta_{j+\frac{1}{2}}/r^-_{j+\frac{1}{2}} < \frac{2}{1-\nu}, \quad \theta_{j+\frac{1}{2}}/r^-_{j+\frac{1}{2}} - \theta_{j-\frac{1}{2}} < \frac{2}{\nu}. \quad (3.77)$$

(2) If $\nu < 0$, *then rewrite* (3.76) *as*

$$u^{n+1}_j = u^n_j + C^n_{j+\frac{1}{2}}\Delta^+_x u^n_j,$$

where

$$C^n_{j+\frac{1}{2}} = |\nu|\left[1 - \frac{1}{2}(1-|\nu|)\theta_{j+\frac{1}{2}} + \frac{1}{2}(1-|\nu|)\theta_{j-\frac{1}{2}}/r^+_{j-\frac{1}{2}}\right],$$

Similarly, the inequality $0 \le C^n_{j+\frac{1}{2}} \le 1$ *holds if*

$$\theta_{j+\frac{1}{2}} - \theta_{j-\frac{1}{2}}/r^+_{j-\frac{1}{2}} < \frac{2}{1-|\nu|}, \quad \theta_{j-\frac{1}{2}}/r^+_{j-\frac{1}{2}} - \theta_{j+\frac{1}{2}} < \frac{2}{|\nu|}. \quad (3.78)$$

If assuming that θ *and* θ/r *are always nonnegative, then the "optimal" sufficient conditions for satisfying four conditions in* (3.77) *and* (3.78) *are*

$$\theta_{j+\frac{1}{2}} \le \frac{2}{1-|\nu|}, \quad \theta_{j+\frac{1}{2}}/r^\pm_{j+\frac{1}{2}} \le \frac{2}{|\nu|}.$$

A choice of $\theta_{j+\frac{1}{2}}$ *is*

$$\theta_{j+\frac{1}{2}} = \theta_{j+\frac{1}{2}}(r^-_{j+\frac{1}{2}}, r^+_{j+\frac{1}{2}}) = \varphi(r^-_{j+\frac{1}{2}}) + \varphi(r^+_{j+\frac{1}{2}}) - 1,$$

where φ *is required to satisfy the conditions* $0 < \varphi(r) < 1$ *and* $0 < \varphi(r)/r < 2$. *An example of* φ *satisfying those conditions is*

$$\varphi(r) = minmod(1, r).$$

Other choices of $\theta_{j+\frac{1}{2}}$ *are*

$$\begin{aligned}
\theta(r^-, r^+) =& minmod(1, r^-) + minmod(1, r^+) - 1,\\
\theta(r^-, r^+) =& minmod(1, r^-, r^+),\\
\theta(r^-, r^+) =& minmod(2, 2r^-, 2r^+, 0.5(r^- + r^+)),\\
\theta(r^-, r^+) =& (r^- + |r^-|)/(1 + r^-) + (r^+ + |r^+|)/(1 + r^+) - 1.
\end{aligned}$$

∎

The following gives an extension of the flux limiter scheme (3.120) to quasilinear equation (1.3) [70]

$$\begin{aligned}
u_j^{n+1} =& u_j^n - \lambda(\hat{f}^E_{j+\frac{1}{2}} - \hat{f}^E_{j-\frac{1}{2}}) - \lambda\Delta_x^- \Big\{\varphi(r_j^+)\alpha^+_{j+\frac{1}{2}}(\Delta f_{j+\frac{1}{2}})^+\\
&- \varphi(r^-_{j+1})\alpha^-_{j+\frac{1}{2}}(\Delta f_{j+\frac{1}{2}})^-\Big\}, \qquad (3.79)
\end{aligned}$$

where $\hat{f}^E(u_j, u_{j+1})$ is the numerical flux of the E scheme[1], satisfying

$$\text{sign}(u_{j+1}-u_j)(\hat{f}^E_{j+\frac{1}{2}}-f(u)) \leq 0, \forall u \in [\min\{u_j, u_{j+1}\}, \max\{u_j, u_{j+1}\}],$$

and

$$\alpha^{\pm}_{j+\frac{1}{2}} := \frac{1}{2}\left(1 \mp \nu^{\pm}_{j+\frac{1}{2}}\right), r^+_j := \frac{\alpha^+_{j-\frac{1}{2}}(\Delta f_{j-\frac{1}{2}})^+}{\alpha^+_{j+\frac{1}{2}}(\Delta f_{j+\frac{1}{2}})^+}, r^-_j := \frac{\alpha^-_{j+\frac{1}{2}}(\Delta f_{j+\frac{1}{2}})^-}{\alpha^-_{j-\frac{1}{2}}(\Delta f_{j-\frac{1}{2}})^-}.$$

$$\nu^{\pm}_{j+\frac{1}{2}} := \lambda\frac{(\Delta f_{j+\frac{1}{2}})^{\pm}}{\Delta^+_x u_j}, \ (\Delta f_{j+\frac{1}{2}})^+ := f_{j+1} - \hat{f}^E_{j+\frac{1}{2}}, \ (\Delta f_{j+\frac{1}{2}})^- := \hat{f}^E_{j+\frac{1}{2}} - f_j.$$

[1]The semi-discrete E scheme satisfies the entropy condition, but is only first-order accurate [54].

One example of the E-scheme is the Engquist-Osher scheme which has numerical flux

$$\hat{f} = f_j^+ + f_{j+1}^- + f(u_0), f_j^+ = \int_{u_0}^{u_j} \max\{f'(u), 0\}du, f_j^- = \int_{u_0}^{u_j} \min\{f'(u), 0\}du,$$

where u_0 is the sonic point of $f(u)$, satisfying $f'(u_0) = 0$.

The fully discrete explicit E scheme

$$u_j^{n+1} = u_j^n - \lambda(\hat{f}^E_{j+\frac{1}{2}} - \hat{f}^E_{j-\frac{1}{2}}), \tag{3.80}$$

may be cast into form (3.34) with the incremental coefficients

$$C^E_{j+\frac{1}{2}} = -\lambda\frac{\hat{f}^E_{j+\frac{1}{2}} - f_j}{u_{j+1} - u_j}, \; D^E_{j-\frac{1}{2}} = \lambda\frac{f_j - \hat{f}^E_{j-\frac{1}{2}}}{u_j - u_{j-1}},$$

which are nonnegative and satisfy

$$C^E_{j+\frac{1}{2}} + D^E_{j+\frac{1}{2}} = \lambda\frac{f_{j+1} + f_j - 2\hat{f}^E_{j+\frac{1}{2}}}{u_{j+1} - u_j} = \nu^+_{j+\frac{1}{2}} - \nu^-_{j+\frac{1}{2}}.$$

For the above Engquist-Osher scheme

$$\nu^+_{j+\frac{1}{2}} - \nu^-_{j+\frac{1}{2}} = \frac{\lambda}{\Delta^+_x u_j} \int_{u_j}^{u_{j+1}} |f'(s)|\, ds \le \lambda \max\{|f'(u)|\}.$$

Here we assume that the general discrete E-scheme (3.80) is TVD under a CFL condition

$$\lambda \max_u\{|f'(u)|\} \le \mu \le 1,$$

in other words, the general discrete E-scheme (3.80) considered here satisfies

$$\nu^+_{j+\frac{1}{2}} - \nu^-_{j+\frac{1}{2}} \le \lambda \max_u\{|f'(u)|\}. \tag{3.81}$$

Lemma 3.30 *Assume the numerical flux of E-scheme in* (3.79) *satisfies* (3.81). *The flux limiter scheme* (3.79) *is TVD under*

the condition

$$\lambda \max_u\{|f'(u)|\} \le \Big(\frac{2}{2+\theta}\Big)\mu,$$

where θ *is given in* (3.72) *and satisfies* $1 \le \theta \le 2$.

Proof: The scheme (3.79) can be cast into the incremental form (3.34) with

$$C_{j+\frac{1}{2}} = -\nu^-_{j+\frac{1}{2}}\left\{1+\alpha^-_{j+\frac{1}{2}}\left[\frac{\varphi(r^-_j)}{r^-_j}-\varphi(r^-_{j+1})\right]\right\},$$

$$D_{j-\frac{1}{2}} = \nu^+_{j-\frac{1}{2}}\left\{1+\alpha^+_{j-\frac{1}{2}}\left[\frac{\varphi(r^+_j)}{r^+_j}-\varphi(r^+_{j-1})\right]\right\}.$$

Because $\nu^+_{j+\frac{1}{2}} \ge 0$, $\nu^-_{j+\frac{1}{2}} \le 0$, and $\theta \le 2$, $C_{j+\frac{1}{2}} \ge 0$, $D_{j+\frac{1}{2}} \ge 0$,

and

$$
\begin{aligned}
C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} &\leq \nu^{+}_{j+\frac{1}{2}} \left\{ 1 + \alpha^{+}_{j+\frac{1}{2}} \theta \right\} - \nu^{-}_{j+\frac{1}{2}} \left\{ 1 + \alpha^{-}_{j+\frac{1}{2}} \theta \right\} \\
&= (\nu^{+}_{j+\frac{1}{2}} - \nu^{-}_{j+\frac{1}{2}})(1 + \frac{\theta}{2}) - \frac{\theta}{2} \left[(\nu^{+}_{j+\frac{1}{2}})^2 + (\nu^{-}_{j+\frac{1}{2}})^2 \right] \\
&\leq (\nu^{+}_{j+\frac{1}{2}} - \nu^{-}_{j+\frac{1}{2}})(1 + \frac{\theta}{2}).
\end{aligned}
$$

Moreover, the lemma hypothesis implies

$$
\nu^{+}_{j+\frac{1}{2}} - \nu^{-}_{j+\frac{1}{2}} \leq \lambda \max\{|f'(u)|\} \leq \left(\frac{2}{2+\theta} \right) \mu.
$$

Thus, $C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} \leq \mu \leq 1$. The proof is completed. ∎

Remark 3.14 *The accuracy of the scheme* (3.79) *is discussed*

here. Let $\varphi = 1$, *then* (3.79) *may be rewritten as*

$$\begin{aligned} u_j^{n+1} =& u_j^n - \frac{1}{2}(f_{j+1}^n - f_{j-1}^n) \\ & - \frac{\lambda^2}{2}\Big\{ \frac{f_{j+1} - \hat{f}_{j+\frac{1}{2}}}{u_{j+1} - u_j}(f_{j+1} - \hat{f}_{j+\frac{1}{2}}) + \frac{\hat{f}_{j+\frac{1}{2}} - f_j}{u_{j+1} - u_j}(\hat{f}_{j+\frac{1}{2}} - f_j) \\ & - \frac{f_j - \hat{f}_{j-\frac{1}{2}}}{u_j - u_{j-1}}(f_j - \hat{f}_{j-\frac{1}{2}}) - \frac{\hat{f}_{j-\frac{1}{2}} - f_{j-1}}{u_j - u_{j-1}}(\hat{f}_{j-\frac{1}{2}} - f_{j-1})\Big\}. \end{aligned} \tag{3.82}$$

Taking the partial derivatives of $\hat{f}(u, u) = f(u)$ *with respect to* u *gives* $\hat{f}_1 + \hat{f}_2 = f'(u) = a(u)$. *Using the Taylor series expansion*

gets

$$f_{j+1} - \hat{f}_{j+\frac{1}{2}} = (\hat{f}_1')_{j+1}(u_{j+1} - u_j) - \frac{1}{2}(\hat{f}_1'')_{j+1}(u_{j+1} - u_j)^2 + \cdots,$$

$$\hat{f}_{j+\frac{1}{2}} - f_j = (\hat{f}_2')_j(u_{j+1} - u_j) + \frac{1}{2}(\hat{f}_2'')_j(u_{j+1} - u_j)^2 + \cdots.$$

Thus, it is not difficult to show that (3.82) *is second-order accurate, in comparison to the LW scheme.*

It is interesting to check whether the scheme (3.79) *satisfies the discrete entropy condition?* ∎

3.4.5 Modified flux method

This section introduces the modified flux method of Harten [25], which is based on Theorem 3.6.

The modified equation of the $(2l+1)$-point conservative monotone scheme (3.27) for scalar conservation law (1.3) is rewritten

as follows

$$u_t + f(u)_x = \left(\frac{1}{\lambda} g\right)_x, \tag{3.83}$$

where

$$g = \tau\lambda\beta(u,\lambda)u_x = \frac{h}{2}\left[\sum_{k=-l}^{l} k^2 H_k(u,\cdots,u) - \lambda^2 a^2(u)\right] u_x = \mathcal{O}(h). \tag{3.84}$$

If the first-order accurate scheme (3.27) is applied to the PDE

$$u_t + f^{\mathrm{M}}(u)_x = 0, \tag{3.85}$$

where $f^{\mathrm{M}} = f + \frac{1}{\lambda} g$ is the modified flux, then the concrete form of the scheme is

$$u_j^{n+1} = u_j^n - \lambda\big(\hat{h}(u_{i-l+1}^n,\cdots,u_{j+l}^n) - \hat{h}(u_{i-l}^n,\cdots,u_{j+l-1}^n) =: \overline{H}(u^n; j), \tag{3.86}$$

where $\hat{h}(u,\cdots,u) = f^{\mathrm{M}}(u)$, and the modified equation of (3.86) is

$$u_t + f_x^{\mathrm{M}} = \frac{1}{\lambda}\left\{\left[\frac{h}{2}\sum_{k=-l}^{l} k^2 \overline{H}_k(u,\cdots,u) - \lambda^2 \bar{a}^2(u)\right] u_x\right\}_x,$$

where $\bar{a} = df^{\mathrm{M}}/du = a + \mathcal{O}(h)$. Substituting g in (3.84) into the above equation gives

$$\begin{aligned} & u_t + \left(f + \frac{1}{\lambda}\left\{\frac{h}{2}\left[\sum_{k=-l}^{l} k^2 H_k(u,\cdots,u) - \lambda^2 a^2(u)\right] u_x\right\}\right)_x \\ & = \frac{1}{\lambda}\left[\frac{h}{2}\left(\sum_{k=-l}^{l} k^2 H_k(u,\cdots,u) - \lambda^2 a^2(u) + \mathcal{O}(h)\right) u_x\right]_x, \end{aligned}$$

which results in

$$u_t + f_x = \mathcal{O}(h^2).$$

Thus the scheme (3.86) may be considered as a second-order accurate scheme for the equation (1.3), and is TVD if (3.86) is a TVD approximation of (3.85) .

The remaining tasks is construction of the numerical flux approximating g in (3.84). Consider the first-order accurate, three-point, conservative TVD scheme

$$u_j^{n+1} = u_j^n - \lambda \left(\hat{f}(u_j^n, u_{j+1}^n) - \hat{f}(u_{j-1}^n, u_j^n) \right), \quad \lambda = \frac{\tau}{h}, \tag{3.87}$$

for (1.3), where

$$\hat{f}_{j+\frac{1}{2}} = \frac{1}{2} \left[f(u_j) + f(u_{j+1}) - \frac{1}{\lambda} Q(\nu_{j+\frac{1}{2}}) \Delta_x^+ u_j \right], \tag{3.88}$$

here $Q(\nu)$ is numerical viscosity coefficient, $\nu_{j+\frac{1}{2}} = \lambda a_{j+\frac{1}{2}}$, and $a_{j+\frac{1}{2}}$ is defined in (3.6). Eq. (3.87) can also be rewritten in the

viscous form

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right) + \frac{Q(\nu_{j+\frac{1}{2}}^n)}{2}\Delta_x^+ u_j^n - \frac{Q(\nu_{j-\frac{1}{2}}^n)}{2}\Delta_x^+ u_{j-1}^n, \tag{3.89}$$

which contains the following schemes

$$\begin{aligned} \text{LF scheme:} \qquad & Q(\nu) = 1, \\ \text{Upwind scheme:} \qquad & Q(\nu) = |\nu|, \\ \text{LW scheme:} \qquad & Q(\nu) = \nu^2. \end{aligned}$$

Lemma 3.31 *Assume* $0 \le |\nu| \le \mu \le 1$. *If* $Q(\nu)$ *in* (3.89) *satisfies the inequality*

$$|\nu| \le Q(\nu) \le 1, \tag{3.90}$$

then the scheme (3.89) *is TVD under the CFL-like restriction*

$$\lambda \max_u |f'(u)| \le \mu \le 1. \tag{3.91}$$

Proof: Rewrite (3.89) in the incremental form

$$\begin{aligned} u_j^{n+1} &= u_j^n + \frac{1}{2}\left[Q(\nu_{j+\frac{1}{2}}^n) - \nu_{j+\frac{1}{2}}^n\right]\Delta_x^+ u_j^n - \frac{1}{2}\left[Q(\nu_{j-\frac{1}{2}}^n) + \nu_{j-\frac{1}{2}}^n\right]\Delta_x^+ u_{j-1}^n \\ &= u_j^n + C_{j+\frac{1}{2}}^n \Delta_x^+ u_j^n - D_{j-\frac{1}{2}}^n \Delta_x^+ u_{j-1}^n. \end{aligned}$$

Under the hypothesis, the incremental coefficients satisfy the TVD conditions of Theorem 3.13 and the proof is completed. ∎

Example 3.21 *Use the first-order accurate upwind scheme ($Q = |\nu|$) and the modified LF scheme ($Q = 1/4\lambda$) to solve the initial value problem of Burgers' equation (1.3) and (3.32) with*

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

The results are shown in Fig. 13 where the solid line denotes the exact solution. We see that the upwind scheme gives the

incorrect solution, even if $\max\{\tau, h\} \to 0$. *Therefore, the TVD scheme does not necessarily satisfy the entropy condition.*

Lemma 3.32 *If numerical viscosity* $Q(\nu)$ *satisfies* (3.90), *then the three-point scheme* (3.87)-(3.88) *or* (3.89) *is at most only first-order accurate, and the modified equation is* (3.83) *with*

$$\beta(u, \lambda) = \frac{1}{2\lambda^2}\big(Q(\nu) - \nu^2\big) \ \text{ or } g = \frac{h}{2}\big(Q(\nu) - \nu^2\big)u_x. \tag{3.92}$$

Proof: Use $u_j = u(x_j, t)$ to denote the smooth solution of (1.3), and assume $0 < h < 1$. The numerical flux $\hat{f}_{j+\frac{1}{2}}$ of any second-order accurate scheme should satisfy

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}^{\mathrm{LW}}_{j+\frac{1}{2}} + \mathcal{O}(h^2), \tag{3.93}$$

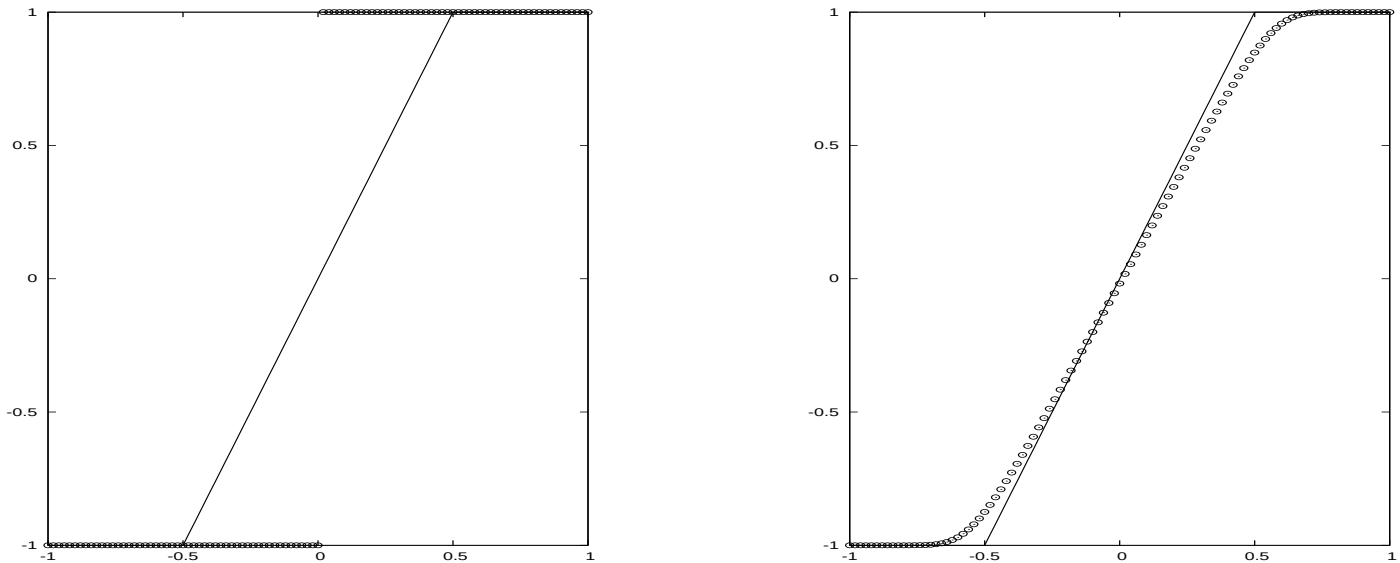


Figure 13: The solutions at $t = 0.5$ obtained with 101 mesh points. Left: first-order accurate upwind scheme, CFL=0.9; Right: modified LF scheme, CFL=0.45. The discrete initial data are $u_j^0 = -1$ $(j \leq 0)$; $u_j^0 = 1$ $(j > 0)$.

where

$$\hat{f}^{\mathrm{LW}}_{j+\frac{1}{2}} = \frac{1}{2}\left(f_j + f_{j+1} - \frac{\nu^2_{j+\frac{1}{2}}}{\lambda}\Delta^+_x u_j\right).$$

Comparing the numerical flux of the three-point scheme (3.87)-(3.88) with that of the LW scheme gives

$$\begin{aligned}\left|\hat{f}^{\mathrm{LW}}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\right| &= \left|\frac{1}{2\lambda}\left[Q(\nu_{j+\frac{1}{2}}) - \nu^2_{j+\frac{1}{2}}\right]\Delta^+_x u_j\right| \\ &\geq \frac{1}{2\lambda}\left||\nu_{j+\frac{1}{2}}| - |\nu^2_{j+\frac{1}{2}}|\right| \cdot \left|\Delta^+_x u_j\right| \\ &= \mathcal{O}(h) > \mathcal{O}(h^2),\end{aligned}$$

thus the three-point scheme (3.87)-(3.88) is at most only first-order accurate. ∎

The above results implies that in order to construct a second-order accurate TVD scheme satisfying (3.90), it does at least use

five points such as $\{u_{j-l}, \cdots, u_{j+l}\}$ $(l \geq 2)$ and the scheme is nonlinear. If the three-point first-order accurate TVD scheme (3.87)-(3.88) is applied to (3.85), then the numerical flux in (3.86) may be written in the form

$$\hat{f}^M_{j+\frac{1}{2}} = \frac{1}{2}(f_j + f_{j+1}) + \frac{1}{2\lambda}(g_j + g_{j+1}) - \frac{1}{2\lambda}Q(\nu_{j+\frac{1}{2}} + \gamma_{j+\frac{1}{2}})\Delta_x^+ u_j, \tag{3.94}$$

where $\gamma_{j+\frac{1}{2}} = \frac{\Delta_x^+ g_j}{\Delta_x^+ u_j}$, $f_j = f(u_j)$, $g_j = g(u_{j-1}, u_j, u_{j+1})$. Obviously, when $Q(\nu)$ satisfies (3.90), the scheme (3.86) with the numerical flux (3.94) may be TVD.

Lemma 3.33 *If $Q(\nu)$ is Lipschitz continuous, and the Taylor expansion of g in (3.94) satisfies*

$$g = \frac{h}{2}\big(Q(\nu) - \nu^2\big)u_x + \mathcal{O}(h^2), \tag{3.95}$$

then the scheme (3.86) *with the numerical flux* (3.94) *is a second-order accurate approximation of* (1.3).

Lemma 3.34 *If define*

$$\bar{g}_j := s \cdot \max\{0, \min(\sigma_{j+\frac{1}{2}} |\Delta_x^+ u_j|, s\sigma_{j-\frac{1}{2}} \Delta_x^- u_j)\}, \tag{3.96}$$

where $s = sign(\Delta_x^+ u_j)$, $\sigma_{j+\frac{1}{2}} = \frac{1}{2}[Q(\nu) - \nu^2]_{j+\frac{1}{2}}$, *then* $\bar{g}_j$ *and* $\bar{\gamma}_{j+\frac{1}{2}}$ *satisfy*

$$\bar{g} = h\sigma(\nu) u_x + \mathcal{O}(h^2), \tag{3.97}$$

and

$$|\bar{\gamma}_{j+\frac{1}{2}}| = \left| \frac{\Delta_x^+ \bar{g}_j}{\Delta_x^+ u_j} \right| \leq \sigma(\nu_{j+\frac{1}{2}}). \tag{3.98}$$

Proof: (1). Under the CFL condition $|\nu| \leq 1$ and the TVD condition (3.90), we have $\sigma(\nu) \geq 0$.

If $\Delta_x^+ u_j \Delta_x^+ u_{j-1} > 0$, then the definition of $\bar{g}_j$ gives

$$\begin{aligned}
\bar{g}_j &= s \cdot \min(\sigma_{j+\frac{1}{2}} |\Delta_x^+ u_j|, \sigma_{j-\frac{1}{2}} |\Delta_x^+ u_{j-1}|) \\
&= \frac{s}{2} \left(\sigma_{j-\frac{1}{2}} |\Delta_x^+ u_{j-1}| + \sigma_{j+\frac{1}{2}} |\Delta_x^+ u_j| - \left| \sigma_{j-\frac{1}{2}} |\Delta_x^+ u_{j-1}| - \sigma_{j+\frac{1}{2}} |\Delta_x^+ u_j| \right| \right) \\
&= \sigma_{j\pm\frac{1}{2}} \Delta_x^{\pm} u_j - \frac{1}{2} \left[s | \sigma_{j+\frac{1}{2}} \Delta_x^+ u_j - \sigma_{j-\frac{1}{2}} \Delta_x^- u_j | \pm (\sigma_{j+\frac{1}{2}} \Delta_x^+ u_j - \sigma_{j-\frac{1}{2}} \Delta_x^- u_j) \right] \\
&= \sigma_{j\pm\frac{1}{2}} \Delta_x^{\pm} u_j + \mathcal{O}(h^2).
\end{aligned}$$

If $\Delta_x^+ u_j \Delta_x^+ u_{j-1} < 0$ and assume that $\left(\frac{\partial u}{\partial x}\right)_{x_j} = 0$, then (3.96) tells us that $\bar{g}_j = 0$. Because $u_{j\pm1} - u_j = \pm h \frac{\partial u}{\partial x}\Big|_j + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}\Big|_j + \mathcal{O}(h^3) = \mathcal{O}(h^2)$, $\bar{g}_j$ can still be rewritten as $0 = \bar{g}_j = \sigma_{j\pm\frac{1}{2}} \Delta_x^{\pm} u_j + \mathcal{O}(h^2)$. Thus $\bar{g}_j$ is an approximation of g and satisfies (3.97).

(2). The definition of $\bar{g}_j$ tells us that $\bar{g}_j \bar{g}_{j+1} > 0$, and

$$\begin{aligned}|\bar{g}_{j+1} - \bar{g}_j| &\le \max\Big\{|\bar{g}_j|, |\bar{g}_{j+1}|\} \le \max\{\min(\sigma_{j+\frac{1}{2}}|\Delta_x^+ u_j|, \\ &\qquad \sigma_{j-\frac{1}{2}}|\Delta_x^- u_j|), \min(\sigma_{j+\frac{3}{2}}|\Delta_x^+ u_{j+1}|, \sigma_{j+\frac{1}{2}}|\Delta_x^+ u_j|)\Big\} \\ &\le \sigma_{j+\frac{1}{2}}|\Delta_x^+ u_j|,\end{aligned}$$

which implies (3.98). ∎

Theorem 3.35 *If g and $Q(\nu)$ in the scheme (3.86) with flux (3.94) satisfy (3.98) and (3.90), respectively, then under the CFL condition*

$$\max_j \left|\nu_{j+\frac{1}{2}} + \gamma_{j+\frac{1}{2}}\right| \le 1, \tag{3.99}$$

the scheme (3.86) with flux (3.94) is TVD, and second-order accurate away from the extreme point where $u_x = 0$.

Proof: Thanks to Lemma 3.31, the scheme (3.86) with flux (3.94) is TVD under the CFL condition

$$\max \left| \nu_{j+\frac{1}{2}} + \gamma_{j+\frac{1}{2}} \right| \leq \mu \leq 1,$$

which is implied by the original CFL condition (3.91). In fact, using (3.98) and (3.90) gives

$$\begin{aligned} \left| \nu_{j+\frac{1}{2}} + \gamma_{j+\frac{1}{2}} \right| \leq & |\nu_{j+\frac{1}{2}}| + \sigma_{j+\frac{1}{2}} \leq |\nu_{j+\frac{1}{2}}| + \frac{1}{2}[Q - \nu^2]_{j+\frac{1}{2}} \\ \leq & |\nu_{j+\frac{1}{2}}| + \frac{1}{2}[1 - \nu^2]_{j+\frac{1}{2}} = 1 - \frac{1}{2}(|\nu_{j+\frac{1}{2}}| - 1)^2 \leq 1, \end{aligned}$$

whenever $|\nu| \leq Q(\nu) \leq 1$. ∎

Generally, it is possible to choose

$$g = \bar{g} + \hat{g}, \qquad \gamma_{j+\frac{1}{2}} = \bar{\gamma}_{j+\frac{1}{2}} + \hat{\gamma}_{j+\frac{1}{2}},$$

where $\bar{g}$ is defined by (3.96), while $\hat{g} = \mathcal{O}(h^2)$ is Lipschitz continuous, and

$$\hat{\gamma}_{j+\frac{1}{2}} = \frac{\Delta_x^+ \hat{g}_j}{\Delta_x^+ u_j},$$

is uniformly bounded.

3.4.6 ENO and WENO schemes

This section introduces the ENO and WENO schemes, which are motivated by the accuracy analysis of the two-dimensional TVD scheme that "*any conservative scheme for solving scalar conservation laws in two space dimensions, which is total variation diminishing, is at most first-order accurate*" [23].

Both ENO and WENO schemes use the idea of adaptive stencils to automatically achieve high order accuracy and non-oscillatory property near discontinuities. The ENO schemes use

the "smoothest" stencil among several candidates to approximate the functions (fluxes or solutions) to a high order accuracy, while WENO schemes perform a nonlinear weighted combination of several candidates to obtain a higher order approximation. The first ENO scheme is constructed in 1987 by Harten, Osher, Engquist, and Chakravarthy in the form of cell averages [29], while the first WENO scheme is developed in 1994 by Liu, Osher and Chan for a third-order accurate finite volume version [50]. Multi-dimensional, third- and fifth-order finite difference WENO schemes are constructed in 1996 by Jiang and Shu [37] with a general framework for the design of smoothness indicators and nonlinear weights. The readers are referred to [27, 28] and the review paper [66].

3.4.6.1 ENO reconstruction

The ENO reconstruction problem is set as follows: Give a mesh and the cell averages of a piecewise smooth function $u(x)$, then reconstruct $R(x;\bar{u})$, a piecewise polynomial function of x of uniform polynomial degree r, that satisfies

(1) rth-order accurate approximation

$$R(x;\bar{u}) = u(x) + O(h^r), \tag{3.100}$$

wherever $u(x)$ is smooth.

(2) conservation in the sense of

$$\frac{1}{h_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} R(x;\bar{u})\ dx = \bar{u}_j. \tag{3.101}$$

(3) ENO property in sense that

$$TV\big(R(\cdot;\bar{u})\big) \le TV(u) + O(h^r), \tag{3.102}$$

which excludes a Gibbs-like phenomenon but allows for the production of spurious oscillations on the level of the truncation error.

(i) ENO reconstruction via primitive function Assume that a non-uniform mesh $\{x_{j+\frac{1}{2}}, j \in \mathbb{Z}\}$ is given and the cell averages of a piecewise smooth function $u(x)$ are calculated by

$$\bar{u}_j = \frac{1}{h_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\, dx,$$

where $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. The above ENO reconstruction problem of $u(x)$ may be converted to an interpolation problem of the

primitive function $w(x) = \int_{x_{\frac{1}{2}}}^{x} u(\xi)\, d\xi$ via the interpolation data $\{(x_{j+\frac{1}{2}}, w_{j+\frac{1}{2}})\}$. Such technique has been used in Section 3.4.3, where $w_{j+\frac{1}{2}-\ell}$, $\cdots$, $w_{j+\frac{1}{2}-\ell+r}$ are used to interpolate $w(x)$ and give an approximation of u in the cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. How to set the value of ℓ? If $u(\cdot, t) \in C^r$, the interpolation of w for any ℓ $(1 \le \ell \le r)$ will give an approximation with rth order accuracy near x_j, that is,

$$q_{r,j}(x; w) = w(x) + O(h^{r+1}), \quad x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}},$$

and

$$\frac{d}{dx} q_{r,j}(x; w) = u(x) + O(h^r), \quad x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}}.$$

However, a high resolution shock-capturing method must be able to cope with the possibility of non-smooth data. Near the discon-

tinuity, the high-order accuracy is not expected, but a stencil denoted by $S_\ell(j,r)$ consisting of the nodes $\{x_{j+\frac{1}{2}-\ell}, \cdots, x_{j+\frac{1}{2}-\ell+r}\}$ has to be chosen to avoid numerical oscillation. It is well-known that the high-order interpolation polynomial may admit strong oscillation (so-called Runge phenomenon) even though for the smooth data, and becomes more serious for non-smooth data.

In the piecewise linear reconstruction of Section 3.4.2, the slope limiter, e.g. minmod slope limiter, has been used to limit the possible oscillation by comparing two linear polynomials with the slope of $\Delta_x^- \bar{u}_j / h_{j-\frac{1}{2}}$ and $\Delta_x^+ \bar{u}_j / h_{j+\frac{1}{2}}$ respectively, and then choosing the one with smaller slope in the absolute value, that is to say, a linear polynomial within each cell is chosen to be not steeper. Those linear polynomials within cells give a globally (discontinuous) piecewise linear approximation of the solution which is non-oscillatory in the sense that the total variation is not more

than the original, or no new extreme point is produced. The idea may be extended to higher-degree polynomial interpolation by choosing ℓ or $S_\ell(j,r)$ for each mesh index j such that an interpolation polynomial with smallest oscillation may be obtained from the interpolation data $\{(x_{j+\frac{1}{2}-\ell}, w_{j+\frac{1}{2}-\ell}), \cdots, (x_{j+\frac{1}{2}-\ell+r}, w_{j+\frac{1}{2}-\ell+r})\}$ for $1 \le \ell \le r$. It is the key point of the ENO method of Harten, Engquist, Osher and Chakravarthy [29].

The ENO procedure via primitive function is as follows. Introduce $H_r(x;w)$, a piecewise but continuous polynomial function of x that interpolates w at the points $\{x_{j+\frac{1}{2}}\}$, i.e.

$$H_r(x_{j+\frac{1}{2}}; w) = w(x_{j+\frac{1}{2}}),$$

and

$$H_r(x;w) = q_{r,j}(x;w), \quad \text{for } x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}],$$

where $q_{r,j}(x;w)$ denotes the (unique) polynomial that interpolates $w(x)$ at the $(r+1)$ successive points $\{x_{i_r(j)}, \cdots, x_{i_r(j)+r}\}$ that contains $x_{j\pm\frac{1}{2}}$

$$q_{r,j}(x_i;w) = w(x_i), \quad \text{for } i_r(j) \le i \le i_r(j)+r,\ 1-r \le i_r(j) \le 0. \tag{3.103}$$

For example, $i_r(j) = j + \frac{1}{2} - \ell$, where $1 \le \ell \le r$. Other examples are give below.

Example 3.22 *If $r = 1$, then $0 \le i_1(j) - j \le 0$, which implies only one choice of $i_1(j)$, i.e. $i_1(j) = j$. For this, (3.103) becomes*

$$q_{1,j}(x_i;w) = w(x_i), \quad j \le i \le j+1,$$

and $q_{r,j}(x;w)$ is a 1st degree polynomial in x approximating $w(x)$.

Example 3.23 *If $r = 2$, then $-1 \le i_2(j) - j \le 0$, which implies that there are two different choices of $i_2(j)$: $i_2(j) = j$ or $j-1$. For two different interpolation polynomials with degree of 2, (3.103) becomes*

$$q_{2,j}(x_i; w) = w(x_i), \qquad j \le i \le j + 2,$$

and

$$q_{2,j}(x_i; w) = w(x_i), \qquad j - 1 \le i \le j + 1.$$

Clearly, (3.103) implies that there exist exactly r such polynomials corresponding to r different choices of $i_r(j)$ subject to $1 - r \le i_r(j) \le 0$. This freedom is used to assign the interval $(x_{i_r(j)}, x_{i_r(j)+r})$ covering a stencil of $(r + 1)$ points so that $w(x)$ is "smoothest" in $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ in some asymptotic sense. The

information about smoothness of $w(x)$ is extracted from a table of the Newton divided differences of w, which are recursively defined by

$$\begin{cases} w[x_{j+\frac{1}{2}}] = w(x_{j+\frac{1}{2}}), \\ w[x_{j+\frac{1}{2}}, \cdots, x_{j+\frac{1}{2}+k}] = (w[x_{j+\frac{1}{2}+1}, \cdots, x_{j+\frac{1}{2}+k}] \\ \qquad\qquad -w[x_i j + \frac{1}{2} \cdots, x_{j+\frac{1}{2}+k-1}])/(x_{j+\frac{1}{2}+k} - x_{j+\frac{1}{2}}). \end{cases}$$

It is easy to verify that

$$w[x_{j+\frac{1}{2}}, \cdots, x_{j+\frac{1}{2}+k}] = \frac{1}{k!}\frac{d^k}{dx^k} w(\xi_{j,k}), \quad x_{j+\frac{1}{2}} \le \xi_{j,k} \le x_{j+\frac{1}{2}+k}.$$

if $w \in C^\infty[x_{j+\frac{1}{2}}, x_{j+\frac{1}{2}+k}]$. However, if p-order derivative of w has a jump discontinuity within this interval, then

$$w[x_{j+\frac{1}{2}}, \cdots, x_{j+\frac{1}{2}+k}] = \mathcal{O}(h^{-k+p}[w^{(p)}]),\ 0 \le p \le k.$$

Here $[w^{(p)}]$ denotes the jump in the pth derivative of w. The above both equations show that $|w[x_{j+\frac{1}{2}}, \cdots, x_{j+\frac{1}{2}+k}]|$ provides an asymptotic measure of the smoothness of w in the interval $(x_{j+\frac{1}{2}}, x_{j+\frac{1}{2}+k})$, in the sense that if w is smooth in (x_{i_1}, x_{i_1+k}) but is discontinuous in (x_{i_2}, x_{i_2+k}), then $|w[x_{i_1}, \cdots, x_{i_1+k}]| < |w[x_{i_2}, \cdots, x_{i_2+k}]|$ for h sufficiently small. Hence the problem of choosing a stencil of points for which w is "smoothest" is basically the same as that of finding an interval in which w has the "smallest divided difference".

Recursive algorithm to evaluate $i_r(j)$

- Set $i_1(j) = j$, i.e. $q_{1,j}$ is the 1st-degree polynomial interpolating w by $q_{1,j}(x_j; w) = w(x_j)$ and $q_{1,j}(x_{j+1}; w) = w(x_{j+1})$.

- Assume that $i_k(j)$ has been defined, i.e. $q_{k,j}$ is the kth

degree polynomial satisfying the interpolation conditions

$$q_{k,j}(x_i; w) = w(x_i),\ i = i_k(j), \cdots, i_k(j) + k$$

- Consider two candidates for $q_{k+1,j}$, which are $(k+1)$th degree polynomials obtained by adding the left and right neighboring points of stencil $\{x_i, i = i_k(j), \cdots, i_k(j) + k\}$ to the left and right of the above stencil respectively; this corresponds to setting $i_{k+1}(j) = i_k(j) - 1$ or $i_{k+1}(j) = i_k(j)$, respectively. We choose the one of two candidates that gives a $(k+1)$th order divided difference with smaller absolute value, that is

$$i_{k+1}(j) = \begin{cases} i_k(j), & |w[x_{i_k(j)-1}, \cdots, x_{i_k(j)+k}]| \geq |w[x_{i_k(j)}, \cdots, x_{i_k(j)+k+1}]|, \\ i_k(j) - 1, & \text{otherwise.} \end{cases}$$

Finally, define

$$R(x; \bar{u}) = \frac{d}{dx} H_r(x, w) = \frac{d}{dx} q_{r,j}(x; w), \quad \text{for } x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}].$$

Example 3.24 *In the case of* $r = 1$*, based on the stencil* $S_1(j,1) = \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}$*, a linear interpolation polynomial may be uniquely obtained by*

$$q_{1,j}(x;w) = w_{j-\frac{1}{2}} + w\left[j-\tfrac{1}{2}, j+\tfrac{1}{2}\right](x - x_{j-\frac{1}{2}}), \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}],$$

where $w\left[j-\frac{1}{2}, j+\frac{1}{2}\right] := \frac{w_{j+\frac{1}{2}} - w_{j-\frac{1}{2}}}{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}}$ *is the first-order divided difference. In this case, there exists only one candidate.*

Example 3.25 *Assume that* $r = 2$*. From the stencils* $S_1(j,2) = \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}\}$ *and* $S_2(j,2) = \{x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}$*, we have two quadratic interpolation polynomials*

$$p_{2,1}(x) := q_{1,j}(x;w) + w\left[j-\tfrac{1}{2}, j+\tfrac{1}{2}, j+\tfrac{3}{2}\right](x - x_{j-\frac{1}{2}})(x - x_{j+\frac{1}{2}}),$$
$$p_{2,2}(x) := q_{1,j}(x;w) + w\left[j-\tfrac{1}{2}, j+\tfrac{1}{2}, j-\tfrac{3}{2}\right](x - x_{j-\frac{1}{2}})(x - x_{j+\frac{1}{2}}),$$

for $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, *where* $w\left[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}\right]$, $w\left[j-\frac{1}{2}, j+\frac{1}{2}, j-\frac{3}{2}\right]$ *are two second-order divided differences. The ENO method chooses*

$$q_{2,j}(x;w) = \begin{cases} p_{2,1}(x), & \text{if } |w\left[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}\right]| \le |w\left[j-\frac{1}{2}, j+\frac{1}{2}, j-\frac{3}{2}\right]|, \\ p_{2,2}(x), & \text{otherwise.} \end{cases}$$

Example 3.26 *Assume that* $r = 3$. *The ENO method is to choose one of three stencils*

$$S_1(j,3) = \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}, x_{j+\frac{5}{2}}\},$$
$$S_2(j,3) = \{x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}\},$$
$$S_3(j,3) = \{x_{j-\frac{5}{2}}, x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}.$$

For this purpose, we perform the following steps.

(a) If $|w\left[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}\right]| \le |w\left[j-\frac{1}{2}, j+\frac{1}{2}, j-\frac{3}{2}\right]|$, *then*

compare the absolute values of two 3rd-order divided differences

$$w\left[j-\frac{1}{2},j+\frac{1}{2},j+\frac{3}{2},j+\frac{5}{2}\right],\ \textit{and}\ w\left[j-\frac{1}{2},j+\frac{1}{2},j+\frac{3}{2},j-\frac{3}{2}\right].$$

If $|w\left[j-\frac{1}{2},j+\frac{1}{2},j+\frac{3}{2},j+\frac{5}{2}\right]|$ *is smaller, then the final stencil is* $S_1(j,3)$ *and*

$$q_{3,j}(x;w)=p_{2,1}(x)+w\left[j-\tfrac{1}{2},j+\tfrac{1}{2},j+\tfrac{3}{2},j+\tfrac{5}{2}\right](x-x_{j-\frac{1}{2}})(x-x_{j+\frac{1}{2}})(x-x_{j+\frac{3}{2}}),$$

otherwise, if $|w\left[j-\frac{1}{2},j+\frac{1}{2},j+\frac{3}{2},j-\frac{3}{2}\right]|$ *is smaller, the final stencil is* $S_2(j,3)$ *and*

$$q_{3,j}(x;w)=p_{2,1}(x)+w\left[j-\tfrac{1}{2},j+\tfrac{1}{2},j+\tfrac{3}{2},j-\tfrac{3}{2}\right](x-x_{j-\frac{1}{2}})(x-x_{j+\frac{1}{2}})(x-x_{j+\frac{3}{2}}).$$

where $p_{2,1}(x)$ *is defined in Example 3.25 and* $x\in[x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}]$.

(b) If $|w\left[j-\frac{1}{2},j+\frac{1}{2},j+\frac{3}{2}\right]|>|w\left[j-\frac{1}{2},j+\frac{1}{2},j-\frac{3}{2}\right]|$, *then compare the absolute values of two third-order divided differences*

$$w\left[j-\frac{1}{2},j+\frac{1}{2},j-\frac{3}{2},j+\frac{3}{2}\right]\ \textit{and}\ w\left[j-\frac{1}{2},j+\frac{1}{2},j-\frac{3}{2},j-\frac{5}{2}\right].$$

If $|w\left[j-\frac{1}{2}, j+\frac{1}{2}, j-\frac{3}{2}, j+\frac{3}{2}\right]|$ *is smaller, then the final stencil is* $S_2(j,3)$ *and*

$$q_{3,j}(x;w) = p_{2,2}(x) + w\left[j-\tfrac{1}{2}, j+\tfrac{1}{2}, j-\tfrac{3}{2}, j+\tfrac{3}{2}\right](x-x_{j-\frac{1}{2}})(x-x_{j+\frac{1}{2}})(x-x_{j-\frac{3}{2}}),$$

otherwise, if $|w\left[j-\frac{1}{2}, j+\frac{1}{2}, j-\frac{3}{2}, j-\frac{5}{2}\right]|$ *is smaller, the final stencil is* $S_3(j,3)$ *and*

$$q_{3,j}(x;w) = p_{2,2}(x) + w\left[j-\tfrac{1}{2}, j+\tfrac{1}{2}, j-\tfrac{3}{2}, j-\tfrac{5}{2}\right](x-x_{j-\frac{1}{2}})(x-x_{j+\frac{1}{2}})(x-x_{j-\frac{3}{2}}).$$

where $p_{2,2}(x)$ *is also defined in Example 3.25 and* $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$.

For general r, the above process can be recursively continued, until a rth degree polynomial is obtained by using some $(r+1)$ points containing $x_{j-\frac{1}{2}}$ and $x_{j+\frac{1}{2}}$. For the primitive function $w(x)$ with point values $w_{j+\frac{1}{2}}$, `ENO procedure` is as follows.

- Compute kth order divided differences of w and present the Newton divided difference table, $0 \le k \le r$, e.g. zeroth-

and first- order divided differences are $w[j+\frac{1}{2}] = w_{j+\frac{1}{2}}$, $w[j-\frac{1}{2}, j+\frac{1}{2}] = \frac{w[j+\frac{1}{2}]-w[j-\frac{1}{2}]}{x_{j+\frac{1}{2}}-x_{j-\frac{1}{2}}} = \bar{u}_j$.

- Compare the second-order divided differences $w[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}]$ and $w[j-\frac{3}{2}, j-\frac{1}{2}, j+\frac{1}{2}]$ in the absolute value. If the absolute value of $w[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}]$ is smaller, then the interpolation stencil will consist of the points $\{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}\}$
- Under the assumption of the last step, compare third-order divided differences

  $w\left[j-\frac{3}{2}, j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}\right]$ and $w\left[j-\frac{1}{2}, j+\frac{1}{2}, j+\frac{3}{2}, j+\frac{5}{2}\right]$,

  in the absolute value.
- Repeat the above process.

Because the 1st order divided difference of w is equal to the cell average $\bar{u}_j$, that is, $\frac{w[j+\frac{1}{2}]-w[j-\frac{1}{2}]}{h} = \bar{u}_j$, so that the $(k+1)$th-order divided difference of w is equal to the "kth-order divided difference" of u in the cell averages $\{\bar{u}_j\}$, $k \geq 0$. Moreover, the term with a degree of 0 in $q_{r,j+\frac{1}{2}}(x;w)$ does not appear in $\frac{d}{dx}q_{r,j+\frac{1}{2}}(x;w)$. Thus, the procedure in Example 3.26 can be directly implemented on the cell averages $\{\bar{u}_j\}$ and avoids calculation of the point values $\{w_{j+\frac{1}{2}}\}$.

For the cell averages of u, `ENO procedure` may be translated into as follows.

- Compute "kth order divided differences" of u and present the "Newton divided difference table", $0 \leq k \leq r-1$, e.g. zeroth- and first- order divided differences are $\bar{u}[j] = \bar{u}_j$, $\bar{u}[j, j+1] = \frac{\bar{u}[j+1]-\bar{u}[j]}{x_{j+1}-x_j}$.

- Take the stencil $S_1 = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ initially.

- Compare the "first-order divided differences" $\bar{u}[j, j+1]$ and $\bar{u}[j-1, j]$ in the absolute value.

  If the absolute value of $\bar{u}[j, j+1]$ is smaller, then the stencil S_1 is extended to $S_2 = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$, otherwise the stencil S_1 is replaced with $S_2 = [x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$.

- Assume that $|\bar{u}[j, j+1]| < |\bar{u}[j-1, j]|$. Compare the "2nd order divided differences" $\bar{u}[j, j+1, j+2]$ and $\bar{u}[j-1, j, j+1]$ in the absolute value to decide that the stencil S_2 can be extended to $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}, x_{j+\frac{5}{2}}]$ or $[x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$,

- Repeat the above process. Finally, we can reconstruct the rth order polynomial $R(x, \bar{u}) := \frac{d}{dx} q_{r,j}(x; w)$ for $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$.

Some examples for the reconstruction via the primitive function $w(x)$ are given below under the assumption of the uniform mesh.

Example 3.27 $r = 1$. *For stencil* $S_1(j,1) = \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}$, *the interpolation polynomial of* w *is*

$$\begin{aligned} q_{1,j}(x,w) &= w_{j-\frac{1}{2}} + (x - x_{j-\frac{1}{2}}) \frac{w_{j+\frac{1}{2}} - w_{j-\frac{1}{2}}}{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}} \\ &= w_{j-\frac{1}{2}} + (x - x_{j-\frac{1}{2}}) \bar{u}_j, \end{aligned}$$

which gives

$$R(x,\bar{u}) = \frac{d}{dx} q_{1,j}(x,w) = \bar{u}_j, \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}].$$

Example 3.28 *In the case of* $r = 2$. *According to Example 3.25, the derivatives of* $p_{2,1}(x)$ *and* $p_{2,2}(x)$ *are*

$$\frac{d}{dx}p_{2,1}(x) = \bar{u}_j + (x - x_j)\frac{\bar{u}_{j+1} - \bar{u}_j}{h}, \quad \frac{d}{dx}p_{2,2}(x) = \bar{u}_j + (x - x_j)\frac{\bar{u}_j - \bar{u}_{j-1}}{h},$$

for $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. *Thus the finally reconstructed polynomial* $R(x, \bar{u})$ *is equal to one of them, which has a smaller absolute value of the slope.*

Example 3.29 $r = 3$. *For* $S_3(j, r) = \{x_{j-\frac{5}{2}}, x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}$, *the reconstructed polynomial* $R(x, \bar{u})$ *is*

$$\begin{aligned}\frac{d}{dx}q_{r,j}(x, w) =& \bar{u}_{j-1} + \frac{\bar{u}_j - \bar{u}_{j-2}}{2h}(x - x_{j-1}) \\ &+ \frac{\bar{u}_j - 2\bar{u}_{j-1} + \bar{u}_{j-2}}{2h^2}\left[(x - x_{j-1})^2 - \frac{h^2}{12}\right].\end{aligned}$$

For the stencil $S_2(j,r) = \{x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}\}$, *the reconstructed polynomial* $R(x,\bar{u})$ *is*

$$\begin{aligned}\frac{d}{dx}q_{r,j}(x,w) =& \bar{u}_j + \frac{\bar{u}_{j+1}-\bar{u}_{j-1}}{2h}(x-x_j) \\ &+ \frac{\bar{u}_{j+1}-2\bar{u}_j+\bar{u}_{j-1}}{2h^2}\left[(x-x_j)^2 - \frac{h^2}{12}\right].\end{aligned}$$

For the stencil $S_1(j,r) = \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}, x_{j+\frac{5}{2}}\}$, *the reconstructed polynomial* $R(x,\bar{u})$ *is*

$$\begin{aligned}\frac{d}{dx}q_{r,j}(x,w) =& \bar{u}_{j+1} + \frac{\bar{u}_{j+2}-\bar{u}_j}{2h}(x-x_{j+1}) \\ &+ \frac{\bar{u}_{j+2}-2\bar{u}_{j+1}+\bar{u}_j}{2h^2}\left[(x-x_{j+1})^2 - \frac{h^2}{12}\right].\end{aligned}$$

For a piecewise smooth function w, Harten et al. have showed that the above interpolation technique $H_r(x;w)$ satisfies

$$\frac{d^\ell}{dx^\ell} H_r(x;w) = \frac{d^\ell}{dx^\ell} w(x) + \mathcal{O}(h^{r+1-\ell}), \quad 0 \le \ell \le r,$$

wherever $w(x)$ is smooth, and $H_r(x;w)$ is an ENO interpolation of w in the sense that

$$TV(H_r(\cdot;w)) \le TV(w) + O(h^{r+1}). \tag{3.104}$$

Those implies

$$\frac{d^\ell}{dx^\ell} R(x;\bar{u}) = \frac{d^\ell}{dx^\ell} u(x) + \mathcal{O}(h^{r-\ell}), \quad 0 \le \ell < r,$$

wherever $u(x)$ is smooth, and

$$\frac{1}{h_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} R(x;\bar{u})\,dx = \frac{1}{h_j}\left[H_r(x_{j+\frac{1}{2}};w) - H_r(x_{j-\frac{1}{2}};w)\right]$$
$$= \frac{1}{h_j}\left[w(x_{j+\frac{1}{2}}) - w(x_{j-\frac{1}{2}})\right] = \bar{u}_j.$$

The non-oscillatory nature (3.102) in the reconstruction follows primarily from that of the interpolation (3.104).

(*ii*) **ENO reconstruction via deconvolution** Assume that the mesh $\{x_j = jh, j \in \mathbb{Z}\}$ is uniform and the cell averages of $u(x)$ are given and denoted by $\{\bar{u}_j\}$. If define the function

$$\bar{u}(x) = \frac{1}{h}\int_{-h/2}^{h/2} u(x+y)\,dy, \tag{3.105}$$

which is globally defined sliding-average function of u, then $\bar{u}_j$ is the point values of $\bar{u}(x)$ at the mesh point x_j, that is, $\bar{u}_j =$

$\bar{u}(x_j)$. With the aid of the interpolation data $\{x_j, \bar{u}(x_j)\}$, the reconstruction problem may be converted into an interpolation problem. It is worth noting that $\bar{u}(x)$ is the convolution of $u(x)$ and $\psi_h(x)$, i.e. $\bar{u}(x) = (u * \psi_h)(x)$, where

$$\psi_h(x) = \begin{cases} 1/h, & |x| < \frac{h}{2}, \\ 0, & |x| \geq \frac{h}{2}. \end{cases}$$

Expanding $u(x+y)$ at $y = 0$ as a Taylor series and substituting it into (3.105) gives

$$\bar{u}(x) = \sum_{k=0}^{\infty} \frac{u^{(k)}}{k!} \int_{-h/2}^{h/2} y^k \, dy = \sum_{k=0}^{\infty} \alpha_k h^k u^{(k)}(x),$$

where

$$\alpha_k = \begin{cases} 0, & k \text{ odd}, \\ 2^{-k}/(k+1)!, & k \text{ even}. \end{cases}$$

Multiplying both sides by $h^\ell \frac{d^\ell}{dx^\ell}$, and then truncating the expansion in the right hand side at $\mathcal{O}(h^r)$ gives

$$h^\ell \bar{u}^{(\ell)}(x) = \sum_{k=0}^{r-\ell-1} \alpha_k h^{k+\ell} u^{(k+\ell)}(x) + \mathcal{O}(h^r).$$

For $\ell = 0, \cdots, r-1$, the above equation may be written in a matrix-vector form

$$\begin{pmatrix} \bar{u}(x) \\ h\bar{u}'(x) \\ h^2\bar{u}''(x) \\ \vdots \\ h^{r-1}\bar{u}^{(r-1)}(x) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \alpha_2 & 0 & \alpha_4 & \cdots & \alpha_{r-1} \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \ddots & \ddots & \alpha_4 \\ & & & \ddots & \ddots & \ddots & 0 \\ & & & & \ddots & \ddots & \alpha_2 \\ & & & & & \ddots & 0 \\ 0 & & & & & & 1 \end{pmatrix} \begin{pmatrix} u(x) \\ hu'(x) \\ h^2u''(x) \\ \vdots \\ h^{r-1}u^{(r-1)}(x) \end{pmatrix}$$

$$+ \mathcal{O}(h^r) =: \boldsymbol{C}[u(x), hu'(x), h^2u''(x), \cdots, h^{r-1}u^{(r-1)}(x)]^T + \mathcal{O}(h^r),$$

which gives the relation between the derivatives $u^{(\ell)}(x)$ and their averaged functions $\bar{u}^{(\ell)}(x)$. It is obvious that the matrix $\boldsymbol{C}$ is upper triangular and diagonally dominant. Multiplying both sides

by $\boldsymbol{C}^{-1}$ from the left gives

$$\begin{pmatrix} u(x) \\ hu'(x) \\ \vdots \end{pmatrix} = \boldsymbol{C}^{-1} \begin{pmatrix} \bar{u}(x) \\ h\bar{u}'(x) \\ \vdots \end{pmatrix} + \mathcal{O}(h^r). \tag{3.106}$$

Based on the interpolation data $\{x_j, \bar{u}_j\}$ of the function $\bar{u}(x)$, we can interpolate $\bar{u}(x)$ by the polynomial $H_m(x; \bar{u})$ with $m \geq r-1$. Since $\bar{u}(x)$ is smoother than $u(x)$, it follows that

$$\frac{d^k}{dx^k} H_m(x; \bar{u}) = \frac{d^k}{dx^k} \bar{u}(x) + \mathcal{O}(h^{m+1-k}),$$

whenever $\bar{u}(x)$ is smooth. Although H_m is only continuous at x_j, the one-sided derivatives at $x_j \pm 0$ do satisfy the relations

$$\frac{d^k}{dx^k} H_m(x_j \pm 0; \bar{u}) = \frac{d^k}{dx^k} \bar{u}(x_j) + \mathcal{O}(h^{m+1-k}).$$

Now define $\bar{D}_{0,j} := \bar{u}_j$ and

$$\bar{D}_{\ell,j} := h^l \text{minmod}\left(\frac{d^\ell}{dx^\ell} H_m(x_j - 0; \bar{u}), \frac{d^\ell}{dx^\ell} H_m(x_j + 0; \bar{u})\right), 1 \le \ell \le r-1,$$

and denote $\bar{\boldsymbol{D}}_j = (\bar{D}_{0,j}, \cdots, \bar{D}_{r-1,j})^T$. It is clear that $\bar{D}_{l,j} = h^l \bar{u}^{(l)}(x_j) + \mathcal{O}(h^r)$. From (3.106), $\boldsymbol{D}_j = \boldsymbol{C}^{-1} \bar{\boldsymbol{D}}_j$ satisfies

$$\boldsymbol{D}_j = \left(u(x_j), hu'(x_j), \cdots, h^{r-1} u^{(r-1)}(x_j)\right)^T + \mathcal{O}(h^r).$$

Finally, the reconstruction polynomial approximating $u(x)$ is defined by

$$R(x; \bar{u}) := \sum_{k=0}^{r-1} \frac{1}{k!} D_{k,j} \left[\frac{(x - x_j)}{h}\right]^k, \quad x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}). \quad (3.107)$$

Because (3.106) may be regarded as a deconvolution to $\mathcal{O}(h^r)$), (3.107) is referred to as **reconstruction via deconvolution**. It

is not difficult to verify

$$\frac{d^\ell}{dx^\ell} R(x,\bar{u}) = \frac{d^\ell}{dx^\ell} u(x) + \mathcal{O}(h^{r-\ell}),$$

$$\frac{1}{h}\int_{-h/2}^{h/2} R(x+y;\bar{u})\, dy = \sum_{k=0}^{r-1} \frac{D_{k,j}}{k!} \frac{1}{h^{k+1}} \int_{-h/2}^{h/2} y^k\, dy$$

$$= D_{0,j} + \sum_{k=1}^{r-1} \alpha_k D_{k,j} = \bar{D}_{0,j} = \bar{u}_j,$$

where we have used the identities

$$\begin{cases} D_{k,j} = \bar{D}_{k,j} - \sum\limits_{\ell=k+1}^{r-1} \alpha_\ell D_{\ell,j}, \quad k = r-2, \cdots, 0, \\ D_{r-1,j} = \bar{D}_{r-1,j}, \end{cases}$$

which obtained by solving $\boldsymbol{D}_j = \boldsymbol{C}^{-1}\bar{\boldsymbol{D}}_j$ by back-substitution. Similarly, the non-oscillatory nature (3.102) in the reconstruction

(3.107) follows primarily from that of the interpolation $H_m(x;\bar{u})$.

3.4.6.2 ENO scheme based on fluxes

This section introduces two kind of ENO schemes based on the fluxes [67, 68]. The first is the ENO FDS (derived based on the point values of the flux and interpolation), while the second is the ENO finite volume scheme (developed based on the cell averages of the flux and reconstruction).

(i) **ENO FDS based on the flux interpolation**

Give the point values of the solution u of (1.3) at the mesh point x_j and calculate the the point values of $f^{\pm}(u)$, defined by

$$f^{\pm}(u)=\frac{1}{2}(f(u)\pm\alpha(u)),\quad \alpha\geq\max_{u}|f'(u)|,\quad \alpha \text{ is constant},\tag{3.108}$$

which implies

$$\frac{df^+}{du} \geq 0, \quad \frac{df^-}{du} \leq 0, \quad f^+(u) + f^-(u) = f(u), \tag{3.109}$$

and the numerical flux of the LF scheme for (1.3) can be expressed as

$$\hat{f}_{j+\frac{1}{2}} = f_j^+ + f_{j+1}^-, \quad f_j^\pm = f^\pm(u_j). \tag{3.110}$$

The Taylor series expansion reveals the following fact.

Lemma 3.36 *There exist constants*

$$a_2 = -\frac{1}{24}, \quad a_4 = \frac{7}{5760}, \quad a_6 = -\frac{31}{967680}, \quad \cdots,$$

such that if

$$\hat{f}_{j\pm\frac{1}{2}} = f_{j\pm\frac{1}{2}} + \sum_{k=1}^{m-1} a_{2k} h^{2k} \left(\frac{\partial^{2k}}{\partial x^{2k}} f \right)_{j\pm\frac{1}{2}} + \mathcal{O}(h^{2r+1}), \tag{3.111}$$

then the scheme

$$u_j^{n+1} = u_j^n - \lambda \left(\hat{f}^n_{j+\frac{1}{2}} - \hat{f}^n_{j-\frac{1}{2}} \right), \tag{3.112}$$

is $2r$th-order accurate in space in the sense that $\mathcal{L}_h(u) = \mathcal{L}(u) + \mathcal{O}(h^{2r})$ if u is smooth, where $\mathcal{L}_h$ and $\mathcal{L}$ denote the difference and differential operators respectively.

Proof: Using the Taylor series expansion gives

$$\begin{aligned}
\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}} =& f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} + \sum_{k=1}^{m-1} \left[\left(\frac{\partial^{2k} f}{\partial x^{2k}} \right)_{j+\frac{1}{2}} - \left(\frac{\partial^{2k} f}{\partial x^{2k}} \right)_{j-\frac{1}{2}} \right] + \mathcal{O}(h^{2m+1}) \\
=& h \frac{\partial f}{\partial x} + \frac{h^3}{2^2(3!)} \frac{\partial^3 f}{\partial x^3} + \frac{h^5}{2^4(5!)} \frac{\partial^5 f}{\partial x^5} + \frac{h^7}{2^6(7!)} \frac{\partial^7 f}{\partial x^7} + \cdots \\
& + \sum_{k=1}^{m-1} a_{2k} h^{2k} \left[h \frac{\partial^{2k+1} f}{\partial x^{2k+1}} + \frac{h^3}{2^2(3!)} \frac{\partial^{2k+3} f}{\partial x^{2k+3}} + \cdots \right] \\
=& h \frac{\partial f}{\partial x} + \left[\frac{1}{2^2(3!)} + a_2 \right] h^3 \frac{\partial^3 f}{\partial x^3} + \left[\frac{1}{2^4(5!)} + a_2 \frac{1}{2^2(3!)} + a_4 \right] h^5 \frac{\partial^5 f}{\partial x^5} \\
& + \left[\frac{1}{2^6(7!)} + a_2 \frac{1}{2^4(5!)} + a_4 \frac{1}{2^2(3!)} + a_6 \right] h^7 \frac{\partial^7 f}{\partial x^7} + \cdots .
\end{aligned}$$

Thus one needs

$$a_2 = -\frac{1}{2^2(3!)}, a_4 = -\frac{1}{2^4(5!)} - a_2 \frac{1}{2^2(3!)}, a_6 = -\frac{1}{2^6(7!)} - a_2 \frac{1}{2^4(5!)} - a_4 \frac{1}{2^2(3!)}, \cdots$$

which may complete the proof. ∎

According to (3.110), it is nature to consider the flux type splitting

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}^{+}_{j+\frac{1}{2}} + \hat{f}^{-}_{j+\frac{1}{2}}, \tag{3.113}$$

where both positive and negative fluxes $\hat{f}^{\pm}_{j+\frac{1}{2}}$ satisfy (3.111), that are

$$\hat{f}^{\pm}_{j+\frac{1}{2}} = f^{\pm}_{j+\frac{1}{2}} + \sum_{k=1}^{r-1} a_{2k} h^{2k} \left(\frac{\partial^{2k}}{\partial x^{2k}} f^{\pm} \right)_{j+\frac{1}{2}} + \mathcal{O}(h^{2r+1}). \tag{3.114}$$

In order to get (3.114), we interpolate $f^{\pm}$ by the polynomials $p^{\pm}_{j+\frac{1}{2}}(x)$ near $x = x_{j+\frac{1}{2}}$

$$p^{\pm}_{j+\frac{1}{2}}(x) = f^{\pm}(u(x)) + \mathcal{O}(h^{2r+1}), \tag{3.115}$$

and then define

$$\hat{f}^{\pm}_{j+\frac{1}{2}} = p^{\pm}_{j+\frac{1}{2}}(x_{j+\frac{1}{2}}) + \sum_{k=1}^{r-1} a_{2k} h^{2k} \left(\frac{\partial^{2k}}{\partial x^{2k}} p^{\pm}_{j+\frac{1}{2}}(x) \right)_{x=x_{j+\frac{1}{2}}} . \tag{3.116}$$

Clearly, if (3.115) is true, then the numerical fluxes obtained by (3.116) satisfy (3.114).

The previous ENO adaptive stencil idea may be employed to derive the polynomials $p^{\pm}_{j+\frac{1}{2}}(x)$ satisfying (3.115), which are polynomials of degree $2r$ interpolating $f^{\pm}(u(x))$ at $(2r+1)$ points near $x_{j+\frac{1}{2}}$ from the smoothest possible region, but starting with the correct one according to (3.110).

The ENO FDS based on the flux interpolation may be implemented as follows [67]. The first part is to get $p^{+}_{j+\frac{1}{2}}(x)$.

(1). Let

$$k_{\min}^{(0)} = k_{\max}^{(0)} = j, \quad Q_{+}^{(0)}(x) = f^{+}(u_j).$$

(2). Inductively, assume that we have $k_{\min}^{(n-1)}$, $k_{\max}^{(n-1)}$, and $Q_{+}^{(n-1)}(x)$, then compute the nth order divided difference of $f^{+}(u(x))$

$$a^{(n)} = f^{+}\left[u(x_{k_{\min}^{(n-1)}}), \cdots, u(x_{k_{\max}^{(n-1)}+1})\right],$$
$$b^{(n)} = f^{+}\left[u(x_{k_{\min}^{(n-1)}} - 1), \cdots, u(x_{k_{\max}^{(n-1)}})\right].$$

(a). If $|a^{(n)}| \geq |b^{(n)}|$, then

$$c^{(n)} := b^{(n)}, \quad k_{\min}^{(n)} = k_{\min}^{(n-1)} - 1, \quad k_{\max}^{(n)} = k_{\max}^{(n-1)}.$$

(b). If $|a^{(n)}| < |b^{(n)}|$, then

$$c^{(n)} := a^{(n)}, \quad k_{\min}^{(n)} = k_{\min}^{(n-1)}, \quad k_{\max}^{(n)} = k_{\max}^{(n-1)} + 1.$$

(c). $Q_+^{(n)}(x) = Q_+^{(n-1)}(x) + c^{(n)} \prod_{k=k_{\min}^{(n-1)}}^{k_{\max}^{(n-1)}} (x - x_k)$, $n = 1, 2, \cdots, 2r$.

(3). Define $p^+_{j+\frac{1}{2}}(x) = Q_+^{(2m)}(x)$.

The second part is to derive $p^-_{j+\frac{1}{2}}(x)$.

(1). Set $k_{\min}^{(0)} = k_{\max}^{(0)} = j+1$, $\quad Q_-^{(0)}(x) = f^-(u_{j+1})$.

(2). Same as (2) above with f^+ replaced by f^- and $Q_+^{(k)}$ replaced by $Q_-^{(k)}$.

(3). Define $p^-_{j+\frac{1}{2}}(x) = Q_-^{(2m)}(x)$.

Finally, compute $\hat{f}^{\pm}_{j+\frac{1}{2}}$ by (3.116).

Example 3.30 *In the case of* $r = 0$,

$$Q_+^{(0)}(x) = f^+(u_j).$$

Example 3.31 *In the case of $r = 1$, we have*

$$a^{(1)} = f^+[u_j, u_{j+1}] = \frac{f^+(u_{j+1}) - f^+(u_j)}{x_{j+1} - x_j},$$

$$b^{(1)} = f^+[u_{j-1}, u_j] = \frac{f_j^+ - f_{j-1}^+}{x_j - x_{j-1}}.$$

If $|a^{(1)}| \le |b^{(1)}|$, then

$$Q_+^{(1)}(x) = f_j^+(u_j) + \frac{f_{j+1}^+ - f_j^+}{x_{j+1} - x_j}(x - x_j).$$

If $|a^{(1)}| > |b^{(1)}|$, then

$$Q_+^{(1)}(x) = f_j^+(u_j) + \frac{f_j^+ - f_{j-1}^+}{x_j - x_{j-1}}(x - x_j).$$

Example 3.32 *In the case of $r = 2$, if $|a^{(1)}| \le |b^{(1)}|$, then calculate*

$$a^{(2)} = f^+[u_j, u_{j+1}, u_{j+2}] = \frac{f^+[u_{j+1}, u_{j+2}] - f^+[u_j, u_{j+1}]}{x_{j+2} - x_j},$$
$$b^{(2)} = f^+[u_{j-1}, u_j, u_{j+1}] = \frac{f^+[u_j, u_{j+1}] - f^+[u_{j-1}, u_j]}{x_{j+1} - x_{j-1}}.$$

If $|a^{(2)}| < |b^{(2)}|$, then we calculate

$$\begin{aligned} Q_+^{(2)}(x) =& f_j^+(u_j) + \frac{f_{j+1}^+ - f_j^+}{x_{j+1} - x_j}(x - x_j) \\ &+ \frac{f^+[u_{j+1}, u_{j+2}] - f^+[u_j, u_{j+1}]}{x_{j+2} - x_j}(x - x_j)(x - x_{j+1}), \end{aligned}$$

else if $|b^{(2)}| \le |a^{(2)}|$, *then we calculate*

$$\begin{aligned} Q_+^{(2)}(x) =& f_j^+(u_j) + \frac{f_{j+1}^+ - f_j^+}{x_{j+1} - x_j}(x - x_j) \\ &+ \frac{f^+[u_j, u_{j+1}] - f^+[u_{j-1}, u_j]}{x_{j+1} - x_{j-1}}(x - x_j)(x - x_{j+1}). \end{aligned}$$

If $|b^{(1)}| < |a^{(1)}|$, *then calculate*

$$a^{(2)} = f^+[u_{j-1}, u_j, u_{j+1}], \quad b^{(2)} = f^+[u_{j-2}, u_{j-1}, u_j],$$

then compare their absolute values and define $Q_+^{(2)}(x)$ *similarly.*

(*ii*) **ENO finite volume scheme based on the flux reconstruction**

Give the cell average values $\{\bar{u}_j\}$ of the solution u of (1.3).

Lemma 3.37 *If the function $g(x)$ satisfies*

$$f(u(x)) = \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} g(\xi)\ d\xi, \tag{3.117}$$

then

$$f(u(x))_x = \frac{1}{h}\left(g(x+\frac{h}{2}) - g(x-\frac{h}{2})\right). \tag{3.118}$$

The proof is trivial. This lemma implies that the numerical flux $\hat{f}_{j+\frac{1}{2}}$ of a high order accurate conservative scheme should approximate $g(x_{j+\frac{1}{2}})$ to a high order. It is not easy to obtain $g(x)$ directly from (3.117), but the previous reconstruction technique via the primitive function can be applied. The point values of the primitive function of $g(x)$

$$G(x) = \int_{-\infty}^{x} g(\xi)\ d\xi, \tag{3.119}$$

at $x_{j+\frac{1}{2}}$ can be explicitly calculated by

$$G(x_{j+\frac{1}{2}}) = \int_{-\infty}^{x_{j+\frac{1}{2}}} g(\xi)\, d\xi = \sum_{k=-\infty}^{j} \int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} g(\xi)\, d\xi = h \sum_{k=-\infty}^{j} f(\bar{u}_k). \tag{3.120}$$

Noticed that the lower limit $-\infty$ in (3.119) is irrelevant to the final algorithm, and thus can be changed to any fixed mesh point $x_{j_0+\frac{1}{2}}$. Using those point values $\{G(x_{j+\frac{1}{2}})\}$, one can derive the interpolation polynomial $P(x)$ in an ENO fashion satisfying the interpolation conditions $P(x_{j+\frac{1}{2}}) = G(x_{j+\frac{1}{2}})$, and then define $\hat{f}_{j+\frac{1}{2}} := P'(x)\big|_{x=x_{j+\frac{1}{2}}}$. Thanks to (3.120), one has

$$f(\bar{u}_j) = \frac{1}{h}(G(x_{j+\frac{1}{2}}) - G(x_{j-\frac{1}{2}})).$$

On the other hand, the 0th order divided differences of G do not need in the final method. Thus the $(k+1)$th-order divided differences of G can be easily obtained by the kth order divided differences of f so that the summation in (3.120) may be avoided.

Two ENO schemes [68] are given below.

Algorithm (ENO-Roe)

(1). Compute the divided difference table of f, and define $(k = 1, 2, \cdots, r)$

$$G[x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}] = f[u(x_l)],$$

$$G[x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}, \cdots, x_{l+k+\frac{1}{2}}] = \frac{1}{k+1} f[u(x_l), \cdots, u(x_{l+k})].$$

(2). If $\bar{a}_{j+\frac{1}{2}} = \frac{f_{j+1}-f_j}{u_{j+1}-u_j} \geq 0$, then $k_{\min}^{(1)} := j$; else $k_{\min}^{(1)} := j+1$.

(3). $Q^{(1)}(x) = G\left[x_{k_{\min}^{(1)}-\frac{1}{2}}, x_{k_{\min}^{(1)}+\frac{1}{2}}\right]\left(x - x_{k_{\min}^{(1)}-\frac{1}{2}}\right)$.

(4). Inductively, if $k_{\min}^{(l-1)}$ and $Q^{(l-1)}(x)$ are both defined, then let

$$a^{(l)} = G\left[x_{k_{\min}^{(l-1)}-\frac{1}{2}}, \cdots, x_{k_{\min}^{(l-1)}+l-\frac{1}{2}}\right],$$

$$b^{(l)} = G\left[x_{k_{\min}^{(l-1)}-1-\frac{1}{2}}, \cdots, x_{k_{\min}^{(l-1)}+l-1-\frac{1}{2}}\right].$$

(i). If $|a^{(l)}| \geq |b^{(l)}|$, then

$$c^{(l)} = b^{(l)}, \quad k_{\min}^{(l)} = k_{\min}^{(l-1)} - 1,$$

otherwise $c^{(l)} = a^{(l)}, \quad k_{\min}^{(l)} = k_{\min}^{(l-1)}$.

(ii). Define $Q^{(l)}(x) = Q^{(l-1)}(x) + c^{(l)} \prod_{k=k_{\min}^{(l-1)}}^{k_{\min}^{(k-1)}+l-1} (x - x_{k-\frac{1}{2}})$.

(5). Take $\hat{f}_{j+\frac{1}{2}} = \frac{d}{dx} Q_{j+\frac{1}{2}}(x)\big|_{x=x_{j+\frac{1}{2}}}$, where $Q_{j+\frac{1}{2}}(x) := Q^{(r+1)}(x)$.

Algorithm (ENO-LLF) Let $\alpha_{j+\frac{1}{2}} = \max\limits_{u_j \le u \le u_{j+1}} \{|f'(u)|\}$.

(1). Compute the divided difference of u and f. For $k = 1, 2, \cdots, r$, calculate

$$G^{\pm}[x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}] = \frac{1}{2}(f[u(x_l)] \pm \alpha_{j+\frac{1}{2}} u[x_l]),$$

where $l = j-r, \cdots, j+r$ for G^+; and $l = j-r+1, \cdots, j+r+1$ for G^-. Further compute

$$G^{\pm}[x_{l-\frac{1}{2}}, \cdots, x_{l+k+\frac{1}{2}}] = \frac{1}{k+1} \cdot \frac{1}{2}(f[u(x_l), \cdots, u(x_{l+k})] \pm \alpha_{j+\frac{1}{2}} u[x_l, \cdots, x_{l+k}]),$$

where for G^+, $l = j-r, \cdots, j+r-k$, while for G^-, $l = j-r+1, \cdots, j+r-k+1$.

(2). Then for G^+, $k_{\min}^{(1)} = j$, repeat steps (3)-(4) in **Algorithm (ENO-Roe)** to obtain

$$Q^+_{j+\frac{1}{2}}(x) := Q_+^{(r+1)}(x).$$

(3). For $G^-, k_{\min}^{(1)} = j+1$, then repeat steps (3)-(4) in **Algorithm (ENO-Roe)** to get

$$Q^-_{j+\frac{1}{2}}(x) := Q_-^{(r+1)}(x).$$

(4). Finally, take $\hat{f}_{j+\frac{1}{2}} = \frac{d}{dx}Q^+_{j+\frac{1}{2}}(x)\Big|_{x=x_{j+\frac{1}{2}}} + \frac{d}{dx}Q^-_{j+\frac{1}{2}}(x)\Big|_{x=x_{j+\frac{1}{2}}}$.

3.4.6.3 WENO interpolation and reconstruction

This section introduces the WENO interpolation and reconstruction.

(i) **WENO interpolation**

For the sake of convenience, consider a uniform mesh $\{\cdots < x_1 < x_2 < x_3 < \cdots\}$, where $h = x_{j+1} - x_j$ is constant and give the point values of $u(x)$ by $u_j = u(x_j)$. The aim is to find a higher-order accurate approximation of $u(x)$ at the point (e.g. $x_{j+\frac{1}{2}}$) different from the grid point $\{x_j\}$.

Example 3.33 *Given the interpolation data $\{x_j, u_j\}$, find a higher-order accurate approximation of $u(x)$ at the point $x_{j+\frac{1}{2}}$.*

Using the traditional polynomial interpolation may give a polynomial of degree at most 2, denoted by $p_j^{(1)}(x)$, to approximate $u(x)$ according to the stencil $S_1(j) = \{x_{j-2}, x_{j-1}, x_j\}$, i.e.

$$p_j^{(1)}(x_i) = u_i, \quad i = j-2, j-1, j.$$

After that, $u_{j+\frac{1}{2}}^{(1)} := p_j^{(1)}(x_{j+\frac{1}{2}})$ approximates $u(x_{j+\frac{1}{2}})$. By a sim-

ple calculation, one gets

$$u^{(1)}_{j+\frac{1}{2}} = \frac{3}{8}u_{j-2} - \frac{5}{4}u_{j-1} + \frac{15}{8}u_j,$$

satisfying

$$u^{(1)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3),$$

if $u(x)$ *is smooth in* $S_1(j)$.

Similarly, using the stencil $S_2(j) = \{x_{j-1}, x_j, x_{j+1}\}$ *to give a polynomial of degree at most 2, denoted by* $p_j^{(2)}(x)$, *approximating* $u(x)$ *and satisfying*

$$p_j^{(2)}(x_i) = u_i, \quad i = j-1, j, j+1.$$

Thus, $u(x_{j+\frac{1}{2}})$ *may be approximated by*

$$u^{(2)}_{j+\frac{1}{2}} := p_j^{(2)}(x_{j+\frac{1}{2}}) = -\frac{1}{8}u_{j-1} + \frac{3}{4}u_j + \frac{3}{8}u_{j+1},$$

satisfying

$$u^{(2)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3),$$

if $u(x)$ *is smooth in* $S_2(j)$.

Moreover, we can also use the stencil $S_3(j) = \{x_j, x_{j+1}, x_{j+2}\}$ *to give a polynomial of degree at most 2, denoted by* $p_j^{(3)}(x)$, *approximating* $u(x)$ *and satisfying*

$$p_j^{(3)}(x_i) = u_i, \quad i = j, j+1, j+2.$$

Thus $u(x_{j+\frac{1}{2}})$ *may also be approximated by*

$$u^{(3)}_{j+\frac{1}{2}} := p_j^{(3)}(x_{j+\frac{1}{2}}) = \frac{3}{8}u_{j-1} + \frac{3}{4}u_j - \frac{1}{8}u_{j+1},$$

with $u^{(3)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3)$ *whenever* $u(x)$ *is smooth in* $S_3(j)$.

If the function $u(x)$ is globally smooth, all three approximations $u^{(i)}_{j+\frac{1}{2}}$, $i = 1, 2, 3$, obtained above are third order accurate. One could choose one of them based on other considerations, for example, to make the coefficient in the truncation error term $\mathcal{O}(h^3)$ as small as possible. If using them to design finite difference approximations for (1.3), the choice of these stencils $S_i(j)$ would also need to be restricted by the linear stability of the resulting scheme. If the big stencil $S(j) = \{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$ is used, then an interpolation polynomial $p(x)$ of degree at most 4 could be obtained to approximate $u(x)$ in $S(j)$, denoted by $p_j(x)$, satisfying $p_j(x_i) = u_i, i = j \pm 2, j \pm 1, j$, and thus $u(x_{j+\frac{1}{2}})$ may be approximated by

$$u_{j+\frac{1}{2}} := p_j(x_{j+\frac{1}{2}}) = \frac{3}{128}u_{j-2} - \frac{5}{32}u_{j-1} + \frac{45}{64}u_j + \frac{15}{32}u_{j+1} - \frac{5}{128}u_{j+2},$$

which is fifth-order accurate in the sense that

$$u_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^5),$$

if $u(x)$ *is smooth in* $S(j)$.

An important observation is that the above fifth-order accurate approximation $u_{j+\frac{1}{2}}$ *of* $u(x_{j+\frac{1}{2}})$ *based on the large stencil* $S(j)$ *can be written as a linear convex combination of three third-order accurate approximations* $u^{(i)}_{j+\frac{1}{2}}$, $i = 1, 2, 3$, *by*

$$u_{j+\frac{1}{2}} = \gamma_1 u^{(1)}_{j+\frac{1}{2}} + \gamma_2 u^{(2)}_{j+\frac{1}{2}} + \gamma_3 u^{(3)}_{j+\frac{1}{2}},$$

where the constant coefficients $\{\gamma_i\}$ *are*

$$\gamma_1 = \frac{1}{16}, \quad \gamma_2 = \frac{5}{8}, \quad \gamma_3 = \frac{5}{16}.$$

satisfying $\gamma_1 + \gamma_2 + \gamma_3 = 1$, *and are usually referred to as the linear weights in the WENO literature.*

The WENO scheme is to choose the final approximation of $u(x_{j+\frac{1}{2}})$ *as a nonlinear convex combination of the three third approximations* $\{u^{(i)}_{j+\frac{1}{2}}\}$ *by*

$$u^{WENO}_{j+\frac{1}{2}} = \omega_1 u^{(1)}_{j+\frac{1}{2}} + \omega_2 u^{(2)}_{j+\frac{1}{2}} + \omega_3 u^{(3)}_{j+\frac{1}{2}},$$

where the nonlinear weights satisfy $\omega_i \geq 0$ *and* $\omega_1 + \omega_2 + \omega_3 = 1$. *If* $u(x)$ *is smooth in the big stencil* $S(j)$, *then* $\omega_i \approx \gamma_i$; *if* $u(x)$ *has a discontinuity in* S_i, *but is smooth in at least one of the other two stencils, then* $\omega_i \approx 0$. *It can be verified [37] that, as long as* $\omega_i = \gamma_i + \mathcal{O}(h^2)$, *the WENO interpolation* $u^{WENO}_{j+\frac{1}{2}}$ *is fifth-order accurate in the sense that*

$$u^{WENO}_{j+\frac{1}{2}} = u(x_{j+\frac{1}{2}}) + \mathcal{O}(h^5),$$

if $u(x)$ is smooth in the large stencil $S(j)$.

The choices of the nonlinear weights ω_i mentioned in Example 3.33 depend on the *smoothness indicators* β_i, which measures the relative smoothness of $u(x)$ in the stencil $S_i(j)$. The larger β_i, the less smooth $u(x)$ in the stencil $S_i(j)$. Usually, the smoothness indicator β_i is chosen as

$$\beta_i = \sum_{\ell=1}^{r-1} h^{2\ell-1} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^\ell}{dx^\ell} p_j^{(i)}(x) \right)^2 \, dx,$$

where the upper limit in sum $(r-1)$ is equal to the degree of polynomial $p_j^{(i)}(x)$.

Example 3.34 *In Example 3.33, $r = 3$ and $\{\beta_i\}$ may be explic-*

itly given by

$$\beta_1 = \frac{1}{3}(4u_{j-2}^2 - 19u_{j-2}u_{j-1} + 25u_{j-1}^2 + 11u_{j-2}u_j - 31u_{j-1}u_j + 10u_j^2),$$

$$\beta_2 = \frac{1}{3}(4u_{j-1}^2 - 13u_{j-1}u_j + 13u_j^2 + 5u_{j-1}u_{j+1} - 13u_ju_{j+1} + 4u_{j+1}^2),$$

$$\beta_3 = \frac{1}{3}(10u_j^2 - 31u_ju_{j+1} + 25u_{j+1}^2 + 11u_ju_{j+2} - 19u_{j+1}u_{j+2} + 4u_{j+2}^2),$$

so that the nonlinear weights $\{\omega_i\}$ *are defined by*

$$\omega_i = \frac{\tilde{\omega}_i}{\tilde{\omega}_1 + \tilde{\omega}_2 + \tilde{\omega}_3}, \qquad \tilde{\omega}_i = \frac{\gamma_i}{(\varepsilon + \beta_i)^2}, \quad i = 1, 2, 3.$$

Here ε *is a small positive number used to avoid the denominator becoming zero and is typically chosen to be* $\varepsilon = 10^{-6}$.

Now assume that $u(x)$ *is piecewise smooth and only discontinuous at isolated points. For such function* $u(x)$, *if the spatial step*

size h is small enough that the large stencil $S(j)$ does not contain two discontinuous points of $u(x)$, then for each mesh index j, we discuss the WENO accuracy case by case.

(1). If $u(x)$ is smooth in $S(j)$, then three third-order accurate approximations $u^{(i)}_{j+\frac{1}{2}}$, $i = 1, 2, 3$, and the fifth-order accurate approximation $u_{j+\frac{1}{2}}$ are available.

(2). If $u(x)$ has a discontinuous point in the interval $[x_{j-2}, x_{j+2}]$, then there is at least one of three stencils $S_i(j)$, $i = 1, 2, 3$, in which $u(x)$ is smooth. That is to say, at least one of the three third-order accurate approximations $u^{(i)}_{j+\frac{1}{2}}$, $i = 1, 2, 3$, is still valid to approximate $u(x_{j+\frac{1}{2}})$.

We remark that there may be situation in which all small stencils $S_i(j)$, $i = 1, 2, 3$, contain the discontinuity of u. For example, this would be the case if two small stencils $S_2(j)$ and $S_3(j)$ are only considered, and $u(x)$ has a discontinuous point

in the interval (x_j, x_{j+1}). *For this interpolation problem, such situation can be avoided if there is a small stencil which does not include* (x_j, x_{j+1}), *for example,* S_1. *Unfortunately, for solving* (1.3) *with the requirement of conservation, it is usually not possible to avoid such situation. It turns out that this seemingly difficult case is actually not problematic. It is because the interpolation polynomial* $p_j(x)$, *as well as* $p_j^{(2)}(x)$ *and* $p_j^{(3)}(x)$ *in such situation are all essentially monotone in the interval* $[x_j, x_{j+1}]$. *That is, in the interval which contains a discontinuity of* $u(x)$, *no spurious overshoot or undershoot would appear. This fact is demonstrated below. For the sake of convenience, assume that* u *is a step function with a discontinuity in the interval* (x_j, x_{j+1}), *defined by*

$$\begin{aligned} \cdots =& u(x_{j-2}) = u(x_{j-1}) = u(x_j) = 1, \\ 0 =& u(x_{j+1}) = u(x_{j+2}) = \cdots, \end{aligned}$$

and $p(x)$ is a polynomial of degree at most 3 interpolating u in the stencil $S(j)$ satisfying the interpolation conditions $p(x_{j-2}) = p(x_{j-1}) = p(x_j) = 1$ and $p(x_{j+1}) = p(x_{j+2}) = 0$. Thus its derivative $p'(x)$ has at least one zero point in the interval (x_{j-2}, x_{j-1}), (x_{j-1}, x_j) and (x_{j+1}, x_{j+2}). But $p'(x)$ is a a polynomial of degree at most 3, so it has at most three different zero points that are all accounted for in the three intervals above. We thus conclude that $p'(x)$ has no zero point in the interval $[x_j, x_{j+1}]$, so that $p(x)$ is monotone in this interval. If $u(x)$ is a general piecewise smooth function, such result still hold, see [28]. Thus in this case, the interpolation of $u(x_{j+\frac{1}{2}})$ is non-oscillatory, even though it may be not accurate.

Generally, using $S_k(j) = \{x_{j+k-r+1}, x_{j+k-r+2}, \cdots, x_{j+k}\}$ and r point values $\{u_{j+k-r+1}, \cdots, u_{j+k}\}$ may give $p_{r,j}^{(k)}(u_{j+k-r+1}, \cdots, u_{j+k})$ by calculating the value of the interpolated polynomial with de-

gree of $(r-1)$ at $x_{j+\frac{1}{2}}$, for $k = 0, 1, \cdots, r-1$. We expect that

$$u(x_{j+\frac{1}{2}}) = p_{r,j}^{(k)}(u_{j+k-r+1}, \cdots, u_{j+k}) + \mathcal{O}(h^r),$$

wherever $u(x)$ is smooth. The final WENO approximation at $x_{j+\frac{1}{2}}$ denoted by $\hat{u}_{j+\frac{1}{2}}$ can be written as a convex combination of those rth-order approximations

$$\hat{u}_{j+\frac{1}{2}}^{\text{WENO}} = \sum_{k=0}^{r-1} \omega_k p_{r,j}^{(k)}(u_{j+k-r+1}, \cdots, u_{j+k}),$$

where ω_k is nonnegative nonlinear weight satisfying $\sum_{k=0}^{r-1} \omega_k = 1$.

Example 3.35 *Consider the WENO interpolation in the case of $r = 2$. Given the point values of $f^+(u(x))$ at x_j, where $f^{\pm}(u)$ are*

defined in (3.108). *The interpolation may give two second-order approximations of* $f^+(u(x))$ *at* $x_{j+\frac{1}{2}}$ *as follows*

$$\hat{f}^{(1),+}_{j+\frac{1}{2}} = -\frac{1}{2}f^+_{j-1} + \frac{3}{2}f^+_j, \quad \hat{f}^{(2),+}_{j+\frac{1}{2}} = \frac{1}{2}f^+_j + \frac{1}{2}f^+_{j+1}.$$

The third-order accurate WENO interpolation for $f^+(u(x))$ *at* $x_{j+\frac{1}{2}}$ *is*

$$\begin{aligned} \hat{f}^+_{j+\frac{1}{2}} =& \omega_1 \hat{f}^{(1),+}_{j+\frac{1}{2}} + \omega_2 \hat{f}^{(2),+}_{j+\frac{1}{2}}, \\ \omega_i =& \frac{\widetilde{\omega}_i}{\widetilde{\omega}_1 + \widetilde{\omega}_2}, \quad \widetilde{\omega}_i = \frac{\gamma_i}{(\epsilon + \beta_i)^2}, \quad i = 1, 2, \end{aligned}$$

with

$$\beta_1 = (f^+_j - f^+_{j-1})^2, \quad \beta_2 = (f^+_{j+1} - f^+_j)^2, \quad \gamma_1 = \frac{1}{3}, \quad \gamma_2 = \frac{2}{3},$$

Similarly, for the negative flux f^-, we have two approximations at $x_{j+\frac{1}{2}}$

$$\hat{f}^{(2),-}_{j+\frac{1}{2}} = \frac{1}{2}f^-_{j+1} + \frac{1}{2}f^-_j, \quad \hat{f}^{(1),-}_{j+\frac{1}{2}} = \frac{3}{2}f^-_{j+1} - \frac{1}{2}f^-_{j+2},$$

and corresponding smoothness indicators are

$$\beta_1 = (f^-_{j+1} - f^-_j)^2, \quad \beta_2 = (f^-_{j+2} - f^-_{j+1})^2.$$

(*ii*) **WENO reconstruction**

Given the cell averages of $u(x)$, $\{\bar{u}_j\}$, and then computed the point values

$$w(x_{j+\frac{1}{2}}) = \sum_{l=0}^{j} h\bar{u}_l,$$

where $w(x)$ is the primitive function of u defined by

$$w(x) = \int_{x_{-\frac{1}{2}}}^{x} u(\xi)\, d\xi.$$

If using $P_j^{(1)}(x)$ to denote a polynomial with a degree at most 3 interpolated by using the point values $w_{i+\frac{1}{2}}$, $i = j-3, j-2, j-1, j$, then $p_j^{(1)}(x) := \frac{d}{dx} P_j^{(1)}(x)$ is a polynomial with a degree at most 2 approximating $u(x)$ satisfying

$$\frac{1}{h} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} p_j^{(1)}(x)\, dx = \bar{u}_i, \quad i = j-2, j-1, j,$$

that is to say, $p_j^{(1)}(x)$ conserves the mass in each cell of $S_1(j) = \{I_{j-2}, I_{j-1}, I_j\}$. After getting the function $p_j^{(1)}(x)$, we can calcu-

late the approximation of $u(x)$ at some point, for example,

$$u^{(1)}_{j+\frac{1}{2}} := p^{(1)}_j(x_{j+\frac{1}{2}}) = \frac{1}{3}\bar{u}_{j-2} - \frac{7}{6}\bar{u}_{j-1} + \frac{11}{6}\bar{u}_j,$$

approximating $u(x_{j+\frac{1}{2}})$ and satisfying $u^{(1)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3)$ if $u(x)$ is smooth within $S_1(j)$.

Similarly, from the stencil $S_2(j) = \{I_{j-1}, I_j, I_{j+1}\}$, a polynomial $p^{(2)}_j(x)$ may be reconstructed to approximate $u(x)$ so that

$$u^{(2)}_{j+\frac{1}{2}} := p^{(2)}_j(x_{j+\frac{1}{2}}) = -\frac{1}{6}\bar{u}_{j-1} + \frac{5}{6}\bar{u}_j + \frac{1}{3}\bar{u}_{j+1}.$$

If $u(x)$ is smooth in $S_2(j)$, then $u^{(2)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3)$. From stencil $S_3(j) = \{I_j, I_{j+1}, I_{j+2}\}$, a polynomial $p^{(3)}_j(x)$ can be re-

constructed to approximate $u(x)$ so that

$$u^{(3)}_{j+\frac{1}{2}} := p^{(3)}_j(x_{j+\frac{1}{2}}) = \frac{1}{3}\bar{u}_j + \frac{5}{6}\bar{u}_{j+1} - \frac{1}{6}\bar{u}_{j+2},$$

satisfying $u^{(3)}_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^3)$ if $u(x)$ is smooth in $S_3(j)$. On the other hand, from the big stencil $S(j) = \{I_{j-2}, I_{j-1}, I_j, I_{j+1}, I_{j+2}\}$, a high-order polynomial may be directly reconstructed to approximate $u(x)$ so that

$$\begin{aligned} u_{j+\frac{1}{2}} :=& p(x_{j+\frac{1}{2}}) = \frac{1}{30}\bar{u}_{j-2} - \frac{13}{60}\bar{u}_{j-1} + \frac{47}{60}\bar{u}_j + \frac{9}{20}\bar{u}_{j+1} - \frac{1}{20}\bar{u}_{j+2} \\ =& \gamma_1 u^{(1)}_{j+\frac{1}{2}} + \gamma_2 u^{(2)}_{j+\frac{1}{2}} + \gamma_3 u^{(3)}_{j+\frac{1}{2}}, \end{aligned}$$

where

$$\gamma_1 = \frac{1}{10}, \quad \gamma_2 = \frac{3}{5}, \quad \gamma_3 = \frac{3}{10}.$$

If $u(x)$ is smooth in $S(j)$, then $u_{j+\frac{1}{2}} - u(x_{j+\frac{1}{2}}) = \mathcal{O}(h^5)$.

The WENO reconstruction is to choose a convex combination of $u^{(i)}_{j+\frac{1}{2}}$ $(i = 1, 2, 3)$

$$u_{j+\frac{1}{2}} = \omega_1 u^{(1)}_{j+\frac{1}{2}} + \omega_2 u^{(2)}_{j+\frac{1}{2}} + \omega_3 u^{(3)}_{j+\frac{1}{2}},$$

where the definition of the nonlinear weights ω_i are similar to those in the WENO interpolation, with the smoothness indicator β_i given by

$$\begin{aligned}
\beta_1 =& \frac{13}{12}(\bar{u}_{j-2} - 2\bar{u}_{j-1} + \bar{u}_j)^2 + \frac{1}{4}(\bar{u}_{j-2} - 4\bar{u}_{j-1} + 3\bar{u}_j)^2, \\
\beta_2 =& \frac{13}{12}(\bar{u}_{j-1} - 2\bar{u}_j + \bar{u}_{j+1})^2 + \frac{1}{4}(\bar{u}_{j-1} - \bar{u}_{j+1})^2, \\
\beta_3 =& \frac{13}{12}(\bar{u}_j - 2\bar{u}_{j+1} + \bar{u}_{j+2})^2 + \frac{1}{4}(3\bar{u}_j - 4\bar{u}_{j+1} + \bar{u}_{j+2})^2,
\end{aligned}$$

which are different from the previous because the reconstructed polynomials $p_j^{(i)}(x)$ are unlike those interpolated.

Generally, using $S_k = \{I_{j+k-r+1}, I_{j+k-r+2}, \cdots, I_{j+k}\}$ and cell averages $\{\bar{u}_{j+k-r+1}, \cdots, \bar{u}_{j+k}\}$ may gives $p_{r,j}^{(k)}(\bar{u}_{j+k-r+1}, \cdots, \bar{u}_{j+k})$ by evaluating the point value of the reconstructed polynomial with degree of $(r-1)$ at $x_{j+\frac{1}{2}}$, for $k = 0, 1, \cdots, r-1$. We expect

$$u(x_{j+\frac{1}{2}}) = p_{r,j}^{(k)}(\bar{u}_{j+k-r+1}, \cdots, \bar{u}_{j+k}) + \mathcal{O}(h^r),$$

wherever $u(x)$ is smooth. The final WENO approximation at $x_{j+\frac{1}{2}}$ denoted by $\hat{u}_{j+\frac{1}{2}}^{\text{WENO}}$ can be written as a convex combination of those rth-order approximations

$$\hat{u}_{j+\frac{1}{2}}^{\text{WENO}} = \sum_{k=0}^{r-1} \omega_k p_{r,j}^{(k)}(\bar{u}_{j+k-r+1}, \cdots, \bar{u}_{j+k}),$$

where ω_k is nonlinear weight satisfying $\sum_{k=0}^{r-1} \omega_k = 1$ and $\omega_k \geq 0$.

Example 3.36 *A semi-discrete fifth-order accurate WENO scheme for* (1.3) *is implemented as follows [38]*

$$h\frac{d\bar{u}_j}{dt} = -(\hat{f}^+_{j+\frac{1}{2}} - \hat{f}^+_{j-\frac{1}{2}}) - (\hat{f}^-_{j+\frac{1}{2}} - \hat{f}^-_{j-\frac{1}{2}}), \tag{3.121}$$

where

$$\hat{f}^+_{j+\frac{1}{2}} = \frac{1}{12}(-f^+_{j-1} + 7f^+_j + 7f^+_{j+1} - f^+_{j+2} - \varphi_N(\Delta f^+_{j-\frac{3}{2}}, \Delta f^+_{j-\frac{1}{2}}, \Delta f^+_{j+\frac{1}{2}}, \Delta f^+_{j+\frac{3}{2}}),$$
$$\hat{f}^-_{j+\frac{1}{2}} = \frac{1}{12}(-f^-_{j-1} + 7f^-_j + 7f^-_{j+1} - f^-_{j+2} - \varphi_N(\Delta f^-_{j+\frac{5}{2}}, \Delta f^-_{j+\frac{3}{2}}, \Delta f^-_{j+\frac{1}{2}}, \Delta f^-_{j-\frac{1}{2}}),$$

and

$$
\begin{aligned}
\varphi_N(a,b,c,d) :=& \frac{1}{3}\omega_0(a-2b+c)+\frac{1}{6}(\omega_2-0.5)(b-2c+d),\\
\omega_0 =& \frac{\alpha_0}{\alpha_0+\alpha_1+\alpha_2}, \qquad \omega_2=\frac{\alpha_2}{\alpha_0+\alpha_1+\alpha_2},\\
\alpha_0 =& \frac{1}{(\epsilon+\beta_0)^2}, \quad \alpha_1=\frac{6}{(\epsilon+\beta_1)^2}, \quad \alpha_2=\frac{3}{(\epsilon+\beta_2)^2},\\
\beta_0 =& 13(a-b)^2+3(a-3b)^2, \quad \beta_1=13(b-c)^2+3(b+b)^2,\\
\beta_2 =& 13(c-d)^2+3(3a-d)^2,
\end{aligned}
$$

here ϵ is a small positive number used to avoid the denominator becoming zero.

The time derivatives in (3.121) may be discretized by using a high-order accurate, explicit, Runge-Kutta method, see an example in Section 4.3.

4 RKDG方法

DG方法最初是Reed和Hill于1973年就定常中子输运方程$\sigma u + \mathrm{div}(\boldsymbol{a}u) = 0$提出的, [59], 其中$\sigma$是一个实数, $\boldsymbol{a}$是个常数向量. A major development of the DG method is carried out by Cockburn, Shu, and their co-workers in a series of papers [8, 10, 11, 12, 13], in which a framework is established to easily solve unsteady quasi-linear hyperbolic conservation laws by using explicit, nonlinearly stable high-order accurate Runge-Kutta time discretization and DG discretization in space with exact or approximate Riemann solvers as interface fluxes and total variation bounded (TVB) limiter to achieve nonoscillatory properties for strong shocks. These schemes are termed as *RKDG methods*. 读者可以参阅综述文章[9].

4.1 Galerkin方法

考虑一个抽象问题, 它在Hilbert空间V里的弱形式如下

$$\text{寻找}u \in V \text{ 使得}\forall v \in V,\ a(u,v) = f(v), \tag{4.1}$$

式中$a(\cdot,\cdot)$是一个双线性形式, f是V上的一个有界的线性泛函. 稍后将指定$a(\cdot,\cdot)$的确切要求。

Galerkin方法是选取一个k子空间$V_h \subset V$, 求解投影问题

$$\text{寻找}u_h \in V_h \text{ 使得}\forall v_h \in V_h,\ a(u_h,v_h) = f(v_h). \tag{4.2}$$

它被称为*Galerkin 方程*. 明显地, the Galerkin equation has remained unchanged and only the spaces have changed, in comparison to the original. The key is that reducing the problem to a finite-dimensional vector subspace allows us to numerically compute u_h as a finite linear combination of the basis vectors in

V_h. An important property of the Galerkin method is that the error $\epsilon_h = u - u_h$ is orthogonal to the chosen subspaces, where u is the solution of the original problem (4.1) and u_h denotes the solution of the Galerkin equation (4.2). Since $V_h \subset V$, we can use v_h as a test vector in the original equation (4.1) and then subtract (4.1) to (4.2) to get the Galerkin orthogonality relation for the error

$$a(\epsilon_h, v_h) = a(u, v_h) - a(u_h, v_h) = f(v_h) - f(v_h) = 0.$$

The analysis of the Galerkin method proceeds in two steps: (1) show that the Galerkin equation is a well-posed problem in the sense of Hadamard and thus admits a unique solution, and (2) study the quality of approximation of the Galerkin solution u_h. The analysis does mostly rely on the boundedness and ellipticity of the bilinear form $a(\cdot, \cdot)$.

Theorem 4.1 (Lax-Milgram theorem) *Let V be a Hilbert space and $a(\cdot,\cdot)$ a bilinear form on V, which is $a(u,v)$: (1) bounded: $a(u,v) \leq c_1\|u\|\,\|v\|$ for all $u,v \in V$, and (2) elliptic or coercive: $a(u,u) \geq c_0\|u\|^2$ for all $u \in V$, where c_0 and c_1 are some positive constants. Then, for any $f \in V'$, being the dual of V, there is a unique solution $u \in V$ to the equation*

$$a(u,v) = f(v),$$

and it holds

$$\|u\| \leq \frac{1}{c_0}\|f\|_{V'}.$$

The Lax-Milgram theorem tells us that the boundedness and ellipticity of the bilinear form $a(\cdot,\cdot)$ imply the well-posedness of the original problem in weak formulation. All norms in the following will be norms for which the above inequalities hold. They are

often called as the energy norm. Since $V_h \subset V$, boundedness and ellipticity of the bilinear form may apply to V_h. Therefore, the well-posedness of the Galerkin problem (4.2) is actually inherited from the well-posedness of the original problem (4.1).

Lemma 4.2 (Quasi-best approximation (Céa's lemma)) *The error $\epsilon_h = u - u_h$ between the original and the Galerkin solutions admits the estimate*

$$\|u - u_h\| \leq \frac{c_1}{c_0} \inf_{v_h \in V_h} \|u - v_h\|.$$

This means that up to the constant c_1/c_0, the Galerkin solution u_h is as close to the original solution u as any other vector in V_h. Particularly, it will be sufficient to study approximation by spaces V_h, completely forgetting about the equation being solved.

4.2 Continuous Galerkin finite element method

In the finite element method, the domain is divided into elements (intervals in one dimension) and the approximate solution is (usually) sought in the space of piecewise polynomials V_h.

Consider the convection equation (2.1), with $\Omega = [0, 1]$ and $u(0, t) = u(1, t)$. Continuous FE method of (2.1) is as follows: Seek u_h in V_h such that for all $v \in V_h$ such that

$$\int_0^1 (\frac{\partial u_h}{\partial t} + a\frac{\partial u_h}{\partial x})v \; dx = \int_0^1 \frac{\partial u_h}{\partial t} v \; dx + \int_0^1 a\frac{\partial u_h}{\partial x} v \; dx$$
$$= \int_0^1 \frac{\partial u_h}{\partial t} v \; dx + (a u_h v)_0^1 - \int_0^1 a u_h \frac{\partial v}{\partial x} \; dx = 0,$$

which leads to a semi-discrete system $\boldsymbol{M}(\boldsymbol{u}_h)_t + \boldsymbol{K}\boldsymbol{u} = 0$, with element-wise matrices $\boldsymbol{M}$ and $\boldsymbol{K}$. The matrix $\boldsymbol{M}^{-1}$ is dense, explicit methods for $\boldsymbol{u}_t = -\boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{u}$ not practical. Also unclear

how to stabilize it by upwinding (but other techniques exist such as streamline upwind/Petrov-Galerkin).

The Petrov-Galerkin method is a Galerkin type method used to obtain approximate solutions of PDEs which contain terms with odd order. An example of differential equation containing a term with odd order is as follows

$$a(x)\frac{\mathrm{d}u}{\mathrm{d}x} + b(x)\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = f(x), \quad x \in (0, L),$$
$$u(0) = u_o, \ \left.\frac{\mathrm{d}u}{\mathrm{d}x}\right|_{x=L} = u'_L.$$

In these type of problems a weak formulation with similar function space for the test and solution functions is not possible. Hence the method is used in case that the test and solution functions belong to different function spaces. If a test function $v(x)$ is used to obtain the weak form, the Galerkin formulation is given

by

$$\int_0^L a(x)v(x)\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}x - \int_0^L b(x)\frac{\mathrm{d}v}{\mathrm{d}x}\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}x + \left[b(x)v\frac{\mathrm{d}u}{\mathrm{d}x}\right]_0^L = \int_0^L v(x)f(x)\,\mathrm{d}x,$$

in which the term with even order (2nd term at the left hand side) is now symmetric, as both the test and solution functions may have same order of differentiation and they both belong to H_0^1, but the first term at the LHS cannot be made in this way as the solution space H_0^1 and test function space L^2 are different.

4.3 Discontinuous Galerkin finite element method

Discontinuous FE methods do not enforce continuity, and allows "jumps" between elements.

The Galerkin formulation of (2.1) for single element $e_1 = [0, h]$ is stated as follows: for all $v \in P^k(e_1)$, the space of polyno-

mials on e_1 of degree at most k, find $u_h \in P^k(e_1)$ such that

$$\int_0^h \left(\frac{\partial u_h}{\partial t} + a\frac{\partial u_h}{\partial x}\right) v \; dx = \int_0^h \frac{\partial u_h}{\partial t} v \; dx + \int_0^h a\frac{\partial u_h}{\partial x} v \; dx$$
$$= \int_0^h \frac{\partial u_h}{\partial t} v \; dx + \hat{f}\big(u_h(h^-), u_h(h^+)\big) v(h)$$
$$- \hat{f}\big(u_h(0^-), u_h(0^+)\big) v(0) - \int_0^h a u_h \frac{\partial v}{\partial x} \; dx = 0. \qquad (4.3)$$

From the conservation and stability (upwinding) considerations, the numerical flux $\hat{f}(u_L, u_R)$ is taken as a single valued monotone numerical flux satisfying $\hat{f}(u, u) = f(u)$ (consistency), $\hat{f}(\uparrow, \downarrow)$ (monotonicity), and Lipschitz continuous with respect to both arguments. An example is

$$\hat{f} = \frac{a}{2}(u_L + u_R) - \frac{|a|}{2}(u_R - u_L). \qquad (4.4)$$

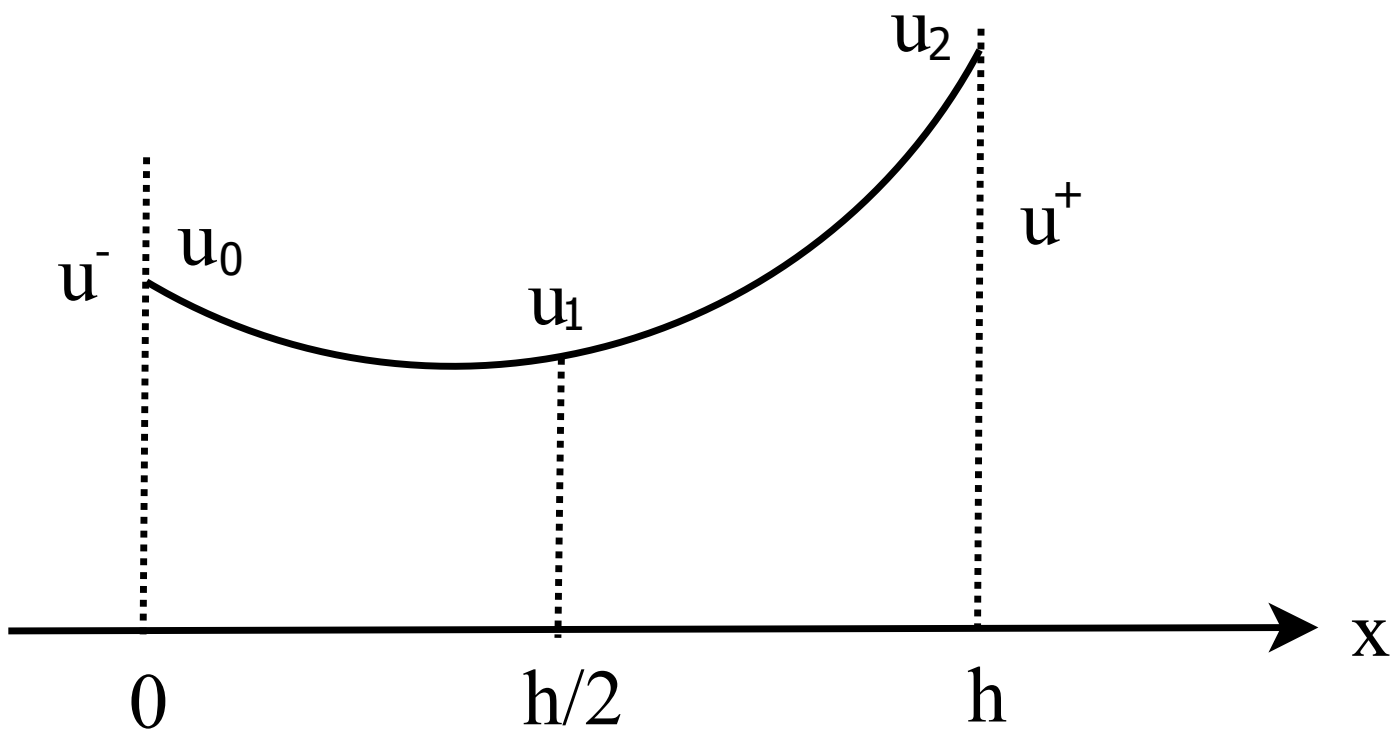


Figure 14: Discontinuous FE solution within cell $[0, h]$ when $k = 2$.

With $a = 1$, the DG formulation (4.3)-(4.4) reduces to

$$\frac{d}{dt}\int_0^h u_h v\, dx + u_h(h^-)v(h) - u_h(0^-)v(0) - \int_0^h u_h \frac{\partial v}{\partial x}\, dx = 0. \tag{4.5}$$

If taking $u_h(x,t) = \sum_{j=0}^{k} u_j(t)\phi_j(x)$ and $v(x) = \phi_i(x)$, $i = 1, 2, \cdots, k$, where $\{\phi_i(x)\}$ are Lagrange polynomials with a degree of k, then (4.5) leads to a system of ordinary differential equations (ODE)

$$M\boldsymbol{u}_t + (-u^-, 0, \cdots, 0, u_k)^T - K\boldsymbol{u} = 0, \tag{4.6}$$

where $\boldsymbol{M} = \left(\int_0^h \phi_i\phi_j dx\right)_{(k+1)\times(k+1)}$ and $\boldsymbol{K} = \left(\int_0^h \phi_i'\phi_j dx\right)_{(k+1)\times(k+1)}$.

If $k = 2$, see Fig. 14, $u_h(x,t) = \sum_{i=0}^{2} u_i(t)\phi_i(x)$, where $\phi_i(x)$

are Lagrange polynomials with a degree of 2 defined by

$$\phi_0(x) = \frac{2}{h^2}(x-\frac{h}{2})(x-h),\ \phi_1(x) = -\frac{4}{h^2}x(x-h),\ \phi_2(x) = \frac{2}{h^2}x(x-\frac{h}{2}).$$

The system (4.6) for $\boldsymbol{u} = (u_0, u_1, u_2)$ becomes

$$\begin{aligned}
\boldsymbol{u}_t =& \boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{u} - \boldsymbol{M}^{-1}(-u^-, 0, u_2)^T \\
=& \boldsymbol{M}^{-1}\big(\boldsymbol{K}\boldsymbol{u} - (0, 0, u_2)^T\big) + \boldsymbol{M}^{-1}(1, 0, 0)^T u^- \\
=& \boldsymbol{M}^{-1}\widehat{\boldsymbol{K}}\boldsymbol{u} + \boldsymbol{M}^{-1}(1, 0, 0)^T u^- \\
=& \frac{1}{h}\begin{pmatrix} -6 & -4 & 1 \\ 2.5 & 0 & -1 \\ -4 & 4 & -3 \end{pmatrix}\begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix} + \frac{1}{h}\begin{pmatrix} 9 \\ -1.5 \\ 3 \end{pmatrix} u^-,
\end{aligned}$$

where

$$\boldsymbol{M} = \frac{h}{30}\begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix}, \quad \boldsymbol{M}^{-1} = \frac{1}{4h}\begin{pmatrix} 36 & -6 & 12 \\ -6 & 9 & -6 \\ 12 & -6 & 36 \end{pmatrix},$$

and

$$\widehat{\boldsymbol{K}} = \boldsymbol{K} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} -3 & -4 & 1 \\ 4 & 0 & -4 \\ -1 & 4 & -3 \end{pmatrix}.$$

The above DG formula is element-wise and local finite difference type stencil, stabilized, and upwind through u^-, but calculation of the matrices $\boldsymbol{M}$ and $\boldsymbol{K}$ in (4.6) as well as $\boldsymbol{M}^{-1}$ is more complicated and time-consuming. To simplify the derivation of the DG formulation, we usually choose the Legendre polynomials $P_l(\xi)$ as local basis functions and exploit sufficiently their

L^2-orthogonality

$$\int_{-1}^{1} P_l(\xi) P_{l'}(\xi)\, d\xi = \frac{2}{2l+1}\delta_{l,l'}, \quad l \le l', \tag{4.7}$$

to obtain a diagonal mass matrix $\boldsymbol{M}$. The first few Legendre polynomials are

$$P_0(\xi) = 1, \quad P_1(\xi) = \xi, \quad P_2(\xi) = \xi^2 - 1/3, \quad P_3(\xi) = \xi^3 - 3\xi/5, \quad \cdots.$$

Consider the DG method for the quasilinear hyperbolic equation in conservative (or divergence) form (1.3). Triangulate the domain Ω into elements $e_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \in T_h$ and denote the length of e_j by h_j. The DG method is to seek solution u_h in space of element-wise polynomials

$$\mathcal{V}_h = \left\{ v \in L^2(\Omega) : v|_{e_j} \in P^k(e_j),\ \forall e_j \in T_h \right\},$$

where $P^k(e_j)$ detentes the space of polynomials with degrees less than or equal to k over the element e_j indexed by j.

Multiply with a test function $v \in \mathcal{V}_h$ and integrate in space over element e_j and integrate by parts

$$\begin{aligned}\int_{e_j} \big(u_t + f(u)_x\big) v \, dx = \int_{e_j} u_t v \, dx - \int_{e_j} f(u) v_x \, dx \\ + \hat{f}\big(u(x_{j+\frac{1}{2}} - 0), u(x_{j+\frac{1}{2}} + 0)\big) v(x_{j+\frac{1}{2}} - 0) \\ - \hat{f}\big(u(x_{j-\frac{1}{2}} - 0), u(x_{j-\frac{1}{2}} + 0)\big) v(x_{j-\frac{1}{2}} + 0) = 0,\end{aligned}$$

with the numerical flux $\hat{f}(a, b)$. Therefore, the Galerkin formulation for (1.3) may be stated as follows: for all $v \in P^k(e_j)$, find

$u_h \in P^k(e_j)$ such that

$$\begin{aligned}
&\frac{d}{dt}\int_{e_j}(u_h)_t v\ dx - \int_{e_j} f(u_h)v_x\ dx \\
&+ \hat{f}\big(u_h(x_{j+\frac{1}{2}}-0), u_h(x_{j+\frac{1}{2}}+0)\big)v(x_{j+\frac{1}{2}}-0) \\
&- \hat{f}\big(u_h(x_{j-\frac{1}{2}}-0), u_h(x_{j-\frac{1}{2}}+0)\big)v(x_{j-\frac{1}{2}}+0) = 0.
\end{aligned}$$

For $x \in \mathbb{R}$, our approximate solution u_h may be expressed by

$$u_h(x,t) = \sum_{\ell=0}^{k} u_j^{(\ell)}(t)\phi_j^{(\ell)}(x) =: u_j(x,t), \quad \text{if } x \in e_j, \tag{4.8}$$

where

$$\phi_j^{(\ell)}(x) = P_\ell(\xi), \quad \xi = \frac{2(x-x_j)}{h_j}, \tag{4.9}$$

and the degrees of freedom $u_j^{(\ell)}(t)$ are the moments defined by

$$u_j^{(\ell)}(t) = \frac{1}{a_\ell} \int_{e_j} u_h(x,t)\phi_j^{(\ell)}(x)\ dx, \quad \ell = 0, 1, \cdots, k,$$

where $a_\ell = \int_{e_j} \big(\phi_j^{(\ell)}(x)\big)^2\ dx$.

Substituting (4.8) into the above formulation and taking v as $\phi_j^{(\ell')}(x)$ gives

$$\frac{d}{dt}u_j^{(\ell')}(t) + \frac{1}{a_{\ell'}}\Big(- \int_{e_j} f(u_h)\frac{d}{dx}\phi_j^{(\ell')}(x)\ dx$$
$$+ \hat{f}\big(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}\big)\phi_j^{(\ell')}(x_{j+\frac{1}{2}} - 0) - \hat{f}\big(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}\big)\phi_j^{(\ell')}(x_{j-\frac{1}{2}} - 0)\Big) = 0, \tag{4.10}$$

for $\ell' = 0, 1, \cdots, k$, where $u^\pm_{j+\frac{1}{2}} = u_h(x_{j+\frac{1}{2}} \pm 0)$. When $k = 0$,

(4.10) reduces to the finite volume method

$$h_j \frac{d}{dt} u_j^{(0)} + \hat{f}\left(u_j^{(0)}, u_{j+1}^{(0)}\right) - \hat{f}\left(u_{j-1}^{(0)}, u_j^{(0)}\right) = 0,$$

which is also most useful for solving hyperbolic conservations laws in their divergence form.

The semi discrete DG scheme (4.10) is cast into the form

$$\boldsymbol{u}_t = \boldsymbol{L}(\boldsymbol{u}),$$

which may be discretized in time by using an explicit, nonlinearly stable high-order accurate Runge-Kutta method, e.g. a third-

order version

$$\begin{aligned}
\boldsymbol{u}^{(1)} =& \boldsymbol{u}^n + \tau \boldsymbol{L}(\boldsymbol{u}^n), \\
\boldsymbol{u}^{(2)} =& \frac{3}{4}\boldsymbol{u}^n + \frac{1}{4}\boldsymbol{u}^{(1)} + \frac{1}{4}\tau \boldsymbol{L}(\boldsymbol{u}^{(1)}), \\
\boldsymbol{u}^{n+1} =& \frac{1}{3}\boldsymbol{u}^n + \frac{2}{3}\boldsymbol{u}^{(2)} + \frac{2}{3}\tau \boldsymbol{L}(\boldsymbol{u}^{(2)}).
\end{aligned}$$

Due to the discontinuous nature of the solution and the test function space and the explicit time marching, the fully-discrete RKDG methods, as described above, has the additional advantages of local communications and easy h- or p-adaptivity, besides easily handling complicated geometry and arbitrary triangulations.

The above RKDG methods can numerically solve (1.3) without further modification, if their solutions are either smooth or have weak shock waves and other weak discontinuities. However,

it will generate significant oscillations and even nonlinear instability if those solutions contain strong discontinuities. To avoid such difficulties, we have to borrow a technique of a slope limiter from the finite volume methodology and use it after each Runge-Kutta inner stage (or after the complete Runge-Kutta time step) to limit the RKDG solution. Such limiting procedure must be designed to control spurious oscillations and at the same time maintain accuracy in the smooth regions of solution in a robust way, which is usually difficult to achieve. Up to now, many such limiters exist in the literature, for example, the minmod-type limiters etc. The TVB minmod limiter may be implemented on $u^{\pm}_{j+\frac{1}{2}}$ in (4.10) as follows.

Let

$$u^{-}_{j+\frac{1}{2}} = u^{(0)}_{j} + \tilde{u}_j, \quad u^{+}_{j+\frac{1}{2}} = u^{(0)}_{j} + \tilde{\tilde{u}}_j.$$

These are modified by either the standard minmod limiter

$$\tilde{u}_j^{\text{mod}} = m(\tilde{u}_j, \Delta_x^+ u_j^{(0)}, \Delta_x^- u_j^{(0)}), \quad \tilde{\tilde{u}}_j^{\text{mod}} = m(\tilde{\tilde{u}}_j, \Delta_x^+ u_j^{(0)}, \Delta_x^- u_j^{(0)}),$$

where $m(a_1, a_2, a_3)$ is defined by

$$m(a_1, \cdots, a_m) = \begin{cases} s \cdot \min_{1 \le i \le m}\{|a_i|\}, & \text{if } \operatorname{sign}(a_1) = \cdots = \operatorname{sign}(a_m) = s, \\ 0, & \text{otherwise,} \end{cases}$$

or by the TVB modified minmod function

$$\tilde{m}(a_1, \cdots, a_m) = \begin{cases} a_1, & \text{if } |a_1| \le Mh^2, \\ m(a_1, \cdots, a_m), & \text{otherwise,} \end{cases}$$

where $M > 0$ is a constant dependent on the solution of problem.

The ENO and WENO methods have been developed and applied to finite volume and difference methods to successfully achieve both sharp and ENO shock transitions and uniform

high order accuracy. The ENO/WENO methodology is more robust than the slope limiter methodology, especially for high order schemes. Thus it would be natural to try to use it as the limiter for the DG methods. An attempt has been made via the following procedure [56, 97]:

(1) identify the "troubled" cells, namely, those cells which might need the limiting procedure.

(2) replace the RKDG solution polynomials in those "troubled" cells with WENO reconstructed polynomials which maintain the original cell averages (conservation) and the accuracy, but have less numerical oscillation.

The readers are referred to Section 5.3.

Before ending this section, we present the numerical results for the 1D Euler equations in gas dynamics obtained by the P^k

based RKDG methods in Fig. 15, where a comparison of the computed solutions ("$\circ$") with the exact solutions at $t = 1.8$ [75, Page 393] here. They well demonstrates the performance of the higher-order accurate RKDG methods.

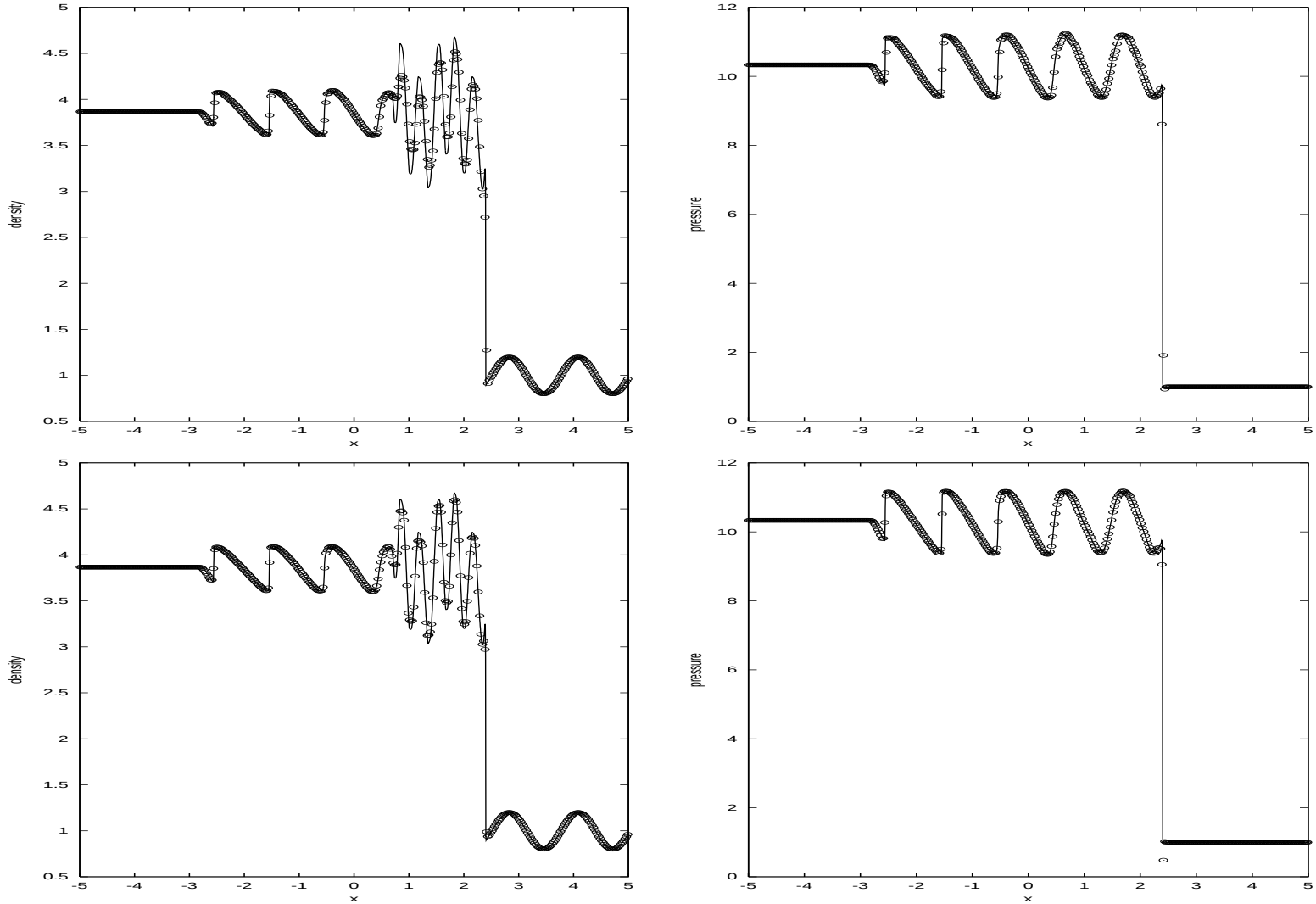


Figure 15: Comparison of the computed solutions ("$\circ$") with the exact solutions at $t = 1.8$. The solutions are obtained by P^k based RKDG method with 400 grid cells, $k = 4$ (top) and $k = 3$ (bottom), respec-

5 Extension to quasilinear system of conservation laws

Some schemes may be extended to quasilinear hyperbolic systems of conservation laws

$$\frac{\partial \boldsymbol{U}}{\partial t} + \frac{\partial \boldsymbol{F}(\boldsymbol{U})}{\partial x} = 0, \quad \boldsymbol{U} \in \mathbb{R}^m, \tag{5.1}$$

component-wisely. But not all schemes are valid in this way, e.g. such as the upwind methods. To approximate a hyperbolic system with so-called upwind differences, we must first establish which way the wind blows. More precisely, we must determine in which direction each of a variety of signals moves through the computational grid. For this purpose, a physical model of the interaction between computational cells is needed. Now there are two models for this purpose [88, 31]. In the first model neighboring

cells interact through discrete, finite amplitude waves. The nature, propagation speed and amplitude of these waves are found by solving, exactly or approximately, local Riemann problem at each cell interface. The numerical technique of distinguishing between the influence of the forward- and the backward-moving waves is called **flux-difference splitting**, see e.g. [61, 20]. It is

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{U}_R) - \boldsymbol{F}(\boldsymbol{U}_L) =& \boldsymbol{A}(\boldsymbol{U}_L, \boldsymbol{U}_R)(\boldsymbol{U}_R - \boldsymbol{U}_L) = \boldsymbol{R}\Lambda\boldsymbol{L}(\boldsymbol{U}_R - \boldsymbol{U}_L) \\ =& \boldsymbol{R}\Lambda^+\boldsymbol{\alpha}(\boldsymbol{U}_L, \boldsymbol{U}_R) + \boldsymbol{R}\Lambda^-\boldsymbol{\alpha}(\boldsymbol{U}_L, \boldsymbol{U}_R) \\ =:& [\boldsymbol{F}(\boldsymbol{U}_R) - \boldsymbol{F}(\boldsymbol{U}_L)]^+ + [\boldsymbol{F}(\boldsymbol{U}_R) - \boldsymbol{F}(\boldsymbol{U}_L)]^-.\end{aligned}$$

Based on this splitting, we may have the upwind type scheme such as

$$\begin{aligned}
\boldsymbol{U}_j^{n+1} =& \boldsymbol{U}_j^n - \frac{\lambda}{2}\left(\boldsymbol{F}(\boldsymbol{U}_{j+1}^n) - \boldsymbol{F}(\boldsymbol{U}_{j-1}^n)\right) \\
&+ \frac{\lambda|\boldsymbol{A}(\boldsymbol{U}_j^n, \boldsymbol{U}_{j+1}^n)|}{2}(\boldsymbol{U}_{j+1}^n - \boldsymbol{U}_j^n) - \frac{\lambda|\boldsymbol{A}(\boldsymbol{U}_{j-1}^n, \boldsymbol{U}_j^n)|}{2}(\boldsymbol{U}_j^n - \boldsymbol{U}_{j-1}^n) \\
=& \boldsymbol{U}_j^n + \frac{\lambda(|\boldsymbol{A}| - \boldsymbol{A})(\boldsymbol{U}_j^n, \boldsymbol{U}_{j+1}^n)}{2}(\boldsymbol{U}_{j+1}^n - \boldsymbol{U}_j^n) \\
&- \frac{\lambda(|\boldsymbol{A}| + \boldsymbol{A})(\boldsymbol{U}_{j-1}^n, \boldsymbol{U}_j^n)}{2}(\boldsymbol{U}_j^n - \boldsymbol{U}_{j-1}^n) \\
=& \boldsymbol{U}_j^n - \lambda\sum_{i=1}^m \boldsymbol{R}^{(i)}\lambda^{(i),-}\alpha^{(i)}(\boldsymbol{U}_j^n, \boldsymbol{U}_{j+1}^n) - \lambda\sum_{i=1}^m \boldsymbol{R}^{(i)}\lambda^{(i),+}\alpha^{(i)}(\boldsymbol{U}_{j-1}^n, \boldsymbol{U}_j^n) \\
=& \boldsymbol{U}_j^n - \lambda[\boldsymbol{F}(\boldsymbol{U}_{j+1}^n) - \boldsymbol{F}(\boldsymbol{U}_j^n)]^- - \lambda[\boldsymbol{F}(\boldsymbol{U}_j^n) - \boldsymbol{F}(\boldsymbol{U}_{j-1}^n)]^+.
\end{aligned}$$

In the second model, the interaction of neighboring cells is accomplished through mixing of pseudo-particles that move in and out of each cell according to a given velocity distribution. We may call this the *Boltzmann approach*. The numerical technique

of distinguishing between the influence of the forward- and the backward-moving particles is called **flux-vector splitting** or simply **flux-splitting**, see e.g. [65, 88, 69]. The detail is as follows

$$\boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{F}^+(\boldsymbol{U}) + \boldsymbol{F}^-(\boldsymbol{U}), \quad \lambda\left(\frac{\partial \boldsymbol{F}^+}{\partial \boldsymbol{U}}\right) \geq 0, \ \lambda\left(\frac{\partial \boldsymbol{F}^-}{\partial \boldsymbol{U}}\right) \leq 0.$$

Based on such flux splitting, we may easily have the upwind type scheme such as

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n - \lambda\big(\boldsymbol{F}^-(\boldsymbol{U}_{j+1}^n) - \boldsymbol{F}^-(\boldsymbol{U}_j^n)\big) - \lambda\big(\boldsymbol{F}^+(\boldsymbol{U}_j^n) - \boldsymbol{F}^+(\boldsymbol{U}_{j-1}^n)\big).$$

5.1 Some flux-difference splitting type schemes

Some examples on flux-difference splitting type scheme are given below.

(i): Roe's scheme [61]

$$\hat{\boldsymbol{F}}(\boldsymbol{U}_j, \boldsymbol{U}_{j+1}) = \frac{\boldsymbol{F}(\boldsymbol{U}_j) + \boldsymbol{F}(\boldsymbol{U}_{j+1})}{2} - \frac{1}{2}|\hat{\boldsymbol{A}}_{j+1/2}|(\boldsymbol{U}_{j+1} - \boldsymbol{U}_j), \tag{5.2}$$

where $|\hat{\boldsymbol{A}}|$ is defined by: $|\hat{\boldsymbol{A}}| = \boldsymbol{R}|\hat{\boldsymbol{\Lambda}}|\boldsymbol{R}^{-1}|$, $|\hat{\boldsymbol{\Lambda}}| = \mathrm{diag}\{|\hat{\lambda}_1|, \cdots, |\hat{\lambda}_m|\}$, $\boldsymbol{R}$ is the right eigenvector matrix of $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}\{\hat{\lambda}_1, \cdots, \hat{\lambda}_m\}$, that is to say, $\boldsymbol{R}^{-1}\hat{\boldsymbol{A}}\boldsymbol{R} = \hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{A}}_{j+1/2}(\boldsymbol{U}_{j+1} - \boldsymbol{U}_j) = \boldsymbol{F}(\boldsymbol{U}_{j+1}) - \boldsymbol{F}(\boldsymbol{U}_j)$.

(ii): Huang's scheme [34]

$$\hat{\boldsymbol{F}}_{j+\frac{1}{2}} = \frac{\boldsymbol{F}(\boldsymbol{U}_j) + \boldsymbol{F}(\boldsymbol{U}_{j+1})}{2} - \frac{1}{2}\mathrm{sign}\left(\boldsymbol{A}\left(\frac{\boldsymbol{U}_j + \boldsymbol{U}_{j+1}}{2}\right)\right)(\boldsymbol{F}(\boldsymbol{U}_{j+1}) - \boldsymbol{F}(\boldsymbol{U}_j)),$$

where

$$\mathrm{sign}(\boldsymbol{A}) = \boldsymbol{R}\mathrm{sign}(\boldsymbol{\Lambda})\boldsymbol{R}^{-1}, \quad \mathrm{sign}(\boldsymbol{\Lambda}) = \mathrm{diag}\{\mathrm{sign}(\lambda_1), \cdots, \mathrm{sign}(\lambda_m)\}. \tag{5.3}$$

(iii): Engquist-Osher scheme [20]

$$\hat{\boldsymbol{F}}(\boldsymbol{U}_j, \boldsymbol{U}_{j+1}) = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}_j) + \boldsymbol{F}(\boldsymbol{U}_{j+1})) - \frac{1}{2}\int_{\boldsymbol{U}_j}^{\boldsymbol{U}_{j+1}} |\boldsymbol{A}(\boldsymbol{U})|\, d\boldsymbol{U}, \tag{5.4}$$

where the integral path is $\Gamma_j = \cup_{i=1}^{m} \Gamma_j^{(i)}$, and the curve $\Gamma_j^{(i)}$ is a line segment parallel to $\boldsymbol{R}^{(i)}$, the ith-right eigenvector of $\boldsymbol{A}$, that is to say, $\Gamma_j^{(i)}$ is defined by

$$\Gamma_j^{(i)}: \quad \begin{cases} \dfrac{d\boldsymbol{U}^{(i)}}{ds} = \boldsymbol{R}^{(i)}(\boldsymbol{U}^{(i)}), \quad 0 \le s \le s_j^{(i)} \text{ or } 0 \ge s \ge s_j^{(i)}, \\ \boldsymbol{U}^{(i)}(s=0) = \boldsymbol{U}^{(i+1)}(s_j^{(i+1)}), \end{cases}$$

here $i = m, \cdots, 1$, $\boldsymbol{U}^{(m+1)}(s_j^{(m+1)}) = \boldsymbol{U}_j$, $\boldsymbol{U}^{(1)}(s_j^{(1)}) = \boldsymbol{U}_{j+1}$. Fig. 16 shows the integral path Γ_j for the Engquist-Osher scheme. For

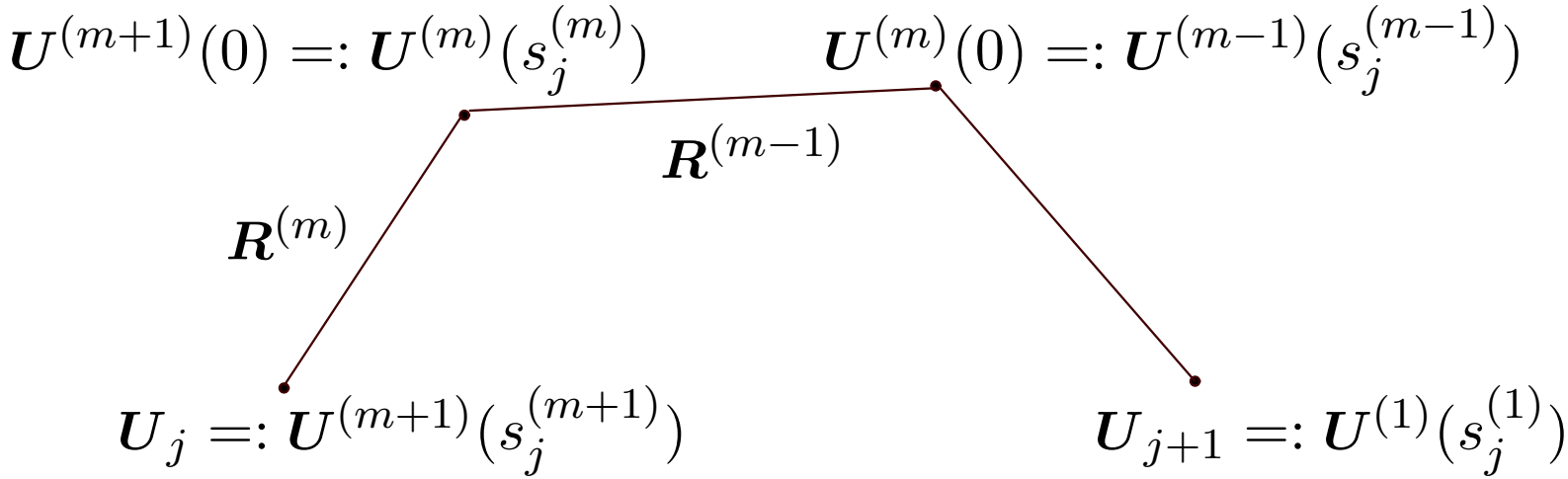


Figure 16: The integral path Γ_j for the Engquist-Osher scheme.

$\Gamma_j^{(i)}$, if $s_j^{(i)} \leq 0$, then it is the shock curve, otherwise rarefaction wave curve.

5.2 Modified flux method

Harten's second-order accurate TVD scheme [25] may be extended to hyperbolic system of conservation laws (5.1) by using the

characteristic decomposition

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n - \lambda(\hat{\boldsymbol{F}}_{j+\frac{1}{2}}^n - \hat{\boldsymbol{F}}_{j-\frac{1}{2}}^n),$$

with

$$\begin{aligned}\hat{\boldsymbol{F}}_{j+\frac{1}{2}} =& \frac{1}{2}[\boldsymbol{F}(\boldsymbol{U}_j) + \boldsymbol{F}(\boldsymbol{U}_{j+1})] \\ &+ \frac{1}{2\lambda}\sum_{k=1}^{m} \boldsymbol{R}_{j+\frac{1}{2}}^{(k)} \left[g_j^{(k)} + g_{j+1}^{(k)} - Q^{(k)}(\nu_{j+\frac{1}{2}}^{(k)} + \gamma_{j+\frac{1}{2}}^{(k)})\alpha_{j+\frac{1}{2}}^{(k)} \right],\end{aligned}$$

where $\nu^{(k)}_{j+\frac{1}{2}} = \lambda a^{(k)}(\boldsymbol{U}_{j+\frac{1}{2}})$, $\alpha^{(k)}_{j+\frac{1}{2}} = \boldsymbol{L}^{(k)}_{j+\frac{1}{2}}(\boldsymbol{U}_{j+1} - \boldsymbol{U}_j)$, and

$$
\begin{aligned}
g^{(k)}_j =& s^{(k)}_{j+\frac{1}{2}} \max\left\{0, \min\{|\tilde{g}^{(k)}_{j+\frac{1}{2}}|, \tilde{g}^{(k)}_{j-\frac{1}{2}} s^{(k)}_{j+\frac{1}{2}}\}\right\}, \quad s^{(k)}_{j+\frac{1}{2}} = \operatorname{sign}(\tilde{g}^{(k)}_{j+\frac{1}{2}}),\\
\tilde{g}^{(k)}_{j+\frac{1}{2}} =& \frac{1}{2}[Q^{(k)}(\nu^{(k)}_{j+\frac{1}{2}}) - (\nu^{(k)}_{j+\frac{1}{2}})^2]\alpha^{(k)}_{j+\frac{1}{2}},\\
\gamma^{(k)}_{j+\frac{1}{2}} =& \begin{cases} (g^{(k)}_{j+1} - g^{(k)}_j)/\alpha^{(k)}_{j+\frac{1}{2}}, & \alpha^{(k)}_{j+\frac{1}{2}} \neq 0,\\ 0, & \alpha^{(k)}_{j+\frac{1}{2}} = 0. \end{cases}
\end{aligned}
$$

Here $a^{(k)}(\boldsymbol{U}_{j+\frac{1}{2}})$ is the kth eigenvalue of the Jacobian matrix $\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{U}}$, and $\boldsymbol{L}^{(k)}$ and $\boldsymbol{R}^{(k)}$ are corresponding left and right eigenvectors.

If $g^{(k)}_j = 0$ for all j, then the above scheme reduces to the

three-point scheme

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n - \lambda(\hat{\boldsymbol{F}}_{j+\frac{1}{2}}^n - \hat{\boldsymbol{F}}_{j-\frac{1}{2}}^n),$$

$$\hat{\boldsymbol{F}}_{j+\frac{1}{2}} = \frac{1}{2}[\boldsymbol{F}(\boldsymbol{U}_j) + \boldsymbol{F}(\boldsymbol{U}_{j+1})] - \frac{1}{2\lambda}\sum_{k=1}^{m} \boldsymbol{R}_{j+\frac{1}{2}}^{(k)} Q^{(k)}(\nu_{j+\frac{1}{2}}^{(k)})\alpha_{j+\frac{1}{2}}^{(k)}.$$

5.3 WENO limiter for DG method

This section will give the details of the limiting procedure by using the WENO reconstruction as a limiter for the P^k-based RKDG methods. For the system case, the WENO schemes are usually based on local characteristic decompositions and flux splitting to avoid spurious oscillatory.

The local orthogonal basis $\{\phi_j^{(\ell)}(x), \ell = 0, 1, \cdots, k\}$ over e_j is adopted as before. The 1D WENO limiter is implemented as follows.

Step 1 (Identify the "troubled" cell):

For given DG approximate solutions $\boldsymbol{U}_j(x,t_n)=\sum_{\ell=0}^{k}\boldsymbol{U}_j^{(\ell)}(t_n)\phi_j^{(\ell)}(x)$ of the hyperbolic system of conservation laws over the element e_j, calculate its limiting values at two end points of e_j by

$$\boldsymbol{U}^-_{j+\frac{1}{2}}:=\boldsymbol{U}_j(x_{j+\frac{1}{2}}-0,t_n),\quad \boldsymbol{U}^+_{j-\frac{1}{2}}:=\boldsymbol{U}_j(x_{j-\frac{1}{2}}+0,t_n),$$

and the (appropriately linearized) characteristic variables

$$\boldsymbol{W}^-_{j+\frac{1}{2}}:=\boldsymbol{L}_j\boldsymbol{U}^-_{j+\frac{1}{2}},\quad \boldsymbol{W}^+_{j-\frac{1}{2}}:=\boldsymbol{L}_j\boldsymbol{U}^+_{j-\frac{1}{2}},\quad \boldsymbol{W}^{(0)}_{j}:=\boldsymbol{L}_j\boldsymbol{U}^{(0)}_{j},$$

where $\boldsymbol{L}_j:=\boldsymbol{L}(\boldsymbol{U}_j^{(0)}(t_n))$ is a matrix, the rows of which are the left eigenvectors of the Jacobian matrix $\boldsymbol{A}(\boldsymbol{U}_j^{(0)}(t_n))=\frac{\partial\boldsymbol{F}}{\partial\boldsymbol{U}}(\boldsymbol{U}_j^{(0)}(t_n))$.

Apply the TVB modified minmod function $\tilde{m}(a_1,a_2,a_3)$, de-

fined by

$$\tilde{m}(a_1, a_2, a_3) = \begin{cases} a_1, & \text{if } |a_1| \le Mh^2, \\ m(a_1, a_2, a_3), & \text{otherwise,} \end{cases} \tag{5.5}$$

to each component of the characteristic variables $\boldsymbol{W}$, where $M > 0$ is a constant, and

$$m(a_1, a_2, a_3) = \begin{cases} s \cdot \min_{1 \le i \le 3} |a_i|, & \text{if } \operatorname{sign}(a_1) = \cdots = \operatorname{sign}(a_3) = s, \\ 0, & \text{otherwise.} \end{cases}$$

Here for the sake of convenience, we still denote them in a vector form as follows

$$\begin{aligned} \tilde{\boldsymbol{W}}_j^{mod} :=& \tilde{m}(\tilde{\boldsymbol{W}}_j, \Delta_x^+ \boldsymbol{W}^{(0)}, \Delta_x^- \boldsymbol{W}^{(0)}), \\ \tilde{\tilde{\boldsymbol{W}}}_j^{mod} :=& \tilde{m}(\tilde{\tilde{\boldsymbol{W}}}_j, \Delta_+ \boldsymbol{W}^{(0)}, \Delta_- \boldsymbol{W}^{(0)}), \end{aligned}$$

where

$$\Delta_x^+\mathbf{W}^{(0)} =\mathbf{W}_{j+1}^{(0)} - \mathbf{W}_j^{(0)},\ \Delta_x^-\mathbf{W}^{(0)} = \mathbf{W}_j^{(0)} - \mathbf{W}_{j-1}^{(0)},$$
$$\tilde{\boldsymbol{W}}_j =\boldsymbol{W}_{j+\frac{1}{2}}^- - \boldsymbol{W}_j^{(0)},\quad \tilde{\tilde{\boldsymbol{W}}}_j = -\boldsymbol{W}_{j-\frac{1}{2}}^+ + \boldsymbol{W}_j^{(0)}.$$

If one component of $\tilde{\boldsymbol{W}}_j^{mod}$ is not equal to that of $\tilde{\boldsymbol{W}}_j$, or one component of $\tilde{\tilde{\boldsymbol{W}}}_j^{mod}$ is not equal to that of $\tilde{\tilde{\boldsymbol{W}}}_j$, then the cell e_j is marked as a "troubled" cell and denoted by e_j^{tc}.

Step 2 (WENO limiter for "troubled" cell):

For each "troubled" cell e_j^{tc}, from the cell average values of the RKDG solutions (the characteristic variables) in the neighboring cells as well as the "troubled" cells, i.e. some data $\{\boldsymbol{W}_j^{(0)}\}$, one uses the $(2k+1)$ order accurate WENO reconstruction [66] to give the approximate values of the characteristic variables $\boldsymbol{W}(x)$ at Gauss points x_m^{G} in the cell e_j^{tc}, denoted by $\boldsymbol{W}_m^{\mathrm{G}}$, $m = 1, 2, \cdots, q$,

e.g. $q = k + 1$ in the 1D case, and then recover the conservative variables $\boldsymbol{U}_m^{\mathrm{G}} = \boldsymbol{R}_j \boldsymbol{W}_m^{\mathrm{G}}$, where $\boldsymbol{R}_j := \boldsymbol{R}(\boldsymbol{U}_j^{(0)}(t_n))$ is a matrix, the columns of which are the right eigenvectors of the Jacobian matrix, constituting a complete system which is bi-orthonormal to the system of left eigenvectors, i.e.,

$$\boldsymbol{L}_j \boldsymbol{R}_j = \boldsymbol{I}.$$

After that, the RKDG solution $\boldsymbol{U}_j(x, t_n)$ in the "troubled" cell e_j^{tc} will be replaced with

$$\boldsymbol{U}_j^{\mathrm{WENO}}(x, t_n) = \boldsymbol{U}_j^{(0)} \phi_j^{(0)}(x) + \sum_{\ell=1}^{k} \boldsymbol{U}_j^{(\ell),\mathrm{WENO}} \phi_j^{(\ell)}(x), \qquad x \in e_j^{tc},$$

where the reconstructed WENO moments $\boldsymbol{U}_j^{(\ell),\mathrm{WENO}}$, $\ell = 1, 2, \cdots, k$, are calculated based on the reconstructed point values $\boldsymbol{U}_\ell^{\mathrm{G}}$ at the

Gauss points $x_\ell^{\rm G}$ in the cell e_j^{tc} and the numerical quadrature, i.e.

$$\boldsymbol{U}_j^{(\ell),\rm WENO} := \frac{1}{a_\ell}\int_{e_j^{tc}} \boldsymbol{U}_h^{\rm WENO}(x,t_n)\phi_j^{(\ell)}\,dx \approx \frac{h_j}{a_\ell}\sum_{m=1}^{q}\omega_m \boldsymbol{U}_m^{\rm G}\phi_j^{(\ell)}(x_m^{\rm G}),\ 1\le \ell\le k,$$

here ω_m, $m=1,2,\cdots,q$, are the Gaussian quadrature weights for the Gaussian point $x_m^{\rm G}$, and $a_\ell = \int_{e_j}\left(\phi_j^{(\ell)}(x)\right)^2\,dx$ are the normalization constants since the basis is not orthonormal.

Remark 5.1 *According to the analysis in [8], the numerical quadratures should be exact at least up to $O(h^{2k+2})$ for the P^k-based RKDG methods. Similar to [56] we may use the WENO reconstruction with $(2k+1)$ order of accuracy as limiters for the P^k-based RKDG methods, in which the two-point Gauss quadrature points $x_{j-\sqrt{3}/6} := x_j - \sqrt{3}/6h_j$ and $x_{j+\sqrt{3}/6} := x_j + \sqrt{3}/6h_j$ are used in practical computations for the P^1 case; third-point Gauss quadrature points $x_{j-\sqrt{15}/10}$, x_j, $x_{j+\sqrt{15}/10}$ are used for the P^2*

case; and the four-point Gauss quadrature points

$$x_{j-\sqrt{525+70\sqrt{30}}/70}, x_{j-\sqrt{525-70\sqrt{30}}/70}, x_{j+\sqrt{525-70\sqrt{30}}/70}, x_{j+\sqrt{525+70\sqrt{30}}/70},$$

are used for the P^3 *case.*

6 Several advanced topics

This section will introduce several advanced topics which are the sonic point glitch [72], the local oscillation in monotone schemes [48], the adaptive moving mesh method [73], and some discussions on the implicit schemes.

6.1 Sonic point glitch

Although much attention has been paid to capture shock waves and contact discontinuities, the proper resolution of rarefaction

wave has also proved to be difficult. For example, even in the smooth flow region, such as a high speed expansion wave passing through a corner, the sonic glitch or the so-called "dog-leg" phenomena, see Fig. 17, has occasionally been observed by Woodward & Colella in [94], when they solved an inviscid flow in a channel containing a forward step. There are other cases, e.g. diffraction of shock waves in [53, 57] and supersonic flows past a circular cylinder [58] where the sonic glitch can arise. Glitches do not really occur along sonic lines in multidimensional flows but only where the normal speed component is sonic.

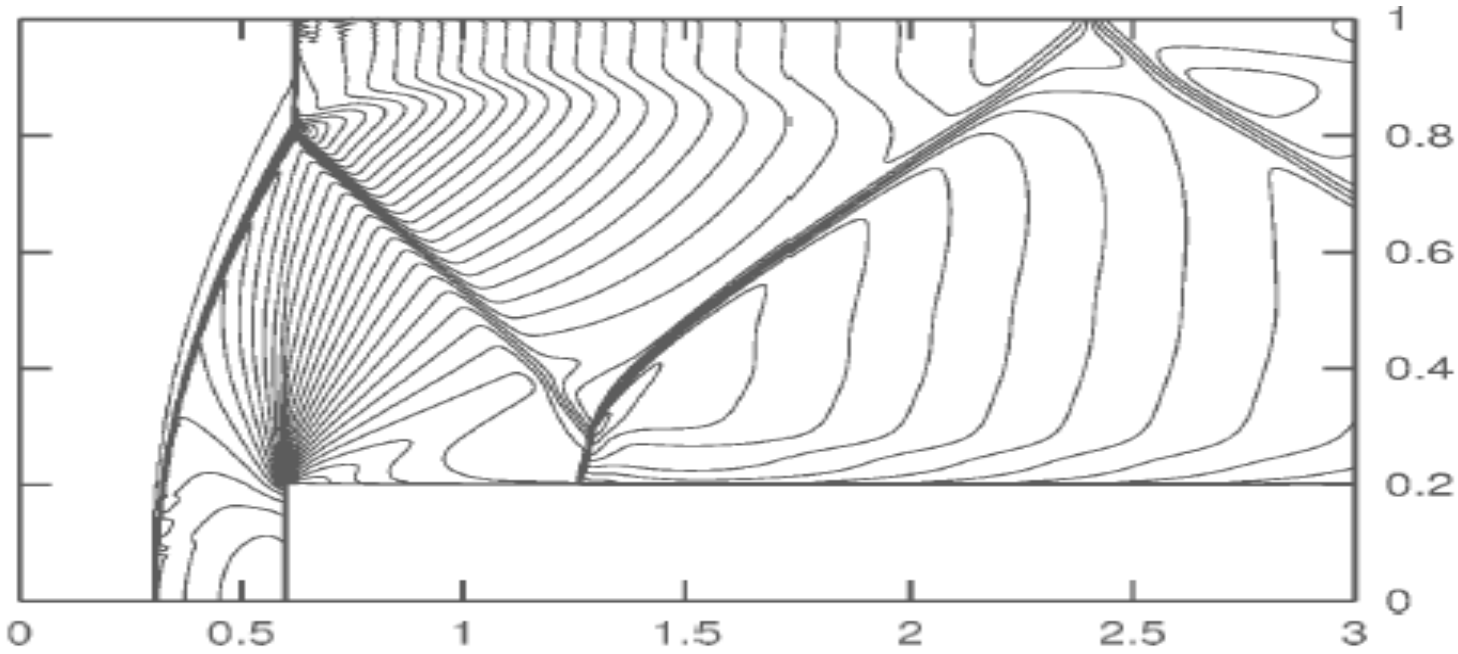


Figure 17: Density contours with 32 equally spaced contour lines for the Mach 3 wind tunnel problem [94] calculated by using first-order accurate Godunov scheme to solve two-dimensional Euler equations in gas dynamics on a uniform grid with spatial step sizes $h_x = h_y = 1/200$ in x and y directions. The so-called "dog-leg" phenomena has been observed near the sonic line above the corner of the forward step, i.e. the point $(0.6, 0.2)$.

The sonic glitch arises only in the presence of sonic rarefaction waves, and is a small non-physically discontinuous jump or any visible error around the sonic point generated by numerical methods within a sonic rarefaction wave.

Consider the initial value problem of Burgers' equation

$$u_t + \left(u^2/2\right)_x = 0 \tag{6.1}$$

with the initial data $u(x,0) = u_0(x)$, where $u_0(x)$ is a given function, $x \in \mathbb{R}$ and $t > 0$. The simplest and useful initial value problem is the so-called Riemann problem with the initial data

$$u_0(x) = \begin{cases} u_L, & x < 0, \\ u_R, & x > 0, \end{cases} \tag{6.2}$$

where u_L and u_R are two constants. Its solution will be a fundamental component of Godunov type scheme. The solution $u(x,t)$

to the Riemann problem of Burgers' equation with (6.2) can be given in an explicit form. It is a rarefaction wave solution

$$u(x,t) = \begin{cases} u_L, & x < u_L t, \\ \frac{x}{t}, & u_L t \le x \le u_R t, \\ u_R, & x > u_R t, \end{cases} \tag{6.3}$$

if $u_L < u_R$, or a shock wave solution

$$u(x,t) = \begin{cases} u_L, & x < st, \\ u_R, & x > st, \end{cases} \tag{6.4}$$

if $u_L > u_R$, where s denotes speed of shock wave satisfying

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

The sonic point corresponds to a point with $f'(u) \equiv u = 0$, and the location of this point is fixed in space due to its diminishing wave speed. Thus, for the Riemann problem of Burgers'

equation with (6.2), if $u_L < 0 < u_R$ or $u_L > 0 > u_R$, the solution given in (6.3) or (6.4) is corresponding to a transonic solution. Generally, the sonic glitch does not arise in any transonic compression region. To confirm it, we do a numerical experiment on computation of a 2π-periodic problem of Burgers' equation, with a 2π-periodic initial data

$$u_0(x) = 0.5 + \sin(x), \quad x \in [0, 2\pi).$$

Fig. 18 shows the numerical solutions with a resolution of 100 grid cells, calculated by using three first-order accurate schemes, the Godunov, Roe, and LF schemes, respectively. There is a transonic compression wave in the solution at $t = 0.8$ in the interval $[2.8, 4.2]$ before the shock is formed. Obviously, the sonic glitch is not observed in the computed solutions. The reason for that will be analyzed later in this section.

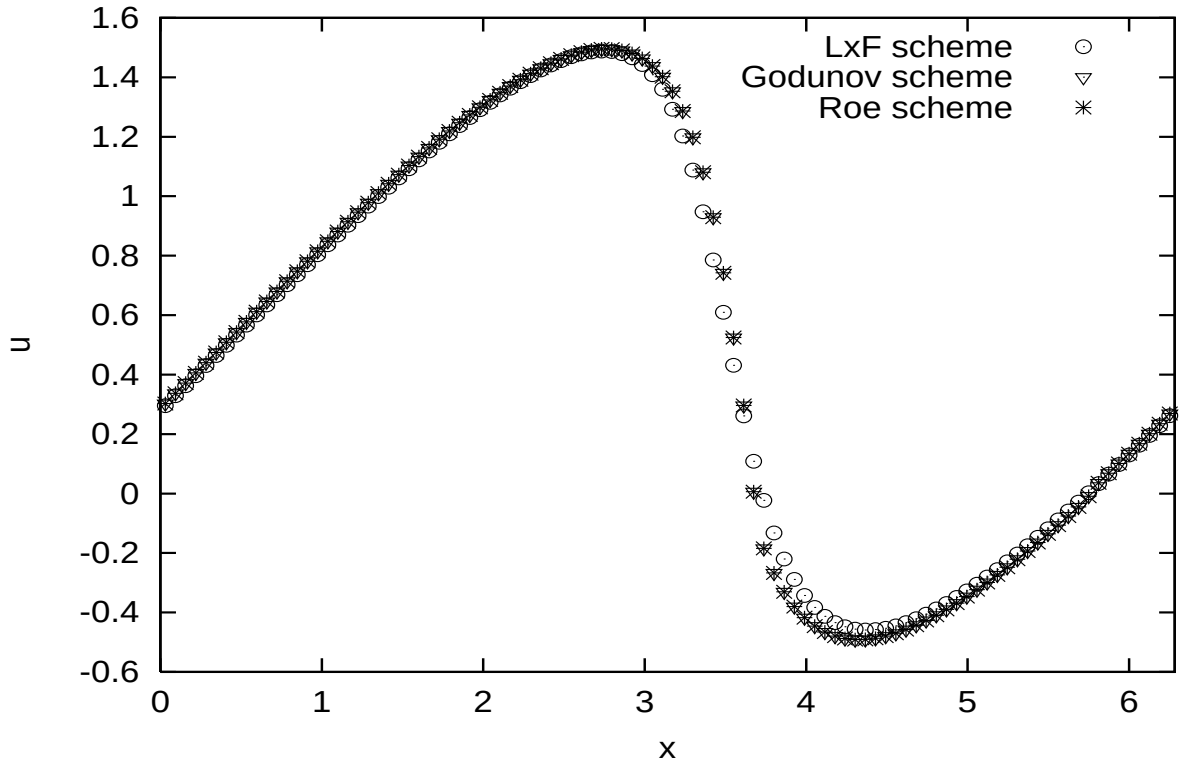


Figure 18: The computed solutions of Burgers' equation at $t = 0.8$ containing a transonic compression wave. They are calculated by using the LF , Godunov (6.7), and Roe schemes, respectively.

6.1.1 Godunov scheme

This section investigates numerical evolution of the sonic rarefaction wave (6.3) with $u_L < 0 < u_R$. Unless stated otherwise, we will take $u_R = -u_L = 1$. Give a uniform partition of the physical domain $\mathbb{R}$, $x_j = jh$, where h denotes a cell size in space, and $j \in \mathbb{Z}$. The initial value function $u(x,0) = u_0(x)$ will be approximated by the cell average over each cell $I_j = \{x | x_j - \frac{h}{2} < x < x_j + \frac{h}{2}\}$, i.e.

$$u_h(x,0) = \frac{1}{h}\int_{I_j} u_0(x)\, dx =: u_j^0 \quad \text{for } x \in I_j. \tag{6.5}$$

Under the assumption that $u_L < u_R$, the solution at time t_n to the Riemann problem of Burgers' equation (6.1) with the

initial data (6.2) is given exactly by

$$u(x,t_n)=\begin{cases} u_L, & x<u_Lt_n,\\ \frac{x}{t_n}, & u_Lt_n\le x\le u_Rt_n,\\ u_R, & x>u_Rt_n. \end{cases} \tag{6.6}$$

In the following we assume that the time step size τ satisfies the CFL condition

$$\lambda\max_u\{|f'(u)|\}\le\mu<1,\quad \lambda=\frac{\tau}{h}.$$

In order to evolve numerically the sonic rarefaction wave (6.6), we should project the initial data (6.6) onto the given uniform grid, following (6.5). Thus, we can represent numerically the initial data (6.6) as a constant state inside each cell I_j. For example, within the expansion fan but away from its two corners, the tail and the head, the initial data inside the cell I_j are

approximated by

$$u_j^n = \frac{1}{h}\int_{I_j} \frac{x}{t_n}\, dx \equiv \frac{x_j}{t_n}.$$

These constant states at the right hand side of the sonic point satisfy $u_j > 0$ and $u_j < 0$ at the left. At a later time $t_{n+1} = t_n + \tau$, the exact solution becomes

$$u(x, t_{n+1}) = \begin{cases} u_L, & x < u_L t_{n+1}, \\ \frac{x}{t_n+\tau}, & u_L t_{n+1} \le x \le u_R t_{n+1}, \\ u_R, & x > u_R t_{n+1}, \end{cases}$$

and corresponding exact cell-averaged value is

$$u_j^e = \frac{x_j}{t_n+\tau} = \frac{x_j}{t_n}\left(\frac{1}{1+\tau/t_n}\right) = \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n} + (\frac{\tau}{t_n})^2 - (\frac{\tau}{t_n})^3 + \cdots\right),$$

where the superscript e means the exact solution.

For the in-viscid Burgers' equation (6.1), the numerical flux of the Godunov method (3.17) becomes

$$\hat{f}^G(u_j^n, u_{j+1}^n) = \max\left\{\frac{1}{2}(u_j^+)^2, \frac{1}{2}(u_{j+1}^-)^2\right\}, \tag{6.7}$$

where $u^+ = \max\{u, 0\}$ and $u^- = \min\{u, 0\}$. Therefore, on the right hand side of the sonic point, $t_n - h > x_j > h$, the flow variable is updated by using the Godunov method through

$$\begin{aligned} u_j^{n+1} =& u_j^n + \lambda\left(\widehat{f}_{j-\frac{1}{2}} - \widehat{f}_{j+\frac{1}{2}}\right) = \frac{x_j}{t_n} + \lambda\left(\frac{1}{2}(\frac{x_{j-1}}{t_n})^2 - \frac{1}{2}(\frac{x_j}{t_n})^2\right) \\ =& \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right) + \frac{\tau h}{2(t_n)^2}, \end{aligned} \tag{6.8}$$

and on the left side, $-t_n + h < x_j < -h$, it is

$$\begin{aligned} u_j^{n+1} =& u_j^n + \lambda\left(\widehat{f}_{j-\frac{1}{2}} - \widehat{f}_{j+\frac{1}{2}}\right) = \frac{x_j}{t_n} + \lambda\left(\frac{1}{2}(\frac{x_j}{t_n})^2 - \frac{1}{2}(\frac{x_{j+1}}{t_n})^2\right) \\ =& \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right) - \frac{\tau h}{2(t_n)^2}. \end{aligned} \tag{6.9}$$

If we define

$$\widetilde{u}_j^e = \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right) = u_j^e + \mathcal{O}(\tau^2),$$

then we can find from (6.8) and (6.9) that after one time step, the numerical solution u_j^{n+1} is bigger than the "exact" one $\widetilde{u}_j^e$ in the region $t_n - h > x_j > h$, while smaller than $\widetilde{u}_j^e$ in the region $-t_n + h < x_j < -h$. Unfortunately, the shift appeared in the approximate solution has opposite signs in the two regions:

$x > h$ and $x < -h$. Therefore, a jump with the magnitude of $\frac{\tau h}{(t_n)^2}$ will appear at the sonic point $x = 0$ after one evolution time step. As a result, the approximate solution will move upward (or downward) in comparison with the "exact" one in the right (or left) region at the next time step. Moreover, the strength of this jump will tend to zero, as the space size h tends to zero.

Remark 6.1 *In the above, $\widetilde{u}_j^e$ is a second-order accurate approximation of the exact solution u_j^e in time. It is possible to improve it by using a higher-order Runge-Kutta time discretization. But, the magnitude of the relative error $u_j^{n+1} - \widetilde{u}_j^e$ is $O(h)$.*

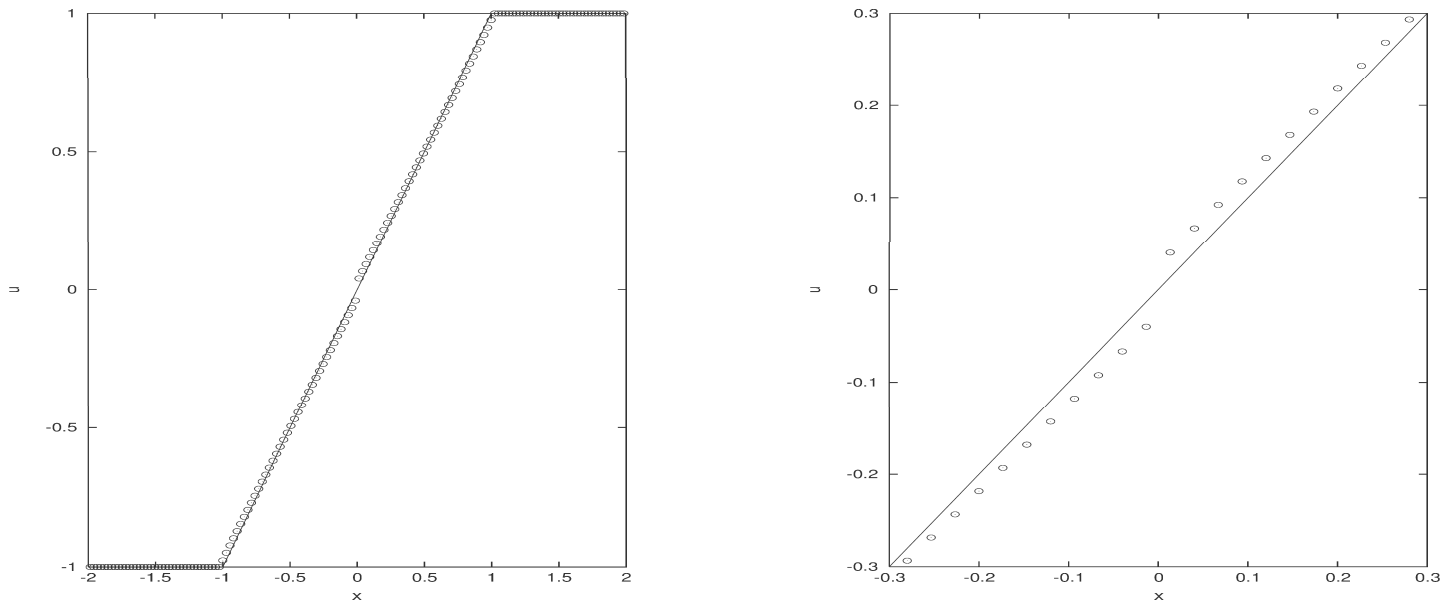


Figure 19: The computed solutions ("circle") of Burgers' equation at $t = t_n + 0.9$ are given by using the Godunov scheme (3.17) with (6.7) with 150 grid cells and $t_n = 0.1$. The solid line denotes the exact solution. Left: the solution within the global domain $[-2, 2]$; right: close-up of the solution in the vicinity of the sonic point.

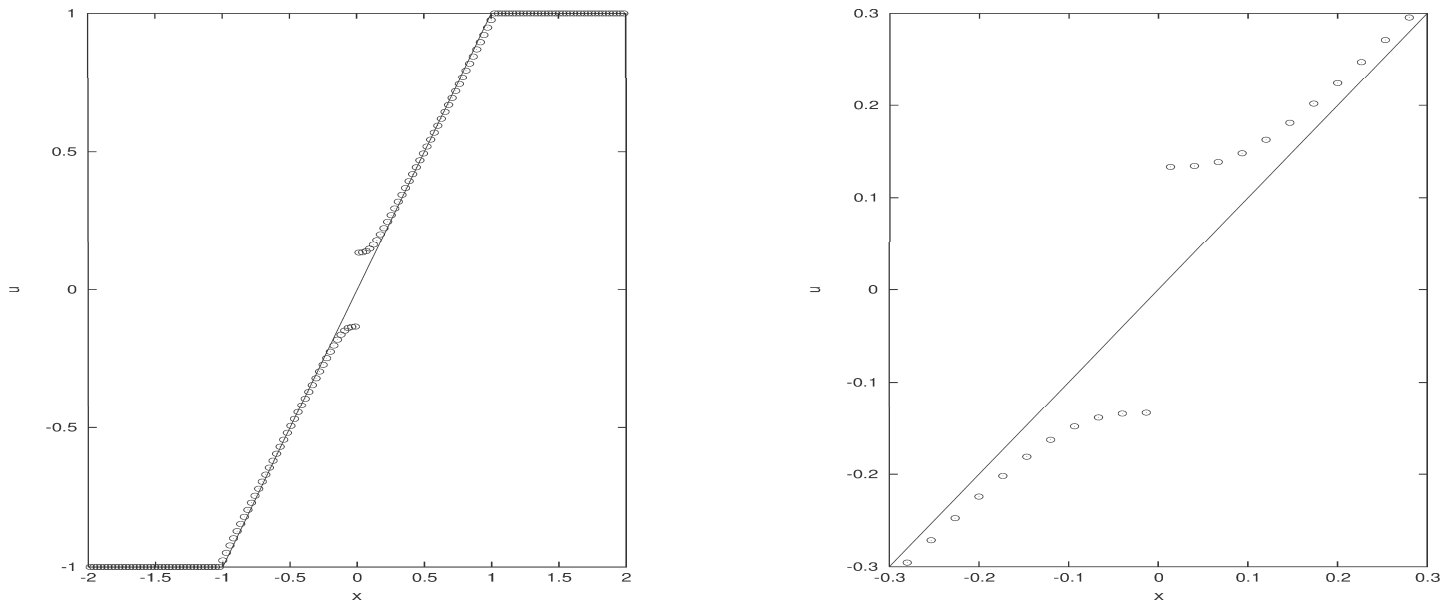


Figure 20: Same as Fig. 19, except for the Roe scheme.

An alternative way to understand the sonic glitch formation can be the following. Because the solution is smooth within the expansion fan, Burgers' equation (6.1) can be rewritten in a non-

conservative form as follows:

$$\frac{\partial u}{\partial t} + a(u)\frac{\partial u}{\partial x} = 0, \quad a(u) = f'(u) \equiv u. \tag{6.10}$$

Hence, the exact wave speed at x_j equals to $u_j^n = \frac{x_j}{t_n}$ when t belongs to the time interval $[t_n, t_n + \tau)$. On the other hand, (6.8) and (6.9) can also be rewritten in a non-conservative form as a finite difference approximation of (6.10),

$$u_j^{n+1} = \begin{cases} \frac{x_j}{t_n} + \lambda \frac{x_j + x_{j-1}}{2t_n} \left(\frac{x_{j-1}}{t_n} - \frac{x_j}{t_n} \right), & \text{for (6.8)}, \\ \frac{x_j}{t_n} + \lambda \frac{x_j + x_{j+1}}{2t_n} \left(\frac{x_j}{t_n} - \frac{x_{j+1}}{t_n} \right), & \text{for (6.9)}. \end{cases} \tag{6.11}$$

Due to the upwind flux (6.7) at a cell interface, the numerical wave propagation speed at x_j becomes $\frac{x_j + x_{j-1}}{2t_n}$ for $x_j > h$, and $\frac{x_j + x_{j+1}}{2t_n}$ for $x_j < -h$. Comparing them with the exact wave

speed, we can find that the magnitude of the numerical wave speed is smaller than the exact wave speed in both regions, that is to say, $0 < \frac{x_j+x_{j-1}}{2t_n} < \frac{x_j}{t_n}$ in the region $x > h$ and $\frac{x_j}{t_n} < \frac{x_j+x_{j+1}}{2t_n} < 0$ in the region $x < h$. It is worth noting that the speed difference, $\frac{h}{2t_n}$, is independent of the distance to the sonic point. So, the propagation of the approximate solution undergoes a delay around the sonic point and generates a wiggle there due to slower wave speeds. Using exact value of the wave speeds to replace the numerical ones in (6.11), we may expect to avoid appearance of the sonic glitch. In fact, we have:

$$u_j^{n+1} = \begin{cases} \frac{x_j}{t_n} + \lambda \frac{x_j}{t_n}\left(\frac{x_{j-1}}{t_n} - \frac{x_j}{t_n}\right) = \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right), & \text{for } u_j^n \geq 0, \\ \frac{x_j}{t_n} + \lambda \frac{x_j}{t_n}\left(\frac{x_j}{t_n} - \frac{x_{j+1}}{t_n}\right) = \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right), & \text{for } u_j^n < 0. \end{cases}$$

To validate the theoretical analysis, some numerical experiments are conducted here. We take $t_n = 0.1$, $\mu = 0.9$, and 150

cells. Figs. 19 and 20 show the numerical solutions ("circle") at $t = t_n + 0.9$ calculated by using the Godunov method and the Roe scheme, respectively. For comparison, the exact solution ("solid line") is also given there. Obviously, the sonic glitch is formed around the sonic point $x = 0$. It is in accordance with the theoretical one. The jump around the sonic point in Fig. 20 becomes very large such that the numerical solution is unacceptable. Fig. 21 gives the corresponding actual error plots. The result shows that because local maximum or minimum errors are formed around the sonic point, the error distributions and the numerical solutions are not monotone with respect to x/t within the expansion fan, i.e. the interval $(-1, 1)$. These errors should tend to zero, as $h \to 0$. Fig. 22 plots the local maximum error around the sonic point versus the space size in a log scale. Its slope equals approximately to 0.987 that is a numerical measure of the scheme accuracy. For a rarefaction wave without a son-

ic point in it, such as a wave with u going from $u_L = 0.1$ to $u_R = 1.1$, the uniform upward shift $u_j^{n+1} - u_j^e$ in this case will not generate any glitch in this rarefaction wave, see Fig. 23.

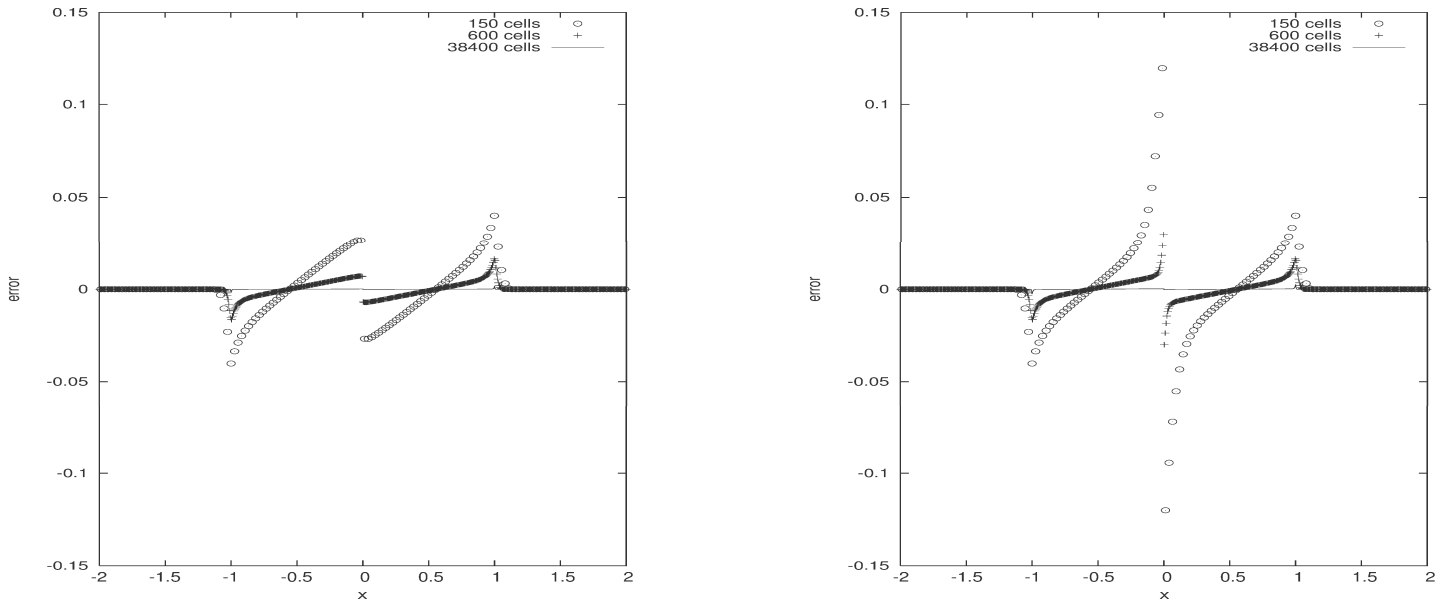


Figure 21: The error distributions, $u(x_j, t_n) - u_j^n$, for Burgers' equation. Left: the Godunov scheme; right: the Roe scheme.

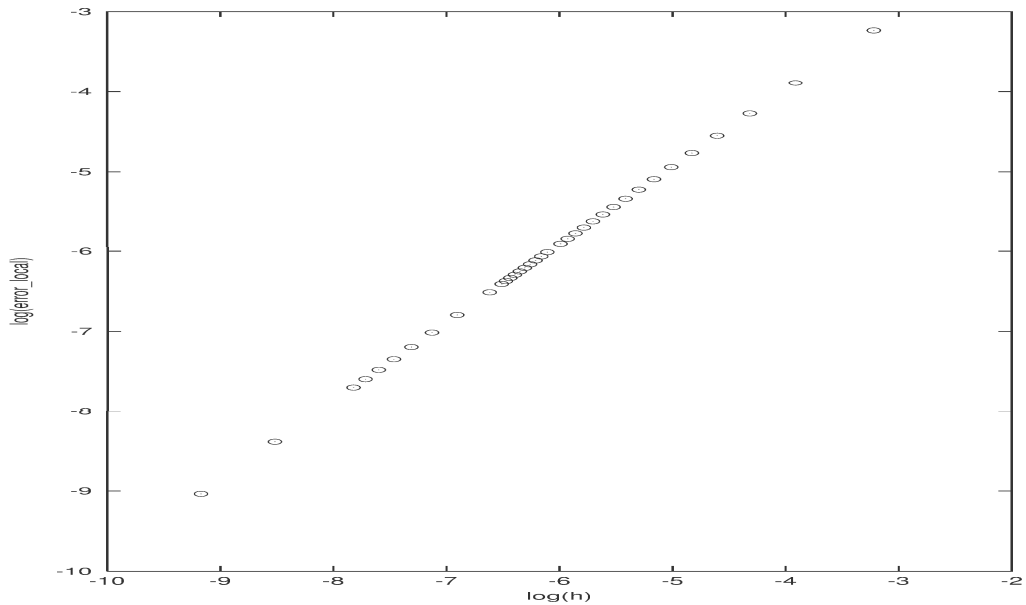


Figure 22: The local maximum error around the sonic point versus the space size in a log scale for Burgers' equation.

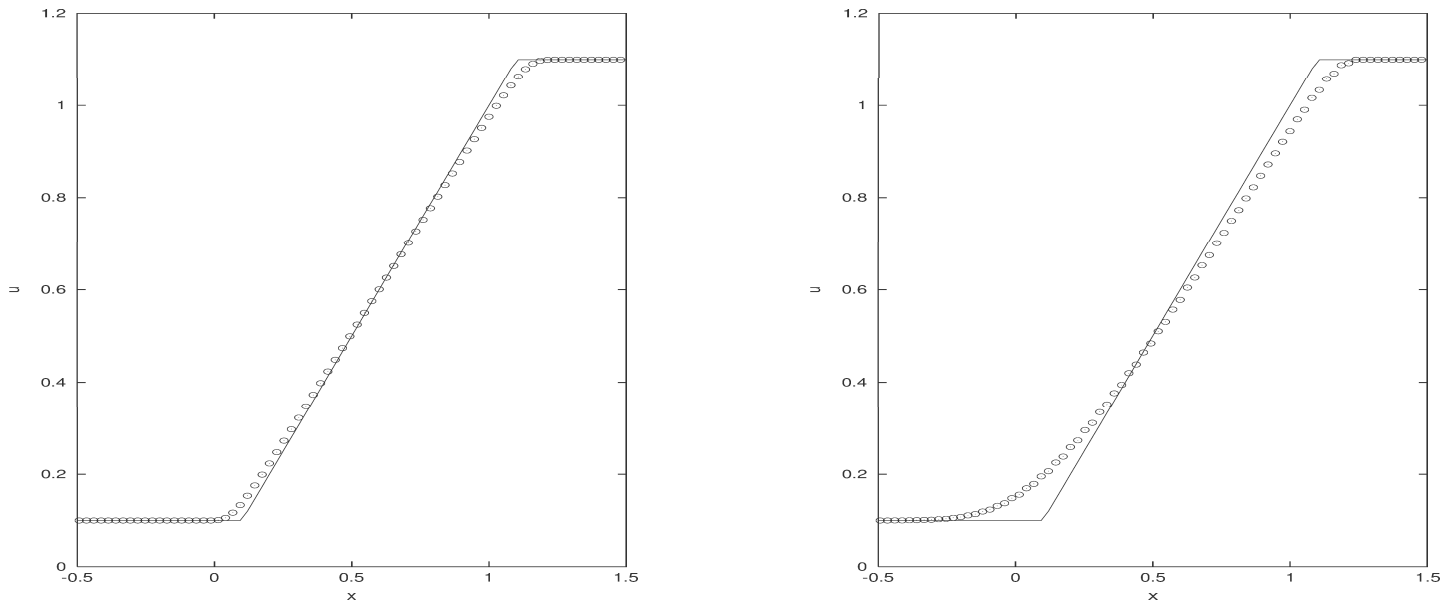


Figure 23: The supersonic solutions ("circle") of Burgers' equation at $t = 0.9$ are computed by using the Godunov method (left) and the LF scheme (right) with 150 grid cells and $t_n = 0.1$ for $u_L = 0.1$ and $u_R = 1.1$. The solid line denotes the exact solution.

Remark 6.2 *The above analysis could be applied to the transonic compression problem of Burgers' equation. For convenience,*

we take initial data (at $t = t_n$) as

$$u(x, t_n) = \begin{cases} u_L, & x < x_L, \\ -\frac{x}{t_n}, & x_L \le x \le x_R, \\ u_R, & x > x_R, \end{cases} \tag{6.12}$$

where $u_L = -x_L/t_n > 0 > -x_R/t_n = u_R$, $t_n > 0$. Assuming that $-x_L$ and x_R are big enough such that the solution to the Cauchy problem of Burgers' equation with (6.12) is smooth within the sub-domain $[t_n, t_n + \tau] \times (x_L, x_R)$, we may write Burgers' equation and the Godunov scheme in a non-conservative form, respectively. Using the Godunov scheme, we have

$$u_j^{n+1} = \begin{cases} \frac{-x_j}{t_n} + \sigma \frac{-x_j - x_{j-1}}{2t_n} \left(\frac{-x_{j-1}}{t_n} - \frac{-x_j}{t_n} \right), & \text{for } x \in (x_L, 0), \\ \frac{-x_j}{t_n} + \sigma \frac{-x_j - x_{j+1}}{2t_n} \left(\frac{-x_j}{t_n} - \frac{-x_{j+1}}{t_n} \right), & \text{for } x \in (0, x_R). \end{cases}$$

Due to the upwind flux at a cell interface, the numerical wave propagation speed at x_j *becomes* $\frac{-x_j - x_{j-1}}{2t_n}$ *for* $x_j < -h$*, and* $\frac{-x_j - x_{j+1}}{2t_n}$ *for* $x_j > h$*. Comparing them with the exact wave speed* $\frac{-x_j}{t_n}$*, we can find that the numerical wave speeds are faster than the exact wave speeds in both regions, that is to say,* $\frac{-x_j - x_{j-1}}{2t_n} > \frac{-x_j}{t_n} > 0$ *in the region* $(x_L, -h)$ *and* $0 > \frac{-x_j}{t_n} > \frac{-x_j - x_{j+1}}{2t_n}$ *in the region* (h, x_R)*. Thus, the propagation of the approximate solution undergoes a advance around the sonic point and the compression is accelerated. Due to it, the discontinuous jump around the sonic point cannot be observed in a transonic compression region. For comparison, we show wave propagation speeds for the sonic compression and rarefaction waves in Fig. 24, where the solid and dotted lines denote the exact and numerical wave propagation speeds, respectively.*

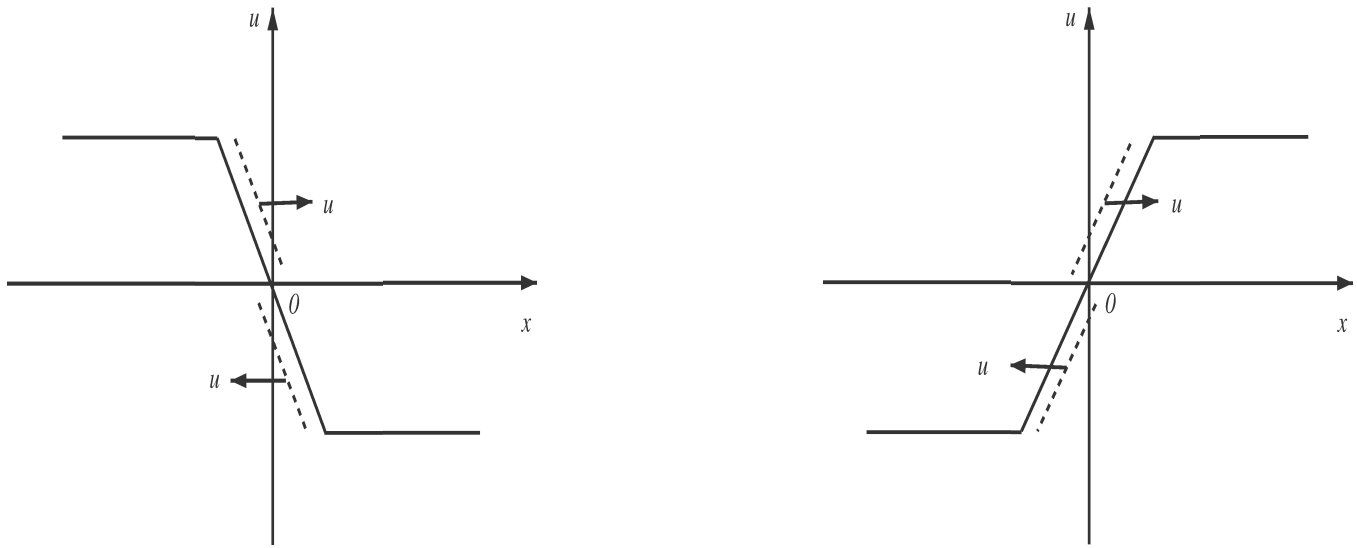


Figure 24: The wave propagation speeds. Left: a sonic compression wave; right: a sonic rarefaction wave.

6.1.2 Central-difference schemes

As we have known, for most upwind schemes, the numerical dissipation goes to a minimum value around the sonic point [33]. *Is the sonic glitch formed due to a smaller numerical viscosity or any upwind scheme*? To answer this question, we check the value of u_j^{n+1} calculated by using the LF scheme and the LW scheme. The LF scheme is of a relatively large numerical viscosity, while the LW scheme is of a smaller viscosity than that of the Godunov

scheme. For the LF scheme, the update of the variable u_j^{n+1} is

$$\begin{aligned} u_j^{n+1} =& \frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right) + \frac{\lambda}{2}\left(f(u_{j-1}^n) - f(u_{j+1}^n)\right) \\ =& \frac{1}{2}\left(\frac{x_{j+1}}{t_n} + \frac{x_{j-1}}{t_n}\right) + \frac{\lambda}{2}\left(\frac{1}{2}(\frac{x_{j-1}}{t_n})^2 - \frac{1}{2}(\frac{x_{j+1}}{t_n})^2\right) \\ =& \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right) = u_j^e + \mathcal{O}(\tau^2), \end{aligned} \tag{6.13}$$

in the region $t_n - h > x_j > -t_n + h$. It is worth noting that the variable u_j^{n+1} calculated by the unstable central scheme:

$$u_j^{n+1} = u_j^n + \frac{\lambda}{2}\left(f(u_{j-1}^n) - f(u_{j+1}^n)\right),$$

equals to $\frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n}\right)$ too, although its numerical viscosity is zero. Moreover, the gradients of their solutions equal to $\left(1 - \frac{\tau}{t_n}\right)/t_n$

within the rarefaction wave region that is a second-order accurate approximation to the exact one in time. If let $\tau \to 0$, then the gradient of the numerical solution computed by the LF scheme should tend to the gradient of the exact solution at any time $t > t_n$.

For the Godunov scheme, from (6.8) and (6.9), we may derive the approximate gradient at $t = t_n + \tau$ within the rarefaction fan as follows:

$$\frac{u_{j+1}^{n+1} - u_j^{n+1}}{h} = \begin{cases} \frac{1}{t_n+\tau} + \widetilde{C}\tau + \mathcal{O}(\tau^2), & \text{near the sonic point,} \\ \frac{1}{t_n+\tau} + \mathcal{O}(\tau^2), & \text{otherwise,} \end{cases}$$

where $\widetilde{C} = \mathcal{O}(1)$. Therefore, the approximate gradient at $t = T \equiv t_n + N\tau$ within the rarefaction fan becomes

$$\frac{u_{j+1}^{n+N} - u_j^{n+N}}{h} = \begin{cases} \frac{1}{T} + \mathcal{O}(1) + \mathcal{O}(\tau), & \text{near the sonic point,} \\ \frac{1}{T} + \mathcal{O}(\tau), & \text{otherwise,} \end{cases}$$

or

$$\lim_{\tau,h\to 0} \frac{u_{j+1}^{n+N} - u_j^{n+N}}{h} = \begin{cases} \frac{1}{T} + \mathcal{O}(1), & \text{near the sonic point,} \\ \frac{1}{T}, & \text{otherwise,} \end{cases}$$

where $\mathcal{O}(1)$ depends on T or N, but does not on the cell number. Fig. 25 shows the gradients of the numerical and exact solutions for the sonic rarefaction problem. The numerical solutions are computed by two difference schemes, the LF scheme with 10000 cells and the Godunov scheme with 10000 cells as well as 38440 cells. These numerical results coincide with the above theoretical results.

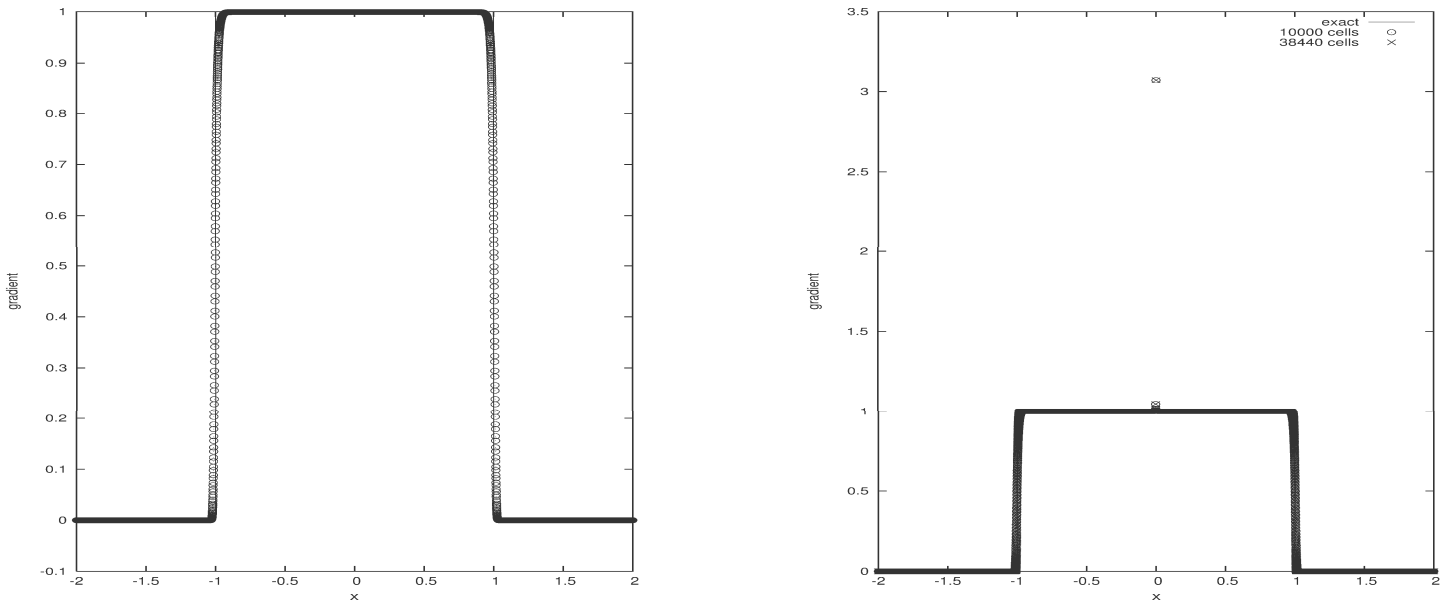


Figure 25: The gradients of the numerical solutions for the sonic rarefaction problem computed by using the LF scheme with 10000 cells (left) and the Godunov scheme with 10000 cells as well as 38440 cells.

Similarly, the variable u_j^{n+1} can also be updated by using the

LW scheme as

$$
\begin{aligned}
u_j^{n+1} =& u_j^n + \frac{\lambda}{2}\left(f(u_{j-1}^n) - f(u_{j+1}^n)\right) + \frac{\left(\lambda a(u_{j+\frac{1}{2}})\right)^2}{2}\left(u_{j+1}^n - u_j^n\right) \\
& - \frac{\left(\lambda a(u_{j-\frac{1}{2}})\right)^2}{2}\left(u_j^n - u_{j-1}^n\right) \\
=& \frac{x_j}{t_n} + \frac{\lambda}{2}\left(\frac{1}{2}(\frac{x_{j-1}}{t_n})^2 - \frac{1}{2}(\frac{x_{j+1}}{t_n})^2\right) + \frac{\lambda^2\left(u_{j+1}^n + u_j^n\right)^2}{8}\left(\frac{x_{j+1}}{t_n} - \frac{x_j}{t_n}\right) \\
& - \frac{\lambda^2\left(u_j^n + u_{j-1}^n\right)^2}{8}\left(\frac{x_j}{t_n} - \frac{x_{j-1}}{t_n}\right) \\
=& \frac{x_j}{t_n}\left(1 - \frac{\tau}{t_n} + (\frac{\tau}{t_n})^2\right) = u_j^e + \mathcal{O}(\tau^3),
\end{aligned}
\tag{6.14}
$$

in both region $t_n - h > x_j > -t_n + h$. From (6.13) and (6.14), it is not difficult to find that there is neither additional term $\tau h/2t_n^2$, nor the sonic glitch. The gradient of the solution of the LW scheme equals to $\left(1 - \frac{\tau}{t_n} + \left(\frac{\tau}{t_n}\right)^2\right)/t_n$ within the rarefaction

wave region that is a third-order accurate approximation to the exact one in time. When $\tau \to 0$, it should tend to the gradient of the exact solution too. The numerical solutions shown in Fig. 26 are consistent with this theoretical analysis. They are obtained by using the LF scheme (left) and the LW scheme (right), respectively. From Fig. 27, we may observe that there is no local extremum in the actual error profile near the sonic point.

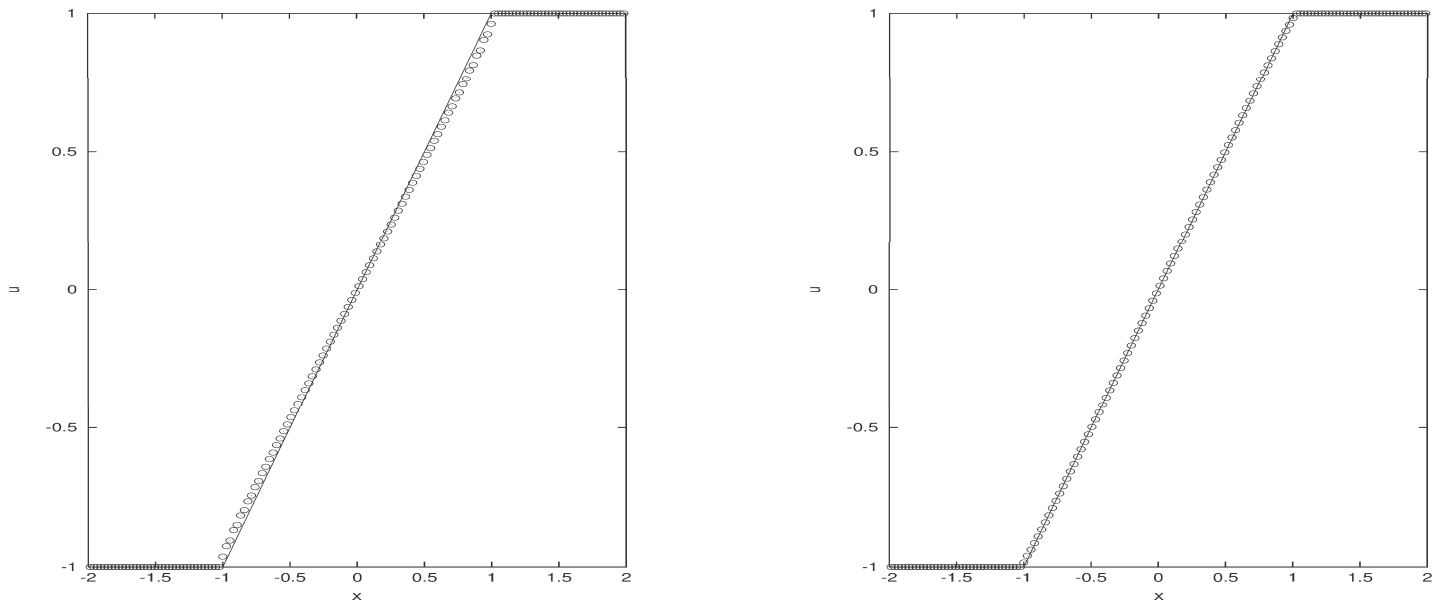


Figure 26: Same as Fig. 19, except for the LF scheme (left) and the LW scheme (right), respectively.

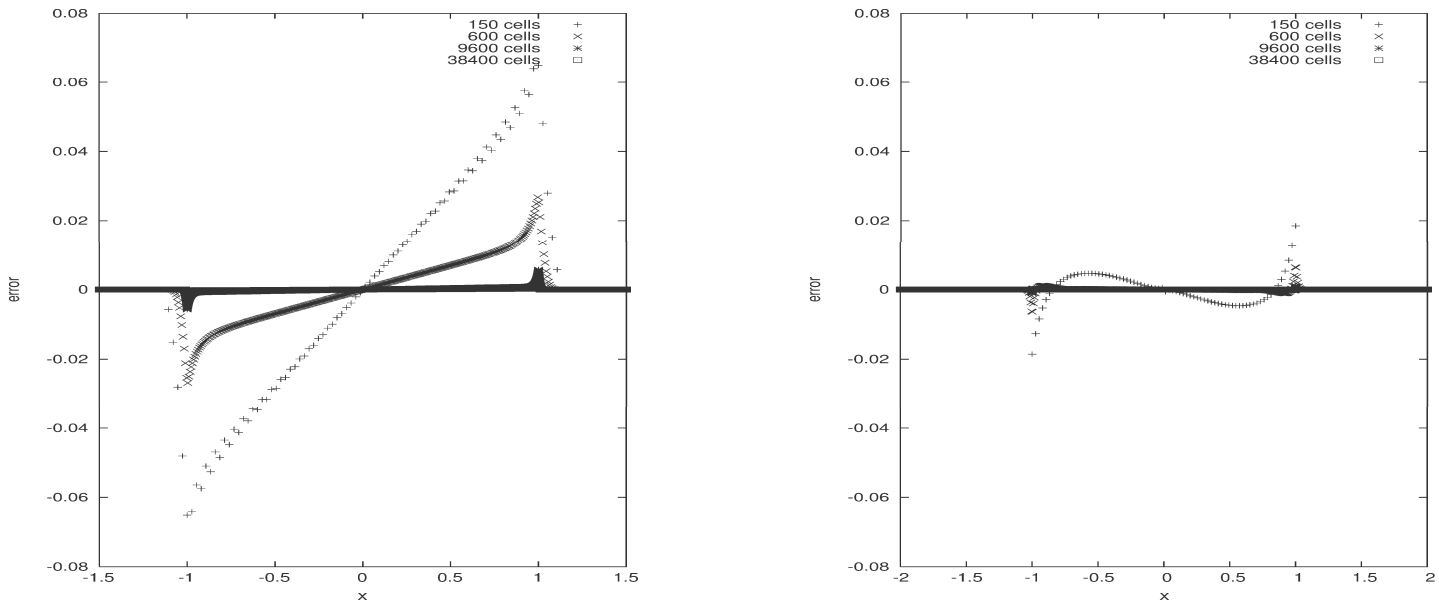


Figure 27: Same as Fig. 21, except for the LF scheme (left) and the LW scheme (right), respectively.

6.1.3 Other schemes

We may further show that the glitch is closely associated with the upwind flux, even though the numerical viscosity is big enough. Consider a general 3-point scheme in the viscous form (3.89). According to the above analysis, if we take $Q_{j+\frac{1}{2}} = \frac{1}{2}(Q^G_{j+\frac{1}{2}} + Q^L_{j+\frac{1}{2}})$ or $\frac{1}{2}(Q^R_{j+\frac{1}{2}} + Q^L_{j+\frac{1}{2}})$, then a shift, $\frac{\tau h}{2(t_n)^2}$, will be also formed when the variable u_j^{n+1} is updated, where $Q^G_{j+\frac{1}{2}}$, $Q^R_{j+\frac{1}{2}}$, and $Q^L_{j+\frac{1}{2}}$ denote the numerical viscosities of the Godunov method, the Roe scheme, and the LF scheme, respectively. We have used these two weighted average schemes to solve the above problem with same grid points and parameters. Fig. 28 only shows close-up of the solutions and error given by the weighted average scheme with $Q = \frac{1}{2}(Q^R + Q^L)$. Although the error around the sonic point is very small and almost invisible, local extrema still exist and the

solution is not monotone with respect to x/t along the transonic expansion fan.

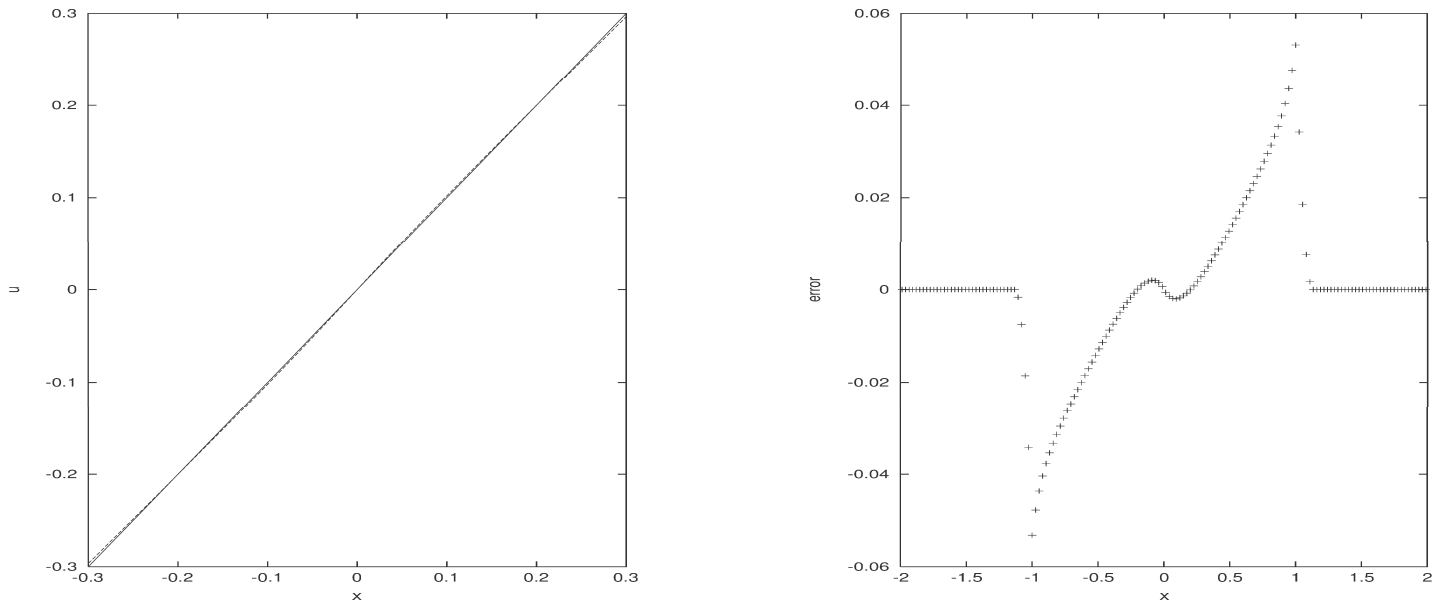


Figure 28: Close-up of the computed solution (left) and error (right) for Burgers' equation are calculated by using the weighted average scheme with $Q = \frac{1}{2}(Q^R + Q^L)$.

In the literature, to cure the sonic glitch, an entropy fix is

usually added to modify numerical viscosity of a FDS, see e.g. [89]. For the traditional upwind scheme, we may use Harten's entropy fix [25], i.e.,

$$Q(x) = \begin{cases} |x|, & \text{if } |x| > \epsilon, \\ \frac{x^2+\epsilon^2}{2\epsilon}, & \text{if } |x| \leq \epsilon, \end{cases} \tag{6.15}$$

where ϵ is a given positive constant. In fact, after doing that, it is seen that the numerical viscosity of the scheme does not depend on the characteristic direction if $|x| \leq \epsilon$. But, there still exists the shift between the approximate solution and the "exact" solution if $t_n - h > |x| > \epsilon + h$. In other words, we can smear out the non-physical jump or give a smooth transition around the sonic point by adding an entropy fix, but the error within the sonic rarefaction wave cannot be completely eliminated in theory, except that the parameter ϵ is big enough such that $\epsilon + h \geq$

$t_n - h$. It means that the error around the sonic point may become (almost) invisible, if we take a relatively large ϵ. But, resolution of the tail and head of the rarefaction wave will also be decreased. Moreover, numerical solutions will also suffer possibly a loss of monotonicity with respect to x/t along the rarefaction waves. To demonstrate it, in Figs. 29, 30, and 31, we give the numerical solutions, corresponding errors, and the gradients of the numerical solutions calculated by using the Roe scheme with the entropy fix (6.15) for $\epsilon = 0.1$ (left) and $\epsilon = 0.5$ (right), respectively. Obviously, the numerical solutions with an entropy fix have been improved in comparison with one in Fig. 20, in particular, the error around the sonic point is almost invisible in the case of $\epsilon = 0.5$. But numerical dissipation of the scheme has also become larger than one in Fig. 20 such that resolution of the tail and head of the rarefaction waves is decreased. We refer the reader to compare our error plots in Figs. 21 and 30, and

figures given in [80] for more computations.

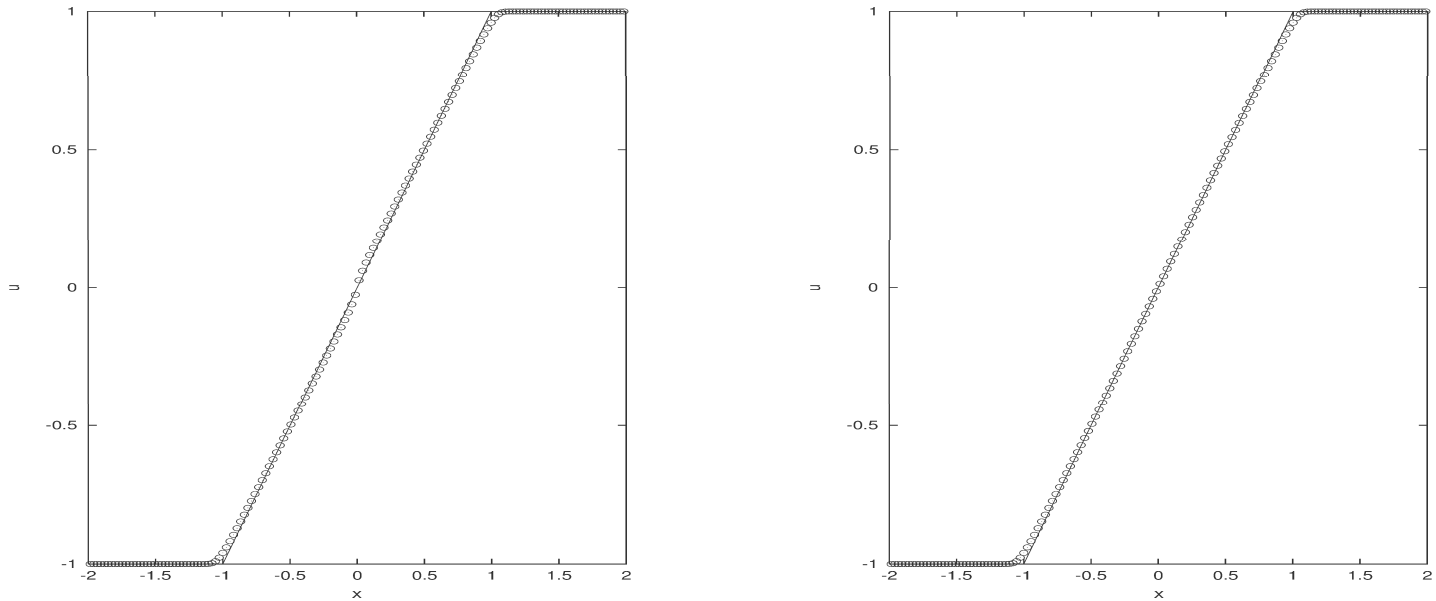


Figure 29: Same as Fig.20, except with an entropy fix. Left: $\epsilon = 0.1$; right: $\epsilon = 0.5$.

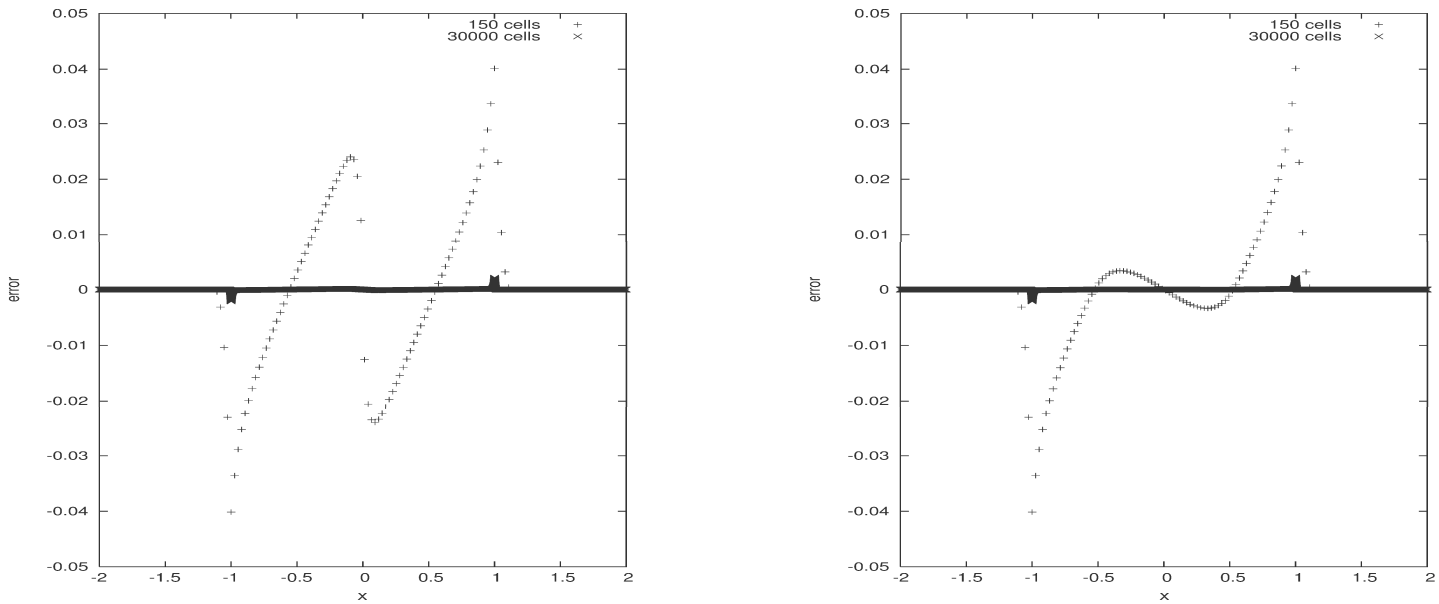


Figure 30: The numerical errors corresponding to Fig. 29. Left: $\epsilon = 0.1$; right: $\epsilon = 0.5$.

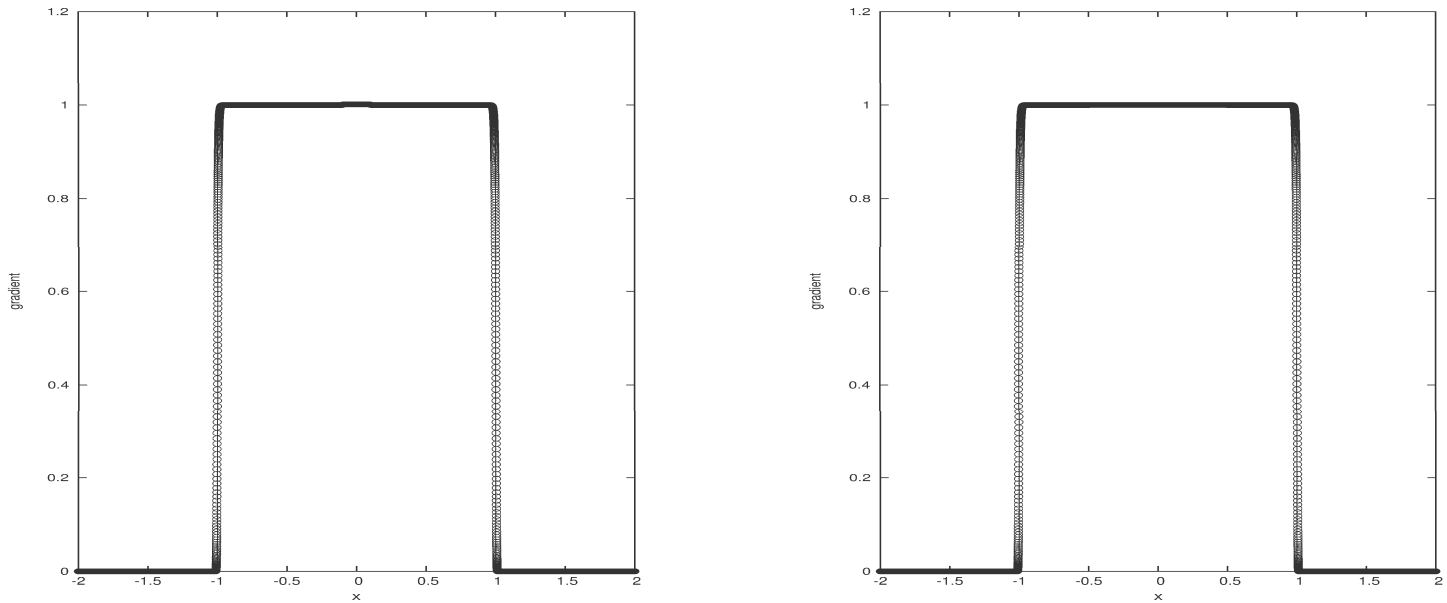


Figure 31: The numerical gradients corresponding to Fig. 29. Left: $\epsilon = 0.1$; right: $\epsilon = 0.5$.

Finally, we consider evolution of the transonic expansion fan by using the MUSCL scheme of van Leer [87] and the TVD scheme of Harten [25].

Under our assumption, the MUSCL scheme advances the solution via the equation

$$u_j^{n+1} = u_j^n - \lambda \left(\hat{f}(u_{j+\frac{1}{2}}^{n,L}, u_{j+\frac{1}{2}}^{n,R}) - \hat{f}(u_{j-\frac{1}{2}}^{n,L}, u_{j-\frac{1}{2}}^{n,R} \right),$$

$$u_{j+\frac{1}{2}}^{n,L} = u_j^n + \frac{h}{2} S_j^n, \quad u_{j+\frac{1}{2}}^{n,R} = u_{j+1}^n - \frac{h}{2} S_{j+1}^n,$$

where $\hat{f}_{j+\frac{1}{2}}$ is any numerical flux of three-point conservative schemes, and $S_j^n \approx \frac{\partial u}{\partial x}$. Here, we take $\hat{f}_{j+\frac{1}{2}}$ as the numerical flux of the Godunov scheme, and S_j^n is defined by

$$S_j^n = \begin{cases} \frac{1}{h} s_{j+\frac{1}{2}} \min\{ s_{j+\frac{1}{2}} (u_j - u_{j-1}), |u_{j+1} - u_j| \}, & (u_j - u_{j-1})(u_{j+1} - u_j) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where $s_{j+\frac{1}{2}} = \text{sign}(u_{j+1} - u_j)$. Within the rarefaction fan away

from its two corners, we have

$$S_j^n = \frac{1}{t_n}.$$

Hence, (6.8) and (6.9) are replaced by

$$u_j^{n+1} = \frac{x_j}{t_n} + \lambda \left(\frac{1}{2} (\frac{x_{j-1}}{t_n} + \frac{h}{2t_n})^2 - \frac{1}{2} (\frac{x_j}{t_n} + \frac{h}{2t_n})^2 \right) = \widetilde{u}_j^e,$$

and

$$u_j^{n+1} = \frac{x_j}{t_n} + \lambda \left(\frac{1}{2} (\frac{x_j}{t_n} - \frac{h}{2t_n})^2 - \frac{1}{2} (\frac{x_{j+1}}{t_n} - \frac{h}{2t_n})^2 \right) = \widetilde{u}_j^e,$$

respectively. It is very obvious that the additional terms in (6.8) and (6.9) have been eliminated. In fact, the numerical viscosity terms in the MUSCL scheme approximating the Burgers' equation have become zero within the transonic expansion fan now.

The actual error of the MUSCL scheme is presented in Fig. 32. We can find that local maximum of $|u_j^{n+1} - u_j^e|$ is only located at two corners of the rarefaction wave.

Harten's second-order accurate TVD scheme in Section 3.4.5 for the inviscid Burgers' equation can be written (3.1)-(3.2) with

$$
\begin{aligned}
\hat{f}_{j+\frac{1}{2}} =& \frac{1}{4}\Big((u_j)^2 + (u_{j+1})^2\Big) + \frac{1}{2\lambda}\Big(g_j + g_{j+1} - |\lambda u + \gamma|_{j+\frac{1}{2}}(u_{j+1} - u_j)\Big), \\
g_j =& s_{j+\frac{1}{2}} \min\{s_{j+\frac{1}{2}}\widetilde{g}_{j-\frac{1}{2}}, |\widetilde{g}_{j+\frac{1}{2}}|\}, \ s_{j+\frac{1}{2}} = \mathrm{sign}(\widetilde{g}_{j+\frac{1}{2}}), \\
\widetilde{g}_{j+\frac{1}{2}} =& \frac{1}{2}\Big(|u| - \lambda u^2\Big)_{j+\frac{1}{2}}(u_{j+1} - u_j), \\
\gamma_{j+\frac{1}{2}} =& \begin{cases} (g_{j+1} - g_j)/(u_{j+1} - u_j), & u_{j+1} \neq u_j, \\ 0, & u_{j+1} = u_j. \end{cases}
\end{aligned}
$$

Within the rarefaction fan away from its two corners, we have

$$g_j = \frac{h}{2\lambda t_n}\min\{\left(|\lambda u| - (\lambda u)^2\right)_{j+\frac{1}{2}}, \left(|\lambda u| - (\lambda u)^2\right)_{j-\frac{1}{2}}\},$$

which depends strongly on the characteristic direction and is highly nonlinear. Thanks to this, it is difficult to give exactly a simple representation of the solution u_j^{n+1}. Here, we only plot the computed error $u_j^{n+1} - u_j^e$ in Fig. 32. The sonic glitch has been formed by the second-order accurate TVD scheme of Harten.

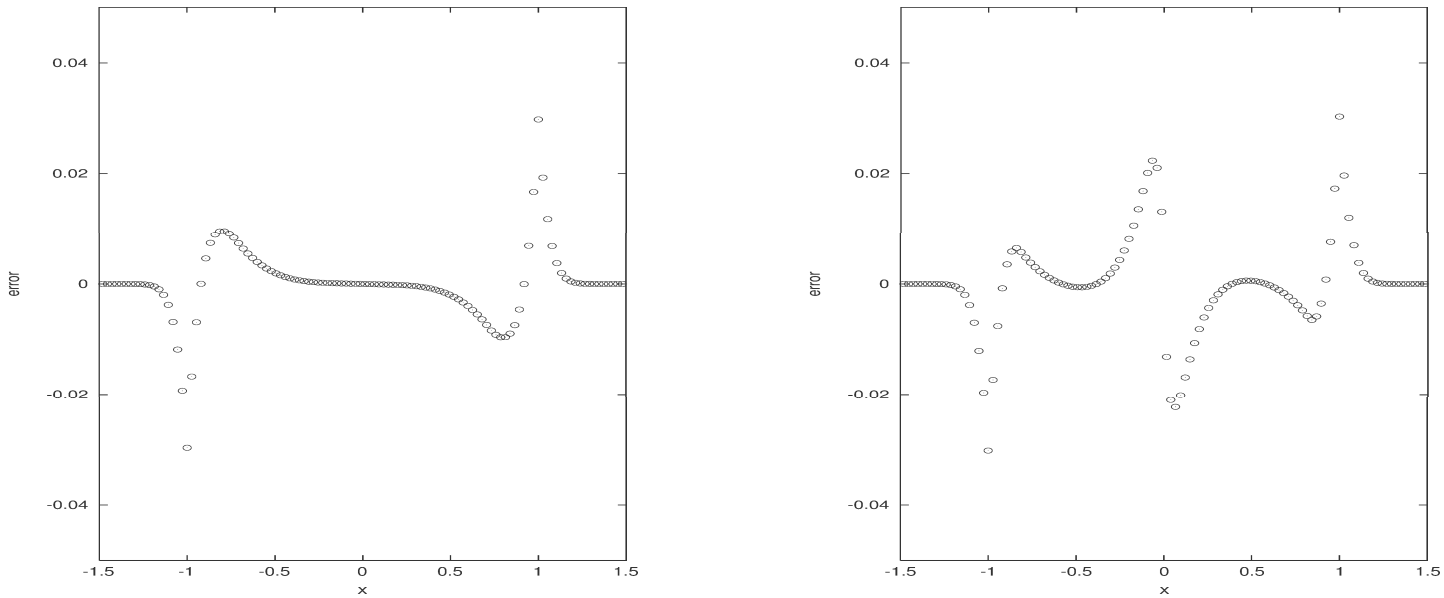


Figure 32: Same as Fig. 20, except for the MUSCL scheme of van Leer (left) and Harten's TVD scheme [25] (right).

From the above analysis, we can conclude that the sonic glitch is not left from any initial discontinuous data of a shock tube problem. Even started from the continuous exact rarefaction

wave, a glitch can still be formed. The sonic glitch affects the upwind flux and not the central difference method, such as the LF scheme and the LW scheme as well as the central scheme with zero viscosity, because they do not depend on the characteristic direction. The sonic glitch has no any direct connection with the violation of the entropy condition or the size of numerical viscosity of a FDS. Glitch may also be formed by several "good" schemes, such as the Godunov scheme and the Engquist-Osher scheme as well as the high resolution scheme of Harten. The sonic glitch is mainly coming from the disparity wave speeds across the sonic point in a transonic expansion fan. If a viscous governing equation is solved by a central difference scheme, there will not have a clear upwind wave propagation direction at a cell interface; and we may expect that the glitch is not formed. A suitable entropy fix can be used to give a smooth transition around the sonic point, but the disparity wave speeds are still existing gen-

erally and it will decrease resolution of two corners of rarefaction wave. The initial data reconstruction technique of van Leer can be used to eliminate the sonic glitch when we solve the inviscid Burgers' equation (1.3) and (3.32).

6.2 Local oscillation in monotone schemes

It is expected that monotone schemes give stable approximations to the scalar hyperbolic conservation law (1.3). The resulting solution satisfies the TVD property, the maximum principle and the entropy inequality as well as monotonicity preserving, see [46]. The monotonicity preserving property only implies that if the initial data is monotone then the solution should have the same property for all time $t > 0$. Theoretically it does not exclude the onset of oscillations at local extrema. It was a common understanding and observation in practice that the numerical vis-

cosity of such schemes may sufficiently offset relative phase errors and suppress oscillations completely. However, to the contrary, local oscillations were observed in [5, 6, 45, 74] and analyzed in [5, 6] by checking the formation of new local extrema in solutions. The purpose of this paper is to analyze and understand this seemingly paradoxical phenomenon by applying the method of discretized Fourier analysis and the modified equation analysis, which will help us to further understand local oscillations caused individually by low and high frequency modes.

Our attention is limited to 3-point FDS for (1.3) in a viscous form (3.89) where the *mesh ratio* $\lambda = \tau/h$ is assumed to be a constant, τ and h are step sizes in time and space, respectively, u_j^n denotes an approximation of $u(jh, n\tau)$, the terms $Q_{j+1/2} \in]0, 1]$ are the *coefficients of numerical viscosity*. If $Q_{j+1/2}^n = 1, \lambda|a_{j+1/2}^n|$, or $(\lambda a_{j+1/2}^n)^2$, (3.89) is namely the LF scheme, the CIR scheme,

or a version of the LW scheme, respectively, where

$$a^n_{j+1/2} = \begin{cases} \frac{f(u^n_{j+1})-f(u^n_j)}{u^n_{j+1}-u^n_j}, & u^n_{j+1} \neq u^n_j, \\ f'(u^n_j), & u^n_{j+1} = u^n_j. \end{cases}$$

The CIR scheme, see Tadmor [71, Page 371], is a generalization of the simplest upwind scheme for linear convection equation ($f(u) = au$ and $Q^n_{j+1/2} \equiv |\lambda a|$) to the case of nonlinear flux functions. In case the coefficients $Q_{j\pm1/2}$ do not depend on the numerical solution u^n_j the condition $\lambda|f'(u^n_{j\pm1})| \leq Q^n_{j\pm1/2} \leq 1$ implies that the scheme (3.89) is monotone. The LF scheme is monotone. For the linear convection equation $f(u) = au$ the upwind scheme is also monotone. On the other hand the LW scheme is non-monotone. If $Q_{j+1/2} = q \in]0, 1[$ is constant with the time step restriction $\max\{\nu|f'(u)|\} \leq q$, the scheme (3.89) is usually called a generalized LF scheme and it is also monotone

under this condition. The special case $q = 1/2$ is the modified LF scheme, see Tadmor [71]. This scheme turns out to be a special case in the analysis below, see also [6].

Numerical oscillations caused by the LW scheme are quite well understood, see [52, Page 100]. Assume that $f(u) = au$, i.e. $a^n_{j+1/2} = a$, and $Q^n_{j+1/2} \equiv q$ are constant, for all $j \in \mathbb{Z}$. The relative phase error of low frequency modes $u^n_j = \lambda^n_k e^{i2\pi kjh} = \lambda^n_k e^{i\xi j}$, for kh or scaled wave number $\xi = 2\pi kh$ small and $i = \sqrt{-1}$, is of order $\mathcal{O}(kh) = \mathcal{O}(\xi)$, and cannot be offset by the small dissipation of order $\mathcal{O}((kh)^2) = \mathcal{O}(\xi^2)$. As is well-known, the LW scheme is second order accurate both in time and space. In contrast, for first order accurate monotone schemes, the numerical viscosity is enough to control numerical oscillations caused by the relative phase errors of low frequency modes because the dissipation, i.e. amplitude error, has a magnitude of order higher than that of

the relative phase errors. However, for high frequency modes, the situation is quite different. the highest frequency mode, *a checkerboard mode*, is taken as the initial data

$$u_j^0 = (-1)^j, \tag{6.16}$$

and catch a glimpse of the resolution of high frequency modes by (3.89). The solution of the 3-point scheme (3.89) for constant q (i.e., the generalized LF scheme) at $t = n\tau$ is easily seen to be

$$u_j^n = (1 - 2q)^n (-1)^j. \tag{6.17}$$

This shows that the solution of (3.89) at time $t = n\tau$ for $n \in \mathbb{N}$ is still a checkerboard mode, except for the modified LF scheme with $q = 1/2$. The amplitude of the solution is diminishing if $0 < q < 1$ but keeps invariant for the LF scheme with $q = 1$ or for the unstable central scheme with $q = 0$. For $q = 1/2$ it is

wiped out immediately. The typical checkerboard mode usually contaminates the solution and causes unwanted oscillations in computations. The above rough analysis motivates us to investigate the interrelation among oscillations, relative phase errors and numerical dissipations, i.e. numerical viscosity and numerical damping, systematically.

We see that numerical oscillations in the second order accurate LW solution are caused by the relative phase error of low frequency modes; while local observable oscillations in first order accurate monotone schemes are caused by high frequency modes, particularly of the form (6.16). Numerical viscosity is not sufficient to offset the oscillations by the high frequency modes. In nonlinear cases the nonlinearity of the flux function $f(u)$ does not have any effect on this matter.

Due to the role of the frequency of the Fourier modes in the

oscillations, we consider the discretization of general initial data

$$u(x,0) = u_0(x), \tag{6.18}$$

in different ways, which possibly contain the checkerboard mode (6.16). It becomes obvious from the analysis of the discretization of a single square signal with an odd or even number of grid points: The former contains the highest frequency checkerboard mode (6.16) and the resulting solution displays the oscillatory phenomenon; but the latter displays exactly the expected non-oscillatory behavior of a monotone scheme, cf. [5, 6].

In order to understand the resolution of Fourier modes of different frequencies by the scheme (3.89), we proceed to use the discrete Fourier analysis to find that the relative phase error of high frequency modes is of order $\mathcal{O}(1)$ and causes severe oscillations unless there exists a strong numerical dissipation to suppress these errors of the high frequency modes. To further

understand this observation, we will respectively consider the smooth solution $(U^s)_j^n$ corresponding to low frequency modes and the oscillatory solution $(U^0)_j^n$ related to high frequency modes, and check their respective modified equations. As in [52], for the smooth part, the modified equation has the familiar form

$$\partial_t U^s + a\partial_x U^s = \alpha_1^s(q,a,h)\tau\partial_x^2 U^s + \alpha_2^s(q,a,h)\tau^2\partial_x^3 U^s + \cdots .$$

While for high frequency modes, the modified equation is of the form

$$\partial_t U^o + a\partial_x U^o = \frac{\ln|2q-1|}{\tau}U^o + \alpha_1^o(q,a,h)\tau\partial_x^2 U^o + \alpha_2^o(q,a,h)\tau^2\partial_x^3 U^o + \cdots ,$$

where $q \neq 1/2$. Here α_j^s, α_j^o, $j = 1, 2, \cdots$, are all uniformly bounded functions. The zero order term $\frac{\ln(|2q-1|)}{\tau}U^o$ is numerical damping and has the order $\mathcal{O}(1)$. This term displays a stronger dissipative effect, compared with the numerical viscosity

$\alpha_1^o(q,a,h)\tau\partial_x^2 U^o$. Here it exerts dominant effects of dissipation on the high frequency modes. For the LF scheme ($q = 1$) the damping term vanishes and numerical viscosity, the second order term, $\alpha_1^o(q,a,h)\tau\partial_x^2 U^o$ takes effect of dissipation of order $\mathcal{O}(\xi)$, thus it is not sufficient to offset the relative phase error of order $\mathcal{O}(1)$.

Thus we can clarify the oscillations in the solutions: *The oscillations are caused by large phase errors and they are not offset by the numerical dissipation of the same order. For low frequency modes, the relative phase errors are offset by the numerical viscosity of lower order. For high frequency modes, the error is of order $\mathcal{O}(1)$, and thus the numerical viscosity is not enough to dampen the resulting oscillations and numerical damping is important in this aspect.*

This section is organized into several parts. Section 6.2.1 presents two examples to display oscillations. Section 6.2.2 takes

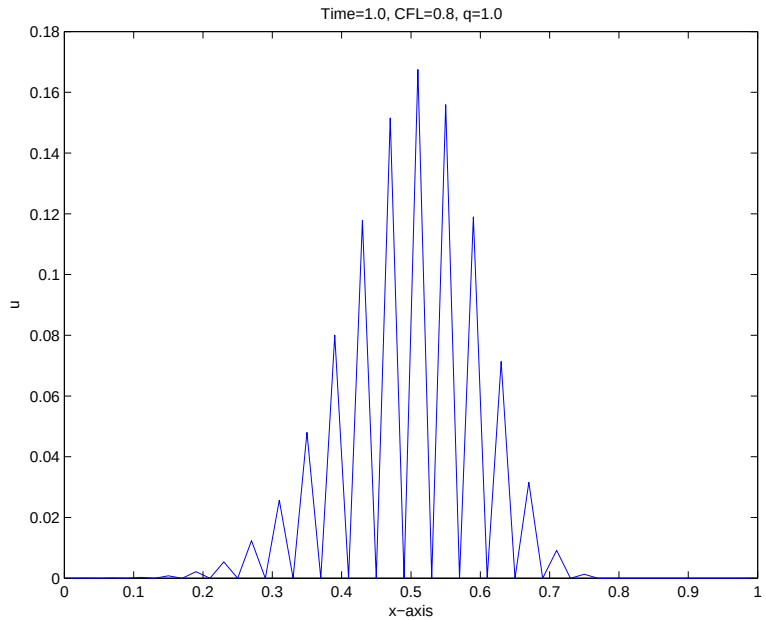
two different ways to discretize initial data and find checkerboard modes in the presence of initial discrete values. Then the generalized LF scheme is used to glimpse at the resolution of Fourier modes in Section 6.2.5. The main analysis is made in Section 6.2.6 to investigate the dissipative and dispersive mechanisms by using linear convection equation.

6.2.1 Local oscillations in generalized LF schemes

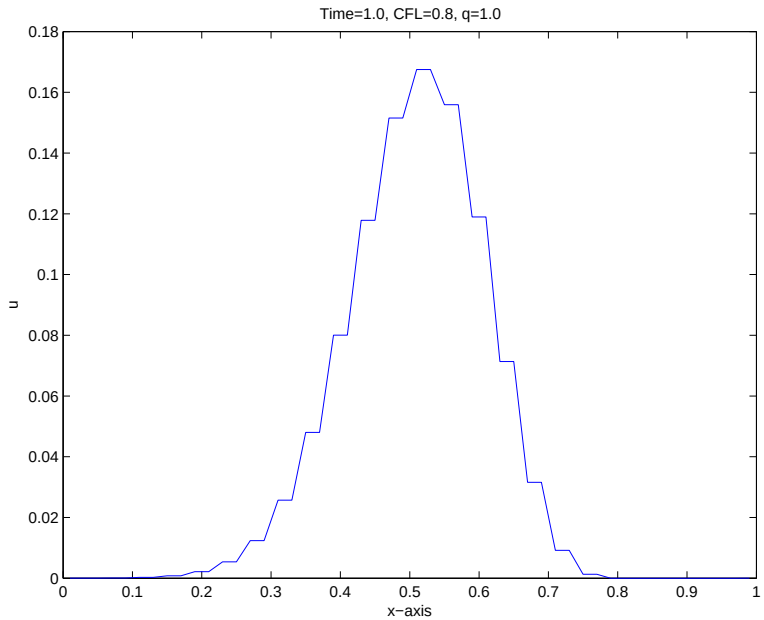
The fact that the second-order accurate LW scheme produces oscillations is well-known. This section presents two examples to display local oscillations in the solution of (1.3) by the first-order accurate generalized LF schemes, as observed in [5, 6, 74], although the schemes are monotone and under a certain restriction. From these examples we see that the different ways of initial discretization lead to distinct solution behaviors. This motivates

the analysis of the discretization of initial data in the next section.

Example 6.1 (Linear convection equation) *Consider the equation* (2.1) *with* $a = 1$ *over the region* $x \in [0, 1]$ *by using the LF scheme. We take the grid points* $M = 50$, $\nu = \tau/h = 0.8$, *and use periodic boundary condition just for simplicity, and first look at the impulsive initial data* $u_j^0 = 1$ *for* $j = M/2$, *and* $u_j^0 \equiv 0$ *otherwise. The numerical solution is shown in Fig.* 33*(a) to display clear oscillations. Note that the total variation keeps invariant* 2. *Then we investigate the case that distributed square pulse initial data* $u_j^0 = 1$ *for* $j = M/2,\ M/2 + 1$ *and* $u_j^0 \equiv 0$ *otherwise are taken. The numerical solution displays exactly the opposite behavior. No oscillation is present, see Fig.* 33*(b). However the total variation is decreasing to be* 0.3398 *at time* $t = 1$.



(a) $u_j^0 = 1$ for $j = 25$ and $u_j^0 = 0$ otherwise

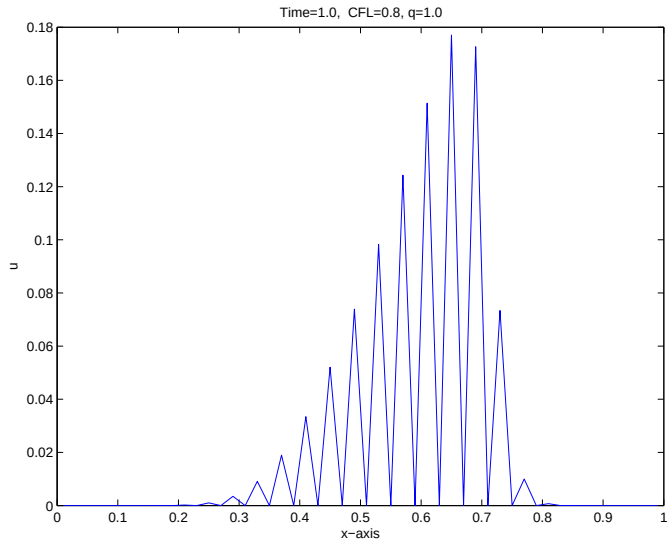


(b) $u_j^0 = 1$ for $j = 25$, 26, and $u_j^0 = 0$ otherwise

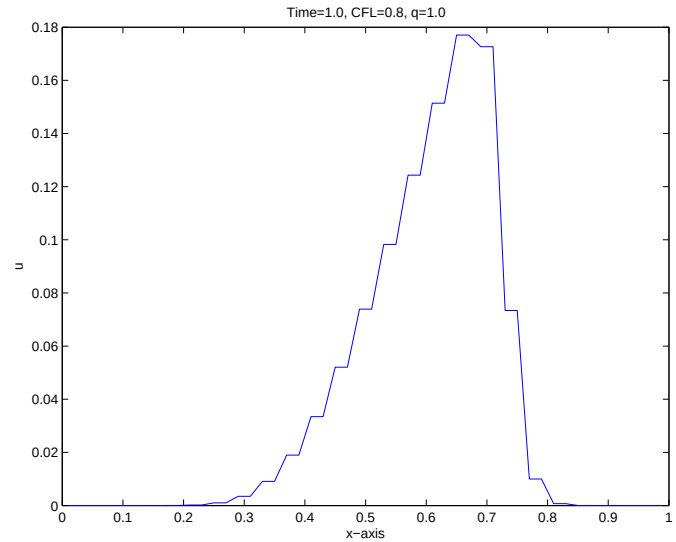
Figure 33: Numerical solutions for Example 6.1 with 50 grid points and $\lambda = 0.8$.

Example 6.2 (Burgers' equation) *In order to see if the above*

oscillatory phenomenon is strongly influenced by the nonlinearity, we use the inviscid Burgers' equation (6.1) *as a nonlinear example. The same initial data are taken as for the above linear case. Similar numerical results are displayed in Fig. 34. Therefore, the oscillations are not connected to the nonlinearity, just as we pointed out for checkerboard mode in the introduction. This result is in a sharp contrast with the fact that the nonlinearity introduces an additional effect of dissipation at discontinuities, such as the step function data, that can be observed when comparing solutions to linear convection and the Burgers' equation.*



(a) $u_j^0 = 1$ for $j = 25$ and $u_j^0 = 0$ otherwise



(b) $u_j^0 = 1$ for $j = 25$, 26, and $u_j^0 = 0$ otherwise

Figure 34: Numerical solutions for Example 6.2 with 50 grid points.

We observe that the presence of oscillations is not only related to the ways in which the initial data are approximated, but

also to the numerical viscosity coefficient q. For the LF scheme, the local oscillations are very strong although the total variation property is not violated due to a strong decay of numerical solution amplitude.

6.2.2 Checkerboard modes in the initial discretization

As observed in the last section and also in [5], the numerical solutions display very distinct behaviors if the initial data are discretized in different ways. It turns out that checkerboard modes are present and affect the solutions if the initial data contains a square signal and are discretized with an odd number of grid points. This motivates us to discuss the discretization of initial data

$$u(x,0) = u_0(x), \quad x \in [0,1], \tag{6.19}$$

with M grid points and $h = 1/M$. For simplicity, we assume that M is even, and $u_0(0) = u_0(1)$. The numerical solution value at the grid point x_j is denoted by u_j^0. This grid point value u_j^0 is expressed by using the usual discrete Fourier sums, as in [79, Page 120]. We use the *scaled wave number* $\xi = 2\pi kh$ and obtain

$$u_j^0 = \sum_{k=-M/2+1}^{M/2} c_k^0 e^{i\xi j}, \quad i^2 = -1, \quad j = 0, 1, \cdots, M-1, \quad (6.20)$$

where the coefficients c_k^0 are, in turn, expressed as

$$c_k^0 = \frac{1}{M} \sum_{j=0}^{M-1} u_j^0 e^{-i\xi j}, \qquad k = -M/2+1, \cdots, M/2. \quad (6.21)$$

A special attention is paid to the particular case that

$$c_k^0 = \begin{cases} 1, & \text{if } k = M/2, \\ 0, & \text{otherwise,} \end{cases}$$

i.e. the initial data are taken to be just the highest Fourier mode which is a single checkerboard mode oscillation

$$u_j^0 = e^{i2\pi \frac{M}{2} jh} = e^{i\pi j} = (-1)^j.$$

6.2.3 Single square signal case

We start with the simple initial data of a square signal, i.e. an initial function of the following type

$$u(x,0) = \begin{cases} 1, & 0 < x^{(1)} < x < x^{(2)} < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{6.22}$$

The following two ways are taken to approximate the step function (6.22) as a grid function: One uses an odd number of grid points to take the value one of the square signal and the other uses the next smaller even number of grid points. They could be seen as approximations to some given fixed interval with end points not represented on the mesh, which we do not explicitly specify here. The discrete Fourier sum is used to clarify the difference.

(i) Discretization with an odd number of grid points. Take $j_1, j_2 \in \mathbb{N}$ such that $j_1 + j_2$ is an even number. We set $x^{(1)} = (\frac{M}{2} - j_1)h$ and $x^{(2)} = (\frac{M}{2} + j_2)h$, and first discretize the square signal (6.22) with $p := j_1 + j_2 + 1$ nodes, i.e. an odd number of grid points, such that

$$u_j^0 = \begin{cases} 1, & \text{if } \ j = M/2 - j_1, \cdots, M/2 + j_2, \\ 0, & \text{otherwise.} \end{cases} \tag{6.23}$$

Substituting them into (6.21) gives by simple calculation,

$$c_k^0 = h \sum_{j=0}^{M-1} u_j^0 e^{-i\xi j} = \begin{cases} \frac{(-1)^k e^{i\xi j_1}(1-e^{-i\xi p})}{M(1-e^{-i\xi})}, & \text{for } k \neq 0, \\ ph, & \text{for } k = 0. \end{cases} \tag{6.24}$$

Special attention is paid to the term

$$c_{M/2}^0 = (-1)^{j_1+M/2} h,$$

since M is even and p is odd. Hence the initial data (6.23) can be expressed in the form

$$u_j^0 = (-1)^{j+j_1+M/2} h + ph + \sum_{k\neq 0,M/2} \frac{(-1)^k e^{i\xi(j+j_1)}(1-e^{-i\xi p})}{M(1-e^{-i\xi})}. \tag{6.25}$$

(ii) Discretization with an <u>even</u> number of grid points. Rather than (i) above, we use $p := j_1 + j_2$ even number of grid

points to express the square signal in (6.22) as follows

$$u_j^0 = \begin{cases} 1, & \text{if } j = M/2 - j_1 + 1, \cdots, M/2 + j_2, \\ 0, & \text{otherwise.} \end{cases}$$

Then we substitute these initial data into (6.21) to obtain

$$c_k^0 = \begin{cases} \frac{(-1)^k e^{i\xi(j_1-1)}[1-e^{-i\xi(p-1)}]}{M(1-e^{-i\xi})}, & \text{for } k \neq 0, \\ (p-1)h, & \text{for } k = 0, \end{cases}$$

and $c_{M/2}^0 = 0$. The initial data can be written by using the discrete Fourier sums as

$$u_j^0 = 0 \times (-1)^j + (p-1)h + \sum_{k \neq 0, M/2} \frac{(-1)^k e^{i\xi(j+j_1-1)}[1 - e^{-i\xi(p-1)}]}{M(1 - e^{-i\xi})}. \tag{6.26}$$

Comparing (6.25) with (6.26), we observe an essential difference lies in the fact that a checkerboard mode $(-1)^{j+j_1+M/2}$ is present in (6.25), but it is filtered out in (6.26). This is closely related to the oscillatory phenomenon observed in [5, 6, 74].

6.2.4 Step function initial data case

For more general piecewise constant initial data (6.19), there are analogous discrete Fourier sum expressions. The computational domain $[0, 1]$ is divided into L subintervals I_l $(l = 1, 2, \cdots, L)$, $\bigcup_{l=1}^{L} I_l = [0, 1]$, the number of the discrete points of a subinterval I_l is M_l, $M_1 + M_2 + \cdots + M_L = M$, and the initial data (6.19) are expressed as

$$u_j^0 = \sum_{l=1}^{L} \bar{U}_0^l \cdot \chi_l(j), \tag{6.27}$$

where $\bar{U}_0^l$ are constants, and $\chi_l(j)$ is the characteristic function on I_l,

$$\chi_l(j) = \begin{cases} 1, & \text{if } x_j \in I_l, \\ 0, & \text{otherwise .} \end{cases}$$

Note that (6.27) can be regarded as the superposition of several single square signals of the form (6.22). Then we express (6.27) as a discrete Fourier sum of the form (6.20) with c_k^0. For $k \neq 0$,

$M/2$, we have

$$
\begin{aligned}
c_k^0 =& \frac{1}{M} \sum_{l=1}^{L} \bar{U}_0^l \left[\sum_{j=0}^{M-1} \chi_l(j) e^{-i\xi j} \right] \\
=& \frac{1}{M} \left[\bar{U}_0^1 \sum_{j=0}^{M_1-1} e^{-i\xi j} + \bar{U}_0^2 \sum_{j=M_1}^{M_1+M_2-1} e^{-i\xi j} + \cdots + \bar{U}_0^L \sum_{j=p_{L-1}}^{M-1} e^{-i\xi j} \right] \\
=& \frac{1}{M(1-e^{-i\xi})} \sum_{l=1}^{L} \bar{U}_0^l (e^{i\xi M_l} - 1) e^{-i\xi p_l}, \qquad (6.28)
\end{aligned}
$$

where $p_l = M_1 + \cdots + M_l$; and for $k = 0$, $M/2$, we have

$$
\begin{aligned}
c_0^0 &= \tfrac{1}{M} (\bar{U}_0^1 M_1 + \bar{U}_0^2 M_2 + \cdots + \bar{U}_0^L M_L), \\
c_{M/2}^0 &= \tfrac{1}{M} \sum_{l=1}^{L} \bar{U}_0^l \left[\sum_{j=0}^{M-1} \chi_l(j) (-1)^j \right].
\end{aligned}
$$

Thus, the initial data are expressed as

$$u_j^0 = \frac{1}{M}\sum_{l=1}^{L}\bar{U}_0^l\left[\sum_{m=0}^{M-1}\chi_l(m)(-1)^m\right](-1)^j + \frac{1}{M}(\bar{U}_0^1 M_1 + \bar{U}_0^2 M_2 + \cdots + \bar{U}_0^L M_L)$$
$$+ \frac{1}{M}\sum_{k\neq 0,M/2}\frac{1}{(1-e^{-i\xi})}\sum_{l=1}^{L}\bar{U}_0^l(e^{i\xi M_l}-1)e^{-i\xi p_l}. \qquad (6.29)$$

Similar to the case of a single square signal, it depends on $c_{M/2}^0$ whether there is a checkerboard mode in the discrete initial data. Therefore, we have three cases here too.

(i) If the number M_l of the discrete points in each I_l is odd, $c_{M/2}^0$ is

$$c_{M/2}^0 = \frac{1}{M}\sum_{l=1}^{L}\bar{U}_0^l\left[\sum_{j=0}^{M-1}\chi_l(j)(-1)^j\right] = \frac{1}{M}\sum_{l=1}^{L}\bar{U}_0^l(-1);$$

(ii) If the number M_l of the discrete points in each I_l is even, $c^0_{M/2}$ vanishes,

$$c^0_{M/2} = \frac{1}{M} \sum_{l=1}^{L} \bar{U}^l_0 \left[\sum_{j=0}^{M-1} \chi_l(j)(-1)^j \right] = 0.$$

(iii) If the number of the discrete points in some I_l is odd while in the others it is even, the summation in the even case is zero and in the odd case is 1 or -1. That is $c^0_{M/2} = \frac{1}{M} \sum_{l=1}^{L} \bar{U}^l_0 \phi(l)$, where

$$\phi(l) = \begin{cases} 0, & \text{if } I_l \text{ is in the even case,} \\ 1 \text{ or } (-1), & \text{if } I_l \text{ is in the odd case.} \end{cases}$$

This case is just the superposition of (i) and (ii) above.

It is observed that there is no checkerboard mode for Case (ii). For Case (i), this summation may be zero when the factors cancel. However, since this summation is taken in the global sense and the checkerboard mode exists in each subinterval, the solution may still contain oscillations due to the finite propagation speed property of the scheme. To some extent, this is analogous to the fact that the scheme (3.89) is TVD under a certain restriction, but that local oscillations are still observed. All of the above analysis are summarized in the following proposition.

Proposition 6.1 *Suppose that the initial data* (6.19) *are given as, or approximated by, a step function. They can be approximated as the superposition of several single square signals. For each square signal we have two different types of discretization. If they are discretized with an odd number of grid points, the checkerboard, i.e. highest frequency, mode is present. In contrast, if they*

are discretized with an even number of grid points, this mode is suppressed.

6.2.5 A glimpse of checkerboard mode propagation

This section simply looks at the (generalized) LF scheme for the linear convection equation (2.1)

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{q}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \quad (6.30)$$

and catch a glimpse of the resolution of high frequency modes, where q under the condition $|\lambda a| \leq q \leq 1$ is a parameter for monotone schemes. If $q = 1$ then (6.30) is the LF scheme; taking $q = |\lambda a|$ it is the upwind scheme; and for $q = a^2\lambda^2$ being smaller than the monotonicity range it is the non-monotone LW scheme.

As usual for stability analysis, the solution to (6.30) is expressed analogously to (6.20) in the standard form of a discrete

Fourier sum using $\xi = 2\pi kh$

$$u_j^n = \sum_{k=-M/2+1}^{M/2} c_k^n e^{i\xi j}. \tag{6.31}$$

The coefficients c_k^n are obtained successively and expressed as follows

$$c_k^n = (1 + q(\cos\,\xi - 1) - i\lambda a \sin\,\xi)^n\, c_k^0. \tag{6.32}$$

In correspondence with the two kinds of discretization of a single square signal in Section 6.2.2, the Fourier coefficients c_j^0 have different expressions, and the solutions are expressed, respectively, as follows.

(i) Odd discretization case. With the initial data (6.25), the

solution of (6.30) is

$$u_j^n = \frac{1}{M}(1-2q)^n(-1)^{j+j_1+M/2} + \sum_{k=-M/2+1}^{M/2-1} c_k^n e^{i\xi j}. \qquad (6.33)$$

(ii) Even discretization case. With the initial data (6.26), we have

$$u_j^n = 0 \times (1-2q)^n(-1)^j + \sum_{k=-M/2+1}^{M/2-1} c_k^n e^{i\xi j}. \qquad (6.34)$$

We emphasize that the Fourier coefficients c_k^n have different expressions in correspondence with the odd and even number of nodes taken for the discretization of the signal in the initial data. By comparing (6.33) and (6.34), we see that in the odd case the checkerboard mode does not vanish if it exists initially, unless we

have the modified LF scheme $q = 1/2$, although it decays with a rate of $|2q - 1|$ at each time step, see Fig. 35. Therefore, a proper discretization of initial data (6.19) would be important to suppress these oscillations in the numerical solution of (1.3). This may not be feasible in real flow applications.

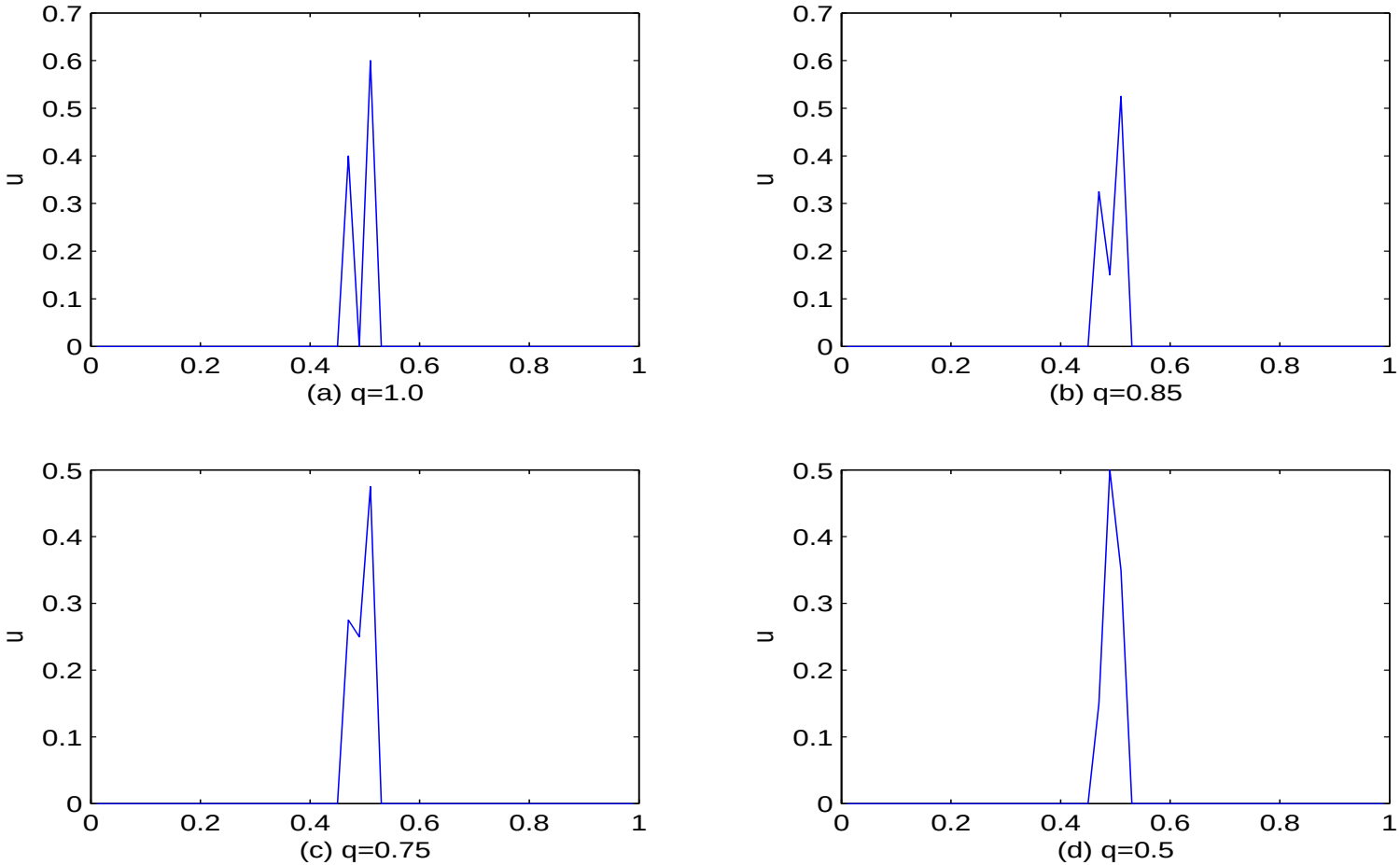


Figure 35: The decay of oscillations as the parameter q decreases: The initial data are the same as for Fig. 33(a), the CFL number is 0.2 and only one time step is taken.

Remark 6.3 *(1) For the LF scheme with $q = 1$ the solution $u_j^n = (-1)^{n+j}$ oscillates between 1 and -1 alternately if the checkerboard mode initial data are taken. The large numerical dissipation does not have any effect.*

(2) In case $q \neq 1$, i.e. $|1 - 2q| < 1$, the checkerboard mode is damped out quickly. In particular, for the modified LF scheme $q = 1/2$ the checkerboard mode is eliminated and has no influence on the solution at all. We will analyze this in Section 6.2.6.

Remark 6.4 *In [74] a $(2l+1)$-point scheme of the following type was investigated*

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2l}\big(f(u_{j+l}^n) - f(u_{j-l}^n)\big) + \frac{\tau\alpha}{2lh}(u_{j+l}^n - 2u_j^n + u_{j-l}^n),$$

where $\alpha = \max_u\{|f'(u)|\}$ and l is a positive integer. This scheme with $l = 1$ is regarded as a generalized LF scheme. Especially in

the linear case when $f(u) = au$ this is the upwind scheme. Using the similar analysis as above, we obtain the following solution if we take the square pulse initial data,

$$u_j^n = \left\{1 + \frac{\alpha\lambda}{l}\left[(-1)^l - 1\right]\right\}^n (-1)^j + \sum_{k=-M/2+1}^{M/2-1} c_k^n e^{i\xi j},$$

where $c_k^n = [1+\frac{\alpha\lambda}{l}(\cos(\xi l)-1)-ia\lambda\sin(\xi l)]c_j^0$. There is a checkerboard mode in the solution and naturally the solution displays such oscillations.

6.2.6 Numerical dissipation and phase error

This section attempts to analyze the numerical dissipation and phase error mechanisms of (6.30), particularly on high frequency modes and explain the phenomenon of oscillations caused by high

frequency modes. The discrete Fourier analysis and the method of modified equation analysis are applied. Both of them give complimentary results, which are consistent. The results show that as $0 \leq q \leq 1$ is away from $1/2$, the damping on high frequency modes becomes weak. In particular, there is no damping effect in the LF scheme (i.e. $q = 1$) and the unstable central scheme (i.e. $q = 0$). It is noted that in [6] the dependence of oscillatory properties on the numerical viscosity was also investigated.

6.2.7 Discrete Fourier analysis

This section uses the method of discrete Fourier analysis to discuss the dissipation and phase error mechanism of the generalized LF scheme (6.30), which is monotone if $0 < |\lambda a| < q \leq 1$. We are particularly concerned with the phase accuracy of Fourier modes. Denote a Fourier mode using the scaled wave number $\xi = 2\pi kh$

by $e^{i\xi}$. Then using it as initial data for a linear FDS results in the solution at the nth time level

$$u_k^n = \lambda_k^n e^{i\xi} = (\lambda(k))^n \, e^{i\xi}, \quad i^2 = -1, \tag{6.35}$$

where λ_k^n is the amplitude. The modulus of the ratio

$$\lambda(k) = \lambda_k^{n+1}/\lambda_k^n$$

is the *amplitude* of the mode for one time step. For the scheme (6.30) we have with $\lambda = \tau/h$ in particular

$$\lambda(k) = 1 + q(\cos\xi - 1) - i\lambda a \sin\xi \tag{6.36}$$

and

$$\begin{aligned} |\lambda(k)|^2 =& \big(1 + q(\cos\xi - 1)\big)^2 + (\lambda a)^2 \sin^2\xi \\ =& 1 + 4(a^2\lambda^2 - q)\sin^2(\xi/2) + 4(q^2 - a^2\lambda^2)\sin^4(\xi/2). \end{aligned} \tag{6.37}$$

For the upwind scheme with $q = \lambda|a|$ the last term on the right hand side vanishes, whereas for the LW scheme with $q = \lambda^2 a^2$ the second term vanishes. Also we see that for the schemes of type (6.30), under the conditions $0 < \lambda^2 a^2 \leq q \leq 1$, we have from (6.37) the estimate

$$|\lambda(k)|^2 = 1+4(a^2\lambda^2-q)\left(\sin^2(\xi/2) - \sin^4(\xi/2)\right)+4q(q-1)\sin^4(\xi/2) \leq 1. \tag{6.38}$$

Thus these conditions imply that the schemes are linearly stable. Conversely, assume that $|\lambda(k)|^2 \leq 1$ then from (6.37) we have $(a^2\lambda^2 - q) + (q^2 - a^2\lambda^2)\sin^2(\xi/2) \leq 0$, which further implies that $(a^2\lambda^2 - q) + [q^2 - a^2\lambda^2 + q(q-1)]\sin^2(\xi/2) \leq 0$ must hold for $\xi \in [0, \pi]$. Now for $\xi = 0$ we obtain $a^2\lambda^2 \leq q$. For $\xi = \pi$ this gives $q(q-1) \leq 0$ or since we have $0 < a^2\lambda^2 < q$ we obtain $q \leq 1$. Therefore, these conditions are necessary and sufficient for stability.

The exact solution of the Fourier mode $e^{i\xi}$ for $x = h$ after one time step τ is $e^{i(\xi - 2\pi ak\tau)} = e^{-i2\pi ak\tau}e^{i\xi} = \lambda_{exact}(k)e^{i\xi}$. The exact amplitude $\lambda_{exact}(k)$ has modulus 1. We see from (6.37) that the *amplification error*, i.e. the error in amplitude modulus, is of order $\mathcal{O}(\xi)$ for the monotone schemes and order $\mathcal{O}(\xi^2)$ for the LW scheme. If the modulus of $\lambda(k)$ is less than one, the effect of the multiplication of a solution component with $\lambda(k)$ is called *numerical dissipation* and then the amplification error is called *dissipation error*. If the modulus is larger than 1, this leads to the amplification of the Fourier mode, i.e. instability of any solution containing it. Further, comparing the exponents of $\lambda(k)$ and $\lambda_{exact}(k)$ there is a *phase error* $\arg\lambda(k) - (-2\pi ak\tau)$. The *relative phase error* is then defined as

$$E_p(k) := \frac{\arg\lambda(k)}{-2\pi ak\tau} - 1 = -\frac{\arg\lambda(k)}{\lambda a\xi} - 1.$$

A mode is a low frequency mode if $\xi \approx 0$ and a high frequency mode if $\xi \approx \pi$. We first look at the low frequency modes

$$(U^s)_j^n := \lambda_k^n e^{i\xi j}, \quad \xi \approx 0.$$

For $k = \xi = 0$ we have $\lambda(k) = 1$. From (6.37) we obtain

$$\frac{d(|\lambda(k)|^2)}{dq} = 2(1 + q(\cos\xi - 1))(\cos\xi - 1) < 0, \tag{6.39}$$

for fixed $\xi \in]0, \pi/2]$. This implies that the dissipation becomes weaker as q decreases. The LF scheme with $q = 1$ has the largest numerical dissipation for low frequency modes.

The phase of the low frequency modes is approximated by Taylor expansion at $\xi = 0$

$$\arg\lambda = \arctan\left(\frac{-\lambda a \sin\xi}{1 + q[\cos\xi - 1]}\right) \approx -\lambda a\xi\left(1 + \frac{3q - 1 - 2\lambda^2 a^2}{6}\xi^2 + \cdots\right). \tag{6.40}$$

This phase has a relative error $E_p(k)$ of order $\mathcal{O}(\xi^2)$. For the LW scheme, this phase error causes oscillations, which cannot be suppressed by the weaker dissipation of order $\mathcal{O}(\xi^2)$, compared to the dissipation error $\mathcal{O}(\xi)$ of the upwind scheme.

For high frequency modes (6.35), $\xi \approx \pi$, the situation is very different. We introduce the decomposition $\xi = \pi + \xi'$, i.e. $\xi' = 2\pi k' h$ with $kh = 1/2 + k'h$, and thus $\xi' \approx 0$. The modes is written in the form

$$(U^h)_j^n = \lambda_k^n e^{i\xi j} = \lambda_k^n e^{i(\pi+\xi')j} = (-1)^{j+n} \lambda_{k'}^n e^{i\xi' j}, \tag{6.41}$$

with $\lambda_{k'}^n = (-1)^{j+n} e^{i\pi j} \lambda_k^n$ and set

$$(U^o)_j^n := \lambda_{k'}^n e^{i\xi' j}.$$

The factor $(U^o)_j^n$ can be regarded as a perturbation amplitude of the checkerboard modes $(e^{i\pi})^{j+n} = (-1)^{j+n}$. The dissipation

(amplitude error) depends only on $\lambda_{k'}^n$. Then substituting $(U^h)_j^n$ into (6.30) yields

$$\lambda' := \lambda_{k'}^{n+1}/\lambda_{k'}^n = -1 + q(1+\cos\xi') - i\lambda a \sin\xi'.$$

Therefore, we have

$$\begin{aligned} |\lambda'|^2 =& (1 - q(1+\cos\xi'))^2 + \lambda^2 a^2 \sin^2\xi' \\ =& 1 + 4(a^2\lambda^2 - q)\cos^2(\xi'/2) + 4(q^2 - a^2\lambda^2)\cos^4(\xi'/2). \end{aligned} \tag{6.42}$$

This is consistent with a shift of π in the variable ξ in (6.37). Regarding the high frequency modes, for all schemes the amplitude error is $\mathcal{O}(1)$. At $\xi' = 0$ we have

$$|\lambda'|^2 = 1 - 4q(1-q), \tag{6.43}$$

so we have the lowest amplitude error for $q = 1$ or near zero, the highest for the modified LF scheme with $q = 1/2$. Obviously, for small ξ', $|\lambda'|^2$ is an increasing function of q if $q > 1/2$ because

$$d(|\lambda'|^2)/dq = -2[1 - q(1 + \cos(\xi'))](1 + \cos\xi') > 0. \qquad (6.44)$$

That is, (6.30) becomes much more dissipative for high frequency modes as the parameter q decreases, which is in sharp contrast with the situation for low frequency modes, see (6.39).

Note that for the LW scheme with $q = \lambda^2 a^2$ and the upwind scheme with $q = \lambda|a|$ the situation depends on the choice of the CFL number $\lambda|a|$ or equivalently the step size for the computation. We disregard the singular case of CFL number 1 in which the schemes reproduce the exact solution for linear fluxes. But for CFL numbers near one they are not very dissipative for the high frequency modes, whereas this is a strong dissipation for a CFL number giving $q = 1/2$.

Furthermore, let us look at the relative phase error. We compute

$$\begin{aligned}\arg\lambda' =& \tan^{-1}\left(\frac{-\lambda a\sin\xi'}{-1+q(1+\cos\xi')}\right)\\ =& \frac{-\lambda a\xi'}{2q-1} - \frac{\lambda a}{3(2q-1)^2}\left[\frac{q+1}{2} - \frac{\lambda^2 a^2}{2q-1}\right]\xi'^3 + \mathcal{O}(\xi'^5).\end{aligned} \tag{6.45}$$

Then for the high frequency modes $(U^h)_j^n$, we have by recalling that $\xi = \pi + \xi'$

$$\begin{aligned}(U^h)_j^n =& (-1)^{j+n}\lambda_k'^n e^{i\xi' j}\\ =& |\lambda'|^n e^{in(-\pi+\arg\lambda')}\cdot e^{ij(\pi+\xi')}\\ =& |\lambda'|^n e^{i(j\xi - 2\pi kan\tau)}\cdot e^{in(-\pi+\arg\lambda'+\lambda a\xi)}.\end{aligned}$$

Therefore, the relative phase error of high frequency modes at each time step is using (6.45)

$$E_p(k) = -\frac{-\pi + \arg\lambda' + \lambda a\xi}{\lambda a\xi} = -\frac{\pi(1-\lambda a)}{\lambda a\xi} + \frac{2(q-1)\lambda a\xi'}{(2q-1)\lambda a\xi} - \frac{1}{3(2q-1)^2\xi}\left[\frac{q+1}{2} - \frac{\lambda^2 a^2}{2q-1}\right]\xi'^3 + \mathcal{O}(\xi'^5). \quad (6.46)$$

Note that $\xi \approx \pi$. Therefore the relative phase error has $\mathcal{O}(1)$. This error is huge, and strong numerical dissipation is needed to suppress it.

The above Fourier analysis is summarized in the following theorem.

Theorem 6.1 *We distinguish low and high frequency Fourier modes* $u_j^n = \lambda_k^n e^{ij\xi}$, $\xi = 2\pi kh$, *and they behave differently.*

(i) For the low frequency modes ($\xi \sim 0$), the relative phase error is of order $\mathcal{O}(\xi^2)$, see (6.40), and the amplitude error (dissipation) becomes smaller as the parameter q decreases, see (6.39). The order of amplitude error is $\mathcal{O}(\xi)$ for the monotone schemes and $\mathcal{O}(\xi^2)$ for the LW scheme.

(ii) For the high frequency modes ($\xi \sim \pi$), the relative phase error is of order $\mathcal{O}(1)$, see (6.46), the amplitude error becomes larger as the parameter q is closer to $1/2$, see (6.43).

6.2.8 Modified equation analysis.

As we know, the amplitude error and relative phase error of the Fourier modes have a correspondence with dissipation and phase error mechanisms displayed by related PDEs. Hence we use the method of modified equation analysis to further investigate the mechanisms of dissipation and phase error of (3.89). Particularly,

we want to see how the dissipation offsets the large phase error of high frequency modes. This method of modified equation analysis was originally introduced for low frequency modes [93]. Here it is especially used for high frequency modes, as its usefulness was clearly shown in [52]. This section begins with the linear case, and still use notation $\lambda = \tau/h$. The nonlinear case is left to the next section.

As in [52, Page 173], consider a smooth solution $(U^s)_j^n$ and an oscillatory solution $(U^h)_j^n$, respectively. The oscillatory solution $(U^h)_j^n$ is written as

$$(U^h)_j^n = (-1)^{j+n}(U^o)_j^n, \tag{6.47}$$

where $(U^o)_j^n$ is viewed as the perturbation amplitude of the checkerboard mode.

The smooth solution $(U^s)_j^n$ satisfies (6.30), i.e.

$$(U^s)_j^{n+1} = (U^s)_j^n - \frac{\lambda a}{2}((U^s)_{j+1}^n - (U^s)_{j-1}^n) + \frac{q}{2}((U^s)_{j+1}^n - 2(U^s)_j^n + (U^s)_{j-1}^n). \tag{6.48}$$

Then we derive a modified equation for this solution, and the notation $\widetilde{U}^s$ corresponds to the associated exact solution

$$\partial_t \widetilde{U}^s + a\partial_x \widetilde{U}^s = \frac{1}{2\tau}\left(qh^2 - a^2\tau^2\right)\partial_x^2 \widetilde{U}^s + a\left(-\frac{h^2}{6} + \frac{1}{2}qh^2 - \frac{1}{3}a^2\tau^2\right)\partial_x^3 \widetilde{U}^s + \cdots. \tag{6.49}$$

The process how to derive this equation can be found in [52, Page 169]. It is evident that the numerical viscosity of (6.30) becomes stronger for low frequency modes as q is larger, and vice versa. In particular, for the LW scheme with $q = \lambda^2 a^2$ the dissipation mechanism comes from the fourth order term and therefore is quite weak. This is consistent with the fact observed by the Fourier analysis that the dissipation effect becomes weaker

as the viscosity coefficient q decreases, provided that the scheme is stable.

However, the numerical dissipation of (3.89) is very different for the high frequency modes. Let us discuss the perturbation $(U^o)_j^n$ of the oscillatory solution (6.47). Substituting (6.47) into (6.30) yields

$$\frac{(U^o)_j^{n+1}-(U^o)_j^n}{\tau}=\frac{q-2}{\tau}(U^o)_j^n-\frac{\lambda a}{2\tau}[(U^o)_{j+1}^n-(U^o)_{j-1}^n]+\frac{q}{2\tau}[(U^o)_{j+1}^n+(U^o)_{j-1}^n]. \tag{6.50}$$

Compared to (6.48) for the low frequency modes, (6.50) contains an extra term $\frac{q-2}{\tau}(U^o)_j^n$, which plays the key role of damping on high frequency modes. The notation $\tilde{U}^o(jh,n\tau)$ is used to express $(U^o)_j^n$ inserted into (6.50) and apply the standard approach (see [52, Page 173]). That is, taking the standard Taylor expansion

yields

$$\mathcal{D}_{+t}\widetilde{U}^o + a\partial_x \widetilde{U}^o = \frac{2(q-1)}{\tau}\widetilde{U}^o + \frac{a^2 q\tau}{2\lambda^2}\partial_x^2 \widetilde{U}^o - \frac{a^3\tau^2}{6\lambda^2}\partial_x^3 \widetilde{U}^o + \cdots, \tag{6.51}$$

where D_{+t} is a forward difference operator in time and can be expressed as $D_{+t} = (e^{\tau\partial_t} - 1)/\tau$. Note that in (6.50) the term $\frac{q-2}{\tau}(\widetilde{U}^o)_j^n$ is unusual compared to classical modified equation analysis. Next we write (6.51) as

$$(e^{\tau\partial_t} - 1)\widetilde{U}^o = \beta\widetilde{U}^o - a\tau\partial_x \widetilde{U}^o + \frac{a^2 q h^2}{2}\partial_x^2 \widetilde{U}^o - \frac{a^3\tau h^2}{6}\partial_x^3 \widetilde{U}^o + \cdots$$

where $\beta = 2(q-1)$. Note the following basic facts, namely the formal operator expansion

$$\tau\partial_t = \sum_{m=1}^{\infty}(-1)^{m+1}\frac{(e^{\tau\partial_t} - 1)^m}{m},$$

and the well known power series

$$\frac{1}{(1+z)^2} = \sum_{m=0}^{\infty} (-1)^m (m+1) z^m, \ \frac{1}{(1+z)^3} = \sum_{n=0}^{\infty} (-1)^m \frac{(m+1)(m+2)}{2} z^m.$$

Let $C_m^\ell = \frac{m!}{(m-\ell)!\ell!}$ denote the binomial coefficients for $\ell \leq m$. For $z \in]-1,1[$ and $0 < q \leq 1$ with $q \neq 1/2$ we obtain, by ignoring

terms of orders higher than three, that

$$
\begin{aligned}
\tau\partial_t\widetilde{U}^o &= \sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\beta^m\widetilde{U}^o + \sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}C_m^1\beta^{m-1}(-a\tau\partial_x)\widetilde{U}^o \\
&\quad + \left\{\sum_{m=2}^{\infty}\frac{(-1)^{m+1}}{m}C_m^2\beta^{m-2}a^2\tau^2\partial_x^2 + \sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}C_m^1\beta^{m-1}\frac{qh^2}{2}\partial_x^2\right\}\widetilde{U}^o \\
&\quad + \left\{\sum_{m=3}^{\infty}\frac{(-1)^{m+1}}{m}C_m^3\beta^{m-3}(-a\tau\partial_x)^3 + \sum_{m=2}^{\infty}\frac{(-1)^{m+1}}{m}C_m^1C_{m-1}^1\beta^{m-2}(-a\tau\partial_x)\right. \\
&\quad \left.\cdot\frac{qh^2}{2}\partial_x^2 + \sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}C_m^1\beta^{m-1}\left(-\frac{a\tau h^2}{6}\partial_x^3\right)\right\}\widetilde{U}^o + \cdots \\
&= \ln|\beta+1|\widetilde{U}^o - \frac{a\tau}{1+\beta}\partial_x\widetilde{U}^o + \left\{-\frac{a^2\tau^2}{2(1+\beta)^2} + \frac{qh^2}{2(1+\beta)}\right\}\partial_x^2\widetilde{U}^o \\
&\quad + \left\{\frac{-a^3\tau^3}{3(1+\beta)^3} + \frac{a\tau}{(1+\beta)^2}\frac{qh^2}{2} - \frac{1}{1+\beta}\cdot\frac{a\tau h^2}{6}\right\}\partial_x^3\widetilde{U}^o + \cdots \\
&= \ln|2q-1|\widetilde{U}^o - \frac{a\tau}{2q-1}\partial_x\widetilde{U}^o + \frac{1}{2}\frac{[q(2q-1)h^2 - a^2\tau^2]}{(2q-1)^2}\partial_x^2\widetilde{U}^o \\
&\quad + \frac{a\tau}{6}\frac{[(q+1)(2q-1)h^2 - 2a^2\tau^2]}{(2q-1)^3}\partial_x^3\widetilde{U}^o + \cdots .
\end{aligned}
$$

Thus the modified equation for the oscillatory part is derived as follows

$$\partial_t \widetilde{U}^o + \frac{a}{2q-1}\partial_x \widetilde{U}^o = \frac{\ln|2q-1|}{\tau}\widetilde{U}^o + \frac{h^2}{2\tau}\frac{[q(2q-1)-\lambda^2 a^2]}{(2q-1)^2}\partial_x^2 \widetilde{U}^o + \frac{ah^2}{6}\frac{[(q+1)(2q-1)-2\lambda^2 a^2]}{(2q-1)^3}\partial_x^3 \widetilde{U}^o + \cdots .$$

We introduce for $q \neq 1/2$ a rescaling that becomes singular for $q = 1/2$, namely we set $x' = x(2q-1)$ and omit the use of a primed variable. This gives,

$$\partial_t \widetilde{U}^o + a\partial_x \widetilde{U}^o = \frac{\ln|2q-1|}{\tau}\widetilde{U}^o + \frac{h^2}{2\tau}[q(2q-1)-\lambda^2 a^2]\partial_x^2 \widetilde{U}^o + \frac{ah^2}{6}[(q+1)(2q-1)-2\lambda^2 a^2]\partial_x^3 \widetilde{U}^o + \cdots . \tag{6.52}$$

Unlike the modified equation (6.49) for the low frequency modes the numerical dissipation comes from two terms: Zero order term $\frac{\ln|2q-1|}{\tau}\widetilde{U}^o$ and the second order term $\frac{a^2\tau}{2\lambda^2a^2}[q(2q-1)-\lambda^2a^2]\partial_x^2\widetilde{U}^o$. The former exerts more dominant dissipation than the latter, which can be explained from the following well-known fact that linear source terms decays exponentially in time, whereas second order diffusive terms as in the heat equation have a much lower algebraic decay. Consider

$$\begin{cases} v_t = \alpha v + \mu v_{xx}, \quad \alpha < 0, \quad \mu > 0, \\ v(x,0) = v_0(x). \end{cases} \tag{6.53}$$

The solution expression is

$$v(x,t) = \frac{e^{\alpha t}}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} v_0(y) e^{-\frac{(x-y)^2}{4\mu t}} dy. \tag{6.54}$$

From this solution we clearly see the decay property in time.

The zero order term in (6.52) $\frac{\ln|2q-1|}{\tau}\widetilde{U}^o$ is called as *a numerical damping* term and the second order term $\frac{h^2}{2\tau}[q(2q-1)-\lambda^2a^2]\partial_x^2\widetilde{U}^o$ as a *numerical viscosity*. They play distinct dissipation roles in controlling the amplitude of high frequency modes. Furthermore, two more remarks are in order.

(i) For the LF scheme with $q=1$ the modified equation for the perturbation part $\widetilde{U}^o$ is

$$\partial_t\widetilde{U}^o+a\partial_x\widetilde{U}^o=\frac{a^2\tau}{2\lambda^2a^2}[1-\lambda^2a^2]\partial_x^2\widetilde{U}^o+\frac{a^3\tau^2}{3\lambda^2a^2}[1-\lambda^2a^2]\partial_x^3\widetilde{U}^o+\cdots. \tag{6.55}$$

Although this part is dissipated through the numerical viscosity term if $|\lambda a|<1$, this dissipation is still weak in comparison with the numerical damping term of the form $\frac{\ln(|2q-1|)}{\tau}\widetilde{U}^o$. In particular, $\widetilde{U}^0\equiv 1$ is a solution of (6.55) if the constant unity initial data are given, which implies that the checkerboard mode is not

perturbed and therefore not damped at all. This explains why the oscillations in the LF scheme are observed. This was already highlighted above through the discrete Fourier analysis.

(ii) As $0 < q < 1$, the strong damping term $\frac{\ln(|2q-1|)}{\tau}\widetilde{U}^o$ suppresses the oscillations well no matter how the viscosity term behaves. In particular, if $q = 1/2$, the oscillation is damped out immediately, by noting that

$$\lim_{q \to 1/2+0} \ln(|2q-1|) = -\infty. \tag{6.56}$$

So the damping becomes infinite for $q = 1/2$. This is consistent with two previous observations. The first is that in the solution (6.17) to the checkerboard initial data (6.16) the amplitude vanishes in this case. The second is the computational evidence shown in Fig. 35(d).

The high frequency modes are dissipated very quickly if $0 <$

$q < 1$, unlike the LF scheme with $q = 1$ or the unstable central scheme with $q = 0$, even through there is a checkerboard mode initially when we have $\widetilde{U}^0 = 1$ as initial data for (6.52). The solution decays algebraically, i.e.,

$$|2q - 1|^{\frac{1}{\tau}} \widetilde{U^o} = \mathcal{O}(1). \tag{6.57}$$

Therefore the checkerboard mode is damped algebraically. This explains why the oscillations are less visible for $0 < q < 1$ than those in the standard LF scheme for $q = 1$ or the unstable central scheme for $q = 0$.

Before ending this section, we present several remarks on the local oscillations in the monotone schemes. The previous discussion on the local oscillations in the generalized LF scheme has general implications into more extensive (monotone) schemes and multidimensional cases. Indeed, the generalized LF scheme is monotone and thus TVD under a certain restriction. The TVD

property is proposed to describe the global property of solutions of (1.3). The phenomenon of oscillations we are investigating is *local* and does not contradict to this *global* TVD property. As far as hyperbolic problems are concerned, local properties should be paid more attention because of finite wave propagation. We have individually analyzed the resolution of the low and high frequency modes $u_j^n = \lambda_k^n e^{ij\xi}$, $\xi = 2\pi kh$ in numerical solutions. Our approach is the discrete Fourier analysis and the modified equation analysis, which are applied to investigating the numerical dissipative and dispersive mechanisms as well as relative phase errors. Our results are summarized as follows.

(1) For the low frequency modes, the error is of order $\mathcal{O}(\xi^2)$, while for high frequency modes the error is of order $\mathcal{O}(1)$ after each time step, which is generally independent of the parameter q.

(2) For the low frequency modes, the dissipation is usually of

order $\mathcal{O}(\xi)$ for the scheme (6.30), which closely depends on the parameter q. As $q = \lambda^2 a^2$, (6.30) becomes the LW scheme and it has the amplitude error $\mathcal{O}(\xi^2)$. For high frequency modes, the scheme usually has the numerical damping of order $\mathcal{O}(1)$ that becomes stronger as q is closer to $1/2$, unless it vanishes for the limit case ($q = 1$ or 0), in which the amplitude is dissipated via the numerical viscosity of second order.

Thus we conclude that *the relative phase errors should be at least offset by the numerical dissipation of the same order. Otherwise the oscillation could be caused.* In the second order accurate LW scheme the oscillations are caused by the relative phase error of low frequency modes, while in the first order LF schemes, oscillations are caused by the relative phase error of high frequency modes. In order to control the oscillations by high frequency modes, the strong numerical damping (zero order term) is necessary to add.

The presence of high frequency modes results from initial /boundary conditions. Section 6.2.2 shows that the discretization may produce the checkerboard modes $(-1)^{j+n}$. As discussed in [52], the classical box scheme has serious difficulty in controlling such modes, and a weighted time-average technique is used to cure it. Such a technique is indeed to introduce an *artificial numerical damping*, with sacrificing the accuracy. Compared to the LF scheme ($q = 1$), the scheme (6.30) ($0 < q < 1$) introduces the numerical damping as well, which is stronger as q is closer to $1/2$. Then (6.30) becomes the modified LF scheme [71]. Hence, in order to control the oscillations caused by high frequency modes, *the numerical damping plays an important role*.

6.3 Implicit schemes

In practice, the time step size constraint may arise from the requirement of nonlinear stability and convergence etc. The main disadvantage of the implicit scheme is that at each time step, a (linear or nonlinear, algebraic) system has to be solved by an iteration method, and thus it is very time-consuming. Even for the same time step size as that used in the explicit scheme, unsteady solutions of the implicit schemes for nonlinear hyperbolic conservation laws are less accurate [41, 78]. However, the implicit scheme is very attractive in simulating steady state solutions [96]. Moreover, an implicit treatment is usually necessary for the governing equations with stiff source terms or (linear) higher-order derivative terms in order to reduce the constraint of the time step size.

Some possible directions in the implicit scheme are listed here.

(1) Jacobian-free Newton-Krylov (JFNK) methods [39] are synergistic combinations of Newton-type methods for super-linearly convergent solution of nonlinear equations and Krylov subspace methods for solving the Newton correction equations. (2) Convergence to a steady state may be accelerated by the use of a variable time step determined by the local Courant number [36]. (3) For explicit scheme, the spatial step sizes and the "signal" speeds are the two main elements to limit a choice of the time step size. Varying time stepsize discretization may be adapted, see [76, 77]. (4) Analytical integration of some stiff source terms, and special discretization with a large time step, such as semi-Lagrangian /Lagrangian method, can be considered. (5) It is to use Beam and Warming's implicit scheme [1], which is derived

by the following procedure. Using Taylor's expansion gives

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n + \frac{\tau}{2}\left[\left.\frac{\partial \boldsymbol{U}}{\partial t}\right|_j^n + \left.\frac{\partial \boldsymbol{U}}{\partial t}\right|_j^{n+1}\right] + O(\tau^3),$$

and

$$\left.\frac{\partial \boldsymbol{U}}{\partial t}\right|_j^{n+1} = \left.\frac{\partial \boldsymbol{U}}{\partial t}\right|_j^n + \tau \frac{\partial}{\partial t}\left(\frac{\partial \boldsymbol{U}}{\partial t}\right)\Big|_j^n + O(\tau^2).$$

Thanks to (5.1), we have

$$\frac{\partial}{\partial t}\left(\frac{\partial \boldsymbol{U}}{\partial t}\right)\Big|_j^n = -\frac{\partial}{\partial t}\left(\frac{\partial \boldsymbol{F}(\boldsymbol{U})}{\partial x}\right)\Big|_j^n = -\frac{\partial}{\partial x}\left(\frac{\partial \boldsymbol{F}(\boldsymbol{U})}{\partial t}\right)\Big|_j^n = -\frac{\partial}{\partial x}\left(\boldsymbol{A}(\boldsymbol{U})\frac{\partial \boldsymbol{U}}{\partial t}\right)\Big|_j^n.$$

Combining them gives

$$\frac{\boldsymbol{U}_j^{n+1} - \boldsymbol{U}_j^n}{\tau} = -\left.\frac{\partial \boldsymbol{F}}{\partial x}\right|_j^n - \frac{1}{2}\frac{\partial}{\partial x}\left[\boldsymbol{A}(\boldsymbol{U})(\boldsymbol{U}^{n+1} - \boldsymbol{U}^n)\right]_j^n. \quad (6.58)$$

Using the central differences to approximate the spatial derivatives results in a block tri-diagonal system

$$-\frac{\tau}{4h}\left(\boldsymbol{A}_{j-1}^{n}\boldsymbol{U}_{j-1}^{n+1}\right)+\boldsymbol{U}_{j}^{n+1}+\frac{\tau}{4h}\left(\boldsymbol{A}_{j+1}^{n}\boldsymbol{U}_{j+1}^{n+1}\right)=\boldsymbol{U}_{j}^{n}-\frac{\tau}{2h}\left(\boldsymbol{F}_{j+1}^{n}-\boldsymbol{F}_{j-1}^{n}\right)$$
$$+\frac{\tau}{4h}\left(\boldsymbol{A}_{j+1}^{n}\boldsymbol{U}_{j+1}^{n}-\boldsymbol{A}_{j-1}^{n}\boldsymbol{U}_{j-1}^{n}\right),$$

which can be solved by using the Thomas algorithm (a tridiagonal matrix algorithm). To resolve the shock wave, the dissipation term is required for nonlinear hyperbolic equations such as

$$\boldsymbol{D}=-\varepsilon(\boldsymbol{U}_{j+2}^{n}-4\boldsymbol{U}_{j+1}^{n}+6\boldsymbol{U}_{j}^{n}-4\boldsymbol{U}_{j-1}^{n}+\boldsymbol{U}_{j-2}^{n}).$$

which is explicitly added to the right hand side. This is always used for successful computation where high-frequent oscillations are observed and must be suppressed. Based on (6.58), the

"linearized" implicit conservative scheme may be derived

$$\frac{\boldsymbol{U}_j^{n+1}-\boldsymbol{U}_j^n}{\tau}=-\frac{1}{h}\left(\hat{\boldsymbol{F}}_{j+\frac{1}{2}}^n-\hat{\boldsymbol{F}}_{j-\frac{1}{2}}^n\right)-\frac{1}{2h}\boldsymbol{A}_{j+\frac{1}{2}}^n(\boldsymbol{U}_{j+\frac{1}{2}}^{n+1}-\boldsymbol{U}_{j+\frac{1}{2}}^n)$$
$$+\frac{1}{2h}\boldsymbol{A}_{j-\frac{1}{2}}^n(\boldsymbol{U}_{j-\frac{1}{2}}^{n+1}-\boldsymbol{U}_{j-\frac{1}{2}}^n),$$

or

$$\delta\boldsymbol{U}_j^n+\Delta_x^-(\boldsymbol{A}^+\delta\boldsymbol{U}^n)_j+\Delta_x^+(\boldsymbol{A}^-\delta\boldsymbol{U}^n)_j+\lambda\left(\hat{\boldsymbol{F}}_{j+\frac{1}{2}}^n-\hat{\boldsymbol{F}}_{j-\frac{1}{2}}^n\right)=0,$$

where $\boldsymbol{U}_{j\pm\frac{1}{2}}=\frac{1}{2}(\boldsymbol{U}_{j\pm1}+\boldsymbol{U}_j)$, $\delta\boldsymbol{U}^n:=\boldsymbol{U}^{n+1}-\boldsymbol{U}^n$, and

$$\boldsymbol{A}^{\pm}=\frac{1}{2}(\boldsymbol{A}\pm r_A\boldsymbol{I}),\qquad r_A\geq\max\{|\lambda_A|\}.$$

The last is diagonally dominant.

Acknowledgements

The work was partially supported by the National Natural Science Foundation of China (Nos. 91330205, 11421101, 10925101). The authors would like to thank his co-workers who have made the contribution to this note.

References

[1] R.M. Beam and R.F. Warming, An implicit finite-difference algorithm for hyperbolic systems in conservation law form, *J. Comput. Phys.*, 22(1976), 87-110.

[2] M. Ben-Artzi and J. Falcovitz, *Generalized Riemann Problems in Computational Fluid Dynamics*, Cambridge Univ. Press, 2003.

[3] J.P. Boris, A fluid transport algorithm that works, in: *Computing as a language of physics, International Atomic Energy Commission*, 1971, 171-189.

[4] J.P. Boris and D.L. Book, Flux-corrected transport I: SHASTA, a fluid transport algorithm that works, *J. Comput. Phys.*, 11(1973), 38-69.

[5] M. Breuss, The correct use of the Lax-Friedrichs method, *M2AN Math. Model. Numer. Anal.* 38(2004), 519-540.

[6] M. Breuss, An analysis of the influence of data extrema on some first and second order central approximations of hyperbolic conservation laws, *M2AN Math. Model. Numer. Anal.*, 39(2005), 965-993.

[7] S. Chakravarthy and S. Osher, High resolution applications of the Osher upwind scheme for the Euler equations, *AIAA paper presented at 6th CFD conference*, 1983.

[8] B. Cockburn and C.W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework, *Math. Comp.*, 52(1989), 411-435.

[9] B. Cockburn and C.W. Shu, Runge-Kutta discontinuous Galerkin methods for convection-dominated problems, *J. Sci. Comput.*, 16(2001), 173-261.

[10] B. Cockburn, S. Hou, and C.W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case, *Math. Comp.*, 54(1990), 545-581.

[11] B. Cockburn, S.Y. Lin, and C.W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-dimensional systems, *J. Comput. Phys.*, 84(1989), 90-113.

[12] B. Cockburn and C.W. Shu, The Runge-Kutta local projection P^1-discontinuous Galerkin finite element method for scalar conservation laws, *RAIRO Model. Math. Anal. Numer.*, 25(1991), 337-361.

[13] B. Cockburn and C.W. Shu, The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems, *J. Comput. Phys.*, 141(1998), 199-224.

[14] R. Courant, K. Friedrichs, and H. Lewy, On the partial difference equations of mathematical physics, *Math. Annal.*, 100(1928), 32-74.

[15] P. Colella, A direct Eulerian MUSCL scheme for gas dynamics, *SIAM J. Sci. Stat. Comput.*, 6(1985), 104-117.

[16] P. Colella and P.R. Woodward, The piecewise parabolic method (PPM) for gas-dynamical simulations, *J. Comput. Phys.*, 54(1984), 174-201.

[17] R. Courant, E. Isaacson, and M. Rees, On the solution of nonlinear hyperbolic differential equations by finite differences, *Comm. Pure Appl. Math.* 5(1952), 243-255.

[18] M.G. Crandall and A. Majda, Monotone difference approximations for scalar conservation laws, *Math. Comput.*, 34(1980), 1-21.

[19] B. Engquist and S. Osher, Stable and entropy satisfying approximations for transonic flow calculations, *Math. Comput.*, 34(1980), 45-75.

[20] B. Engquist and S. Osher, One-sided difference approximations for nonlinear conservation laws, *Math. Comput.*, 36(1981), 321-352.

[21] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, 1996.

[22] S.K. Godunov, A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations, *Math. Sbornik*, 47(1959), 271-306.

[23] J.B. Goodman and R.J. LeVeque, On the accuracy of stable schemes for 2D scalar conservation laws, *Math. Comput.*, 45(1985), 15-21.

[24] J. B. Goodman and R. J. LeVeque, A geometric approach to high resolution TVD schemes, *SIAM J. Numer. Anal.*, 25(1988), 268-284.

[25] A. Harten, High resolution schemes for hyperbolic conservation laws, *J. Comput. Phys.*, 49(1983), 357-393.

[26] A. Harten, On a class of high resolution total-variation-stable finite-difference schemes, *SIAM J. Numer. Anal.*, 21(1984), 2-23.

[27] A. Harten, On high-order accurate interpolation for non-oscillatory shock capturing schemes, in "*Oscillation Theory, Computation, and Methods of Compensated Compactness*", edited by C. Dafermos, J.L. Ericksen, D. Kinderlehrer, and M. Slemrod, Springer, 1986, 71-105.

[28] A. Harten, S. Osher, B. Engquist, and S. Chakravarthy, Some results on uniformly high order accurate essentially non-oscillatory schemes, *Appl. Numer. Math.*, 2 (1986), 347-377.

[29] A. Harten, B. Engquist, S. Osher, and S.R. Chakravarthy, Uniformly high order accurate essentially non-oscillatory schemes III, *J. Comput. Phys.*, 71(1987), 231-303; 131(1997), 3-47.

[30] A. Harten, J. M. Hyman, and P.D. Lax, On finite-difference approximation and entropy conditions for shocks, *Comm. Pure Appl. Math.*, 29(1976), 297-321.

[31] A. Harten, P. Lax, and B. van Leer, On upstream differencing and Godunov type methods for hyperbolic conservation laws, *SIAM Rev.*, 25(1983), 35-61.

[32] A. Harten and G. Zwas, Self-adjusting hybrid schemes for shock calculations, *J. Comput. Phys.*, 9(1972), 568-583.

[33] C. Hirsch, *Numerical Computation of Internal and External Flows*, Vol. 1 and 2, Wiley, 1990.

[34] L.C. Huang, Pseudo-unsteady difference schemes for discontinuous solutions of steady-state, one-dimensional fluid dynamics problems, *J. Comput. Phys.*, 42(1981), 195-211.

[35] A. Jameson and P.D. Lax, Conditions for the construction of multi-point total variation diminishing difference schemes, *Appl. Numer. Math.*, 2(1986), 335-345.

[36] A. Jameson, W. Schmidt, and E. Turkel, Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes, *AIAA Paper No. 81-1259*, 1981.

[37] G.S. Jiang and C.W. Shu, Efficient implementation of weighted ENO schemes, *J. Comput. Phys.*, 26(1996), 202-228.

[38] G.S. Jiang and C.C. Wu, A high-order WENO finite difference scheme for the equations of ideal magnetohydrodynamics, *J. Comput. Phys.*, 150(1999), 561-594.

[39] D.A. Knoll and D.E. Keyes, Jacobian-free Newton-Krylov methods: a survey of approaches and applications, *J. Comput. Phys.*, 193(2004), 357-397.

[40] V.P. Kolgan, Application of the principle of minimizing the derivative to the construction of finite-difference schemes for computing discontinuous solutions of gas dynamics, *J. Comput. Phys.*, 230(2011), 2384-2390.

[41] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley, John & Sons, 1997.

[42] P.D. Lax, Weak solutions of non-linear hyperbolic equations and their numerical computation, *Comm. Pure Appl. Math.*, 7(1954), 159-193.

[43] P.D. Lax and R.D. Richtmyer, Survey of the stability of linear finite difference equations, *Comm. Pure Appl. Math.*, 9(1956), 267-293.

[44] P.D Lax and B. Wendroff, Systems of conservation laws, *Commun. Pure Appl Math.*, 13(1960), 217-237.

[45] P.D. LeFloch and J.G. Liu, Generalized monotone schemes, discrete and paths of extrema, and discrete entropy conditions, *Math. Comp.*, 68(1999), 1025-1055.

[46] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser Verlag, Berlin, 1990.

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

[48] J.Q. Li, H.Z. Tang, G. Warnecke, and L.M. Zhang, Local oscillations in finite difference solutions of hyperbolic conservation laws, *Math. Comp.*, 78(2009), 1997-2018.

[49] V.D. Liseikin, *Grid Generation Methods*, 2nd ed., Springer, 2010.

[50] X.D. Liu, S. Osher, and T. Chan, Weighted essentially non-oscillatory schemes, *J. Comput. Phys.*, 115(1994), 200-212.

[51] R.W. MacCormack, The effect of viscosity in hypervelocity impact cratering, *AIAA Paper No. 69-354*, 1969.

[52] K.W. Morton and D.F. Mayers, *Numerical Solution of Partial Differential Equations*, 2nd edition, Cambridge University Press, 2005.

[53] J.M. Moschetta and J. Gressier, A cure for the sonic point glitch, *Int. J. Comput. Fluid Dyn.*, 13(2000), 143-159.

[54] S. Osher, Riemann solvers, the entropy condition, and difference approximations, *SIAM J. Numer. Anal.*, 21(1984), 217-235.

[55] S. Osher, Convergence of generalized MUSCL schemes, *SIAM J. Numer. Anal.*, 22(1985), 947-961.

[56] J.X. Qiu and C.W. Shu, Runge-Kutta discontinuous Galerkin method using WENO limiters, *SIAM J. Sci. Comput.*, 26(2005), 907-929.

[57] J.J. Quirk, A contribution to the great Riemann solver debate, *Int. J. Numer. Methods Fluids*, 18(1994), 555-574.

[58] S. Osher and S. Chakravarthy, Upwind schemes and boundary conditions with applications to Euler equations in general geometries, *J. Comput. Phys.*, 50(1983), 447-481.

[59] W.H. Reed and T.R. Hill, Triangular mesh methods for the neutron transport equation, *Tech. Report LA-UR-73-479*, Los Alamos Scientific Laboratory, 1973.

[60] R.D. Richtmyer and K.W. Morton, *Difference Methods for Initial-value Problems*, Interscience, New York, 1967.

[61] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes, *J. Comput. Phys.*, 43(1981), 357-372; 135(1997), 250-258.

[62] P.L. Roe, Numerical algorithms for the linear wave equation, *Royal Aircraft Establishment Technical Report 81047*, 1981.

[63] P.L. Roe, Generalized formulation of TVD Lax Wendroff schemes, ICASE Report No.84-53, 1984.

[64] P.L. Roe, Characteristic-based schemes for the Euler equations, *Ann. Rev. Fluid Mech.*, 18(1986), 337-365.

[65] R.H. Sanders and K.H. Prendergast, The Possible Relation of the 3-KILOPARSEC Arm to Explosions in the Galactic Nucleus, *Astrophys. J.*, 188(1974), 489-500.

[66] C.-W. Shu, High order weighted essentially non-oscillatory schemes for convection dominated problems, *SIAM Rev.*, 51(2009), 82-126.

[67] C.W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, *J. Comput. Phys.*, 77(1988), 439-471.

[68] C.W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, II, *J. Comput. Phys.*, 83(1989), 32-78.

[69] J.L. Steger and R.F. Warming, Flux-vector splitting of the inviscid gas dynamic equations with application to finite difference methods, *J. Comput. Phys.*, 40(1981), 263-293.

[70] P.K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation law, *SIAM J. Numer. Anal.*, 21(1984), 995-1011.

[71] E. Tadmor, Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comp.*, 43(1984), 369-381.

[72] H.Z. Tang, On the sonic point glitch, *J. Comput. Phys.*, 202(2005), 507-532.

[73] H.Z. Tang and T. Tang, Adaptive mesh methods for one- and two-dimensional hyperbolic conservation laws, *SIAM J. Numer. Anal.*, 41(2003), 487-515.

[74] H.Z. Tang and G. Warnecke, A note on $(2K+1)$-point conservative monotone schemes, *M2AN Math. Model. Numer. Anal.*, 38(2004), 345-357.

[75] H.Z. Tang and G. Warnecke, A Runge-Kutta discontinuous Galerkin method for the Euler equations, *Computers & Fluids*, 34(2005), 375-398.

[76] H.Z. Tang and G. Warnecke, A class of high resolution difference schemes for nonlinear Hamilton-Jacobi equations with varying time and space grids, *SIAM J. Sci. Comput.*, 26(2005), 1415-1431.

[77] H.Z. Tang and G. Warnecke, High resolution schemes for conservation laws and convection-diffusion equations with varying time and space grids, *J. Comput. Math.*, 24(2006), 121-140.

[78] H.Z. Tang and G. Warnecke, On convergence of a domain decomposition method for hyperbolic conservation laws, *SIAM J. Numer. Anal.*, 45(2007), 1453-1471.

[79] J.W. Thomas, *Numerical Partial Differential Equations: Finite Difference Methods*, Springer-Verlag, 1995.

[80] E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dyanmics, A Practical Introduction*, 3rd ed., Springer, 2009.

[81] L.N. Trefethen, Group velocity in finite difference scheme, *SIAM Rev.*, 24(1982), 113-136.

[82] G.D. van Albada, B. van Leer, and W.W. Jr. Roberts, A comparative study of computational methods in cosmic gas dynamics, *Astron. Astrophys.*, 108(1982), 76-84.

[83] B. van Leer, Towards the ultimate conservative difference scheme I: The quest of monotonicity, *Lecture Notes in Physics*, 18(1973), 163-168.

[84] B. van Leer, Towards the ultimate conservative difference scheme II: Monotonicity and conservation combined in a second-order scheme, *J. Comput. Phys.*, 14(1974), 361-370.

[85] B. van Leer, Towards the ultimate conservative difference scheme III: Upstream-centered finite-difference schemes for ideal compressible flow, *J. Comput. Phys.*, 23(1977), 263-275.

[86] B. van Leer, Towards the ultimate conservative difference scheme IV: A new approach to numerical convection, *J. Comput. Phys.*, 23(1977), 276-299.

[87] B. van Leer, Towards the ultimate conservative difference scheme V: A second order sequel to Godunov's method, *J. Comput. Phys.*, 32(1979), 101-136.

[88] B. van Leer, Flux-vector splitting for the Euler equations, *Proceedings of the 8th Inter. Conf. on Numer. Methods in Fluid Dynamics*, Springer-Verlag, 1982, 507-512.

[89] B. van Leer, On the relation between the upwind-differencing schemes of Godunov, Engquist-Osher and Roe, *SIAM J. Sci. Stat. Comput.*, 5(1984), 1-20.

[90] B. van Leer, A historical oversight: Vladimir P. Kolgan and his high resolution scheme, *J. Comput. Phys.*, 230(2011), 2378-2383.

[91] J. von Neumann and R. D. Richtmyer, A method for the numerical calculation of hydrodynamic shocks, *J. Appl. Phys.*, 21(1950), 232-237.

[92] J.H. Wang and G. Warnecke, On entropy consistency of large time step schemes II. Approximate Riemann solvers, *SIAM J. Numer. Anal.*, 30(1993), 1252-1267.

[93] R.F. Warming and B.J. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods, *J. Comput. Phys.*, 14(1974), 159-179.

[94] P. Woodward and P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks, *J. Comput. Phys.*, 54(1984), 115-173.

[95] H.M. Wu and S.L. Yang, MmB–A new class of accurate high resolution schemes for conservation laws in two dimensions, *IMPACT Comput. Sci. & Engrg.*, 1(1989), 217-259.

[96] H.C. Yee, R.F. Warming, and A. Harten, Implicit total variation diminishing (TVD) schemes for steady-state calculations, *J. Comput. Phys.*, 57(1985), 327-360.

[97] J. Zhao and H.Z. Tang, Runge-Kutta discontinuous Galerkin methods with WENO limiter for the special relativistic hydrodynamics, *J. Comput. Phys.*, 242(2013), 138-168.