

**Numerical methods for conservation laws  
and related equations**

Siddhartha Mishra



## About these notes

These notes present numerical methods for conservation laws and related time-dependent nonlinear partial differential equations. The focus is on both simple scalar problems as well as multi-dimensional systems.

The MATLAB package `Compack` (Conservation Law MATLAB Package) has been developed as an educational tool to be used with these notes. All the numerical experiments in the lecture notes have been done in `Compack`. The scripts used to generate figures and tables are all in the `+Notes` sub-package. For instance, to generate the plots in Figure 2.3, run `Notes.Chapter2.central()` from the `Compack` base folder. Figures are saved to the `output` folder. `Compack` can be downloaded from <http://www.sam.math.ethz.ch/~ulrikf/compack.zip>.



# Contents

About these notes	iii
Chapter 1. Introduction	1
1.1. Examples for conservation laws.	2
1.2. Content and scope of these notes	5
Chapter 2. Linear Transport Equations	7
2.1. Method of characteristics	7
2.2. Finite difference schemes for the transport equation	9
2.3. An upwind scheme	12
2.4. Stability for the upwind scheme	14
Chapter 3. Scalar conservation laws	17
3.1. Characteristics for Burger's equation	19
3.2. Weak solutions	21
3.3. Entropy solutions	27
Chapter 4. Finite volume schemes for scalar conservation laws	33
4.1. Finite volume scheme	33
4.2. Approximate Riemann Solvers	39
4.3. Comparison of different finite volume schemes	44
4.4. Convergence Analysis	45
4.5. A note on boundary conditions	54
Chapter 5. Second-order (high-resolution) finite volume schemes	57
5.1. Order of accuracy	59
5.2. The REA algorithm	63
5.3. The minmod limiter	67
5.4. Other limiters	70
5.5. Flux limiters and the TVD property.	72
5.6. High-resolution methods for nonlinear problems.	73
5.7. Second-order semi-discrete schemes.	75
5.8. Time stepping	76
5.9. High-resolution algorithm	77
5.10. Numerical experiments	77
Chapter 6. Linear hyperbolic systems in one space dimension	81
6.1. Examples for linear systems	81
6.2. Hyperbolicity and characteristic decomposition	83
6.3. Solutions of Riemann problems, waves	84

6.4. Finite volume schemes	85
6.5. Numerical experiments	88
6.6. High-order finite volume schemes	91
6.7. Numerical experiments	93
Appendix A. Results from real analysis	95
Bibliography	97

## CHAPTER 1

### Introduction

Many interesting problems in the physical, biological, engineering and social sciences are modeled by a simple paradigm: Consider a domain  $\Omega \subset \mathbb{R}^n$  and a quantity of interest  $\mathbf{U}$ , defined for all points  $\mathbf{x} \in \Omega$ . The quantity of interest  $\mathbf{U}$  may be the temperature of a rod, the pressure of a fluid, the concentration of a chemical or a group of cells or the density of a human population. The evolution (in time) of this quantity of interest  $\mathbf{U}$  can be described by a simple phenomenological observation:

*The time rate of change of  $\mathbf{U}$  in any fixed sub-domain  $\omega \subset \Omega$  is equal to the total amount of  $\mathbf{U}$  produced or destroyed inside  $\omega$  and the flux of  $\mathbf{U}$  across the boundary  $\partial\omega$ .*

The above observation says that the change in  $\mathbf{U}$  is due to two factors: the *source* or *sink*, representing the quantity produced or destroyed, and the *flux*, representing the amount of  $\mathbf{U}$  that either goes in or comes out of the sub-domain, see Figure 1.1. This observation is mathematically rendered as

$$(1.1) \quad \frac{d}{dt} \int_{\omega} \mathbf{U} \, d\mathbf{x} = - \underbrace{\int_{\partial\omega} \mathbf{F} \cdot \nu \, d\sigma(\mathbf{x})}_{\text{flux}} + \underbrace{\int_{\omega} \mathbf{S} \, d\mathbf{x}}_{\text{source}},$$

where  $\nu$  is the unit outward normal,  $d\sigma(\mathbf{x})$  is the surface measure, and  $\mathbf{F}$  and  $\mathbf{S}$  are the flux and the source respectively. The minus sign in front of the flux term is for convenience. Note that (1.1) is an *integral* equation for the evolution of the total amount of  $\mathbf{U}$  in  $\omega$ .

We simplify (1.1) by using integration by parts (or the Gauss divergence theorem) on the surface integral to obtain

$$(1.2) \quad \frac{d}{dt} \int_{\omega} \mathbf{U} \, d\mathbf{x} + \int_{\omega} \operatorname{div}(\mathbf{F}) \, d\mathbf{x} = \int_{\omega} \mathbf{S} \, d\mathbf{x}.$$

Since (1.2) holds for all sub-domains  $\omega$  of  $\Omega$ , we can use an infinitesimal  $\omega$  to obtain the following *differential* equation:

$$(1.3) \quad \mathbf{U}_t + \operatorname{div}(\mathbf{F}) = \mathbf{S} \quad \forall (\mathbf{x}, t) \in (\Omega, \mathbb{R}_+).$$

The differential equation (1.3) is often termed as a *balance law* as it is a statement of the fact that the rate of change in  $\mathbf{U}$  is a balance of the flux and the source. Frequently, the only change in  $\mathbf{U}$  is from the fluxes and the source is set to zero. In such cases, (1.3) reduces to

$$(1.4) \quad \mathbf{U}_t + \operatorname{div}(\mathbf{F}) = 0 \quad \forall (\mathbf{x}, t) \in (\Omega, \mathbb{R}_+).$$

Equation (1.4) is termed as a *conservation law*, as the only change in  $\mathbf{U}$  comes from the quantity entering or leaving the domain of interest.

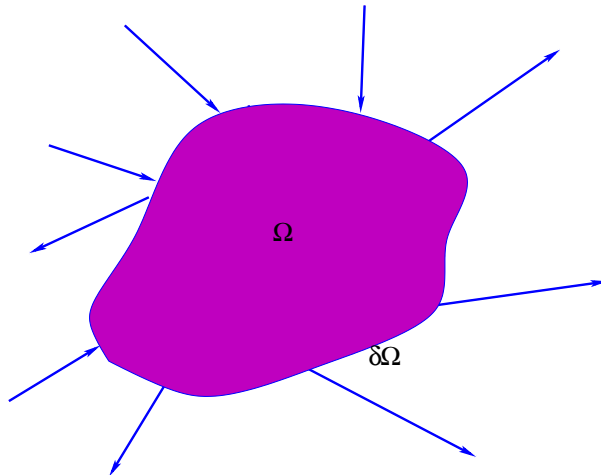


FIGURE 1.1. An illustration of conservation in a domain with the change being determined by the net flux.

The discussion so far is very general. We have not yet specified the explicit forms of  $\mathbf{U}$ ,  $\mathbf{F}$  and  $\mathbf{S}$ . In fact, the conservation law (1.4) and the balance law (1.3) are generic to a very large number of models. Explicit forms of the quantity of interest, flux and source depend on the specific model being considered. The modeling of the flux  $\mathbf{F}$  is the core function of a physicist, biologist, engineer or other domain scientists. We will provide several examples to illustrate conservation laws.

### 1.1. Examples for conservation laws.

For simplicity of the exposition, we begin with scalar examples, i.e, the quantity of interest  $\mathbf{U}$  is a scalar  $U$ .

**Scalar transport equation.** Let  $\mathbf{U} = U$  denote the concentration of a chemical (for example, a pollutant in a river). Assume that the river flows with a velocity field  $\mathbf{a}(\mathbf{x}, t)$  and we know the velocity field at all points in the river. The pollutant will clearly be transported in the direction of the velocity and so the flux in this case is  $\mathbf{F} = \mathbf{a}U$ . Since there is no production or destruction of the pollutant during the flow, the source term in (1.3) is set to zero. Consequently, the conservation law (1.4) takes the form

$$(1.5) \quad U_t + \operatorname{div}(\mathbf{a}(\mathbf{x}, t)U) = 0.$$

This equation is linear. In the simple case of one space dimension and a constant velocity field  $\mathbf{a}(\mathbf{x}, t) \equiv a$ , (1.5) reduces to

$$(1.6) \quad U_t + aU_x = 0.$$

The scalar one-dimensional equation (1.6) is often referred to as the transport or *advection* equation.

**The heat equation.** Another illustrative example of a conservation law is provided by heat conduction. Assume that a hot material (like a metal block) is heated at one end and is left to cool afterwards, without providing any additional source of heat. It is a common observation that the heat spreads or *diffuses* out



and the temperature of the material becomes uniform after some time. Let  $U$  be the temperature of the material. Diffusion of heat is governed by Fourier's or Fick's law

$$\mathbf{F}(U) = -\mathbf{k}\nabla U.$$

Here,  $\mathbf{k}$  is the conductivity tensor for the medium. The minus sign is due to the fact that heat flows from hotter to cooler zones. Substituting Fourier's law into the conservation law (1.4), we obtain the *heat equation*

$$(1.7) \quad U_t - \operatorname{div}(\mathbf{k}\nabla U) = 0.$$

If the conductivity is assumed to be unity and the material is one-dimensional (like a rod), (1.7) reduces to the well-known one-dimensional heat equation

$$(1.8) \quad U_t - U_{xx} = 0.$$

The scalar transport equation (1.5) and the heat equation (1.7) are both linear equations and deal with the evolution of a single scalar quantity. As nature is too complicated to be described by scalar linear equations, their utility is limited. Next, we present a *nonlinear system* of conservation laws.

**Euler equations of gas dynamics.** A gas (as an example consider air) consists of a large number of molecules. The motion of each molecule can be tracked individually. This description is termed as the particle description and leads to a very large number of ODEs. The resulting system of ODEs is too large to be computationally feasible. Instead, a more *macroscopic description* is used. In a macroscopic model, the key variables of interest are: the density  $\rho$ , the velocity field  $\mathbf{u}$  and the gas pressure  $p$ . All these quantities can be measured experimentally. The relevant conservation laws are

- *Conservation of mass:* It is well-known in fluid dynamics that the total mass of the gas is conserved. Mathematically, using Kelvin's theorem, this translates into

$$\rho_t + \operatorname{div}(\rho\mathbf{u}) = 0.$$

- *Conservation of momentum:* By Newton's second law of motion, the rate of change of momentum equals force. In the absence of external forces, the gas pressure is the only force acting on the gas. The resulting conservation law is

$$(\rho\mathbf{u})_t + \operatorname{div}(\rho\mathbf{u} \otimes \mathbf{u}) + \nabla p = 0.$$

Note that the above conservation laws implies that the rate of change of the *advective* (material) derivative of the momentum equals the gradient of pressure. This is a consequence of the following observation: Gas flows from high to low pressure.

The symbol  $\otimes$  is the tensor product, i.e, for any two vectors  $\mathbf{a} = (a_1, a_2, a_3)$  and  $\mathbf{b} = (b_1, b_2, b_3)$ , we have

$$\mathbf{a} \otimes \mathbf{b} = \begin{pmatrix} a_1b_1 & a_1b_2 & a_1b_3 \\ a_2b_1 & a_2b_2 & a_2b_3 \\ a_3b_1 & a_3b_2 & a_3b_3 \end{pmatrix}.$$

- *Conservation of energy:* The total energy of a gas is a sum of its kinetic and internal (potential) energy. The kinetic energy has the standard expression

$$E_k = \frac{1}{2} \rho |\mathbf{u}|^2,$$

whereas the internal energy is determined by an equation of state. If the gas is an ideal gas, then the equation of state is

$$E_i = \frac{p}{\gamma - 1},$$

where  $\gamma$  is the gas constant. It takes the values 5/3 and 7/5 for monoatomic and diatomic gases, respectively. Hence, the total energy of an ideal gas is

$$(1.9) \quad E = \frac{p}{\gamma - 1} + \frac{1}{2} \rho |\mathbf{u}|^2.$$

The rate of change of total energy is computed as:

$$E_t + \operatorname{div}((E + p)\mathbf{u}) = 0.$$

All the three conservation laws are combined together and written in divergence form to obtain the *Euler equations* of gas dynamics:

$$(1.10) \quad \begin{aligned} \rho_t + \operatorname{div}(\rho \mathbf{u}) &= 0, \\ (\rho \mathbf{u})_t + \operatorname{div}(\rho \mathbf{u} \otimes \mathbf{u} + p \mathbf{I}) &= 0, \\ E_t + \operatorname{div}((E + p)\mathbf{u}) &= 0, \end{aligned}$$

where  $\mathbf{I}$  denotes the  $3 \times 3$  identity matrix. The above system is an example of a multi-dimensional *nonlinear system* of conservation laws. This derivation of the Euler equations was very brief and details can be found in fluid dynamics textbooks like [6]. We ignore fluid viscosity effects and heat conduction in the gas while deriving (1.10).

The above examples already reveal a multitude of diverse physical phenomena that can be modeled in terms of conservation laws. The flux  $\mathbf{F}$  in (1.4) is often a function of  $\mathbf{U}$  and its derivatives,

$$\mathbf{F} = \mathbf{F}(\mathbf{U}, \nabla \mathbf{U}, \nabla^2 \mathbf{U}, \dots)$$

For simplicity of the analysis, it is common to neglect the role of the higher than first-order derivatives. Hence, the flux is of the form:

$$\mathbf{F} = \mathbf{F}(\mathbf{U}, \nabla \mathbf{U}).$$

If it is of the form  $\mathbf{F} = \mathbf{F}(\mathbf{U})$ , then the conservation law (1.4) is a first-order PDE. It is usually classified as *hyperbolic*. The notion of hyperbolicity will be described in detail in the sequel. The scalar transport equation (1.5) and the Euler equations of gas dynamics (1.10) are examples for hyperbolic equations.

If we have  $\mathbf{F} = \mathbf{F}(\nabla \mathbf{U})$ , then the conservation law (1.4) is a second-order PDE and is often classified as *parabolic*. The heat equation (1.7) is an example of a parabolic equation. When the flux  $\mathbf{F}$  depends on both the function  $\mathbf{U}$  and its first derivative, the conservation law (1.4) is termed as a *convection-diffusion* equation. In these notes, we will consider hyperbolic equations and convection-diffusion equations with the convection dominating the diffusion.

**1.1.1. Other examples.** Examples for conservation laws of both the hyperbolic and convection-diffusion type abound in nature. In these notes, we will consider the scalar Burgers equation, the Buckley-Leverett equation (modeling flows in oil and gas reservoirs), the wave equation, the shallow water equations of meteorology and oceanography, the equations for linear and nonlinear elastic waves that arise in materials science and the equations of magnetohydrodynamics (MHD) from plasma physics.

## 1.2. Content and scope of these notes

The reason for studying conservation laws extensively is obvious: They arise in many models in the sciences, ranging from the design of aircraft (Euler equations) to the study of supernovas in astrophysics (MHD equations). Since interesting conservation laws like the Euler equations are nonlinear, it is not possible to obtain explicit solution formulas. Hence, numerical methods need to be developed for *approximating* or *simulating* the solutions of conservation laws. The design and implementation of efficient numerical methods is the main focus of these notes.

In order to design efficient numerical methods, we need to understand the analytical structure of the solutions of conservation laws. Therefore, we will briefly discuss theoretical properties of the solutions that are relevant for the design and analysis of numerical schemes.

We begin with the study of one-dimensional scalar problems. Both linear and nonlinear equations are considered, and efficient numerical schemes are described for them. Then, the focus shifts to linear and nonlinear systems like the Euler equations of gas dynamics. Finally, we consider the multi-dimensional versions of systems of conservation laws and describe efficient numerical schemes for them.



## CHAPTER 2

### Linear Transport Equations

In this chapter we consider the one-dimensional version of the linear transport equation,

$$(2.1) \quad U_t + a(x, t)U_x = 0 \quad \forall (x, t) \in \mathbb{R} \times \mathbb{R}_+.$$

The simplest case of the scalar transport equation arises when the velocity field is constant, that is,  $a(x, t) \equiv a$ . The resulting transport equation is

$$(2.2) \quad U_t + aU_x = 0.$$

The rather simple equation (2.2) has served as a crucible for designing highly efficient schemes for much more complicated systems of equations. We concentrate on it for the rest of this chapter.

#### 2.1. Method of characteristics

The initial value problem (or Cauchy problem) for (2.1) consists of finding a solution of (2.1) that also satisfies the initial condition

$$(2.3) \quad U(x, 0) = U_0(x) \quad \forall x \in \mathbb{R}.$$

It is well known that the solution of the initial value problem can be constructed by using the *method of characteristics*. The idea underlying this method is to reduce a PDE like (2.1) to an ODE by utilizing the structure of the solutions. As an ansatz, assume that we are given some curve  $x(t)$ , along which the solution  $U$  is constant. This means that

$$\begin{aligned} 0 &= \frac{d}{dt}U(x(t), t) && \text{(as } U \text{ is constant along } x(t)) \\ &= U_t(x(t), t) + U_x(x(t), t)x'(t) && \text{(chain rule).} \end{aligned}$$

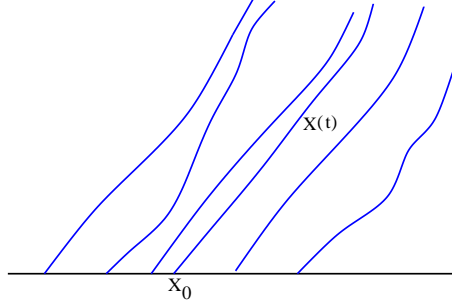
We also know that  $U_t(x(t), t) + U_x(x(t), t)a(x(t), t) = 0$ , since  $U$  is assumed to be a solution of (2.1). Therefore, if  $x(t)$  satisfies the ODE

$$(2.4) \quad \begin{aligned} x'(t) &= a(x(t), t) \\ x(0) &= x_0, \end{aligned}$$

then  $x(t)$  is precisely such a curve. The solution  $x(t)$  of this equation is called a *characteristic curve*. From ODE theory, we know that solutions of (2.4) exist provided that  $a$  is Lipschitz continuous in both arguments. It may or may not be possible to find an explicit solution formula for (2.4).

The importance of characteristic curves lies in the property that  $U$  is constant along them:

$$U(x(t), t) = U(x(0), 0) = U_0(x_0).$$

FIGURE 2.1. Characteristics curves  $x(t)$  for (2.1)

The initial data  $U_0(x)$  is already known, so if we can find characteristic curves that go through all points  $(x, t) \in \mathbb{R} \times \mathbb{R}_+$ , then we have found the solution  $U$  at all points in the plane. (See Figure 2.1) for an illustration.)

In the simple case of a constant velocity field  $a(x, t) \equiv a$ , the characteristic equation (2.4) is explicitly solved as

$$x(t) = x_0 + at.$$

Therefore, given some point  $(x, t)$ , the unique characteristic that goes through  $(x, t)$  (so that  $x(t) = x$ ) has initial value  $x_0 = x - at$ . Hence, the solution of (2.2) is

$$(2.5) \quad U(x, t) = U_0(x_0) = U_0(x - at)$$

for any  $(x, t) \in \mathbb{R} \times \mathbb{R}_+$ . The solution formula (2.5) implies that the initial data is transported with the velocity  $a$ .

In the more general case of (2.1), the characteristic equation (2.4) may not be possible to solve explicitly. Hence, it is essential that we obtain some information about the structure of solutions of (2.1) from the equation itself. This is done by means of the following *a priori* energy estimate:

**Lemma 2.1.** *Let  $U(x, t)$  be a smooth solution of (2.1) which decays to zero at infinity, i.e.,  $\lim_{|x| \rightarrow \infty} U(x, t) = 0$  for all  $t \in \mathbb{R}_+$ , and assume that  $a \in C^1(\mathbb{R}, \mathbb{R}_+)$ .*

*Then  $U$  satisfies the energy bound*

$$(2.6) \quad \int_{\mathbb{R}} U^2(x, t) dx \leq e^{\|a\|_{C^1} t} \int_{\mathbb{R}} U_0^2(x) dx$$

*for all times  $t > 0$ .*

**PROOF.** The proof of the estimate (2.6) is based on multiplying (2.1) with  $U$  on both sides:

$$UU_t + a(x, t)UU_x = 0 \quad (\text{multiplying (2.1) by } U)$$

$$\left(\frac{U^2}{2}\right)_t + a(x, t)\left(\frac{U^2}{2}\right)_x = 0 \quad (\text{chain rule})$$

$$\left(\frac{U^2}{2}\right)_t + \left(a(x, t)\frac{U^2}{2}\right)_x = a_x(x, t)\frac{U^2}{2} \quad (\text{product rule})$$

$$\frac{d}{dt} \int_{\mathbb{R}} \left(\frac{U^2}{2}\right) dx + \int_{\mathbb{R}} \left(a(x, t)\frac{U^2}{2}\right)_x dx = \int_{\mathbb{R}} a_x(x, t)\frac{U^2}{2} dx \quad (\text{integrating over space})$$

$$\begin{aligned} \frac{d}{dt} \int_{\mathbb{R}} \left( \frac{U^2}{2} \right) dx &= \int_{\mathbb{R}} a_x(x, t) \frac{U^2}{2} dx \quad (\text{decay to zero at infinity}) \\ &\leq \|a\|_{C^1} \int_{\mathbb{R}} \frac{U^2}{2} dx \quad (\text{regularity of } a). \end{aligned}$$

The last inequality can be used together with Gronwall's inequality (Theorem A.1) to obtain the bound (2.6).  $\square$

The quantity  $\int U^2/2$  is commonly called the *energy* of the solution. The above lemma shows that the energy of the solutions to the transport equation (2.1) are bounded. The energy estimate is going to be used for designing robust schemes for the transport equation. We remark that the restriction that  $U$  decays to zero at infinity may be relaxed by considering a different energy functional.

## 2.2. Finite difference schemes for the transport equation

It may not be possible to obtain an explicit formula for the solution of the characteristic equation (2.4). For example, the velocity field  $a(x, t)$  might have a complicated nonlinear expression. Hence, we have to devise numerical methods for approximating the solutions of (2.1). For simplicity, we consider  $a(x, t) \equiv a > 0$  and solve (2.2). It is rather straightforward to extend the schemes to the case of a more general velocity field.

**Discretization of the domain.** The first step in any numerical method is to discretize both the spatial and temporal parts of the domain. Since  $\mathbb{R}$  is unbounded, we have to truncate the domain to some bounded domain  $[x_l, x_r]$ . This truncation implies that suitable boundary conditions need to be imposed. We discuss the problem of boundary conditions later on.

For the sake of simplicity, the domain  $[x_l, x_r]$  is discretized uniformly with a mesh size  $\Delta x$  into a sequence of  $N + 1$  points  $x_j$  such that  $x_0 = x_l$ ,  $x_N = x_r$  and  $x_{j+1} - x_j = \Delta x$  for all  $j$ . A non-uniform discretization can readily be considered.

For the temporal discretization, we choose some terminal time  $T$  and divide  $[0, T]$  into  $M$  points  $t^n = n\Delta t$  ( $n = 0, \dots, M$ ). The space-time mesh is shown in Figure 2.2. Our aim is obtain an approximation of the form  $U_j^n \approx U(x_j, t^n)$ . To get from the initial time step  $t^0$  to the terminal time step  $t^M$ , we first set the initial data  $U_j^0 = U_0(x_0)$  for all  $j$ . Then the solution  $U_j^1$  at the next time step is computed

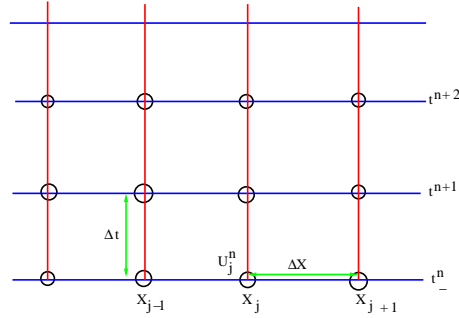


FIGURE 2.2. A representation of the mesh in space-time

using some update formula, again for all  $j$ . This process is reiterated until we arrive at the final time step  $t^M = T$  with our final solution  $U_j^M$ .

**A simple centered finite difference scheme.** On the mesh, we need to approximate the transport equation (2.2). We do so by replacing both the spatial and temporal derivatives by finite differences. The time derivative is replaced with a forward difference and the spatial derivative with a central difference. This combination is standard (see schemes for the heat equation in standard textbooks like [7]). The resulting scheme is

$$(2.7) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a(U_{j+1}^n - U_{j-1}^n)}{2\Delta x} = 0 \quad \text{for } j = 1, \dots, N-1.$$

Some special care must be taken when defining the boundary values. We have a consistent discretization of (2.2) that is very simple to implement. We test it on the following numerical example.

**A numerical example.** Consider the linear transport equation (2.2) in the domain  $[0, 1]$  with initial data

$$(2.8) \quad U_0(x) = \sin(2\pi x).$$

Since the data is periodic, it is natural to assume periodic boundary conditions. We implement this numerically by letting

$$U_0^n = U_{N-1}^n, \quad U_N^n = U_1^n.$$

The exact solution is calculated by (2.5) as  $U(x, t) = \sin(2\pi(x - at))$ . We set  $a = 1$  and compute the solutions with the central scheme (2.7) with 500 mesh points, and plot the solution at time  $t = 0.3$  in Figure 2.3. The figure clearly shows that, despite being a consistent approximation, the scheme is unstable, with very large oscillations.

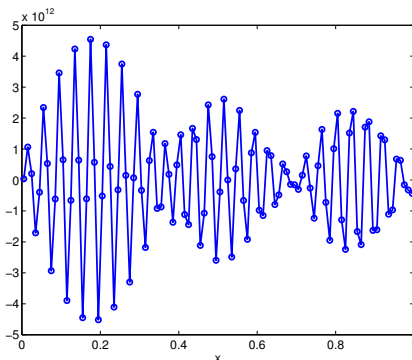


FIGURE 2.3. Approximate solution for (2.2) with the central scheme (2.7) at time  $t = 3$  with 100 mesh points. [central.m]



**A physical explanation.** Why do the solutions computed with the central scheme (2.7) blow up? After all, the central scheme seems a reasonable approximation of the transport equation. A *physical* explanation can be deduced from the following argument: The exact solution moves to the right (as  $a > 0$ ) with a fixed speed. Therefore, information goes from left to right. However, the central scheme (see Figure 2.4) takes information from both the left and the right, violating the physics. Consequently, the solutions are unstable. This explanation seems intuitive but has to be backed by solid mathematical arguments. We proceed to do so below.

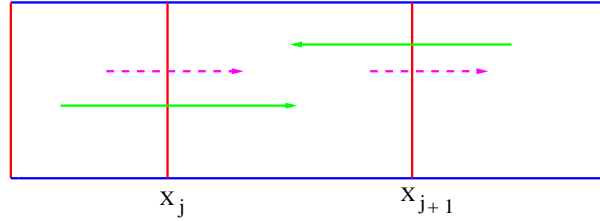


FIGURE 2.4. The central scheme (2.7). Green arrows indicate numerical propagation and magenta arrows physical propagation.

**A mathematical explanation.** The observed instability of the central scheme can be explained mathematically in terms of estimates. We recall that the exact solutions have a bounded energy (see estimate (2.6)). It is reasonable to require that the scheme is *energy stable* like the exact solution, that is, a discrete version of energy remains bounded. For a given  $\Delta x$ , we define the discrete version of energy as

$$(2.9) \quad E^n = \frac{1}{2} \Delta x \sum_j (U_j^n)^2.$$

Note that the integral in the energy for the continuous problem has been replaced with a Riemann sum.

**Lemma 2.2.** *Let  $U_j^n$  be the solutions computed with the central scheme (2.7). Then the following estimate holds:*

$$(2.10) \quad E^{n+1} = E^n + \frac{\Delta x}{2} \sum_j (U_j^{n+1} - U_j^n)^2.$$

*Consequently, the energy grows at every time step for any choice of  $\Delta x, \Delta t$ , and so the scheme is unconditionally unstable.*

**PROOF.** We mimic the steps of continuous energy estimate (Lemma 2.1) and multiply both sides of the scheme (2.7) by  $U_j^n$  to obtain

$$(2.11) \quad U_j^n (U_j^{n+1} - U_j^n) + \frac{a\Delta t}{\Delta x} (U_j^n U_{j+1}^n - U_j^n U_{j-1}^n) = 0.$$

We have the following elementary identity:

$$(2.12) \quad d_2(d_1 - d_2) = \frac{(d_1)^2}{2} - \frac{(d_2)^2}{2} - \frac{1}{2}(d_1 - d_2)^2$$

for any two numbers  $d_1, d_2$ . We denote

$$H_{j+1/2} = a \frac{U_j^n U_j^{n+1}}{2}$$

to reduce (2.11) to

$$(2.13) \quad \frac{(U_j^{n+1})^2}{2} = \frac{(U_j^n)^2}{2} + \frac{1}{2}(U_j^{n+1} - U_j^n)^2 - \frac{\Delta t}{\Delta x}(H_{j+1/2} - H_{j-1/2}).$$

Summing (2.13) over all  $j$  and using zero (or periodic) boundary conditions, the flux term  $H$  vanishes by cancellation and we obtain the estimate (2.10).  $\square$

Although we assumed zero or periodic boundary conditions in the proof of this lemma, a variant of the lemma holds for more general boundary conditions, as for the continuous setting in Lemma 2.1.

The above lemma provides a mathematical justification for our physical intuition. The central scheme leads to a growth of energy at every time step and is unstable. We need to find schemes that possess a discrete version of the energy estimate. This use of rigorous mathematical tools like energy analysis to justify physical reasoning will be an essential ingredient of these notes.

### 2.3. An upwind scheme

The central scheme (2.7) does not respect the direction of propagation of information for the transport equation (2.2). Hence, we must include the correct direction of information propagation and hope that it stabilizes the scheme. This entails using one-sided differences instead of a central difference to approximate the linear transport equation (2.2).

If  $a > 0$  and the direction of information propagation is from left to right, then we can use a backward difference in space to obtain the scheme

$$(2.14) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a(U_j^n - U_{j-1}^n)}{\Delta x} = 0 \quad \text{for } j = 1, \dots, N-1,$$

and if  $a < 0$ , we can use the forward difference to obtain:

$$(2.15) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a(U_{j+1}^n - U_j^n)}{\Delta x} = 0 \quad \text{for } j = 1, \dots, N-1.$$

Using the notation

$$a^+ = \max\{a, 0\}, \quad a^- = \min\{a, 0\}, \quad |a| = a^+ - a^-,$$

(2.14) and (2.15) can be written together as

$$(2.16) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a^+(U_j^n - U_{j-1}^n)}{\Delta x} + \frac{a^-(U_{j+1}^n - U_j^n)}{\Delta x} = 0.$$

The above scheme takes into account the direction of propagation of information – information is “carried with the wind”. Hence, this scheme is termed as the *upwind* scheme.

Using the definition of the absolute value and some simple algebraic manipulations, the upwind scheme (2.16) can be recast as

$$(2.17) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a(U_{j+1}^n - U_{j-1}^n)}{2\Delta x} = \frac{|a|}{2\Delta x}(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$

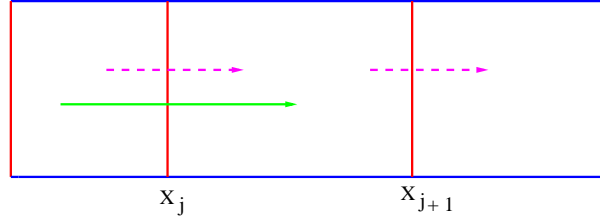


FIGURE 2.5. The upwind scheme (2.16). Green arrows indicate numerical propagation and magenta arrows physical propagation.

(compare to (2.7)). Note that in the above form, the spatial derivatives are the central term and a *diffusion* term. The right hand side of (2.17) approximates  $\frac{\Delta x |a|}{2} U_{xx}$ . Hence, the upwind scheme (2.17) adds *numerical viscosity* or *diffusion* to the unstable central scheme (2.7). Numerical viscosity is going to play a crucial role later on.

Since the upwind scheme incorporates the correct direction of propagation of information (see Figure 2.5), we expect it to be more stable than the central scheme. This is endorsed by the numerical experiment with initial data (2.8). We take  $a = 1$  and compute approximate solutions for the linear transport equation (2.2) on a uniform mesh with 100 mesh points up to  $t = 1$ . We use two different timesteps:  $\Delta t = 1.3\Delta x$  and  $\Delta t = 0.9\Delta x$ . As seen in Figure 2.6, the results with  $\Delta t = 1.3\Delta x$  are still oscillatory and the scheme continues to be unstable. In spite of the upwinding, stability still seems to be elusive. However, results with  $\Delta t = 0.9\Delta x$  are stable. The approximation appears to be good in this case. Much better results are obtained by refining the mesh, while keeping the ratio  $\Delta t/\Delta x$  fixed, as is presented in Figure 2.7.

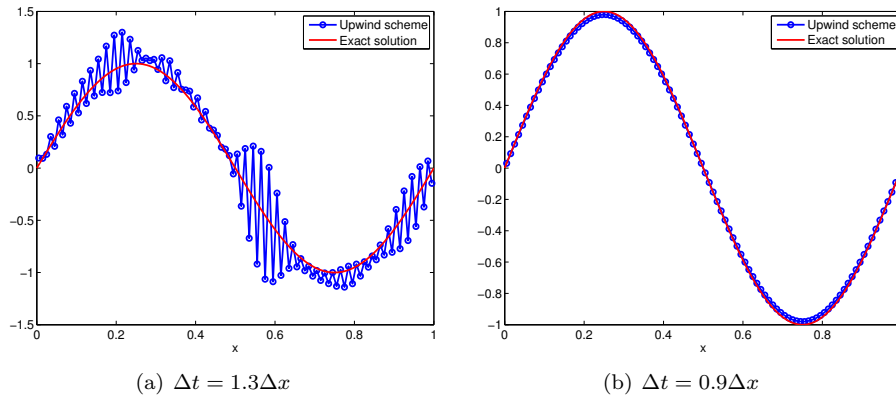


FIGURE 2.6. Solution with initial data (2.8) at  $t = 1$ . The ratio  $\Delta t/\Delta x$  is important for stability. [upwind.cfl.m]

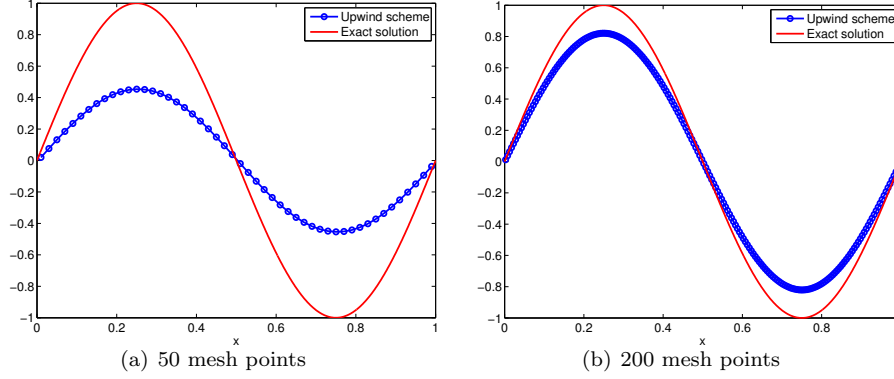


FIGURE 2.7. Solution with initial data (2.8) at  $t = 10$ . Refining the mesh gives a more accurate solution. [upwind\_refinement.m]

#### 2.4. Stability for the upwind scheme

The numerical results indicate that stability for the upwind scheme is subtle. It is not unconditionally unstable as the central scheme (2.7); instead, stability depends on the parameters  $\Delta x, \Delta t$ . Numerical results indicate the crucial role played by the ratio  $\frac{\Delta t}{\Delta x}$ . It seems that one must not only take into account the correct direction of propagation, but also the correct magnitude.

The quantification of stability will involve energy analysis as in the last section. We have the following stability result:

**Lemma 2.3.** *Let the mesh parameters satisfy the condition*

$$(2.18) \quad |a| \frac{\Delta t}{\Delta x} \leq 1.$$

*Then solutions computed with the upwind scheme (2.17) satisfy the energy estimate*

$$(2.19) \quad E^{n+1} \leq E^n,$$

*where the energy is defined as in (2.9). The upwind scheme is thus conditionally stable.*

**PROOF.** For the sake of simplicity, we assume that  $a > 0$ . Hence the upwind scheme (2.17) reduces to

$$(2.20) \quad \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{a(U_{j+1}^n - U_{j-1}^n)}{2\Delta x} = \frac{a}{2\Delta x}(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

It is also equivalent to the scheme (2.14). As in the proof of the estimate (2.10) we multiply both sides of the scheme (2.20) by  $U_j^n$  to obtain

$$(2.21) \quad \begin{aligned} U_j^n(U_j^{n+1} - U_j^n) &= -\frac{a\Delta t}{2\Delta x}(U_j^n U_{j+1}^n - U_j^n U_{j-1}^n) \\ &\quad + \frac{a\Delta t}{2\Delta x}(U_j^n(U_{j+1}^n - U_j^n)) + \frac{a\Delta t}{2\Delta x}(U_j^n(U_{j-1}^n - U_j^n)). \end{aligned}$$

Now we use elementary identity (2.12) a couple of times and rewrite (2.21) as

$$\begin{aligned}
 (2.22) \quad \frac{(U_j^{n+1})^2}{2} &= \frac{(U_j^n)^2}{2} + \frac{(U_j^{n+1} - U_j^n)^2}{2} - \frac{a\Delta t}{2\Delta x}(U_j^n U_{j+1}^n - U_j^n U_{j-1}^n) \\
 &+ \frac{a\Delta t}{4\Delta x}((U_{j+1}^n)^2 - (U_j^n)^2) - \frac{a\Delta t}{4\Delta x}((U_j^n)^2 - (U_{j-1}^n)^2) \\
 &- \frac{a\Delta t}{4\Delta x}(U_{j+1}^n - U_j^n)^2 - \frac{a\Delta t}{4\Delta x}(U_j^n - U_{j-1}^n)^2.
 \end{aligned}$$

Denoting

$$K_{j+1/2} = \frac{a}{2}(U_j^n U_{j+1}^n) - \frac{a}{4}((U_{j+1}^n)^2 - (U_j^n)^2),$$

we may rewrite (2.22) as

$$\begin{aligned}
 (2.23) \quad \frac{(U_j^{n+1})^2}{2} &= \frac{(U_j^n)^2}{2} + \frac{(U_j^{n+1} - U_j^n)^2}{2} - \frac{a\Delta t}{\Delta x}(K_{j+1/2} - K_{j-1/2}) \\
 &- \frac{a\Delta t}{4\Delta x}(U_{j+1}^n - U_j^n)^2 - \frac{a\Delta t}{4\Delta x}(U_j^n - U_{j-1}^n)^2.
 \end{aligned}$$

Summing (2.23) over all  $j$  and using the definition of discrete energy (2.9) and either zero or periodic boundary conditions, we obtain

$$(2.24) \quad E^{n+1} \leq E^n + \frac{\Delta x}{2} \sum_j (U_j^{n+1} - U_j^n)^2 - \frac{a\Delta t}{2} \sum_j (U_j^n - U_{j-1}^n)^2.$$

Using the definition of the upwind scheme (2.14) in (2.24) yields

$$(2.25) \quad E^{n+1} \leq E^n + \left( \frac{a^2 \Delta t^2}{2\Delta x} - \frac{a\Delta t}{2} \right) \sum_j (U_j^n - U_{j-1}^n)^2.$$

Since the term in the sum in (2.25) is positive, we obtain the energy bound (2.19), provided

$$\frac{a^2 \Delta t^2}{\Delta x} \leq a\Delta t,$$

which is precisely the condition (2.18).  $\square$

The stability condition (2.18) is termed the *CFL condition* after Courant, Friedrichs and Lewy who first proposed it. The conditional stability of the upwind scheme is confirmed in numerical experiments.

**Numerical experiment: Discontinuous data.** Consider the transport equation (2.2) with  $a = 1$  in the domain  $[0, 1]$  and initial data

$$(2.26) \quad U_0(x) = \begin{cases} 2 & \text{if } x < 0.5 \\ 1 & \text{if } x > 0.5. \end{cases}$$

The initial data and consequently the exact solution (2.5) are discontinuous. We compute with the upwind scheme using 50 and 200 mesh points and display the results in Figure 2.8. The results show that the upwind scheme approximates the solution quite well, at least at a fine resolution. However the errors on a coarse mesh are somewhat large. This issue will be addressed in later sections.

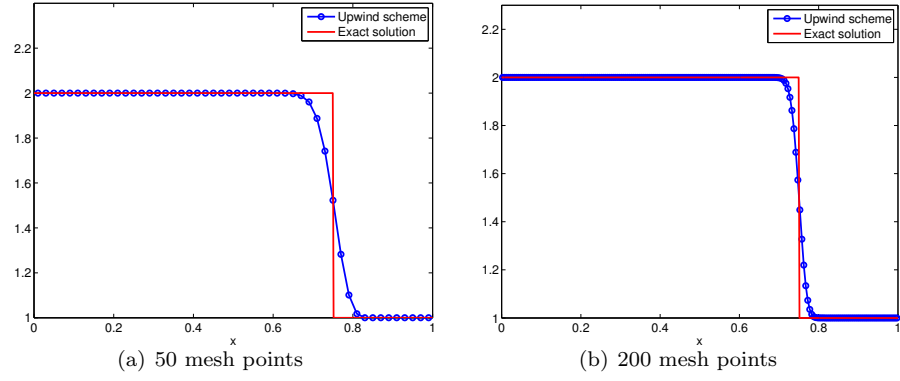


FIGURE 2.8. The upwind scheme (2.14) for advection (2.2) with discontinuous initial data (2.26). Both results are at time  $t = 0.25$ .  
[upwind\_disc\_refinement.m]

## CHAPTER 3

### Scalar conservation laws

In the previous section, we considered the scalar transport equation

$$(3.1) \quad U_t + a(x, t)U_x = 0.$$

This equation is linear as the velocity field  $a$  is a given function. However, most natural phenomena are nonlinear. In such models, the linear velocity field must be replaced with a field that depends on the solution itself. The simplest example of such a field is

$$a(x, t) = U(x, t).$$

Hence, the transport equation (3.1) becomes

$$(3.2) \quad U_t + UU_x = 0.$$

The transport equation (3.2) can be written in the *conservative form*

$$(3.3) \quad U_t + \left( \frac{U^2}{2} \right)_x = 0.$$

This is the inviscid *Burgers* equation. It serves as a prototype for *scalar conservation laws*, which in general take the form

$$(3.4) \quad U_t + f(U)_x = 0,$$

where  $U$  is the unknown and  $f$  is the flux function. Apart from Burgers' equation, scalar conservation laws arise in a wide variety of models. We consider a couple of examples below.

**Traffic flow model.** For simplicity, consider a one-dimensional highway and denote the density of cars (number of cars per square meter) as  $U(x, t)$ . Assume that the cars are moving at a *macroscopic* velocity (the speed of a traffic column)  $V(x, t)$ . A simple requirement of conservation of the number of cars lead to the following equation:

$$(3.5) \quad U_t + (UV)_x = 0.$$

The velocity  $V$  remains to be modeled. One very simple model is based on a couple of observations. First, there exists a maximum velocity at which an individual car can drive, for example specified by the speed limit. Second, the velocity of cars is inversely proportional to the car density. If there are a large number of cars, each individual driver will drive slowly. However, on a remote stretch of the highway, each driver speeds up. These simple observations are combined to yield the velocity

$$V = V_{\max}(1 - U),$$

where  $V_{\max}$  is the maximum velocity for the cars. We use the convention that the maximum density or *road carrying capacity* is 1. Hence, the traffic flow equation is

$$(3.6) \quad U_t + (V_{\max}U(1 - U))_x = 0.$$

**Enhanced oil recovery.** Oil is generally found in sub-surface reservoirs, inside permeable rocks. The primary stage of oil recovery consists of drilling into the rocks and extracting oil by applying pressure. Only 20 to 30 percent of the available oil can be extracted in this manner. The secondary stage of oil recovery consists of injecting water into the rock bed. The water displaces the oil (as water is heavier) and the oil can then be extracted. This complex process is modeled by using two-phase flow (water and oil) in a porous media (rock).

For simplicity, we assume that the reservoir is one-dimensional. The quantities of interest are the oil and water volume fractions or *saturations*  $S^o$  and  $S^w$ , respectively. Being volume fractions, they satisfy

$$(3.7) \quad S^o + S^w \equiv 1.$$

Furthermore, the phases evolve according to the conservation laws

$$(3.8) \quad \begin{cases} S_t^o + V_x^o = 0 \\ S_t^w + V_x^w = 0. \end{cases}$$

The phase velocities  $V^o, V^w$  are modeled by Darcy's law:

$$(3.9) \quad \begin{cases} V^o = -\lambda^o \frac{dP^o}{dx} \\ V^w = -\lambda^w \frac{dP^w}{dx}, \end{cases}$$

where  $\lambda$  and  $P$  are the phase mobility and the phase pressure, respectively. In the above constitutive relation, we have neglected the role of gravity. Furthermore, we can assume that there is no capillary pressure:

$$P^o = P^w.$$

Adding the phase saturation equations (3.8) for each phase and using the requirement (3.7), we obtain

$$(V^o + V^w)_x \equiv 0 \quad \Rightarrow \quad V^o + V^w = q,$$

for some constant  $q$  called the total flow rate. Substituting Darcy's law (3.9) in the above identity and using  $P^w = P^o = P$ , we obtain

$$\frac{dP}{dx} = -\frac{q}{\lambda^o + \lambda^w}.$$

Applying this identity in the evolution of the oil saturation (3.9) and (3.8) yields

$$(3.10) \quad S_t^o + \left( \frac{q\lambda^o}{\lambda^w + \lambda^o} \right)_x = 0.$$

The mobilities generally take the form

$$\lambda^o = (S^o)^2, \quad \lambda^w = (S^w)^2 = (1 - S^o)^2.$$

Hence, the evolution of the oil saturation is governed by the scalar conservation law

$$(3.11) \quad S_t^o + \left( \frac{q(S^o)^2}{(S^o)^2 + (1 - S^o)^2} \right)_x = 0.$$

The above examples demonstrate that scalar conservation laws do occur in many interesting models in physics and engineering. Furthermore, the shape of the flux function  $f$  in (3.4) can be very general. Note that it is convex for Burgers' equation, concave for the traffic flow problem (3.6) and is neither convex nor concave (contains inflection points) for the oil reservoir equation (3.11).



In this section, we embark on a systematic study of scalar conservation laws (3.4) from a theoretical perspective.

### 3.1. Characteristics for Burger's equation

We start with Burgers' equation (3.3) and attempt to construct solutions to the initial value problem associated with it. As for the linear transport equation (3.1), we will use the method of characteristics for this purpose. Since (3.2) and (3.3) are equivalent whenever  $U$  is smooth, the characteristics  $x(t)$  for Burgers' equation are given by

$$(3.12) \quad \begin{aligned} x'(t) &= U(x(t), t) \\ x(0) &= x_0. \end{aligned}$$

Note that these characteristics are different from the linear case (2.4) in that the velocity depends on the solution. We consider initial data

$$(3.13) \quad U_0(x) = \begin{cases} U_l & \text{if } x < 0 \\ U_r & \text{if } x > 0. \end{cases}$$

Data of this form is quite simple and consists of constants separated by a discontinuity at the origin. The initial value problem for a conservation law (3.4) with initial data of the form (3.13) is called a *Riemann problem*.

By definition, the solution  $U$  is constant along characteristics, that is,  $U(x(t), t) = U_0(x_0)$ . Therefore, the solution of (3.12), (3.13) in constant parts of  $U_0$  is

$$x(t) = U_0(x_0)t + x_0.$$

Let  $U_l = 1$  and  $U_r = 0$  in (3.13). For  $x_0 < 0$  the characteristics have velocity  $U_0(x_0) = 1$ , whereas for  $x_0 > 0$  they have velocity 0; see Figure 3.1. We see that the characteristics intersect almost instantaneously. As observed in the last section, the solution should be constant (in time) along the characteristics. What happens to the solution when the characteristics start to intersect? How can the solution be defined in this case? Adding nonlinearity completely changes the situation from the linear case.

Is the intersection of characteristics on account of discontinuous data (3.13)? Can using smooth data lead to non-intersecting characteristics? It turns out that even smooth initial data can lead to the intersection of characteristics after a small time interval. Consider the visual example in Figure 3.2.

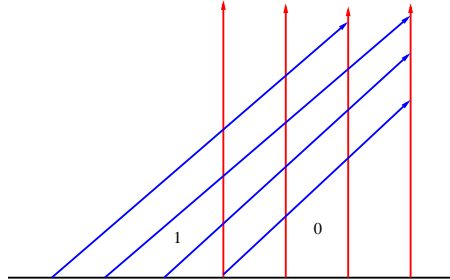


FIGURE 3.1. Characteristics intersecting for the Riemann problem (3.13) with  $(U_l, U_r) = (1, 0)$ .

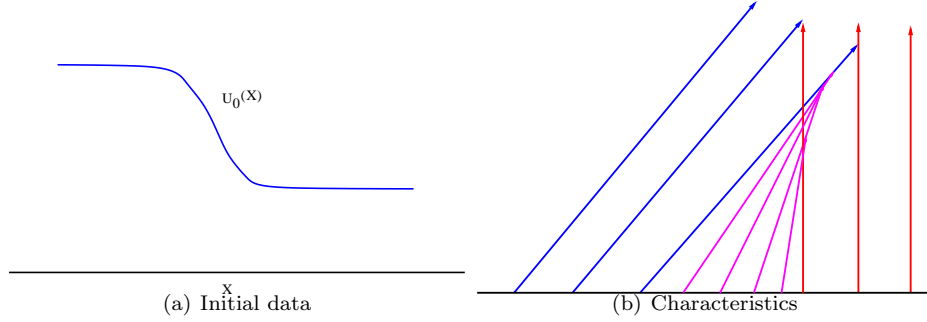


FIGURE 3.2. Characteristics can even intersect for smooth initial data.

**Exercise 3.1.** Let  $U_0(x)$  be differentiable with at least one point  $x$  such that  $U'_0(x) < 0$ . Show that the solution to Burgers' equation with initial data  $U_0$  will develop a discontinuity at time

$$t_{\min} = -\frac{1}{\min_{x \in \mathbb{R}} U'_0(x)}.$$

(Hint: Start with the ansatz that two characteristics  $x(t)$  and  $\tilde{x}(t)$  intersect at some time  $t$ .)

The strange behavior of characteristics indicates that smooth solutions cannot be obtained for the conservation law (3.4), even when the initial data is smooth. Consider the initial data

$$U_0(x) = \sin(\pi x)$$

in the interval  $[-1, 1]$ . A heuristic interpretation of the characteristic equation (3.12) is that the solution at each point  $x$  moves with the velocity  $U_0(x)$ . Hence, the method of characteristics imply that the solution behaves as shown in Figure 3.3. The wave compresses in one part and stretches in another. In particular, the solution can be multi-valued. This is another indication that smooth solutions of (3.4) do not exist. A formal calculation by differentiating (3.2) with respect to  $x$  yields

$$(3.14) \quad V_t + UV_x = -V^2,$$

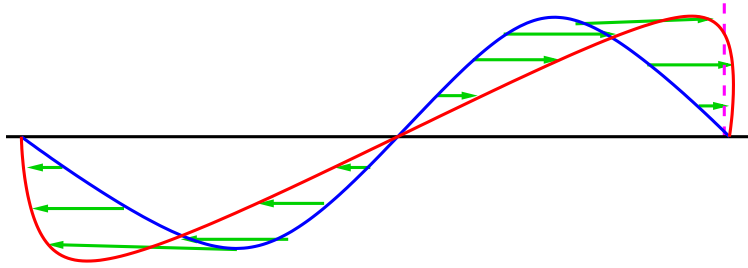


FIGURE 3.3. Smooth initial data leading to multi-valued solution.

where  $V = U_x$ . Hence, along the characteristics  $x(t)$  given by (3.12),  $V$  varies as

$$\frac{d}{dt}V(x(t), t) = -V^2(x(t), t).$$

This is a ODE with quadratic nonlinearity and it is well known that the resulting solution  $V$  can blow up in finite time. Hence, the spatial derivative of the solution to Burgers' equation can blow up, even if the initial derivative is very small. This derivative blowup suggests that smooth solutions to (3.4) may not exist.

### 3.2. Weak solutions

The previous section demonstrates that smooth or classical solutions of the conservation law (3.4) may not exist. However, these models arise in physics and so *some* form of solution does exist. This type of solution is a *weak solution*. To motivate the definition of weak solutions, assume for the moment that smooth solutions of (3.4) exist and multiply both sides by a smooth test function  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+)$ . The space  $C_c^1$  is the space of all continuously differentiable functions with compact support, that is, the functions vanish outside a compact subset of the domain. Using integration by parts, (3.4) reduces to

$$(3.15) \quad \int_{\mathbb{R} \times \mathbb{R}_+} U \varphi_t + f(U) \varphi_x \, dx dt + \int_{\mathbb{R}} U_0(x) \varphi(x, 0) \, dx = 0.$$

This identity holds true for all test functions  $\varphi$ . We base the definition of weak solution for (3.4) on the above identity.

**Definition 3.2** (Weak solution). *A function  $U \in L^1(\mathbb{R} \times \mathbb{R}_+)$  is a weak solution of (3.4) if the identity (3.15) holds for all test functions  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+)$ .*

Note that the identity (3.15) is well-defined as long as  $U \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ .

**Exercise 3.3.** *Show that if a weak solution  $U$  of (3.4) is also differentiable (so  $U \in C^1(\mathbb{R} \times \mathbb{R}_+)$ ), then  $U$  satisfies (3.4) point-wise. Hence, the class of weak solutions contains, but is not restricted to, classical solutions.*

The concept of weak solutions is highly non-standard. Our usual understanding of solutions of PDEs is classical – the solutions must be differentiable functions. However, weak solutions are not necessarily differentiable, not even continuous. This implies that the solutions can contain discontinuities. These discontinuities appear in nature as *shock waves*.

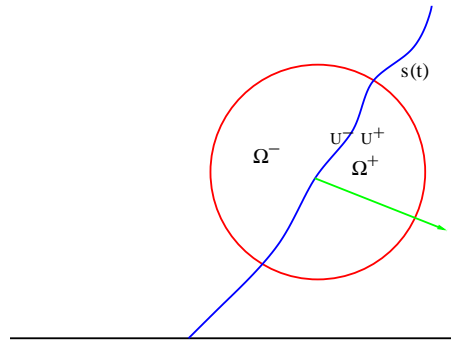


FIGURE 3.4. What happens across a shock?

**3.2.1. The Rankine-Hugoniot condition.** As we will soon find out, shock waves in weak solutions cannot be arbitrary curves in the  $x$ - $t$ -plane, but must satisfy certain conditions. Assume that we are given a weak solution  $U$  consisting of two smooth regions, separated by a shock wave, as depicted in Figure 3.4. Let the shock wave be defined by the curve  $x = \sigma(t)$ .

Let  $\varphi$  be a test function with support in  $\Omega$ . We assume that  $U \in C^1(\Omega^-)$  and  $U \in C^1(\Omega^+)$ ; see Figure 3.4. Integrating (3.15) by parts and using the compact support of the test function, we get

$$\begin{aligned} \int_{\Omega} U \varphi_t + f(U) \varphi_x \, d\Omega &= \int_{\Omega^+} U \varphi_t + f(U) \varphi_x \, d\Omega + \int_{\Omega^-} U \varphi_t + f(U) \varphi_x \, d\Omega \\ &= - \int_{\Omega^+} (U_t + f(U)_x) \varphi \, d\Omega + \int_{\partial\Omega^+} (U^+(t) \nu^t + f(U^+(t)) \nu^x) \varphi \, d\Omega \\ &\quad - \int_{\Omega^-} (U_t + f(U)_x) \varphi \, d\Omega + \int_{\partial\Omega^-} (U^-(t) \nu^t + f(U^-(t)) \nu^x) \varphi \, d\Omega \\ &= 0. \end{aligned}$$

Here,  $U^+(t)$  and  $U^-(t)$  are the trace values of  $U$  on the right and left of the discontinuity  $\sigma$ , and  $\nu$  is the unit outward normal of  $\sigma$  (see Figure 3.4). We have

$$(\nu^t, \nu^x) = (-s(t), 1),$$

where  $s(t) = \sigma'(t)$  is the speed of the shock curve. Since  $U$  is smooth in  $\Omega^-$  and  $\Omega^+$ , the equation (3.4) is satisfied point-wise. Therefore, the above identities imply that

$$\int_{\Omega^- \cup \Omega^+} \underbrace{(U_t + f(U)_x)}_{=0} \varphi \, d\Omega + \int_{\partial\Omega} (s(t) (U^+(t) - U^-(t)) - (f(U^+) - f(U^-))) \varphi \, d\Omega = 0.$$

Since  $\varphi$  is an arbitrary test function, the integrand of the remaining integral must be identically equal to zero. Hence, the shock speed must satisfy

$$(3.16) \quad s(t) = \frac{f(U^+(t)) - f(U^-(t))}{U^+(t) - U^-(t)}.$$

This condition is called the *Rankine-Hugoniot* condition.

**3.2.2. Solutions to Riemann problems.** Consider Burgers' equation (3.3) with the Riemann problem (3.13) with  $U_l = 1$  and  $U_r = 0$ . We recall that the characteristics intersected in this case and a smooth solution couldn't be constructed. We construct a weak solution that consists of two constant states  $U_l$  and  $U_r$ , separated by a shock moving at a speed given by the Rankine-Hugoniot condition (3.16),

$$s(t) = \frac{U_r^2 - U_l^2}{2(U_r - U_l)} \equiv \frac{1}{2}.$$

Hence, the weak solution takes the form

$$(3.17) \quad U(x, t) = \begin{cases} 1 & \text{if } x < \frac{1}{2}t \\ 0 & \text{if } x > \frac{1}{2}t. \end{cases}$$

It is easy to check that (3.17) satisfies (3.15). The structure of the solution (see Figure 3.5) shows that the characteristics *flow into* the shock. As a consequence, there are characteristics covering all points in the plane, and for each point we

can trace a characteristic back to the initial data. Hence, the entire solution is prescribed by the initial data.

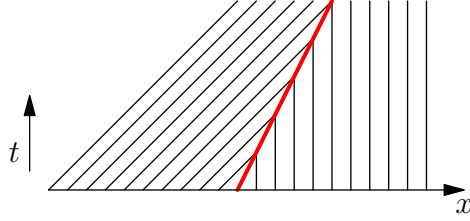


FIGURE 3.5. Characteristics for the Riemann problem (3.17).

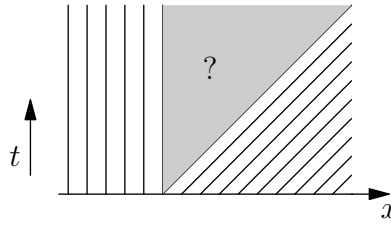


FIGURE 3.6. Characteristics for the Riemann problem (3.18).

Next, we consider another Riemann problem with  $U_l = 0$  and  $U_r = 1$ . If we follow characteristics emanating from the  $x$ -axis, as for the previous problem, we now get an area without characteristics; see Figure 3.6. The "missing" information in this area may be "filled" in several ways. Using the Rankine-Hugoniot condition, we find that one possible weak solution is given by

$$(3.18) \quad U(x, t) = \begin{cases} 0 & \text{if } x < \frac{1}{2}t \\ 1 & \text{if } x > \frac{1}{2}t, \end{cases}$$

see Figure 3.7(a). Note that this solution has one shock curve, drawn in red in the figure. However, this solution is not the only possible weak solution. By adding

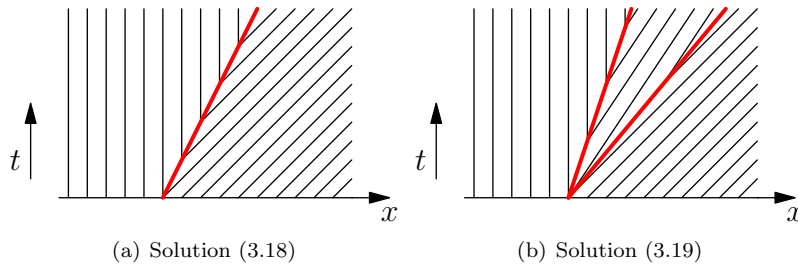


FIGURE 3.7. Characteristics for different weak solutions the Riemann problem (3.18).

an intermediate state with value, say,  $U_m = \frac{2}{3}$  and using the Rankine-Hugoniot condition, we get the weak solution

$$(3.19) \quad U(x, t) = \begin{cases} 0, & \text{if } x < \frac{1}{3}t \\ \frac{2}{3} & \text{if } \frac{1}{3}t < x < \frac{5}{6}t \\ 1 & \text{if } x > \frac{5}{6}t. \end{cases}$$

The characteristics are shown in Figure 3.7(b). In a similar manner one may construct arbitrarily many weak solutions by using the Rankine-Hugoniot condition (3.16) with different intermediate states.

This problem of non-uniqueness is implicit in the definition of weak solutions. These solutions are not necessarily unique, and therefore some extra conditions need to be imposed. For finding these extra criteria, we observe that characteristics for both (3.18) and (3.19) *flow out* from the shock (see Figure 3.7). This is in contrast to the solution (3.17) where the characteristics flow *into* the shock (see Figure 3.5). Characteristics represent the flow of information. For an evolution equation the information should always flow from the initial data. This is clearly the case for the weak solution (3.17). However in the case of weak solutions (3.18) and (3.19), information seems to be created at the shock.

This heuristic requirement, that information is taken from the initial data and is not created at a shock, can be expressed in terms of conditions on the characteristics across a shock. Let  $U^-(t), U^+(t)$  be the states on either side of a shock with speed  $s(t)$ . The requirement that characteristics for Burgers' equation flow into the shock and information is taken from the initial line can be enforced by the condition

$$(3.20) \quad U^-(t) > s(t) > U^+(t).$$

It is simple to generalize (3.20) to the general scalar conservation law (3.4) for convex  $f$ :

$$(3.21) \quad f'(U^-(t)) > s(t) > f'(U^+(t)).$$

This is the *Lax entropy condition*.

Consider the conservation law (3.4) with a convex flux function and Riemann data (3.13). It is easily shown that

$$(3.22) \quad U(x, t) = \begin{cases} U_l & \text{if } x < st \\ U_r & \text{if } x > st, \end{cases}$$

where the shock speed  $s$  is defined by the Rankine-Hugoniot condition, is a weak solution of (3.4). Now, there are two cases: either  $U_l > U_r$ , or  $U_l < U_r$ . It turns out that the Lax entropy condition excludes (3.22) as a solution in the latter case, but not in the former:

**Exercise 3.4.** Assume that  $f$  is strictly convex and that  $U_l > U_r$ . Show that (3.22) is a weak solution that satisfies the entropy condition (3.21). Similarly, if  $U_l < U_r$ , show that (3.22) is a weak solution, but does not satisfy Lax' entropy condition.

It turns out that in the latter case, where  $U_l < U_r$ , a continuous (but not necessarily differentiable) solution exists.

**3.2.3. Rarefaction waves.** For the remainder of this section, assume that the flux function  $f$  is strictly convex. In order to construct a continuous solution to (3.4), we note that replacing  $x, t$  by  $\lambda x, \lambda t$  keeps the equation invariant, in the sense that a solution of one is a solution of the other. Hence, it is natural to assume *self-similarity* – that solutions only depend on the ratio  $x/t$ :

$$(3.23) \quad U(x, t) = V(x/t).$$

Define the similarity variable  $\xi = x/t$ . We substitute the ansatz (3.23) into (3.4) and use the chain rule repeatedly to obtain

$$\begin{aligned} V_t + f(V)_x &= V(\xi)_t + f'(V(\xi))V(\xi)_x \\ &= V_\xi \xi_t + f'(V(\xi))V_\xi \xi_x \\ &= -\frac{x}{t^2}V_\xi + f'(V(\xi))\frac{1}{t}V_\xi = 0 \\ \Rightarrow \quad \left(f'(V(\xi)) - \frac{x}{t}\right)V_\xi &= 0. \end{aligned}$$

In the nontrivial case of  $V_\xi \neq 0$ , the above identity and the fact that  $f'$  is strictly increasing (recall that  $f$  is assumed to be strictly convex) leads to the expression

$$(3.24) \quad V(x/t) = (f')^{-1}(x/t).$$

A self-similar solution of this form is called a *rarefaction wave*.

The rarefaction wave can be employed to construct weak solutions for conservation laws. Consider the Riemann problem (3.4), (3.13). If  $U_l < U_r$ , then the weak solution is given by

$$(3.25) \quad U(x, t) = \begin{cases} U_l & \text{if } x \leq f'(U_l)t \\ (f')^{-1}(x/t) & \text{if } f'(U_l)t < x \leq f'(U_r)t \\ U_r & \text{if } x > f'(U_r)t. \end{cases}$$

Clearly (3.25) is a weak solution that satisfies Lax' entropy condition (3.21). For the particular case of Burgers' equation with Riemann data  $U_l = 0$  and  $U_r = 1$ , the solution (3.25) is shown in Figure 3.8. Note how the characteristics are parallel to the rarefaction wave and contrast this to Figure 3.7.

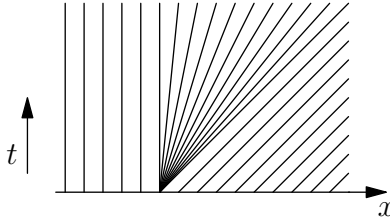


FIGURE 3.8. The rarefaction solution (3.25)

We now have a recipe to construct weak solutions for the Riemann problem (3.13) for a conservation law (3.4) with a strictly convex  $f$ . The solution depends on whether  $U_l < U_r$  or  $U_l > U_r$ . If  $U_l > U_r$ , then the entropy satisfying weak solution (3.22) consists of a shock between the two states. If  $U_l < U_r$ , then the weak solution (3.25) consists of the two states, separated by a rarefaction wave.

In both cases, the wave speed is bounded in absolute value by the maximum of  $|f'(U_l)|$  and  $|f'(U_r)|$ .

It remains to solve the Riemann problem when the flux is not strictly convex.

**3.2.4. Solutions to the Riemann problem in the general case.** We need to generalize the entropy condition (3.20) when the flux is no longer strictly convex. This is done by the following condition.

**Definition 3.5** (Oleinik entropy condition). *Let  $U$  be a weak solution of (3.4), and let  $U_l, U_r$  be the trace values at a shock with speed  $s$ . (All of these quantities depend on  $t$ , but we drop the time dependence for notational convenience.) The solution satisfies the Oleinik entropy condition if*

$$s \leq \frac{f(k) - f(U_l)}{k - U_l}$$

for all  $k$  between  $U_l$  and  $U_r$ .

**Exercise 3.6.** Use the Rankine-Hugoniot condition to show that  $U$  satisfies the Oleinik entropy condition if and only if

$$\frac{f(k) - f(U_r)}{k - U_r} \leq s$$

for all  $k$  between  $U_l$  and  $U_r$ .

A geometric interpretation of the above conditions is shown in Figure 3.9. It amounts to a convexity condition that the straight line joining  $(U_l, f(U_l))$  and  $(U_r, f(U_r))$  must lie below the graph of the function  $f$  between these points when  $U_l < U_r$ , and should lie above the graph if  $U_l > U_r$ . Clearly, the condition (3.20) for convex fluxes is a special case for this condition.

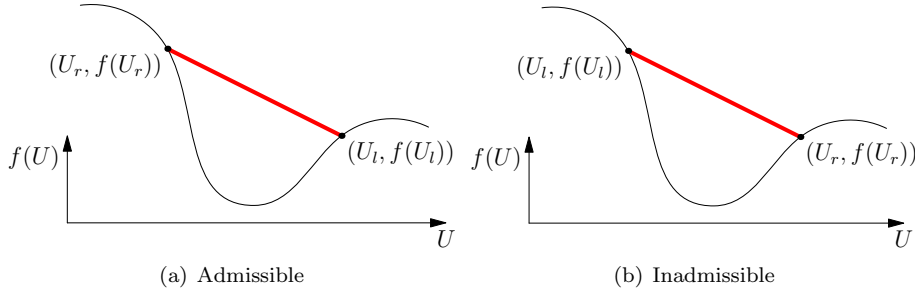


FIGURE 3.9. Admissible and inadmissible shocks under the Oleinik entropy condition.

We have the following recipe for constructing a weak solution for the Riemann problem (3.13) for the conservation law (3.4), that satisfies the Oleinik entropy condition. Assume that the flux  $f$  is not convex but has finitely many inflection points (see Figure 3.10). Without loss of generality, we may assume that  $U_l < U_r$ . In order to satisfy the Oleinik entropy condition, we have to consider the *lower convex envelope*  $f_c$  of  $f$  between  $U_l$  and  $U_r$ . The lower convex envelope of a function  $f$  is the largest convex function (largest in the point-wise sense) that is everywhere



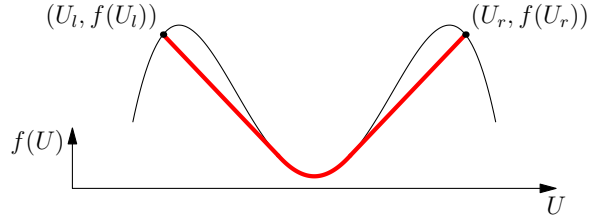


FIGURE 3.10. The solution of the Riemann problem with a non-convex flux. The lower convex envelope is the thick red curve. Solutions are constructed as shocks, followed by rarefactions

smaller than or equal to  $f$  (see Figure 3.10). Analogously, the upper concave envelope of  $f$  is the smallest concave function that is larger than or equal to  $f$ .

The domain  $[U_l, U_r]$  is divided into two sets of regions, one in which  $f_c = f$  and another with  $f_c \neq f$ . In the second region,  $f_c$  is affine. The strategy for constructing an entropy solution is to join  $U_l$  and  $U_r$  by rarefaction waves and shocks. Shocks are used in the affine region and rarefactions in the complement. The solution of the Riemann problem is then (3.25) with  $f$  replaced by  $f_c$ . An illustration is provided in Figure 3.10.

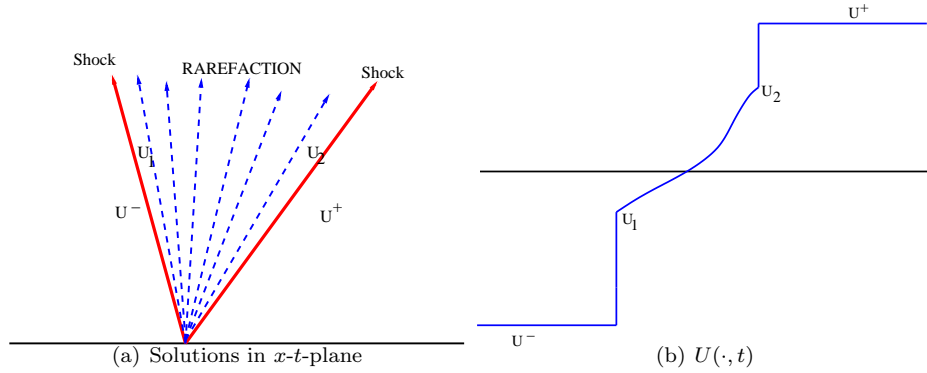


FIGURE 3.11. Entropy solutions for the Riemann problem with a non-convex flux. Left: Solutions in space-time. Right: A snapshot of the solution illustrating compound shocks.

When  $U_l > U_r$ , the upper concave envelope can be analogously used. Details of this construction can be obtained from [1]. A wave consisting of rarefaction, followed immediately by a shock (or vice versa) is termed a *compound shock* (see Figure 3.11).

### 3.3. Entropy solutions

The first step in analyzing solutions to the Cauchy problem for the conservation law (3.4) is to generalize the point-wise Oleinik entropy condition at a shock. This is the main focus of this section.

We consider the following *viscous* approximation to the scalar conservation law:

$$(3.26) \quad U_t^\varepsilon + f(U^\varepsilon)_x = \varepsilon U_{xx}^\varepsilon,$$

where  $\varepsilon > 0$  is a small parameter. Note that (3.26) is a perturbation of the scalar conservation law. The second-order term  $U_{xx}$  is termed the *viscous* term. Adding the viscous term turns the conservation law into a convection-diffusion equation. Such equations are similar to the heat equation (1.8), and as for the heat equation, the solutions to (3.26) are smooth, in fact  $C^\infty$  functions.

Let  $\eta = \eta(U)$  be any strictly convex function, and construct the function

$$q(U) = \int_0^U f'(s)\eta'(s)ds.$$

Note that  $\eta$  and  $q$  satisfy the relation

$$(3.27) \quad q' = \eta' f'.$$

Multiplying  $\eta'(U)$  to both sides of (3.26) and using the chain rule and the relation (3.27), we obtain

$$\begin{aligned} & \eta'(U^\varepsilon)U_t^\varepsilon + \eta'(U^\varepsilon)f'(U^\varepsilon)U_x^\varepsilon = \varepsilon\eta'(U^\varepsilon)U_{xx}^\varepsilon \\ \Rightarrow & \eta'(U^\varepsilon)U_t^\varepsilon + q'(U^\varepsilon)U_x^\varepsilon = \varepsilon\eta'(U^\varepsilon)U_{xx}^\varepsilon \\ \Rightarrow & \eta(U^\varepsilon)_t + q(U^\varepsilon)_x = \varepsilon\eta(U^\varepsilon)_{xx} - \varepsilon\eta''(U^\varepsilon)(U_x^\varepsilon)^2. \end{aligned}$$

Since  $\eta$  is a convex function, we have  $\eta'' \geq 0$ . Therefore, we obtain

$$(3.28) \quad \eta(U^\varepsilon)_t + q(U^\varepsilon)_x \leq \varepsilon\eta(U^\varepsilon)_{xx}.$$

It may be shown with the theory of reaction-diffusion equations that the limit  $U = \lim_{\varepsilon \rightarrow 0} U^\varepsilon$  exists and is a weak solution of the conservation law (3.4). By (3.28),  $U$  satisfies

$$(3.29) \quad \eta(U)_t + q(U)_x \leq 0.$$

The function  $\eta$  is called an *entropy* function and the corresponding function  $q$  is called an *entropy flux*. The pair  $(\eta, q)$  is called an *entropy pair*. The inequality (3.29) is referred to as the *entropy inequality*, and holds for every convex function  $\eta$ . In other words, every convex function is an entropy function for a scalar conservation law.

Since weak solutions are not necessarily differentiable, we need to interpret the entropy inequality (3.29) in a weak manner.

**Definition 3.7.** A function  $U(x, t)$  is an entropy solution of (3.4) if it satisfies the following conditions:

- (i)  $U \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ .
- (ii)  $U$  is a weak solution of (3.4), i.e., it satisfies the identity (3.15) for all test functions  $\varphi \in C^1(\mathbb{R} \times \mathbb{R}_+)$ .
- (iii) For all test functions  $\varphi \in C_c^1(\mathbb{R} \times (0, \infty))$  with  $\varphi \geq 0$ ,  $U$  satisfies

$$(3.30) \quad \int_{\mathbb{R} \times \mathbb{R}_+} \eta(U)\varphi_t + q(U)\varphi_x \geq 0,$$

for all entropy pairs  $(\eta, q)$ .

**Remark 3.8.** Any convex function  $\eta$  serves as an entropy function for a scalar conservation law. Of particular importance are the so-called Kruzhkov entropies:

$$(3.31) \quad \eta = \bar{\eta}(U, k) = |U - k|$$

for constants  $k \in \mathbb{R}$ . The corresponding entropy flux is

$$(3.32) \quad q = q(U, k) = \text{sign}(U - k)(f(U) - f(k)).$$

It is straightforward to check that any smooth convex function  $g(U)$  can be approximated by a linear combination of functions like  $\bar{\eta}(U, k)$  on bounded intervals  $U \in [a, b]$ . More precisely, given  $\delta > 0$ , there exist  $N \in \mathbb{N}$  and  $\alpha_k, c_k \in \mathbb{R}$  with  $1 \leq k \leq N$  such that

$$\left| g(U) - \sum_{k=1}^N \alpha_k |U - c_k| \right| \leq \delta \quad \text{for all } U \in [a, b].$$

Thus, (3.29) holds if and only if

$$(3.33) \quad |U - k|_t + (\text{sign}(U - k)(f(U) - f(k)))_x \leq 0$$

(in the weak sense) for all  $k \in \mathbb{R}$ .

It is straightforward to check that the entropy solution satisfies the Oleinik entropy condition (check [1] for a proof). The concept of entropy solutions is the correct notion of solutions to conservation laws. We will demonstrate the existence and uniqueness of entropy solutions in this section.

The entropy inequality (3.29) can be used to obtain estimates on solutions. To see this, integrate (3.29) over space and integrate by parts to obtain

$$(3.34) \quad \frac{d}{dt} \int_{\mathbb{R}} \eta(U) \leq 0 \quad \Rightarrow \quad \int_{\mathbb{R}} \eta(U(x, t)) dx \leq \int_{\mathbb{R}} \eta(U_0(x)) dx.$$

Since the function  $\eta$  may be any convex function, we can choose  $\eta(U) = \frac{U^2}{2}$  and obtain a bound on the entropy solution in  $L^2$ . This estimate is a *nonlinear* analogue of the energy estimate (2.6) for the linear transport equation. Choosing  $\eta$  as

$$\eta(U) = \frac{|U|^p}{p}$$

will lead to estimates in  $L^p$  spaces for all  $p \geq 1$ . Hence, the entropy inequality serves to yield stability estimates.

A key estimate is the  $L^\infty$  estimate that bounds the maximum and minimum of the solutions of (3.26). The  $L^\infty$  estimate is a consequence of the standard maximum principle for parabolic equations. A formal proof proceeds as follows: Assume that the maximum of the solution  $U$  is attained at a point  $x_0$  at time  $t_0 > 0$ . Clearly  $U_t(x_0, t_0) = U_x(x_0, t_0) = 0$ . Furthermore, as  $(x_0, t_0)$  is a maximum, we can assume that  $U_{xx}(x_0, t_0) < 0$ . Substituting all these relations in (3.26) leads to a contradiction. Hence, the maximum can only be attained at the initial data. A similar estimate holds at minima.

Another crucial stability estimate for (3.26) concerns the control of the derivative. To motivate this estimate, we let  $\eta(U) = \bar{\eta}(U, 0) = |U|$ , the Kruzhkov entropy centered at 0. Formally we have

$$\eta'(U) = \text{sign}(U), \quad \eta''(U) = \delta(U),$$

where  $\delta$  is the Dirac mass or delta function centered at 0. Multiplying (3.26) with  $-\eta'(U_x^\varepsilon)_x$  and using the chain rule and integrating by parts, we obtain

$$\begin{aligned} & - \int_{\mathbb{R}} \eta'(U_x^\varepsilon)_x U_t^\varepsilon dx - \int_{\mathbb{R}} \eta'(U_x^\varepsilon)_x f'(U^\varepsilon) U_x^\varepsilon dx = -\varepsilon \int_{\mathbb{R}} \eta'(U_x^\varepsilon)_x U_{xx}^\varepsilon dx \\ & \int_{\mathbb{R}} \eta'(U_x^\varepsilon) U_{xt}^\varepsilon dx - \int_{\mathbb{R}} \eta'(U_x^\varepsilon)_x f'(U^\varepsilon) U_x^\varepsilon dx = -\varepsilon \int_{\mathbb{R}} \eta'(U_x^\varepsilon)_x U_{xx}^\varepsilon dx \quad (\text{integration by parts}) \\ & \frac{d}{dt} \int_{\mathbb{R}} |U_x^\varepsilon| dx - \underbrace{\int_{\mathbb{R}} \eta''(U_x^\varepsilon) f'(U^\varepsilon) U_x^\varepsilon U_{xx}^\varepsilon dx}_{=0} = -\varepsilon \int_{\mathbb{R}} \eta''(U_x^\varepsilon) (U_{xx}^\varepsilon)^2 dx \quad (\text{chain rule}), \end{aligned}$$

which in the limit  $\varepsilon \rightarrow 0$  gives us

$$\frac{d}{dt} \int_{\mathbb{R}} |U_x| dx \leq 0.$$

These computations are formal but can be made completely rigorous (see [1]).

Of great importance in the context of conservation laws is the concept of total variation. Let  $g$  be a function defined on an interval  $[a, b]$ . The *total variation* of  $g$  is defined as

$$(3.35) \quad \|g\|_{TV} = \sup_{\mathcal{P}} \sum_{j=1}^{N_{\mathcal{P}}-1} |g(x_{j+1}) - g(x_j)|,$$

where the supremum is taken over all partitions  $\mathcal{P} = \{a = x_1 < x_2 < \dots < x_{N_{\mathcal{P}}} = b\}$  of the interval  $[a, b]$ . It is straightforward to check that if  $g$  is differentiable, then

$$\|g\|_{TV} = \int_a^b |g_x| dx.$$

The total variation is only a semi-norm; indeed, the total variation of any constant function is zero. We turn it into a norm by defining

$$(3.36) \quad \|g\|_{BV} = \|g\|_{L^1} + \|g\|_{TV}.$$

We define the space of functions with bounded variation (BV) as

$$(3.37) \quad BV(\mathbb{R}) = \{g \in L^1(\mathbb{R}) : \|g\|_{BV} < \infty\}.$$

The above estimate on  $\int |U_x| dx$  is a BV estimate. It states that the solutions to a conservation law (3.4) are *Total Variation Diminishing* (TVD), that is,

$$\|U(\cdot, t)\|_{BV} \leq \|U_0\|_{BV} \quad \text{for all } t > 0.$$

We are now in a position to state the main well-posedness results for scalar conservation laws:

**Theorem 3.9.** *If  $U_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$ , then (3.4) has a entropy solution  $U$  which satisfies the estimates*

$$(3.38) \quad \|U(\cdot, t)\|_{L^\infty} \leq \|U_0\|_{L^\infty}$$

and

$$(3.39) \quad \|U(\cdot, t)\|_{TV} \leq \|U_0\|_{TV}.$$

Furthermore, if  $U$  and  $V$  are entropy solutions of (3.4) with initial  $U_0$  and  $V_0$ , respectively, then

$$(3.40) \quad \int_{\mathbb{R}} |U(x, t) - V(x, t)| dx \leq \int_{\mathbb{R}} |U_0(x) - V_0(x)| dx \quad \text{for all } t > 0.$$

Hence, entropy solutions are unique.

The proof of this theorem is outside the scope of these notes (consult standard text books like [1]). We only sketch the main ideas of the proof. The existence is based on the viscous approximation (3.26). The entropy conditions yield estimates on the solution in  $L^p$  spaces as discussed before. Formal arguments for establishing the  $L^\infty$  and BV bounds are discussed above. Uniqueness of entropy solutions relying on the  $L^1$  contraction estimate (3.40) is based on the ingenious *doubling of variables* idea of Kruzhkov [5] and uses the Kruzhkov entropies (3.31) extensively.

Summarizing the theoretical discussion of this section, we have the following results:

- Solutions of the conservation law (3.4) may develop discontinuities or shock waves, even for smooth initial data. Consequently, weak solutions are sought. Shock speeds are computed with the Rankine-Hugoniot condition (3.16).
- Weak solutions are not necessarily unique. Entropy conditions like the Oleinik entropy conditions have to be imposed. Self-similar continuous solutions or rarefaction waves have to be considered.
- Explicit solutions for the Riemann problem (even for non-convex fluxes) can be constructed in terms of shocks, rarefaction waves and compound shocks.
- Entropy solutions exist and are unique. Furthermore, the entropy solutions satisfy an  $L^\infty$  estimate,  $L^p$  estimates and are TVD – that is, the BV-norm decreases in time.

Despite the considerable theoretical results, we cannot obtain explicit solution formulas for more complicated initial data. Hence, we have to design efficient numerical methods for computing these solutions.



## CHAPTER 4

### Finite volume schemes for scalar conservation laws

In this chapter we will design efficient schemes for the scalar conservation law

$$(4.1) \quad U_t + f(U)_x = 0.$$

The discussion on the linear transport equation

$$(4.2) \quad U_t + aU_x = 0$$

shows that central differences cannot be used to approximate the conservation law, even in the simplest case of linear transport. For linear transport equations, the crucial step in designing an efficient scheme was to *upwind* it by taking derivatives in the direction of information propagation. For a linear equation with constant coefficients like (4.2), the direction of information propagation is given by the constant velocity field. For a nonlinear conservation law like (4.1), the wave speeds depend on the solution itself and can not be determined a priori. Thus, it is not clear how differences can be upwinded.

Another issue is the very nature of finite difference approximations like (2.16). The key idea underlying finite difference schemes is to replace the derivatives in equations like (4.1) with a finite difference. This procedure requires the solutions to be smooth and the equation to be satisfied point-wise. However, the solutions to the scalar conservation law (4.1) are not necessarily smooth and so the Taylor expansion – essential for replacing derivatives with finite differences – is no longer valid. Hence, the finite difference framework is not suited for approximating conservation laws. Instead, we need to develop a new paradigm for designing numerical schemes for scalar conservation laws.

#### 4.1. Finite volume scheme

The first step in any numerical approximation is to discretize the computational domain in both space and time.

**4.1.1. The grid.** For simplicity, we consider a uniform discretization of the domain  $[x_L, x_R]$ . The discrete points are denoted as  $x_j = x_L + (j + 1/2)\Delta x$  for  $j = 0, \dots, N$ , where  $\Delta x = \frac{x_R - x_L}{N+1}$ . We also define the midpoint values

$$x_{j-1/2} = x_j - \Delta x/2 = x_L + j\Delta x$$

for  $j = 0, \dots, N + 1$ . These values define computational cells or *control volumes*

$$\mathcal{C}_j = [x_{j-1/2}, x_{j+1/2}).$$

As we will see soon, the finite volume method uses the control volumes  $\mathcal{C}_j$  instead of the mesh points  $x_j$ . We use a uniform discretization in time with time step  $\Delta t$ . The time levels are denoted by  $t^n = n\Delta t$ . See Figure 4.1 for an illustration of the grid.

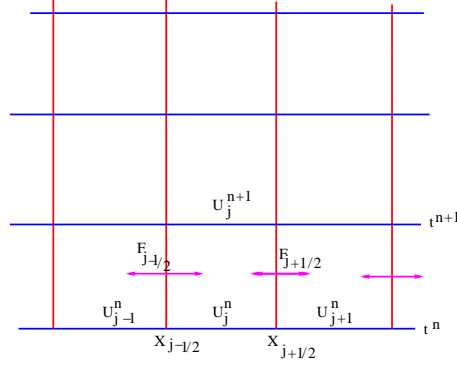


FIGURE 4.1. A typical finite volume grid displaying cell averages and fluxes.

**4.1.2. Cell averages.** A finite difference method is based on approximating the point values of the solution of a PDE. This approach is not suitable for conservation laws as the solutions are not continuous and point values may not make sense. Instead, we change the perspective and use the *cell averages*

$$(4.3) \quad U_j^n \approx \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} U(x, t^n) dx$$

at each time level  $t^n$  as the main object of interest for our approximation.

The cell average (4.3) is well defined for any integrable function, hence also for the solutions of the conservation law (4.1). The aim of the finite volume method is to update the cell average of the unknown at every time step.

**4.1.3. Integral form of the conservation law.** Assume that the cell averages  $U_j^n$  at some time level  $t^n$  are known. How do we obtain the cell averages  $U_j^{n+1}$  at the next time level  $t^{n+1}$ ? A finite volume method computes the cell average at the next time level by integrating the conservation law (4.1) over the domain  $[x_{j-1/2}, x_{j+1/2}) \times [t^n, t^{n+1})$ . This gives

$$\int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} U_t dx dt + \int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} f(U)_x dx dt = 0.$$

Using the fundamental theorem of calculus gives

$$(4.4) \quad \begin{aligned} & \int_{x_{j-1/2}}^{x_{j+1/2}} U(x, t^{n+1}) dx - \int_{x_{j-1/2}}^{x_{j+1/2}} U(x, t^n) dx \\ &= - \int_{t^n}^{t^{n+1}} f(U(x_{j+1/2}, t)) dt + \int_{t^n}^{t^{n+1}} f(U(x_{j-1/2}, t)) dt. \end{aligned}$$

Defining

$$(4.5) \quad \bar{F}_{j+1/2}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(U(x_{j+1/2}, t)) dt$$

and dividing both sides of (4.4) by  $\Delta x$ , we obtain

$$(4.6) \quad U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x} \left( \bar{F}_{j+1/2}^n - \bar{F}_{j-1/2}^n \right).$$



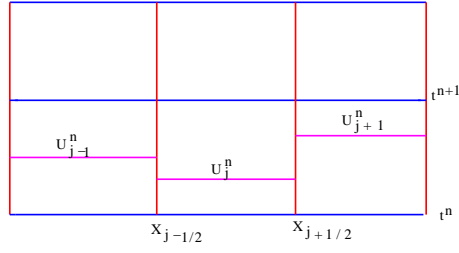


FIGURE 4.2. Cell averages define Riemann problems at every interface.

Equation (4.6) is a statement of conservation: The rate of change of the cell average is given by the difference in fluxes across the boundary of the cell. See Figure 4.1 for an illustration. Note that the relation (4.6) is not explicit as  $\bar{F}$  need a priori knowledge of the exact solution. The main ingredient in a finite volume scheme is a clever procedure to approximate these fluxes.

**4.1.4. Godunov method.** Godunov [2] came up with an ingenious idea for approximating the numerical fluxes in (4.6). We wish to approximate

$$\bar{F}_{j+1/2}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(U(x_{j+1/2}, t)) dt$$

at each interface  $x_{j+1/2}$ . As the cell averages  $U_j^n$  are constant in each cell  $\mathcal{C}_j$  at each time level, Godunov observed that they define at each cell interface  $x_{i+1/2}$  a *Riemann problem*

$$(4.7) \quad \begin{cases} U_t + f(U)_x = 0 \\ U(x, t^n) = \begin{cases} U_j^n & \text{if } x < x_{j+1/2} \\ U_{j+1}^n & \text{if } x > x_{j+1/2}. \end{cases} \end{cases}$$

Thus at every time level, the cell averages define a superposition of Riemann problems of the form (4.7) at each interface (see Figure 4.2). In the previous chapter, we have solved Riemann problems like (4.7) explicitly. The solution consists of shock waves, rarefactions and compound waves. Hence, the Riemann problem at every time level can be solved explicitly in terms of waves, emanating from each interface (Figure 4.3). Furthermore, the solution of each Riemann problem in (4.7) is self-similar, that is, the solution  $\bar{U}_j(x, t)$  of (4.7) can be written as a function  $\bar{U}(\xi)$  of a single variable  $\xi = \frac{x - x_{j+1/2}}{t}$ ,

$$(4.8) \quad \bar{U}_j(x, t) = \bar{U}_j\left(\frac{x - x_{j+1/2}}{t}\right).$$

Waves from neighboring Riemann problems can intersect after some time (Figure 4.3(a)). However, each wave has a finite speed of propagation and the maximum wave speed of any Riemann problem is bounded by

$$\max_j |f'(U_j^n)|$$

(see Chapter 3). Hence, imposing the *CFL condition*

$$(4.9) \quad \max_j |f'(U_j^n)| \frac{\Delta t}{\Delta x} \leq \frac{1}{2}$$

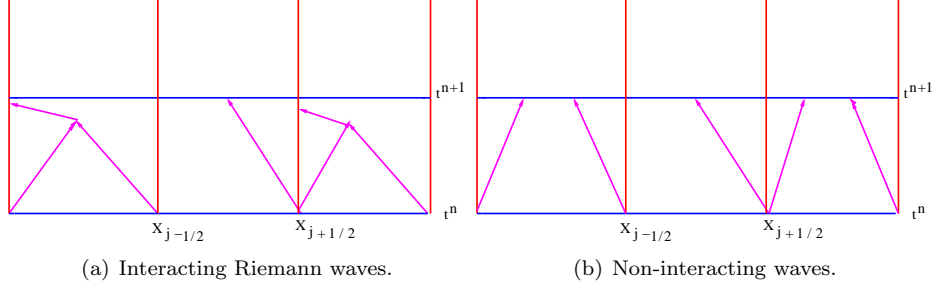


FIGURE 4.3. Left: Waves of Riemann problems from neighboring interface can interact after some time. Right: The waves can be prevented from interacting before time  $\Delta t$  by the CFL condition (4.9)

ensures that waves from neighboring problems do not interact before reaching the next time level (see Figure 4.3(b)). Assume now that this condition is satisfied. By (4.8), the solution is constant when  $\xi$  is constant, so in particular, at the cell interface  $\xi = 0$ , the flux across the interface is given by the constant value

$$f(U(x_{j+1/2}, t)) = f(\bar{U}_j(0)).$$

Along the curve  $\xi = 0$  (corresponding to  $x = x_{j+1/2} \forall t > 0$ ), the function  $\bar{U}_j(\xi)$  is either continuous or discontinuous. If  $\bar{U}$  is continuous at  $\xi = 0$ , we obviously have

$$(4.10) \quad f(\bar{U}_j(0+)) = f(\bar{U}_j(0-)).$$

On the other hand, if  $\bar{U}$  is discontinuous at  $\xi = 0$ , then we have a stationary shock located at the cell interface (discontinuities propagate along the straight lines  $\xi \equiv \text{constant}$ ). Since the discontinuity must satisfy the Rankine-Hugoniot condition (3.16), we have

$$f(\bar{U}_j(0+)) - f(\bar{U}_j(0-)) = 0 \cdot (\bar{U}_j(0+) - \bar{U}_j(0-)) = 0,$$

and so (4.10) holds also in this case. Hence, the term  $f(\bar{U}_j(0))$  is well-defined, and we may define the edge-centered flux value

$$(4.11) \quad F_{j+1/2}^n := f(\bar{U}_j(0+)) = f(\bar{U}_j(0-)).$$

In conclusion, the approximate flux in (4.5) is constant in time and can be explicitly computed as

$$(4.12) \quad \bar{F}_{j+1/2}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(U(x_{j+1/2}, t)) dt = F_{j+1/2}^n,$$

with  $F_{j+1/2}^n$  being the Riemann flux (4.11). Substituting (4.12) in (4.6) leads to the finite volume scheme

$$(4.13) \quad U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x} (F_{j+1/2}^n - F_{j-1/2}^n).$$

The form (4.13) is the standard form of a finite volume scheme for conservation laws. The numerical flux  $F$  is given in terms of the Riemann solution and can be explicitly computed for scalar conservation laws.

**4.1.5. Godunov flux.** It turns out that we can compute explicit formulas for the numerical flux in (4.13). To this end, we need to obtain the value of the flux of the Riemann problem (4.7) at the interface  $x_{j+1/2}$ . A lengthy computation based on a case by case analysis leads to the formula

$$(4.14) \quad F_{j+1/2}^n = F(U_j^n, U_{j+1}^n) = \begin{cases} \min_{U_j^n \leq \theta \leq U_{j+1}^n} f(\theta) & \text{if } U_j^n \leq U_{j+1}^n \\ \max_{U_{j+1}^n \leq \theta \leq U_j^n} f(\theta) & \text{if } U_{j+1}^n \leq U_j^n. \end{cases}$$

This formula is valid also for non-convex flux functions. The *Godunov scheme* is (4.13) with the Godunov flux (4.14).

**Exercise 4.1.** *Computing the flux (4.14) can be complicated, since an optimization problem has to be solved. Show that in the special case where the flux function  $f$  has a single minimum at the point  $\omega$  and no local maxima, the formula (4.14) can be simplified to*

$$(4.15) \quad F_{j+1/2}^n = F(U_j^n, U_{j+1}^n) = \max\left(f(\max(U_j^n, \omega)), f(\min(U_{j+1}^n, \omega))\right).$$

*Note that strictly convex functions have this property. The formulas for the case of a flux with a single maximum and no minima are obtained analogously.*

**Exercise 4.2.** *Show that for the linear transport equation (4.2), the Godunov scheme (4.13), (4.14) is identical to the standard upwind scheme (2.16). Thus, the Godunov scheme can be viewed as a generalization of the upwind scheme to nonlinear scalar conservation laws.*

**4.1.6. Numerical experiments.** Consider Burgers' equation (3.3) with Riemann data

$$(4.16) \quad U(x, 0) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x > 0. \end{cases}$$

In this case, the exact solution is given by a single shock connecting 1 and 0, traveling at speed of  $1/2$  (see Chapter 3). Numerical solutions with the Godunov scheme (4.13), (4.14) with 50 mesh points are plotted in Figure 4.4 (a). The results show that the solution is approximated very well, with the shock being resolved sharply. The numerical solutions do not oscillate or show any anomalies or instabilities.

Next, we test Burgers' equation with initial data

$$(4.17) \quad U(x, 0) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0. \end{cases}$$

The exact solution in this case is given by a rarefaction wave (3.25). The approximate solutions using the Godunov scheme are plotted in Figure 4.4 (b). Again the results demonstrate that the Godunov scheme is stable and robust.

As a final test case, we consider Burgers' equation with initial data

$$(4.18) \quad U(x, 0) = \sin(4\pi x) \quad \text{for } -1 \leq x \leq 1.$$

The initial data is a sine wave and it is much more difficult to write down an explicit formula for the solution. Instead, we compute this configuration with the Godunov scheme using periodic boundary conditions. The results are shown in Figure 4.5. A reference solution computed on a very fine mesh (5000 points) is also shown for the sake of comparison. The behavior of the solution is quite complicated. The initial

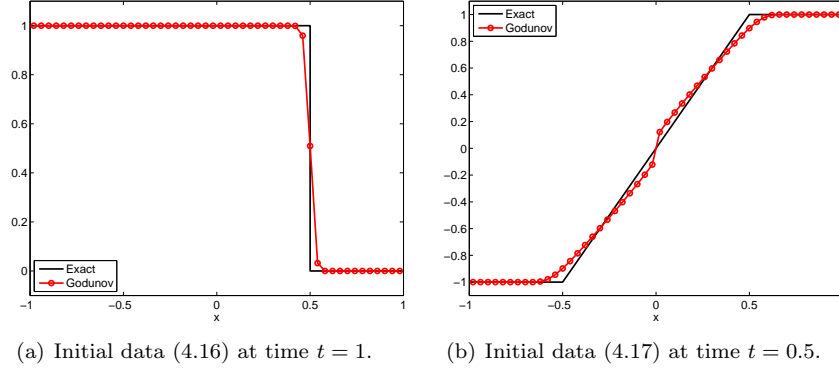


FIGURE 4.4. Approximate solution for Burgers equation with the Godunov scheme with 50 mesh points. [burgers\_disc.m]

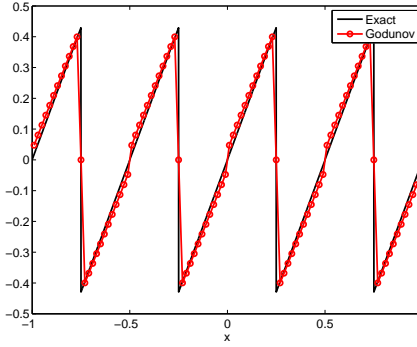


FIGURE 4.5. Approximate solution for Burgers' equation with the Godunov scheme at time  $t = 0.5$  with 50 mesh points with initial data (4.18). [burgers\_godunov\_sine.m]

sine wave compresses in some parts and expands in some other parts, leading to a combination of shocks and rarefactions. The solution finally decays into a so-called *N-wave*. The Godunov scheme provides a good approximation to this complicated solution.

**4.1.7. Beyond the Godunov Scheme.** The Godunov scheme (4.13), (4.14) has many desirable properties as demonstrated by numerical experiments. However, it does present a few problems:

- It relies on the availability of an *explicit* formula for the solutions of the Riemann problem. In the case of scalar conservation law (4.1), we are lucky to have such formulas at hand. However, more complicated systems of conservation laws may not yield such formulas.
- The only information needed in the numerical flux (4.11) is the value of the flux at the interface. Solving the entire Riemann problem for the sake of this value seems unnecessary.

- At the level of implementation, the formula (4.15) provides a simple characterization of the Godunov flux for a large class of flux functions. However, more complicated flux functions with a large number of extremal points need the solution of an optimization problem. Such a problem might be very computationally costly.

These factors encourage the search for alternative numerical fluxes in (4.13).

## 4.2. Approximate Riemann Solvers

Since we are interested in approximating the solutions of the conservation law (4.1), it seems reasonable to replace the exact solutions of the Riemann problem (4.7) (used in the Godunov scheme) with approximate solutions. These approximate solutions can then be used to define the numerical flux  $F$  as in (4.12). Such schemes which replace the exact solutions of the Riemann problem (4.7) with approximations called *approximate Riemann solvers*. We present some of them below.

**4.2.1. Linearized (Roe) solvers.** Our aim is to approximate the solutions of the Riemann problem (4.7). A common method for solving nonlinear equations is to *linearize* them. Linearization entails replacing the nonlinear flux function in (4.1) with a locally linearized version,

$$(4.19) \quad f(U)_x = f'(U)U_x \approx \hat{A}_{j+1/2}U_x,$$

where  $\hat{A} \approx f'$  is a constant state around which the nonlinear flux function is linearized. There are many possible candidates for the linearizing state, one simple choice being

$$\hat{A}_{j+1/2} = f' \left( \frac{U_j^n + U_{j+1}^n}{2} \right),$$

the flux of the arithmetic average of the two constant states. We will use a more sophisticated *Roe average*:

$$(4.20) \quad \hat{A}_{j+1/2} = \begin{cases} \frac{f(U_{j+1}^n) - f(U_j^n)}{U_{j+1}^n - U_j^n} & \text{if } U_{j+1}^n \neq U_j^n \\ f'(U_j^n) & \text{if } U_{j+1}^n = U_j^n. \end{cases}$$

Note that the Roe average also represents a linear approximation of  $f'$ . The numerical flux  $F$  is obtained by replacing the Riemann problem (4.7) with a linearized Riemann problem,

$$(4.21) \quad \begin{cases} U_t + \hat{A}_{j+1/2}U_x = 0 \\ U(x, t^n) = \begin{cases} U_j^n & \text{if } x < x_{j+1/2} \\ U_{j+1}^n & \text{if } x > x_{j+1/2}. \end{cases} \end{cases}$$

This Riemann problem is very simple to solve as it involves a linear transport equation with a constant velocity field. Solving it explicitly we obtain the formula

$$(4.22) \quad F_{j+1/2}^n = F^{\text{Roe}}(U_j^n, U_{j+1}^n) = \begin{cases} f(U_j^n) & \text{if } \hat{A}_{j+1/2} \geq 0 \\ f(U_{j+1}^n) & \text{if } \hat{A}_{j+1/2} < 0. \end{cases}$$

The finite volume scheme (4.13) with the Roe flux (4.22) is termed the *Roe* or *Murman-Roe* scheme. It is simpler to implement when compared to the Godunov scheme as no optimization problem needs to be solved.

Numerical results with the Roe scheme for Burgers' equation with Riemann data (4.16) are shown in Figure 4.6(a). They show that the Roe scheme approximates the shock as accurately as the Godunov scheme.

Figure 4.6 (b) shows numerical results for the Riemann data (4.17). In this case, the Roe scheme fails completely and approximates the wrong stationary shock solution. The same stationary solution persists even when the mesh is refined. Thus, the Roe scheme leads to numerical artifacts for some problems. This failure will be analyzed in detail in the sequel.

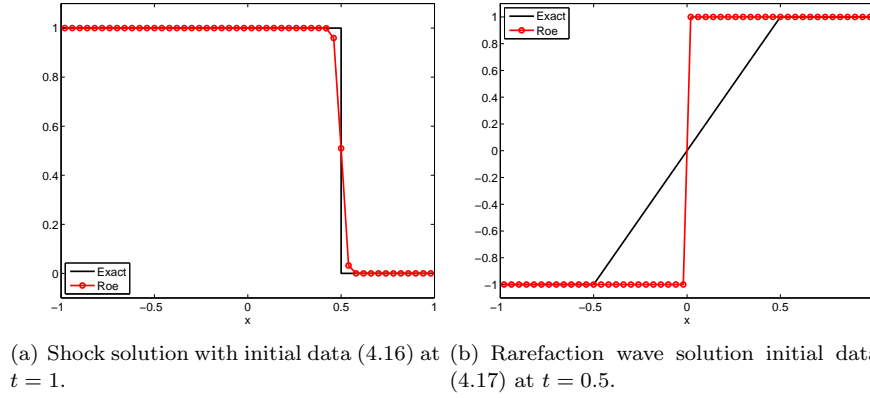


FIGURE 4.6. Approximate solutions for Burgers equation with the Roe scheme with 50 mesh points. [burgers\_disc.m]

**4.2.2. Central schemes.** The Roe scheme fails at resolving rarefactions. Due to linearization, the solution of the approximate Riemann problem (4.21) only consists of a single wave that travels to the right or to the left, depending on the sign of the Roe average  $\hat{A}$ . When the exact solutions of Riemann problems for the conservation law consists of shocks, then the solution is a single, either left- or right-going, wave. However, the situation with rarefactions is very different. The rarefaction wave that solves (4.17) can travel in both directions (see Figure 4.7). As we have seen, the Roe scheme may be unable to capture such behavior.

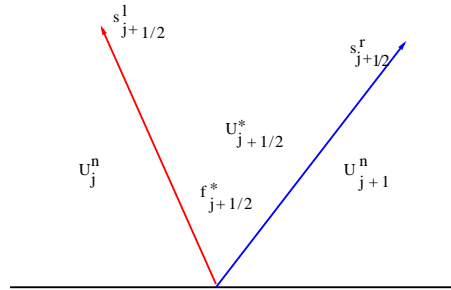


FIGURE 4.7. An approximate Riemann solver with bi-directional waves.

Instead of linearizing the conservation law, we approximate the solutions of the Riemann problem by replacing the exact solution with *two waves*, one traveling to the left of the interface with speed  $s_{j+1/2}^l$  and another to the right with speed  $s_{j+1/2}^r$  (see Figure 4.7). The speeds will be specified later on.

We approximate the solution of (4.7) with

$$(4.23) \quad U(x, t) = \begin{cases} U_j^n & \text{if } x < s_{j+1/2}^l t \\ U_{j+1/2}^* & \text{if } s_{j+1/2}^l t < x < s_{j+1/2}^r t \\ U_{j+1}^n & \text{if } x > s_{j+1/2}^r t. \end{cases}$$

Thus, the exact solution is replaced by two waves separated by a middle state. The middle state can be determined by local conservation using the Rankine-Hugoniot conditions (3.16):

$$(4.24) \quad \begin{aligned} f(U_{j+1}^n) - f_{j+1/2}^* &= s_{j+1/2}^r (U_{j+1}^n - U_{j+1/2}^*), \\ f(U_j^n) - f_{j+1/2}^* &= s_{j+1/2}^l (U_j^n - U_{j+1/2}^*), \end{aligned}$$

where  $f_{j+1/2}^*$  is the intermediate flux (see Figure 4.7). Observe that we require  $f^*$  to be an independent variable. Thus, (4.24) represents a system of two linear equations for two unknowns that can be solved exactly to obtain

$$(4.25) \quad f_{j+1/2}^* = \frac{s_{j+1/2}^r f(U_j^n) - s_{j+1/2}^l f(U_{j+1}^n) + s_{j+1/2}^r s_{j+1/2}^l (U_{j+1}^n - U_j^n)}{s_{j+1/2}^r - s_{j+1/2}^l}.$$

In particular, if we choose the speeds to be equal but of opposite sign, so  $s^r = -s^l = s$ , then (4.25) reduces to

$$(4.26) \quad f_{j+1/2}^* = \frac{f(U_j^n) + f(U_{j+1}^n)}{2} - \frac{s_{j+1/2}}{2} (U_{j+1}^n - U_j^n).$$

In either case, the numerical flux is given by

$$(4.27) \quad F_{j+1/2}^n = F(U_j^n, U_{j+1}^n) = f_{j+1/2}^*.$$

We have yet to specify the local wave speeds  $s^l$ ,  $s^r$ . Different choices of the speeds lead to different schemes; presently we describe three of the most important ones.

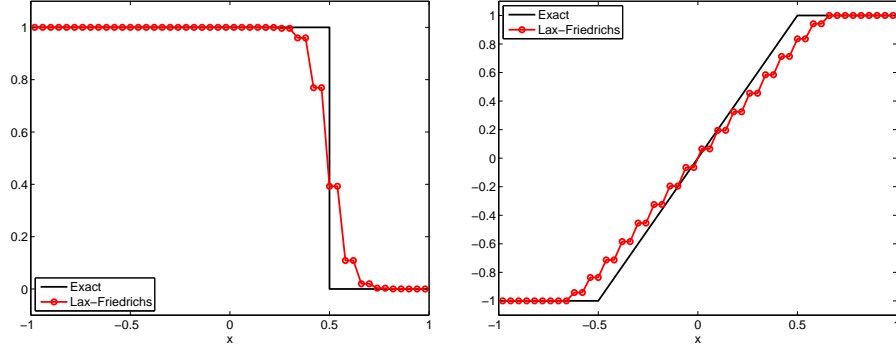
**4.2.3. Lax-Friedrichs scheme.** To ensure that waves from neighboring Riemann problems (4.23) do not interact, the maximum allowed wave speeds are

$$(4.28) \quad s_{j+1/2}^l = -\frac{\Delta x}{\Delta t}, \quad s_{j+1/2}^r = \frac{\Delta x}{\Delta t}.$$

These wave speeds substituted in (4.26) lead to the Lax-Friedrichs flux

$$(4.29) \quad F_{j+1/2}^n = F^{\text{LxF}}(U_j^n, U_{j+1}^n) = \frac{f(U_j^n) + f(U_{j+1}^n)}{2} - \frac{\Delta x}{2\Delta t} (U_{j+1}^n - U_j^n).$$

The Lax-Friedrichs scheme (4.13), (4.29) is very simple to implement. Numerical results for Burgers' equation with initial data (4.16) and (4.17) are shown in Figure 4.8 (compare to figure 4.4). The results show that the approximate solutions are stable and nonoscillatory and approximate the entropy solution, unlike the Roe scheme. However, the computed solutions are *diffusive*. The shocks are smeared to a considerable extent. The numerical results are inferior to those obtained with the Godunov scheme.



(a) Shock solution with initial data (4.16) at  $t = 1$ .  
 (b) Rarefaction wave solution with initial data (4.17) at  $t = 0.5$ .

FIGURE 4.8. Approximate solution for Burgers' equation with the Lax-Friedrichs scheme with 50 mesh points. [burgers.disc.m]

**4.2.4. Rusanov scheme.** The Lax-Friedrichs scheme was quite diffusive around shocks. A possible explanation lies in the choice of the wave speeds (4.28). These speeds were the maximum allowed speeds and did not take into the account the speeds of propagation of the problem under consideration. A better, *locally selected*, choice of speeds is given by

$$(4.30) \quad s_{j+1/2}^r = s_{j+1/2}, \quad s_{j+1/2}^l = -s_{j+1/2},$$

where

$$s_{j+1/2} = \max(|f'(U_j^n)|, |f'(U_{j+1}^n)|).$$

The resulting flux (4.26), called the *Rusanov* (or *Local Lax-Friedrichs*) flux, is given by

$$(4.31) \quad \begin{aligned} F_{j+1/2}^n &= F^{\text{Rus}}(U_j^n, U_{j+1}^n) \\ &= \frac{f(U_j^n) + f(U_{j+1}^n)}{2} - \frac{\max(|f'(U_j^n)|, |f'(U_{j+1}^n)|)}{2} (U_{j+1}^n - U_j^n). \end{aligned}$$

The Rusanov scheme (4.13), (4.31) leads to a considerable improvement in results over the Lax-Friedrichs scheme, as shown in Figure 4.9.

**4.2.5. Engquist-Osher scheme.** A related scheme is the Engquist-Osher scheme, which has flux

$$(4.32) \quad \begin{aligned} F_{j+1/2}^n &= F^{\text{EO}}(U_j^n, U_{j+1}^n) \\ &= \frac{f(U_j^n) + f(U_{j+1}^n)}{2} - \frac{1}{2} \int_{U_j^n}^{U_{j+1}^n} |f'(\theta)| d\theta. \end{aligned}$$

Although it is difficult to write the Engquist-Osher flux as an approximate Riemann solver, it shares several features of approximate Riemann solvers. When the flux function has a single minimum at a point  $\omega$  and no maxima (which is the case for most convex functions), the Engquist-Osher flux can be explicitly computed as

$$(4.33) \quad F^{\text{EO}}(U_j^n, U_{j+1}^n) = f(\max(U_j^n, \omega)) + f(\min(U_{j+1}^n, \omega)) - f(\omega).$$



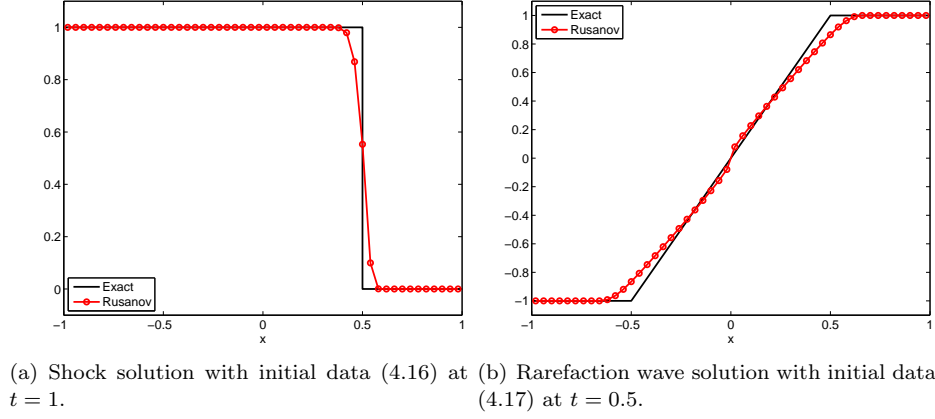


FIGURE 4.9. Approximate solution for Burgers' equation with the Rusanov scheme using 50 mesh points. [burgers\_disc.m]

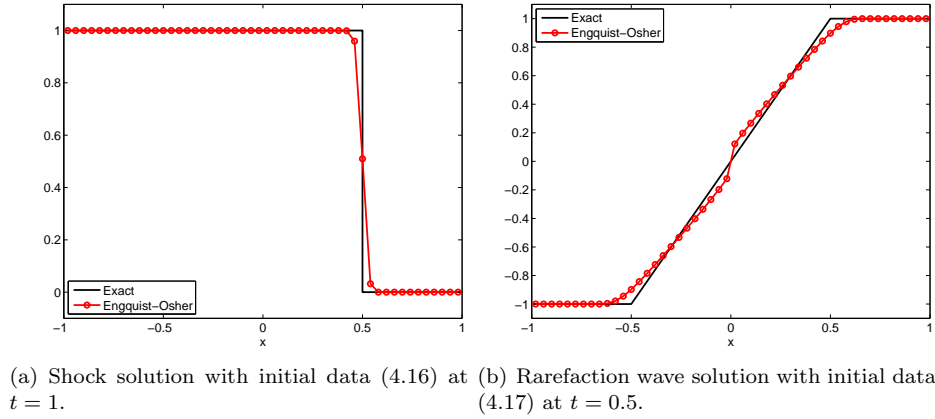


FIGURE 4.10. Approximate solution for Burgers' equation with the Engquist-Osher scheme using 50 mesh points. [burgers\_disc.m]

For convex fluxes with minimum at  $\omega$ , we denote

$$(4.34) \quad f^+(U) = f(\max(U, \omega)), \quad f^-(U) = f(\min(U, \omega)),$$

as the *positive* (increasing) and *negative* (decreasing) parts of  $f$ . As only the flux difference appear in (4.13), we can neglect the constant term  $f(\omega)$  in (4.33) and rewrite the Engquist-Osher scheme for convex fluxes as

$$(4.35) \quad F^{\text{EO}}(U_j^n, U_{j+1}^n) = f^+(U_j^n) + f^-(U_{j+1}^n).$$

Hence, the Engquist-Osher scheme is a *flux splitting scheme*, as it separates the flux into its positive and negative parts and takes the direction of propagation into account.

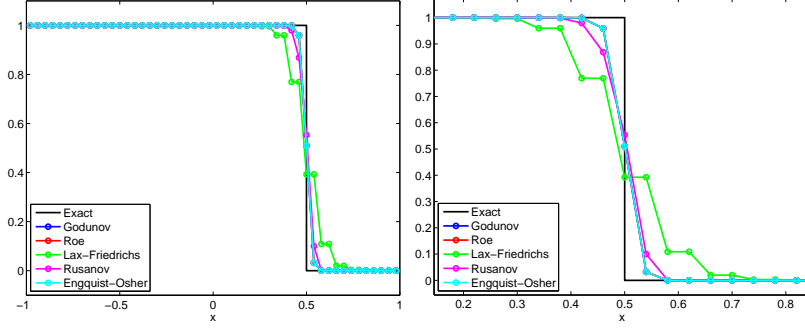


FIGURE 4.11. Approximate solution for Burgers' equation with the Godunov, Lax-Friedrichs, Rusanov, Roe and Engquist-Osher schemes at time  $t = 1.5$  with 50 mesh points for initial data (4.16). [burgers\_disc.m]

**Exercise 4.3.** Prove that the Engquist-Osher flux (4.32) can be written as (4.33) when the flux function has a single minimum at a point  $\omega$ .

#### 4.3. Comparison of different finite volume schemes

We compare all the numerical fluxes presented in this section for two sets of initial data. First, we consider the initial data (4.16) and compare the different numerical fluxes for a mesh consisting of 50 mesh points in Figure 4.11. The results show that the Godunov, Roe and Engquist-Osher schemes agree in this case. In fact, simple calculations show that in this case these three fluxes are equivalent. The Godunov scheme is clearly more accurate than the Rusanov scheme. The Lax-Friedrichs scheme leads to the largest amount of error as it smears the shock wave. We perform a convergence study of the schemes and present the results in Figure 4.12 (a). The solutions for initial data (4.16) are computed on a sequence of meshes and the error (with respect to the exact solution) in  $L^1$  is computed and plotted with respect to the number of mesh points (decreasing mesh sizes). The plot indicates that all the schemes converge as the mesh is refined. The convergence for the Godunov scheme is faster than the Rusanov and Lax-Friedrichs schemes, although the rate of convergence is similar for all the schemes. The results clearly show that the Godunov scheme is superior to the Rusanov and Lax-Friedrichs scheme. However, we must consider the fact that both the Lax-Friedrichs and Rusanov schemes have faster run times than the Godunov scheme. Hence, a fair comparison requires us to plot the *computational efficiency*. To do so, we compute with all the schemes on a sequence of meshes and plot the  $L^1$  error with respect to the *runtime* for each scheme in Figure 4.12 (b). The figure shows the obvious: Decreasing the mesh size gives more accurate approximations, but also leads to a higher run-time. The Engquist-Osher scheme turns out to be the most efficient in this case, at least for coarser meshes. We recall that the Godunov, Engquist-Osher and Roe schemes give the same numerical approximation on this problem. However, their run times are different. We point out that the schemes agree in terms of runtimes for highly refined meshes. Despite being the fastest on a given mesh, the Lax-Friedrichs continues to be the most inefficient scheme in this example.

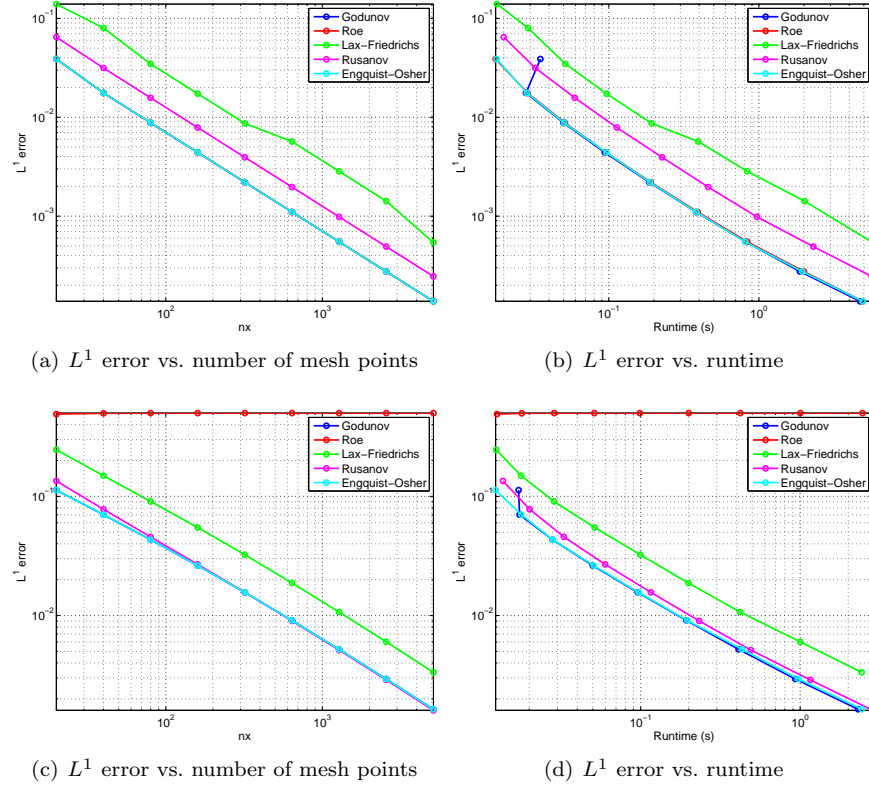


FIGURE 4.12. Convergence study for Burgers' equation with the Godunov, Lax-Friedrichs, Rusanov, Roe and Engquist-Osher schemes. Top row: initial data (4.16); bottom row: initial data (4.17). [burgers\_disc\_error.m]

Next, we repeat the experiments with the *rarefaction* initial data (4.17). The error vs. number of mesh points and error vs. run time is plotted in Figure 4.12 (c) and (d), respectively. The figures show that the Roe scheme does not converge in this case to the entropy solution. The Godunov and Engquist-Osher schemes are equivalent and lead to smaller errors than the Rusanov scheme, at least for coarse meshes. The Lax-Friedrichs scheme leads to the largest errors among the converging schemes. The computational efficiency plot shows that the Rusanov scheme is the most efficient in this case. Thus, the *optimal* scheme is a problem dependent concept.

#### 4.4. Convergence Analysis

The numerical results (particularly convergence results, Figures 4.12) show that most of the schemes of the previous section converge to the entropy solution of the scalar conservation law (4.1). However, the Roe scheme (4.22) may not converge to the entropy solution in some cases. It is easy to design schemes that lead to

incorrect or unstable solutions. One example is the scheme with the flux function

$$(4.36) \quad F_{j+1/2}^n = F^{\text{Cen}}(U_j^n, U_{j+1}^n) = \frac{f(U_j^n) + f(U_{j+1}^n)}{2}.$$

This flux gives the standard central difference scheme for the conservation law (4.1). As shown before, it ends up being unconditionally unstable, even for linear equations. Another possible flux is the one-sided flux

$$(4.37) \quad F_{j+1/2}^n = F(U_j^n, U_{j+1}^n) = f(U_j^n).$$

If we consider Burgers' equation with initial data (4.16), then the one-sided scheme (4.37) reduces to the Godunov scheme and will provide a good approximation of the solution. However, with initial data (4.17), the scheme (4.37) reduces to the Roe scheme and provides an incorrect approximation.

The above examples show that certain fluxes are stable whereas others are unstable. Furthermore, some fluxes converge to the correct entropy solutions whereas others converge to wrong solutions. The natural question is what the criteria are for the schemes to be stable and to converge to the entropy solutions of the scalar conservation law (4.1). We provide answers in this section.

First, we will identify certain features of the schemes presented so far.

**4.4.1. Conservative schemes.** A numerical scheme for approximating (4.1) can be written in the generic form

$$(4.38) \quad U_j^{n+1} = H(U_{j-p}^n, \dots, U_{j+p}^n)$$

for some update function  $H$  and constant  $p \in \mathbb{N}$ . The update function  $H$  depends on  $2p + 1$  points. This set of points  $\{U_{j-p}^n, \dots, U_{j+p}^n\}$  is called the *stencil* of the scheme. All the schemes presented so far have  $p = 1$  and are three-point schemes. More general  $(2p + 1)$ -point schemes will be presented in the sequel.

**Definition 4.4** (Conservative scheme). *Ignoring boundary conditions, a numerical scheme of generic form (4.38) approximating (4.1) is conservative if it satisfies*

$$(4.39) \quad \sum_j U_j^{n+1} = \sum_j U_j^n \quad \text{for all } n.$$

Conservation is a natural requirement from a scheme as the solutions to the continuous problem (4.1) are conservative, in the sense that the integral of the solution is preserved over time.

**Theorem 4.5.** *Assume that  $H(0, \dots, 0) = 0$ . Then (4.38) is conservative if and only if there exists a function  $F_{j+1/2}^n = F(U_{j-p+1}^n, \dots, U_{j+p}^n)$  such that (4.38) can be written in the finite volume form (4.13).*

PROOF. Necessity is straightforward: if (4.38) can be written as (4.13), then

$$\begin{aligned} \sum_j U_j^{n+1} &= \sum_j U_j^n - \frac{\Delta t}{\Delta x} \sum_j \left( F_{j+1/2}^n - F_{j-1/2}^n \right) \\ &= \sum_j U_j^n, \end{aligned}$$

since the second sum is a telescoping sum.

For sufficiency, define

$$G(U_{-p}, \dots, U_p) := \frac{\Delta x}{\Delta t} (U_0 - H(U_{-p}, \dots, U_p))$$

(we drop the time dependence for the moment). By conservation, we have

$$\sum_j G(U_{j-p}, \dots, U_{j+p}) = 0.$$

We want to show that there is an  $F_{j+1/2} = F(U_{j-p+1}, U_{j+p})$  such that

$$(4.40) \quad F_{j+1/2} - F_{j-1/2} = G(U_{j-p}, \dots, U_{j+p}).$$

It will suffice to construct  $F_{j+1/2}$  for  $j = 0$ . Select first  $U_j$  such that  $U_j = 0$  whenever  $j \leq -p$  or  $j > p$ . Then

$$\begin{aligned} 0 &= \sum_j G(U_{j-p}, \dots, U_{j+p}) \\ &= G(0, \dots, 0, U_{-p+1}) + \dots + G(0, U_{-p+1}, \dots, U_p) \\ &\quad + G(U_{-p+1}, \dots, U_p, 0) + \dots + G(U_p, 0, \dots, 0) \\ &= F_{1/2} + B \end{aligned}$$

where we have defined  $F_{1/2} := G(0, \dots, 0, U_{-p+1}) + \dots + G(0, U_{-p+1}, \dots, U_p)$  and  $B := G(U_{-p+1}, \dots, U_p, 0) + \dots + G(U_p, 0, \dots, 0)$ . Next, select  $U_j$  such that  $U_j = 0$  whenever  $j < -p$  or  $j > p$ . Then

$$\begin{aligned} 0 &= \sum_j G(U_{j-p}, \dots, U_{j+p}) \\ &= G(0, \dots, 0, U_{-p}) + \dots + G(0, U_{-p}, \dots, U_{p-1}) + G(U_{-p}, \dots, U_p) \\ &\quad + G(U_{-p+1}, \dots, U_p, 0) + \dots + G(U_p, 0, \dots, 0) \\ &= F_{-1/2} + G(U_{-p}, \dots, U_p) + B. \end{aligned}$$

Subtracting these two identities, we get the desired identity (4.40).  $\square$

The definition of conservation needs to be modified when boundaries are included in the discussion.

**4.4.2. Consistent schemes.** Another crucial requirement is consistency. Let some  $(2p+1)$ -point scheme (4.38) be given with numerical flux

$$(4.41) \quad F_{j+1/2}^n = F(U_{j-p+1}^n, \dots, U_{j+p}^n).$$

**Definition 4.6** (Consistency). *A finite volume scheme (4.13) with numerical flux function (4.41) is consistent if*

$$(4.42) \quad F(U, \dots, U) = f(U) \quad \text{for all } U \in \mathbb{R}.$$

It is straightforward to check that all the two-point numerical flux functions presented so far are consistent. Consistency is a natural requirement to ensure that the scheme approximates the correct conservation law.

Conservation and consistency are shared by all the schemes discussed till now. However, these criteria do not ensure either stability or convergence; for instance, the central scheme (4.36) is both consistent and conservative, but fails to be stable. Similarly, the Roe scheme (4.22) is consistent and conservative, but does not converge to the entropy solution in certain numerical experiments. We need another criterion.

**4.4.3. Monotone schemes.** The entropy solutions of (4.1) are *monotonicity preserving*, meaning that if  $U$  and  $V$  are entropy solutions of (4.1) with initial data  $U_0$  and  $V_0$ , respectively, then

$$U_0(x) \leq V_0(x) \quad \text{for all } x \quad \Rightarrow \quad U(x, t) \leq V(x, t) \quad \text{for all } x, t.$$

It is desirable for numerical schemes to possess this property. At the discrete level, it is defined by the following.

**Definition 4.7.** *A numerical scheme (4.38) is monotone if the update function  $H$  is non-decreasing in each of its arguments.*

If (4.38) can be written in the finite volume form (4.13) with a two-point flux function (so  $p = 1$  in (4.41)), then a sufficient condition for monotonicity is given by

**Lemma 4.8.** *A finite volume scheme (4.13) with a two-point flux  $F = F(a, b)$  is monotone if  $F$  is a Lipschitz continuous function in both arguments, non-decreasing in its first argument and non-increasing in its second argument. More specifically, the following hold:*

$$(4.43) \quad \begin{aligned} a \mapsto F(a, b) & \text{ is non-decreasing for fixed } b, \\ b \mapsto F(a, b) & \text{ is non-increasing for fixed } a, \end{aligned}$$

and the following CFL-type condition holds:

$$(4.44) \quad \left| \frac{\partial F}{\partial a} \right| + \left| \frac{\partial F}{\partial b} \right| \leq \frac{\Delta x}{\Delta t}.$$

The proof of this lemma is left as an exercise to the reader.

**4.4.4. Examples of monotone schemes.** Assume that the flux function  $f$  in (4.1) is a differentiable function. Then the following schemes are monotone:

- The Lax-Friedrichs scheme (4.29) has a smooth  $C^1$  numerical flux function. Explicit computations result in

$$\begin{aligned} \frac{\partial F}{\partial a} &= \frac{1}{2} \left( f'(a) + \frac{\Delta x}{\Delta t} \right) \\ \frac{\partial F}{\partial b} &= \frac{1}{2} \left( f'(b) - \frac{\Delta x}{\Delta t} \right). \end{aligned}$$

The condition (4.44) then follows from the CFL condition (4.9).

- The Rusanov scheme (4.31) is monotone. The proof follows as above.
- The Engquist-Osher scheme (4.32) has a smooth numerical flux. A direct calculation shows that

$$\begin{aligned} \frac{\partial F}{\partial a} &= \frac{1}{2} (f'(a) + |f'(a)|) \\ \frac{\partial F}{\partial b} &= \frac{1}{2} (f'(b) - |f'(b)|). \end{aligned}$$

Furthermore, the CFL condition (4.44) is a consequence of (4.9). Hence, the Engquist-Osher scheme is monotone.

- The Godunov scheme (4.14) has a Lipschitz continuous flux function. It is cumbersome to check by a case by case analysis that the Godunov scheme is monotone. A heuristic explanation relies on the fact that the both

the evolution and averaging operators used in the deriving the Godunov scheme are monotone operators.

On the other hand, it is easy to check that the central flux (4.36), the one-sided flux (4.37) and the Roe flux (4.22) are not monotone. Thus, monotonicity of the scheme provides a demarcation between robust and potentially non-robust schemes.

**4.4.5. Stability estimates.** The numerical experiments indicate that monotone schemes are stable and converge to the entropy solution, whereas the non-monotone schemes may be unstable or may converge to the incorrect solution. A crucial step in proving convergence is to obtain stability estimates on the approximate solutions, computed by the schemes.

From our study of the continuous problem in the previous section, the entropy solutions of the continuous problem (4.1) satisfy the following stability estimates.

- (i)  $L^\infty$  estimate (3.38) as a consequence of a maximum principle.
- (ii)  $L^p$  estimates (3.30) for  $1 \leq p < \infty$  as a consequence of the entropy inequality (3.29).
- (iii) The BV estimate (3.39) showing that the solutions are TVD.

The central philosophy of numerical analysis is to devise numerical schemes that preserve stability properties of the underlying continuous problem. Hence, it is natural to examine whether the finite volume schemes presented in this section satisfy discrete versions of the above stability estimates.

**$L^\infty$  bounds.** A consistent three point finite volume scheme (4.13) can be written as

$$(4.45) \quad U_j^{n+1} = H(U_{j-1}^n, U_j^n, U_{j+1}^n),$$

with the consistency of the fluxes  $F$  in (4.13) implying that

$$H(U, U, U) = U.$$

The scheme (4.45) is monotone if  $H$  is non-decreasing in each of its three arguments. The Godunov, Engquist-Osher, Lax-Friedrichs and Rusanov schemes (under a CFL condition like (4.9)) are examples of three-point monotone schemes. These schemes satisfy the following *discrete maximum principle*.

**Lemma 4.9.** *Let  $U_j^n$  be the approximate solutions generated by a three-point consistent monotone scheme of the form (4.45). Then the solutions satisfy*

$$(4.46) \quad \min(U_{j-1}^n, U_j^n, U_{j+1}^n) \leq U_j^{n+1} \leq \max(U_{j-1}^n, U_j^n, U_{j+1}^n) \quad \text{for all } n, j.$$

*In particular, we have*

$$(4.47) \quad \min_k U_k^0 \leq U_j^n \leq \max_k U_k^0 \quad \text{for all } n, j.$$

PROOF. Let  $\bar{U}_j^n = \max(U_{j-1}^n, U_j^n, U_{j+1}^n)$ . By definition,

$$U_{j-1}^n, U_j^n, U_{j+1}^n \leq \bar{U}_j^n.$$

Since the scheme (4.45) is monotone, the update function  $H$  is monotone non-decreasing in each of its arguments. Therefore

$$\begin{aligned} U_j^{n+1} &= H(U_{j-1}^n, U_j^n, U_{j+1}^n) \leq H(\bar{U}_j^n, U_j^n, U_{j+1}^n) && \text{(monotonicity)} \\ &\leq H(\bar{U}_j^n, \bar{U}_j^n, U_{j+1}^n) && \text{(monotonicity)} \\ &\leq H(\bar{U}_j^n, \bar{U}_j^n, \bar{U}_j^n) && \text{(monotonicity)} \end{aligned}$$

$$= \bar{U}_j^n \quad (\text{consistency}).$$

The minimum principle follows analogously. Iterating the maximum principle (4.46) over all time levels up to  $n$  yields the  $L^\infty$  bound (4.47).  $\square$

**4.4.6. Entropy inequalities and  $L^p$  bounds.** The key to obtaining  $L^p$  bounds and characterizing the correct entropy solution is to obtain a discrete version of the entropy inequality (3.29). As we have seen in Chapter 3, it is enough to obtain a discrete version of the Kruzhkov entropy inequality (3.33). We proceed to do so below.

For a constant  $k \in \mathbb{R}$ , define the Crandall-Majda numerical entropy flux as

$$(4.48) \quad \begin{aligned} Q_{j+1/2}^n &= Q(U_j^n, U_{j+1}^n) \\ &= F(U_j^n \vee k, U_{j+1}^n \vee k) - F(U_j^n \wedge k, U_{j+1}^n \wedge k), \end{aligned}$$

where we use the notation

$$a \vee b = \max(a, b), \quad a \wedge b = \min(a, b).$$

Note that this numerical entropy flux is consistent with (3.32) whenever the numerical flux  $F$  is consistent, as

$$\begin{aligned} Q(U, U) &= F(U \vee k, U \vee k) - F(U \wedge k, U \wedge k) \\ &= f(U \vee k) - f(U \wedge k) \\ &= \text{sign}(U - k)(f(U) - f(k)) \\ &= Q(U, k). \end{aligned}$$

In the following lemma we obtain a discrete version of (3.33).

**Lemma 4.10** (Crandall-Majda). *Let  $U_j^n$  be an approximate solution computed by a consistent and monotone three-point scheme (4.45). Then  $U$  satisfies the discrete entropy inequality*

$$(4.49) \quad |U_j^{n+1} - k| - |U_j^n - k| + \frac{\Delta t}{\Delta x} (Q_{j+1/2}^n - Q_{j-1/2}^n) \leq 0$$

for all  $n, j$ .

PROOF. By definition of the scheme (4.45) and (4.13), we have

$$\begin{aligned} H(U_{j-1}^n \vee k, U_j^n \vee k, U_{j+1}^n \vee k) \\ = U_j^n \vee k - \frac{\Delta t}{\Delta x} (F(U_j^n \vee k, U_{j+1}^n \vee k) - F(U_{j-1}^n \vee k, U_j^n \vee k)) \end{aligned}$$

and

$$\begin{aligned} H(U_{j-1}^n \wedge k, U_j^n \wedge k, U_{j+1}^n \wedge k) \\ = U_j^n \wedge k - \frac{\Delta t}{\Delta x} (F(U_j^n \wedge k, U_{j+1}^n \wedge k) - F(U_{j-1}^n \wedge k, U_j^n \wedge k)). \end{aligned}$$

Using the definition of  $Q_{j+1/2}^n$ , we obtain

$$(4.50) \quad \begin{aligned} H(U_{j-1}^n \vee k, U_j^n \vee k, U_{j+1}^n \vee k) - H(U_{j-1}^n \wedge k, U_j^n \wedge k, U_{j+1}^n \wedge k) \\ = U_j^n \vee k - U_j^n \wedge k - \frac{\Delta t}{\Delta x} (Q_{j+1/2}^n - Q_{j-1/2}^n) \\ = |U_j^n - k| - \frac{\Delta t}{\Delta x} (Q_{j+1/2}^n - Q_{j-1/2}^n). \end{aligned}$$



By monotonicity and the definition of  $H$  (4.45),

$$(4.51) \quad \begin{aligned} H(U_{j-1}^n \vee k, U_j^n \vee k, U_{j+1}^n \vee k) &\geq H(U_{j-1}^n, U_j^n, U_{j+1}^n) = U_j^{n+1}, \\ H(U_{j-1}^n \wedge k, U_j^n \wedge k, U_{j+1}^n \wedge k) &\leq H(U_{j-1}^n, U_j^n, U_{j+1}^n) = U_j^{n+1}. \end{aligned}$$

Similarly, by monotonicity and consistency of  $H$ ,

$$(4.52) \quad \begin{aligned} H(U_{j-1}^n \vee k, U_j^n \vee k, U_{j+1}^n \vee k) &\geq H(k, k, k) = k, \\ H(U_{j-1}^n \wedge k, U_j^n \wedge k, U_{j+1}^n \wedge k) &\leq H(k, k, k) = k. \end{aligned}$$

By subtracting (4.52) from (4.51) and using (4.50), we obtain

$$|U_j^n - k| - \frac{\Delta t}{\Delta x} (Q_{j+1/2}^n - Q_{j-1/2}^n) \geq |U_j^{n+1} - k|,$$

which proves (4.49).  $\square$

**Remark 4.11.** *The discrete entropy inequality (4.49) can be used with an approximation argument for Kruzhkov entropies (3.31) to obtain discrete entropy inequalities for any convex function  $S$ . This implies  $L^p$  bounds on the solution by summing over all mesh points.*

**4.4.7. TV bounds.** Another essential stability estimate for scalar conservation laws (4.1) is the BV estimate (3.39), which provides control over the oscillations in the solution. This estimate says that the entropy solutions of (4.1) are TVD, meaning that the TV norm does not increase in time. We need to obtain a discrete version of this total variation. The total variation of a piecewise constant function at time level  $t^n$  is given by

$$\sum_j |U_{j+1}^n - U_j^n|.$$

The first step for obtaining a TV bound is rewriting the finite volume scheme in the *incremental form*

$$(4.53) \quad U_j^{n+1} = U_j^n + C_{j+1/2}^n (U_{j+1}^n - U_j^n) - D_{j-1/2}^n (U_j^n - U_{j-1}^n)$$

with incremental coefficients

$$(4.54) \quad C_{j+1/2}^n = \frac{\Delta t}{\Delta x} \frac{f(U_j^n) - F_{j+1/2}^n}{U_{j+1}^n - U_j^n}, \quad D_{j+1/2}^n = \frac{\Delta t}{\Delta x} \frac{f(U_{j+1}^n) - F_{j+1/2}^n}{U_{j+1}^n - U_j^n}.$$

A straightforward calculation shows that the incremental form (4.53) is equivalent to the standard finite volume form (4.13). The advantage of the incremental form lies in the following useful lemma.

**Lemma 4.12** (Harten's Lemma). *Assume that the scheme (4.53) satisfies*

$$(4.55) \quad C_{j+1/2}^n, D_{j+1/2}^n \geq 0 \quad \text{and} \quad C_{j+1/2}^n + D_{j+1/2}^n \leq 1 \quad \text{for all } n, j.$$

*Then solutions computed with (4.53) are TVD, i.e. they satisfy*

$$(4.56) \quad \sum_j |U_{j+1}^{n+1} - U_j^{n+1}| \leq \sum_j |U_{j+1}^n - U_j^n|.$$

PROOF. Using the incremental form (4.53), we obtain

$$\begin{aligned} U_{j+1}^{n+1} - U_j^{n+1} &= (1 - C_{j+1/2}^n - D_{j+1/2}^n) (U_{j+1}^n - U_j^n) \\ &\quad + C_{j+3/2}^n (U_{j+2}^n - U_{j+1}^n) + D_{j-1/2}^n (U_j^n - U_{j-1}^n). \end{aligned}$$

Taking absolute values on both sides of the above equation and using (4.55) yields

$$\begin{aligned} |U_{j+1}^{n+1} - U_j^{n+1}| &= \left(1 - C_{j+1/2}^n - D_{j+1/2}^n\right) |U_{j+1}^n - U_j^n| \\ &\quad + C_{j+3/2}^n |U_{j+2}^n - U_{j+1}^n| + D_{j-1/2}^n |U_j^n - U_{j-1}^n|. \end{aligned}$$

Summing over  $j$  and identifying equal terms, we obtain the TVD estimate (4.56).  $\square$

Harten's lemma allows us to check whether a scheme in incremental form is TVD or not. It turns out that consistent monotone schemes are TVD.

**Lemma 4.13.** *Any consistent monotone three-point finite volume scheme (4.13) is TVD under the CFL condition (4.44).*

PROOF. We need to check if the scheme satisfies the conditions (4.55) of Harten's lemma. By explicit computations we obtain

$$\begin{aligned} C_{j+1/2}^n &= \frac{\Delta t}{\Delta x} \frac{f(U_j^n) - F(U_j^n, U_{j+1}^n)}{U_{j+1}^n - U_j^n} && \text{(definition (4.54))} \\ &= \frac{\Delta t}{\Delta x} \frac{F(U_j^n, U_j^n) - F(U_j^n, U_{j+1}^n)}{U_{j+1}^n - U_j^n} && \text{(consistency)} \\ &\geq 0 && \text{(monotonicity).} \end{aligned}$$

The correct sign of  $D$  can be similarly checked. We also have

$$\begin{aligned} C_{j+1/2}^n &= \frac{\Delta t}{\Delta x} \frac{f(U_j^n) - F(U_j^n, U_{j+1}^n)}{U_{j+1}^n - U_j^n} && \text{(definition (4.54))} \\ &= \frac{\Delta t}{\Delta x} \frac{F(U_j^n, U_j^n) - F(U_j^n, U_{j+1}^n)}{U_{j+1}^n - U_j^n} && \text{(consistency)} \\ &\leq \frac{\Delta t}{\Delta x} \left| \frac{\partial F}{\partial b} \right| && \text{(Lipschitz flux).} \end{aligned}$$

Similarly,

$$D_{j+1/2}^n \leq \frac{\Delta t}{\Delta x} \left| \frac{\partial F}{\partial a} \right|.$$

Therefore,

$$\begin{aligned} 1 - C_{j+1/2}^n - D_{j+1/2}^n &\geq 1 - \frac{\Delta t}{\Delta x} \left( \left| \frac{\partial F}{\partial b} \right| + \left| \frac{\partial F}{\partial a} \right| \right) \\ &\geq 0 \end{aligned}$$

by the CFL condition (4.44).  $\square$

**4.4.8. Convergence result.** In the previous sections, we have established that monotone consistent conservative schemes are stable in  $L^\infty$ , are TVD and satisfy a discrete entropy condition (4.49). These stability estimates pave the way for showing that the approximate solutions computed by any monotone consistent conservative scheme converge to the entropy solution of (4.1). The following famous theorem, due to Lax and Wendroff, says that *if* the approximate solutions  $U_j^n$  computed by a consistent and conservative scheme converge, then the limit is a

weak solution of the conservation law. For ease of notation, we will identify the approximate solutions with a piecewise constant function,

$$(4.57) \quad U^\Delta(x, t) = U_j^n \quad \text{for } (x, t) \in [x_{j-1/2}, x_{j+1/2}) \times [t^n, t^{n+1}).$$

**Theorem 4.14** (Lax-Wendroff). *Let  $U_j^n$  be approximate solutions of (4.1), computed by a conservative and consistent finite volume scheme (4.13) with a differentiable (or Lipschitz) numerical flux function  $F$ . Assume that  $U_0 \in L^\infty(\mathbb{R})$  and that the approximating functions  $U^\Delta$*

- *are uniformly bounded, i.e.,*

$$(4.58) \quad \|U^\Delta\|_{L^\infty(\mathbb{R} \times \mathbb{R}_+)} \leq C$$

*for some constant  $C > 0$ ;*

- *converge to some function  $U$  in  $L^1(\mathbb{R} \times \mathbb{R}_+)$  and almost everywhere.*

*Then  $U$  is a weak solution of (4.1)*

PROOF. Let  $\phi \in C_c^\infty(\mathbb{R} \times [0, \infty))$  be a given test function and denote  $\phi_j^n = \phi(x_j, t^n)$ . By multiplying (4.13) by  $\Delta x \phi_j^n$  and summing over  $j \in \mathbb{Z}$  and  $n \in \mathbb{N}_0$ , we obtain

$$0 = \Delta x \Delta t \sum_{j=-\infty}^{\infty} \sum_{n=0}^{\infty} \left( \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{F_{j+1/2}^n - F_{j-1/2}^n}{\Delta x} \right) \phi_j^n$$

(summation by parts)

$$= -\Delta x \sum_{j=-\infty}^{\infty} U_j^0 \phi_j^0 - \Delta x \Delta t \sum_{j=-\infty}^{\infty} \sum_{n=0}^{\infty} \left( U_j^{n+1} \frac{\phi_j^{n+1} - \phi_j^n}{\Delta t} + F_{j+1/2}^n \frac{\phi_{j+1}^n - \phi_j^n}{\Delta x} \right).$$

Note that since  $\phi$  has compact support, only finitely many of the summands in the above expression are nonzero, and so all the sums are well-defined. As with  $U^\Delta$ , define  $\phi^\Delta(x, t) = \phi_j^n$  for  $(x, t) \in [x_{j-1/2}, x_{j+1/2}) \times [t^n, t^{n+1})$ . Recognizing the above expression as a Riemann sum of step functions, it can be written as

$$(4.59) \quad \begin{aligned} & \int_{\mathbb{R}} U^\Delta(x, 0) \phi^\Delta(x, 0) dx + \int_{\mathbb{R}} \int_0^\infty U^\Delta(x, t + \Delta t) \frac{\phi^\Delta(x, t + \Delta t) - \phi^\Delta(x, t)}{\Delta t} dt dx \\ & + \int_{\mathbb{R}} \int_0^\infty F(U^\Delta(x, t), U^\Delta(x + \Delta x, t)) \frac{\phi^\Delta(x + \Delta x, t) - \phi^\Delta(x, t)}{\Delta x} dt dx = 0. \end{aligned}$$

Both  $U^\Delta$  and  $\phi^\Delta$  converge boundedly a.e. to  $U$  and  $\phi$ , respectively, so by Lebesgue's dominated convergence theorem, the first two terms converge to

$$\int_{\mathbb{R}} U(x, 0) \phi(x, 0) dx + \int_{\mathbb{R}} \int_0^\infty U(x, t) \phi_t(x, t) dt dx$$

as  $\Delta x, \Delta t \rightarrow 0$ . It is slightly more tricky to pass to the limit in the third term in (4.59), as  $U^\Delta$  occurs inside the (possibly nonlinear) function  $F$ . However, using (4.58) and the fact that  $F$  is differentiable, there is some  $\tilde{C} > 0$  such that

$$|F(U^\Delta(x, t), U^\Delta(x + \Delta x, t)) - F(U^\Delta(x, t), U^\Delta(x, t))| \leq \tilde{C} |U^\Delta(x + \Delta x, t) - U^\Delta(x, t)|.$$

Furthermore, by the consistency of  $F$ , we have  $F(U^\Delta(x, t), U^\Delta(x, t)) = f(U^\Delta(x, t))$ . Adding and subtracting  $f(U^\Delta(x, t))$ , we have

$$\begin{aligned} & \int_{\mathbb{R}} \int_0^\infty F(U^\Delta(x, t), U^\Delta(x + \Delta x, t)) \frac{\phi^\Delta(x + \Delta x, t) - \phi^\Delta(x, t)}{\Delta x} dx \\ &= \int_{\mathbb{R}} \int_0^\infty f(U^\Delta(x, t)) \frac{\phi^\Delta(x + \Delta x, t) - \phi^\Delta(x, t)}{\Delta x} dx \\ & \quad + \int_{\mathbb{R}} \int_0^\infty \left( F(U^\Delta(x, t), U^\Delta(x + \Delta x, t)) - f(U^\Delta(x, t)) \right) \frac{\phi^\Delta(x + \Delta x, t) - \phi^\Delta(x, t)}{\Delta x} dx. \end{aligned}$$

The first term converges towards  $\int_{\mathbb{R}} \int_0^\infty f(U(x, t)) \phi_x(x, t) dx$ , while the second term is bounded by

$$\tilde{C} \|\phi_x\|_{L^\infty} \int_{\mathbb{R}} \int_0^\infty |U^\Delta(x + \Delta x, t) - U^\Delta(x, t)| dx,$$

which through approximation by  $C^\infty$  functions may be shown to converge to zero. In conclusion, the left-hand side of (4.59) converges to

$$\int_{\mathbb{R}} \int_0^\infty U \phi_t + f(U) \phi_x dt dx + \int_{\mathbb{R}} U(x, 0) \phi(x, 0) dx,$$

thus proving that  $U$  is a weak solution.  $\square$

The convergence of the approximate solutions  $U^{\Delta x}$  to a limit function  $U$  is a consequence of the  $L^\infty$  and TV bounds (4.56) for a monotone, consistent and conservative scheme. The entropy condition is satisfied as the discrete entropy inequality (4.49) holds. The theorem establishes that a monotone, conservative and consistent scheme like the Godunov, Lax-Friedrichs and Engquist-Osher schemes converge to the entropy solution of a scalar conservation law and justify the results of numerical experiments.

#### 4.5. A note on boundary conditions

The discussion so far has ignored boundary conditions. However, we need to specify boundary conditions as

- The problem itself may be an initial- boundary value problem and physical boundary conditions such as Dirichlet, Neumann or periodic boundary conditions might be specified.
- For an initial value problem on the whole real line, we need to truncate the domain, as a computational domain must be bounded. In this case, we need artificial or numerical boundary conditions.

Without exploring this issue in detail in the current discussion, we provide a recipe for implementing boundary conditions with finite volume schemes (4.13). Let  $[x_L, x_R]$  be the physical or computational domain. Denote  $x_{1/2} = x_L$  and  $x_{N+1/2} = x_R$  (for a mesh with  $N + 1$  mesh points). We need the following *ghost cells*:

$$(4.60) \quad \mathcal{C}_0 = [x_L - \Delta x, x_L), \quad \mathcal{C}_{N+1} = [x_R, x_R + \Delta x).$$

We denote the cell averages of the unknown over  $\mathcal{C}_0$  and  $\mathcal{C}_{N+1}$  at time level  $t^n$  as  $U_0^n$  and  $U_{N+1}^n$  respectively.

**4.5.1. Dirichlet boundary conditions.** Let the conservation law (4.1) be augmented with Dirichlet boundary conditions

$$(4.61) \quad U(x_L, t) = g_L(t), \quad U(x_R, t) = g_R(t).$$

These boundary conditions are implemented by specifying the ghost values

$$(4.62) \quad U_0^n = g_L(t^n), \quad U_{N+1}^n = g_R(t^n).$$

Observe that the boundary condition is implemented weakly.

**4.5.2. Periodic boundary conditions.** These boundary conditions are implemented as

$$(4.63) \quad U_0^n = U_N^n, \quad U_{N+1}^n = U_1^n.$$

**4.5.3. Artificial boundary conditions.** Non-reflecting Neumann type boundary conditions are implemented as

$$(4.64) \quad U_0^n = U_1^n, \quad U_{N+1}^n = U_N^n.$$



## CHAPTER 5

### Second-order (high-resolution) finite volume schemes

The finite volume schemes (4.13) with suitable numerical fluxes like the Godunov, Lax-Friedrichs and Engquist-Osher fluxes have been demonstrated to be quite robust in approximating scalar conservation laws

$$(5.1) \quad U_t + f(U)_x = 0.$$

However, some problems still persist as the schemes may lead to large errors. To illustrate this point, we consider the linear advection equation (2.2) and compute the solutions with the Godunov (upwind) scheme (2.16) for initial data

$$(5.2) \quad U(x, 0) = \sin(4\pi x)$$

in the computational domain  $[0, 1]$  with periodic boundary conditions.

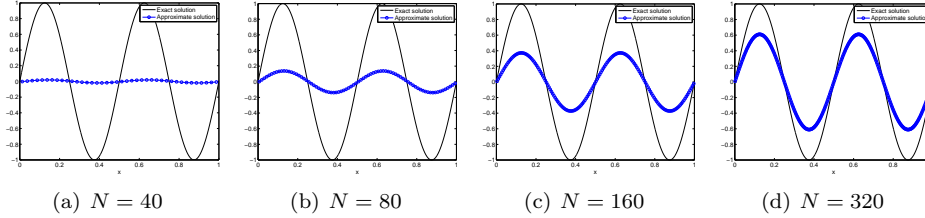


FIGURE 5.1. Linear advection equation (2.2) computed to time  $t = 10$ , using upwind flux with smooth initial conditions (5.2) and periodic boundary conditions at different meshes. `[linAdv_Upw_refine.m]`

The exact solution of the problem (2.2), (5.2) is  $U(x, t) = \sin(4\pi(x - t))$ . Coupled with the periodic boundary condition, it implies that the exact solution returns back to the initial condition (5.2) at time  $t = k$  for all positive integers  $k$ . The computed results on a sequence of meshes are presented in Figure 5.1 and Table 5.1. Table 5.1 shows the relative error in  $L^1$  on a sequence of meshes. The percentage relative error is defined by

$$(5.3) \quad \mathcal{E}^{\Delta x} = 100 \times \frac{\|U^{\Delta x} - U^{\text{ref}}\|_{L^1}}{\|U^{\text{ref}}\|_{L^1}},$$

where  $U^{\Delta x}$  is the approximate solution computed on a mesh with mesh size  $\Delta x$  and  $U^{\text{ref}}$  is a reference (exact) solution of the continuous problem. In Table 5.1, we

No. of Cells	$\mathcal{E}^{\Delta x}$	EOC
20	100	—
40	98.1	0.027
80	86.1	0.188
160	62.7	0.457
320	39	0.688
640	21.9	0.833
1280	11.6	0.914
2560	5.98	0.956

TABLE 5.1. Error and order of convergence for the linear advection equation (2.2) using upwind scheme, with smooth initial data (5.2). [linAdv\_OOC.m]

have also shown the *experimental order of convergence* (EOC),

$$(5.4) \quad \text{EOC}_{\Delta x, \Delta y} = \frac{\log(\mathcal{E}^{\Delta x}) - \log(\mathcal{E}^{\Delta y})}{\log(\Delta x) - \log(\Delta y)}.$$

Here,  $\Delta x, \Delta y$  are two different mesh sizes. In Table 5.1, we have used  $\Delta y = 2\Delta x$  for all the results.

The results of Figure 5.1 indicate that the approximation by the Godunov scheme is quite stable and there are no spurious oscillations or other numerical artifacts. However, the approximation has large errors and the extrema of the solution are clipped. Table 5.1 provides quantitative evidence of these conclusions. The relative errors are quite large, particularly at coarse meshes. An explanation for the large errors is provided in the experimental order of convergence (EOC) column of the table. The observed order of convergence is close to one. This implies that the convergence is slow and the errors are reduced very slowly, impacting the computational efficiency of the scheme.

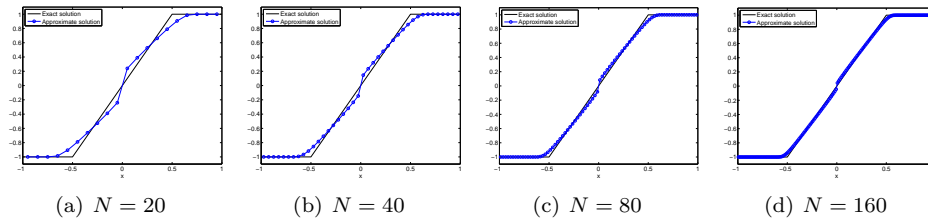


FIGURE 5.2. Burgers' equation (3.3) computed up to time  $t = 0.5$ , using Godunov flux with initial conditions (4.16) and outflow boundary conditions at different meshes. [burgers\_refine.m]

A similar situation is observed with computations for the nonlinear Burgers equation (3.3). We compute approximate solutions for initial data (4.17) with the Godunov scheme (4.15) in the computational domain  $[-1, 1]$  on a sequence of meshes and present the results in Figure 5.2 and Table 5.2. The results show that the Godunov scheme is stable and converges, but slowly. In fact, the order of convergence is less than one.



No. of Cells	$\mathcal{E}^{\Delta x}$	EOC
20	6.86	–
40	4.51	0.605
80	2.84	0.667
160	1.74	0.706
320	1.04	0.745
640	0.606	0.778
1280	0.347	0.804
2560	0.196	0.826

TABLE 5.2. Error and order of convergence for Burgers' equation (3.3) using Godunov's scheme with initial data (4.16). [burgers\_00C.m]

Both numerical experiments indicate that the Godunov type finite volume scheme (4.13) is stable but can lead to large errors, due to its slow convergence. A possible recipe for reducing errors is to increase the rate of convergence. This chapter is devoted to introducing the concepts of truncation errors and order of accuracy of schemes. Furthermore, we will construct second-order accurate versions of the finite volume scheme (4.13) for approximating scalar conservation laws.

### 5.1. Order of accuracy

The order of accuracy of numerical schemes is a very useful concept in numerical analysis.

**Definition 5.1.** Assume that a finite volume scheme for approximating (5.1) can be written in the generic  $(2p+1)$ -point update form (4.38), and that  $\frac{\Delta t}{\Delta x} \equiv \lambda$  for some constant  $\lambda > 0$ . The truncation error of the scheme is defined as

$$(5.5) \quad \mathcal{T}_{j,n} = U(x_j, t^{n+1}) - H(U(x_{j-p}, t^n), \dots, U(x_{j+p}, t^n)),$$

where  $U$  is the exact solution. The scheme is  $q$ -th order accurate if  $q \in \mathbb{N}$  is the largest integer for which

$$(5.6) \quad \mathcal{T}_{j,n} = \mathcal{O}(\Delta t^{q+1}) \quad \text{for all } j, n$$

as  $\Delta t \rightarrow 0$ .

**Lemma 5.2.** Assume that the exact solution  $U(x, t)$  of the scalar conservation law is  $C^2$ , and that a consistent and conservative three-point finite volume method has a  $C^2$  update function  $H$ . Then the scheme is at least first-order accurate.

PROOF. As the solution and the update function  $H$  in (4.45) are at least twice continuously differentiable, we can use Taylor expansions. We write the update function as  $H = H(X, Y, Z)$ . From (4.45) and (4.13) we get

$$(5.7) \quad \begin{aligned} \frac{\partial H}{\partial X}(U_{j-1}^n, U_j^n, U_{j+1}^n) &= \frac{\Delta t}{\Delta x} \frac{\partial F}{\partial a}(U_{j-1}^n, U_j^n), \\ \frac{\partial H}{\partial Z}(U_{j-1}^n, U_j^n, U_{j+1}^n) &= -\frac{\Delta t}{\Delta x} \frac{\partial F}{\partial b}(U_j^n, U_{j+1}^n). \end{aligned}$$

Denoting  $\tilde{U}_j^n = U(x_j, t^n)$  and writing down the Taylor expansion for the truncation error (5.5), we obtain

$$\begin{aligned} \mathcal{T}_{j,n} &= \tilde{U}_j^{n+1} - H(\tilde{U}_{j-1}^n, \tilde{U}_j^n, \tilde{U}_{j+1}^n) \\ &= \tilde{U}_j^n + \Delta t U_t(x_j, t^n) + \mathcal{O}(\Delta t^2) - H(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n) \\ &\quad - \frac{\partial H}{\partial X}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n)(\tilde{U}_{j-1}^n - \tilde{U}_j^n) - \frac{\partial H}{\partial Z}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n)(\tilde{U}_{j+1}^n - \tilde{U}_j^n) + \mathcal{O}(\Delta x^2) \\ &= \Delta t U_t(x_j, t^n) + \mathcal{O}(\Delta t^2) \\ &\quad - \frac{\partial H}{\partial X}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n)(\tilde{U}_{j-1}^n - \tilde{U}_j^n) - \frac{\partial H}{\partial Z}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n)(\tilde{U}_{j+1}^n - \tilde{U}_j^n) + \mathcal{O}(\Delta x^2) \end{aligned}$$

by the consistency of  $H$ . By using the fact that  $\Delta t = \lambda \Delta x$  and Taylor expanding once more, the above relation reduces to

$$\begin{aligned} \mathcal{T}_{j,n} &= \Delta t U_t(x_j, t^n) \\ &\quad + \Delta x \frac{\partial H}{\partial X}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n) U_x(x_j, t^n) - \Delta x \frac{\partial H}{\partial Z}(\tilde{U}_j^n, \tilde{U}_j^n, \tilde{U}_j^n) U_x(x_j, t^n) + \mathcal{O}(\Delta t^2). \end{aligned}$$

Using the partial derivatives (5.7) and differentiating the consistency relation  $F(U, U) = f(U)$  we obtain

$$\begin{aligned} \mathcal{T}_{j,n} &= \Delta t U_t(x_j, t^n) + \Delta t \left( \frac{\partial F}{\partial a}(\tilde{U}_j^n, \tilde{U}_j^n) + \frac{\partial F}{\partial b}(\tilde{U}_j^n, \tilde{U}_j^n) \right) U_x(x_j, t^n) + \mathcal{O}(\Delta t^2) \\ &= \Delta t \underbrace{(U_t(x_j, t^n) + f'(U(x_j, t^n)) U_x(x_j, t^n))}_{= 0 \text{ by (5.1)}} + \mathcal{O}(\Delta t^2) \\ &= \mathcal{O}(\Delta t^2). \end{aligned}$$

Thus (5.6) holds for  $q = 1$ .  $\square$

Note that the conditions of Lemma 5.2 are satisfied by the Lax-Friedrichs and Engquist-Osher schemes. However, the Godunov flux (4.14) is not  $C^2$  and the conclusions of the above lemma do not directly apply to the Godunov flux, except in the linear case. Nevertheless, it may be shown that the Godunov flux is formally first-order accurate.

The fact that the truncation error has a certain decay as  $\Delta t \rightarrow 0$  can be converted to an estimate on the rate of convergence in the linear case. To be more precise, consider the transport equation (2.2) and a first-order scheme (first-order in the sense of (5.6) with  $q = 1$ ). Denote the computed solutions for a mesh size  $\Delta x$  as  $U^{\Delta x}$ , and the exact solution as  $U$ , both evaluated at some time  $t$ . Then it may be shown [7] that

$$\|U^{\Delta x} - U\|_{L^1(\mathbb{R})} \leq C \Delta x.$$

Thus a first-order scheme for a linear equation has a unit rate of convergence to the exact solution. This rate of convergence is demonstrated in the numerical experiments for the linear equation (2.2), presented earlier in this chapter.

However, a similar result does not hold for the nonlinear conservation law (5.1). The truncation error is only defined for smooth solutions, as the Taylor expansion is needed to obtain a truncation error estimate. It is well established by now that solutions of (5.1) are discontinuous and that the Taylor expansion is not valid for such solutions. Hence, the notion of truncation error is *purely formal* for the

nonlinear case. A famous result of Kuznetsov [1] shows that monotone schemes (see Chapter 4 for definitions) satisfy an estimate of the form

$$\|U^{\Delta x} - U\|_{L^1(\mathbb{R})} \leq C\Delta x^{1/2}.$$

Hence, formally first-order schemes like the Godunov, Lax-Friedrichs and related schemes may show a rate of convergence lower than unity. However, the rate of convergence is close to one for most numerical experiments as the ones presented at the beginning of this chapter.

**5.1.1. The Lax-Wendroff scheme.** In order to obtain better resolution of the approximate solutions of (5.1), we need to design numerical schemes that are better than first-order accurate, at least formally. The simplest recipe for obtaining a higher order scheme is to use Taylor expansions more effectively. Let  $U$  be a solution of (5.1) and assume that it is smooth. We then have

$$(5.8) \quad \begin{aligned} U_t &= -f(U)_x, \\ U_{tt} &= -(f(U)_x)_t = -(f(U)_t)_x = -(f'(U)U_t)_x = (f'(U)f(U)_x)_x. \end{aligned}$$

Expanding the exact solution in terms of Taylor expansions and using the identities (5.8), we obtain

$$\begin{aligned} U(x_j, t^{n+1}) &= U(x_j, t^n) + \Delta t U_t(x_j, t^n) + \frac{\Delta t^2}{2} U_{tt}(x_j, t^n) + \mathcal{O}(\Delta t^3) \\ &= U(x_j, t^n) - \Delta t f(U(x_j, t^n))_x \\ &\quad + \frac{\Delta t^2}{2} (f'(U(x_j, t^n))f(U(x_j, t^n))_x)_x + \mathcal{O}(\Delta t^3). \end{aligned}$$

We approximate the spatial derivatives with second-order accurate central differences as

$$\begin{aligned} f(U)_x &\approx \frac{f(U_{j+1}^n) - f(U_{j-1}^n)}{2\Delta x}, \\ (f'(U)f(U)_x)_x &\approx \frac{1}{\Delta x} \left( a_{j+1/2}^n \left( \frac{f(U_{j+1}^n) - f(U_j^n)}{\Delta x} \right) - a_{j-1/2}^n \left( \frac{f(U_j^n) - f(U_{j-1}^n)}{\Delta x} \right) \right), \end{aligned}$$

and obtain the *Lax-Wendroff scheme*

$$(5.9) \quad \begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{2\Delta x} (f(U_{j+1}^n) - f(U_{j-1}^n)) \\ &\quad + \frac{\Delta t^2}{2\Delta x^2} (a_{j+1/2}^n (f(U_{j+1}^n) - f(U_j^n)) - a_{j-1/2}^n (f(U_j^n) - f(U_{j-1}^n))). \end{aligned}$$

Here,

$$a_{j+1/2}^n = f' \left( \frac{U_j^n + U_{j+1}^n}{2} \right)$$

is an approximation to  $f'(U(x_{j+1/2}, t))$ . The Lax-Wendroff scheme can be written in the standard finite volume form (4.13) with the numerical flux function

$$(5.10) \quad F_{j+1/2}^n = F(U_j^n, U_{j+1}^n) = \frac{f(U_j^n) + f(U_{j+1}^n)}{2} - \frac{a_{j+1/2}^n \Delta t}{2\Delta x} (f(U_{j+1}^n) - f(U_j^n)).$$

**Exercise 5.3.** Show that the Lax-Wendroff scheme is second-order accurate in the sense of (5.6).

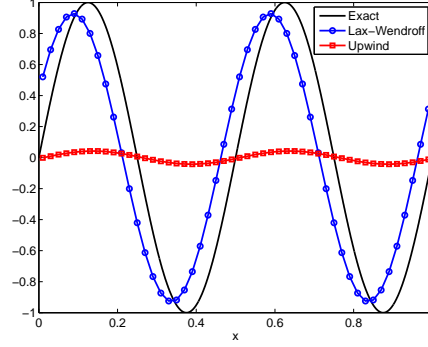


FIGURE 5.3. Linear advection equation computed to time  $t = 10$  on 50 mesh points, using the Lax-Wendroff and upwind schemes with smooth initial conditions (5.2) and periodic boundary conditions. [linAdv\_LW\_Upw\_sine.m]

No. of Cells	$\mathcal{E}^{\Delta x}$	EOC
20	132	—
40	67.3	0.967
80	18.4	1.87
160	4.64	1.98
320	1.16	2
640	0.291	2
1280	0.0727	2
2560	0.0182	2

TABLE 5.3. Error and order of convergence for the linear advection equation using the Lax-Wendroff scheme, with smooth initial data (5.2). [linAdv\_LW\_OOC.m]

**5.1.2. Numerical experiments.** We perform several numerical experiments with the Lax-Wendroff scheme. To begin with, we consider the advection equation (2.2) with initial data (5.2) and periodic boundary conditions. The numerical results are presented in Figure 5.3. For the sake of comparison, we plot the results obtained with the Godunov (upwind) scheme. The Lax-Wendroff scheme is clearly more accurate. Quantitative results in error Table 5.3 attest to the fact that Lax-Wendroff has a rate of convergence equal to 2, thus justifying its derivation as a second-order scheme.

In the next numerical experiment, the Lax-Wendroff scheme is used to compute solutions of the linear advection equation (2.2) with Riemann data (2.26). The results in Figure 5.4(a) show that the discontinuity is resolved much more sharply by the Lax-Wendroff scheme when compared to the Godunov scheme. However, the Lax-Wendroff scheme produces incorrect oscillations in the wake of the shock. These oscillations increase in both amplitude and frequency as the mesh is refined (see Figure 5.4(b)). Thus, increasing the order of accuracy may lead to stability issues, at least with the Lax-Wendroff scheme.

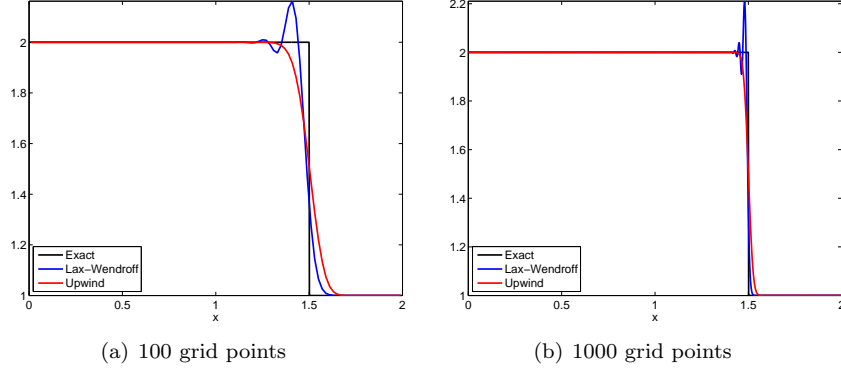


FIGURE 5.4. Linear advection equation computed to time  $t = 1$  with discontinuous initial conditions (2.26) and outflow boundary conditions. Comparison of the Lax-Wendroff and upwind schemes. [linAdv\_LW\_Upw\_disc.m]

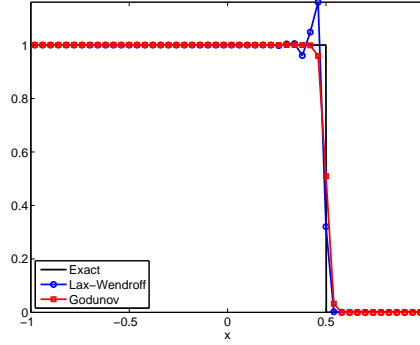


FIGURE 5.5. Burgers' equation computed to time  $t = 0.5$  with discontinuous initial conditions (4.16) and outflow boundary conditions on 50 cells. [burgers\_LW\_God\_disc.m]

As another numerical example, consider the nonlinear Burgers equation (3.3) with Riemann data (4.16). The results, shown in Figure 5.5, show that although the Lax-Wendroff scheme resolves the shock slightly better than the Godunov scheme, the solution is polluted with spurious oscillations. In conclusion, the Lax-Wendroff scheme fails to resolve discontinuous solutions of conservation laws in a stable, monotonous manner.

## 5.2. The REA algorithm

Given the instabilities that result from the Lax-Wendroff scheme, we need to devise a new procedure for obtaining high-order schemes. This forces us to revisit the original derivation of the Godunov scheme, presented in Chapter 4. It turns

out that the entire derivation of the Godunov scheme can be summarized in the following three steps:

**Reconstruction:** At time level  $t^n$ , assume that we know the approximate cell averages  $U_j^n$ . We realize this collection of cell average by a piecewise constant function

$$(5.11) \quad U(x, t^n) = U_j^n \quad \text{for } x_{j-1/2} < x < x_{j+1/2}.$$

**Evolution:** The reconstructed function  $U(x, t^n)$  is evolved in time using either an exact or approximate solution algorithm for the conservation law. This amounts to solving the superposition of Riemann problems (4.7), either exactly or approximately.

**Averaging:** Last, we average the solution at the next time step  $t^{n+1}$  over each control volume  $\mathcal{C}$ .

The three steps of reconstruction, evolution and averaging are branded together as the REA algorithm. Most of the finite volume schemes of the previous chapter can be obtained from the REA algorithm by a suitable approximation of the evolution step.

The main constraint of the REA algorithm so far is that the reconstruction step utilizes only piecewise constant functions (Figure 5.6). For smooth solutions of (5.1), piecewise constant functions are a first-order interpolation, thus resulting in the overall first-order accuracy of the Godunov-type schemes. One possible recipe for constructing high-order accurate schemes lies in employing high-order interpolations in the reconstruction step.

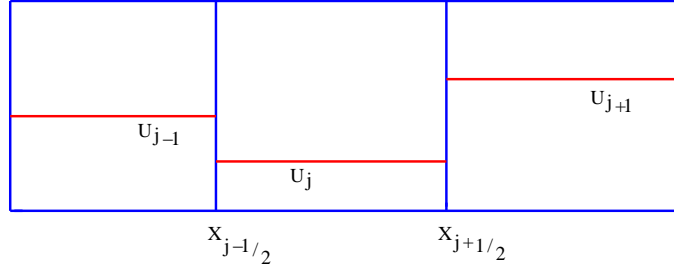


FIGURE 5.6. Representation of cell averages as piecewise constant functions in the REA algorithm.

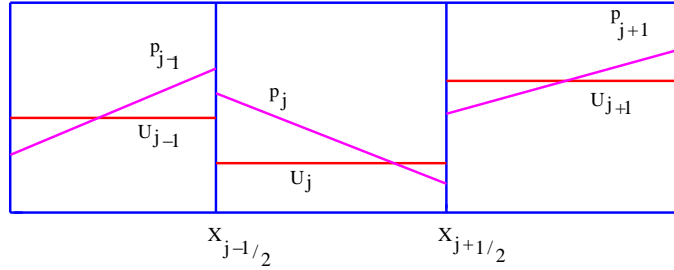


FIGURE 5.7. Piecewise linear functions in the REA algorithm.

**5.2.1. Second-order reconstruction.** The simplest high-order interpolation results from a piecewise linear reconstruction (Figure 5.7). Such a reconstruction will lead to a second-order accurate rendition of the smooth solutions of (5.1). Given the cell averages  $U_j^n$  at time  $t^n$ , there are several possibilities for reconstructing linear functions in each cell  $\mathcal{C}_j$ . We need to narrow down the alternatives by putting constraints on the reconstruction.

One of the key requirements for a stable and convergent approximation of conservation laws is that the scheme is conservative (4.39).

**Exercise 5.4.** *Show that the evolution and averaging steps are conservative by verifying that*

(a) *if  $U(x, t^{n+1})$  is the exact solution at time  $t^{n+1}$ , then*

$$\int_{\mathbb{R}} U(x, t^{n+1}) dx = \int_{\mathbb{R}} U(x, t^n) dx,$$

(b) *if  $U_j^{n+1}$  is the average of  $U(x, t^{n+1})$  in cell  $j$ , then*

$$\Delta x \sum_j U_j^{n+1} = \int_{\mathbb{R}} U(x, t^{n+1}) dx.$$

It is natural to demand that the reconstruction step is also conservative. Denoting the piecewise linear function in the cell  $\mathcal{C}_j$  as  $p_j(x)$ , we require that

$$(5.12) \quad \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} p_j(x) dx = U_j^n.$$

It is readily seen that  $p_j$  must have the form

$$(5.13) \quad p_j(x) = U_j^n + \sigma_j^n(x - x_j),$$

where  $\sigma_j^n$  is a parameter that determines the slope in cell  $\mathcal{C}_j$ . The *local* linear functions  $p_j$  are combined to define the *global* piecewise linear function

$$(5.14) \quad p(x) = p_j(x) \quad \text{for } x_{j-1/2} < x < x_{j+1/2}.$$

This piecewise linear function can be used with the REA algorithm to obtain a higher order scheme.

**5.2.2. Choices of slope.** The slope  $\sigma$  in (5.13) can be determined in a variety of ways. Three of them are

- Central:

$$(5.15) \quad \sigma_j^n = \frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x}.$$

- Backward:

$$(5.16) \quad \sigma_j^n = \frac{U_j^n - U_{j-1}^n}{\Delta x}.$$

- Forward:

$$(5.17) \quad \sigma_j^n = \frac{U_{j+1}^n - U_j^n}{\Delta x}.$$

It is instructive to see what scheme the REA algorithm with the above choices of slope results in. For simplicity of the exposition, we consider the linear advection equation (2.2) with a positive velocity  $a > 0$ .

**Exercise 5.5.** Let (5.14) be the prescribed initial data for the linear advection equation (2.2) at time  $t^n$ . Show that the cell average  $U_j^{n+1}$  of the exact solution at time  $t^{n+1}$  can be written as

$$(5.18) \quad \begin{aligned} U_j^{n+1} &= \frac{a\Delta t}{\Delta x} \left( U_{j-1}^n + \frac{1}{2}(\Delta x - a\Delta t)\sigma_{j-1}^n \right) + \left( 1 - \frac{a\Delta t}{\Delta x} \right) \left( U_j^n - \frac{1}{2}a\Delta t\sigma_j^n \right) \\ &= U_j^n - \frac{a\Delta t}{\Delta x}(U_j^n - U_{j-1}^n) - \frac{1}{2} \frac{a\Delta t}{\Delta x}(\Delta x - a\Delta t)(\sigma_j^n - \sigma_{j-1}^n). \end{aligned}$$

Choosing  $\sigma$  as the downwind slope (5.17), we obtain the following explicit formula for (5.18):

$$(5.19) \quad U_j^{n+1} = U_j^n - \frac{a\Delta t}{2\Delta x}(U_{j+1}^n - U_{j-1}^n) + \frac{a^2\Delta t^2}{2\Delta x^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

This is precisely the Lax-Wendroff scheme (5.9) for the linear advection equation. This conclusion also justifies the fact that a second-order piecewise linear reconstruction leads to an overall second-order accurate scheme like the Lax-Wendroff scheme. However, the Lax-Wendroff scheme produces oscillations, and so the choice of a downwind slope is not desirable.

Next, we consider the upwind slope (5.16). The explicit formula (5.18) then becomes

$$(5.20) \quad U_j^{n+1} = U_j^n - \frac{a\Delta t}{2\Delta x}(3U_j^n - 4U_{j-1}^n + U_{j-2}^n) + \frac{a^2\Delta t^2}{2\Delta x^2}(U_j^n - 2U_{j-1}^n + U_{j-2}^n).$$

This four point scheme is called the Beam-Warming scheme. An approximate

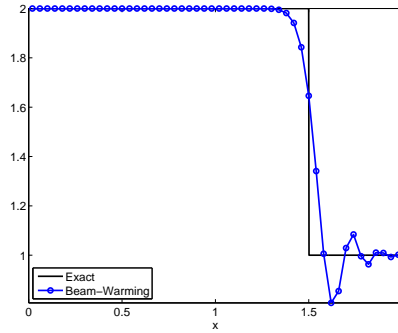


FIGURE 5.8. Linear advection equation computed to time  $t = 1$  with discontinuous initial conditions (2.26) and outflow boundary conditions on 50 cells using the Beam-Warming scheme. `linAdv_beamWarm_disc.m`

solution of the linear advection equation with Riemann data (2.26) computed with the Beam-Warming scheme is shown in Figure 5.8. The results show that the accuracy of the scheme at the shock is comparable to the Lax-Wendroff scheme. However, it also leads to oscillations. Similarly, using the central slope (5.15) leads to oscillatory approximations.



**5.2.3. The source of oscillations.** The exact solution for the linear transport equation (2.2) as well as the nonlinear conservation law (5.1) are TVD and consequently nonoscillatory whenever the initial data is nonoscillatory. It is reasonable to demand (see Chapter 4.4) that stable numerical schemes respect this property. The numerical results presented above demonstrate that second-order schemes like the Lax-Wendroff (5.9) and the Beam-Warming schemes (5.20) violate the requirement that the approximate solutions be TVD. What goes wrong and why do these schemes violate this requirement?

A closer look at the REA algorithm reveals that the evolution step is TVD whenever an exact solution of the Riemann problem is used. Furthermore, suitable approximate Riemann solvers like the Lax-Friedrichs solver are also TVD. It is straightforward to check that the averaging operator respects the TVD property.

The only explanation for the fact that the Lax-Wendroff and Beam-Warming schemes are not TVD is that the second-order reconstruction (5.13) violates the TVD requirement. To illustrate this point, we consider the following simple situation: Let  $J \in \mathbb{Z}$  be fixed and consider cell averages at time level  $t^n$

$$(5.21) \quad U_j^n = \begin{cases} 1 & \text{if } j \leq J \\ 0 & \text{if } j > J. \end{cases}$$

The initial total variation is 1. Computing the downwind slope (5.17), we find that  $\sigma_j^n = 0$  for  $j \neq J$  and  $\sigma_J^n = -0.5$ . Therefore, the piecewise linear reconstruction (5.17) is active only in the cell  $J$  (see Figure 5.9). Furthermore, the reconstructed function has an overshoot with a maximum value of 1.5. The resulting total variation is 1.5, which is greater than the total variation of the initial cell averages. This induction of oscillations is independent of  $\Delta x$  and can lead to much larger oscillations as the solutions are evolved in time. Similar results hold for the upwind and central slopes.

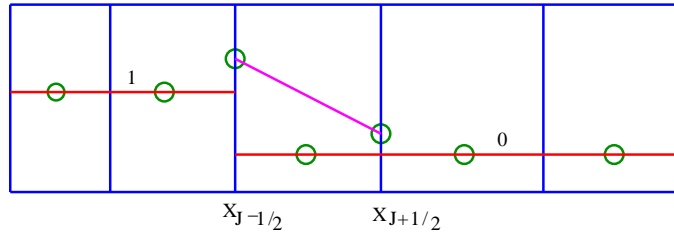


FIGURE 5.9. The Lax-Wendroff reconstruction for (5.21).

Given the nature of the continuous solution, it is reasonable to expect that the reconstruction (5.13) is TVD, i.e.,

$$(5.22) \quad \|p\|_{BV} \leq \|U^{\Delta x}\|_{BV},$$

where  $p$  is the piecewise linear function (5.14) and  $U^{\Delta x}$  is the piecewise constant function (5.11).

### 5.3. The minmod limiter

Straightforward choices of slope like the central (5.15), upwind (5.16) and downwind (5.17) slopes do not satisfy the TVD requirement (5.22). The problem is manifested near discontinuities as shown in Figure 5.9.

A clever choice of the slope in (5.13) that satisfies the TVD property (5.22) is the so-called *minmod* limiter, which is given by

$$(5.23) \quad \sigma_j^n = \text{minmod} \left( \frac{U_{j+1}^n - U_j^n}{\Delta x}, \frac{U_j^n - U_{j-1}^n}{\Delta x} \right).$$

The minmod function is defined as

$$(5.24) \quad \text{minmod}(a_1, \dots, a_n) = \begin{cases} \text{sign}(a_1) \min_{1 \leq k \leq n} (|a_k|) & \text{if } \text{sign}(a_1) = \dots = \text{sign}(a_n), \\ 0 & \text{otherwise.} \end{cases}$$

The minmod limiter compares the upwind slope and the downwind slope and checks if they are of the same sign. If so, it selects the smallest one, and if not, it sets the slope to zero. Thus, the reconstruction is limited based on local gradients.

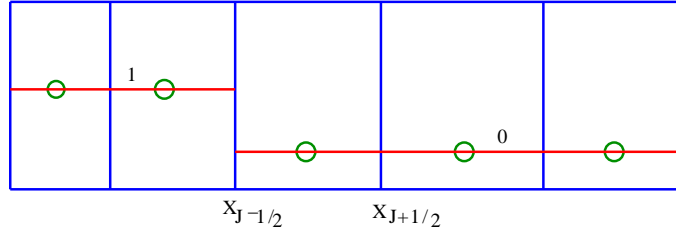


FIGURE 5.10. The minmod reconstruction for (5.21).

To illustrate the effect of the minmod limiter, we take the same example as before and consider initial data (5.21). Using the minmod slope (5.23) in (5.13) results in retaining the initial data (5.21) (see Figure 5.10). Thus, at least in this case, the minmod reconstruction is TVD.

Next, we consider cell averages

$$(5.25) \quad U_j^n = \begin{cases} 1 & \text{if } j \leq J-1 \\ 3/4 & \text{if } j = J \\ 1/4 & \text{if } j = J+1 \\ 0 & \text{if } j \geq J+2. \end{cases}$$

The reconstructed piecewise linear function (5.13) with minmod slopes (5.23) is depicted in Figure 5.11. The reconstruction is clearly TVD.

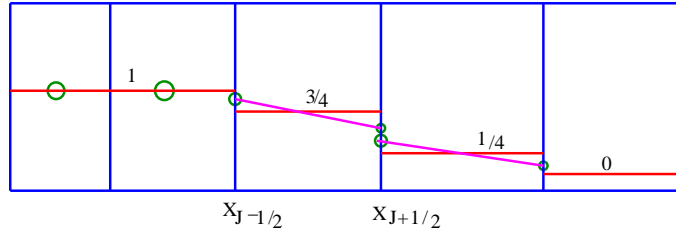


FIGURE 5.11. The minmod reconstruction for (5.25).

**Exercise 5.6.** Show that the minmod reconstruction  $p(x)$  of cell averages  $U_j$  is TVD:  $\|p\|_{\text{TV}} \leq \sum_j |U_{j+1} - U_j|$ .

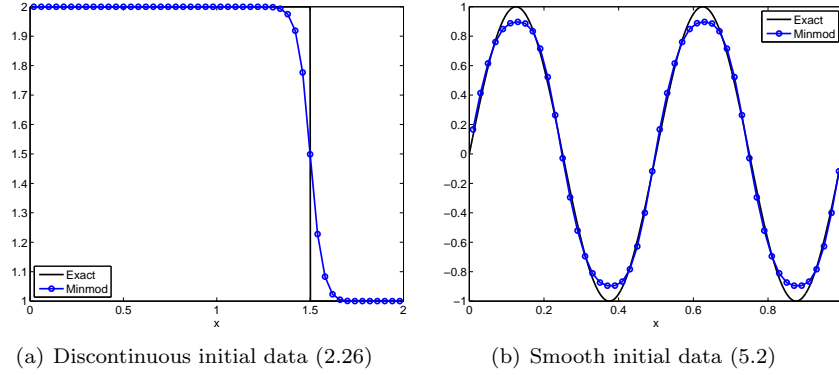


FIGURE 5.12. Linear advection equation computed to time  $t = 1$  50 cells using the minmod limiter. [linAdv\_reconstr.m]

No. of cells	Minmod		MC		Superbee	
	Relative error	EOC	Relative error	EOC	Relative error	EOC
20	94.6	—	68.4	—	49.1	—
40	54.4	0.798	14.7	2.22	12.4	1.98
80	17	1.68	4.93	1.57	3.97	1.64
160	7.7	1.14	1.51	1.71	3.64	0.127
320	2.14	1.85	0.407	1.89	1.44	1.34
640	0.613	1.8	0.104	1.97	0.416	1.79
1280	0.168	1.87	0.0262	1.99	0.112	1.89
2560	0.0449	1.9	0.00657	1.99	0.0292	1.94

TABLE 5.4. Error and order of convergence for the linear advection equation with smooth initial data (5.2). [linAdv\_reconstr\_EOC.m]

**5.3.1. Numerical experiments.** We test the linear advection equation (2.2) with the second-order scheme (5.18) and a minmod slope limiter (5.23). To begin with, we use the initial data (2.26) and see that the resulting scheme is not oscillatory (see Figure 5.12(a)). The scheme improves the accuracy considerably when compared to the first-order scheme, and captures the discontinuity quite sharply.

Next, we test the minmod scheme with smooth initial data (5.2) and show the resulting profiles in Figure 5.12(b). The minmod scheme resolves the solution quite well. Error table 5.4 shows that the scheme is close to second-order accurate. Numerical experiments with nonlinear examples will be presented in the sequel.

#### 5.4. Other limiters

We have demonstrated that the minmod limiter (5.23) improves the quality of the solution considerably. However, other choices of slope limiters exist and might result in better approximations of the underlying PDE. A popular choice is the *superbee* limiter, due to Roe, which has the expression

$$(5.26) \quad \sigma_j^n = \maxmod(\sigma_j^l, \sigma_j^r),$$

where

$$\begin{aligned} \sigma_j^l &= \minmod\left(2\frac{U_j^n - U_{j-1}^n}{\Delta x}, \frac{U_{j+1}^n - U_j^n}{\Delta x}\right), \\ \sigma_j^r &= \minmod\left(\frac{U_j^n - U_{j-1}^n}{\Delta x}, 2\frac{U_{j+1}^n - U_j^n}{\Delta x}\right), \end{aligned}$$

with the maxmod function simply replacing the minimum in (5.24) with a maximum.

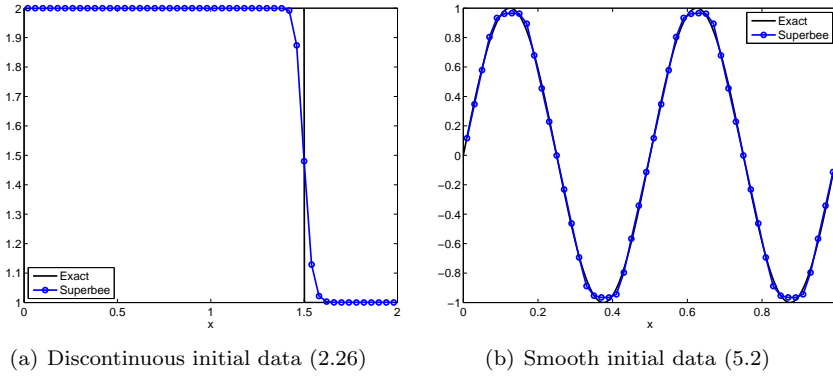


FIGURE 5.13. Linear advection equation computed to time  $t = 1$  50 cells using the superbee limiter. [`linAdv_reconstr.m`]

Using the superbee limiter with the initial data (5.25) results in a piecewise linear function shown in Figure 5.14. The figure indicates that the superbee limiter results in a TVD reconstruction. Furthermore, comparing the superbee limiter with the minmod limiter in 5.11 shows that the superbee limiter results in steeper slopes, while still being TVD. Employing the superbee limiter in the scheme (5.18) to compute approximate solutions for the linear advection equation (2.2), we present results for initial data (2.26) and (5.2) in Figure 5.13 (a) and (b), respectively. The results show that the superbee limiter gives accurate and nonoscillatory approximations. The superbee limiter is more accurate than the minmod limiter (particularly at the extrema). However, it has a tendency to square off extrema, as shown in Figure 5.13(b). Error table 5.4 shows that second-order accuracy is obtained.

Another popular limiter is the MC (monotonized central) limiter of Van-Leer,

$$(5.27) \quad \sigma_j^n = \minmod\left(2\frac{U_{j+1}^n - U_j^n}{\Delta x}, \frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x}, 2\frac{U_j^n - U_{j-1}^n}{\Delta x}\right).$$

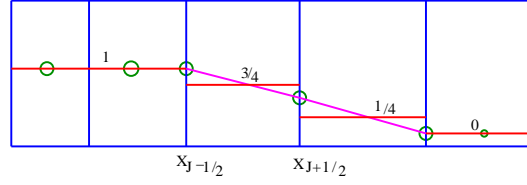


FIGURE 5.14. The superbee reconstruction for (5.25).

Numerical results (see Figure 5.15 and Table 5.4) show that the MC limiter is as accurate as the superbee limiter. It is usually the preferred slope limiter.

**5.4.1. Second-order schemes in the flux form.** So far, we have written the second-order schemes for the linear advection equation (2.2) in the update form (5.18). They can easily be recast into the finite volume flux form (4.13). Assume for the moment that the advection speed  $a > 0$ . Recall that the interface flux  $F_{i+1/2}$  was calculated as the interface integral (4.12). Computing this integral with the exact solution (2.5) and the second-order reconstruction (5.13), we obtain

$$\begin{aligned}
 F_{j+1/2}^n &= \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(U(x_{j+1/2}, t)) dt \\
 &= \frac{a}{\Delta t} \int_{t^n}^{t^{n+1}} U(x_{j+1/2}, t) dt && \text{(as } f(U) = aU \text{)} \\
 &= \frac{a}{\Delta t} \int_{t^n}^{t^{n+1}} p_j(x_{j+1/2} - a(t - t^n)) dt && \text{(from (2.5))} \\
 &= \frac{a}{\Delta t} \int_{t^n}^{t^{n+1}} U_j^n + \sigma_j^n (x_{j+1/2} - a(t - t^n) - x_j) dt && \text{(from (5.13))} \\
 &= aU_j^n + \frac{a}{2\Delta t} \int_{t^n}^{t^{n+1}} \sigma_j^n (\Delta x - 2a(t - t^n)) dt
 \end{aligned}$$

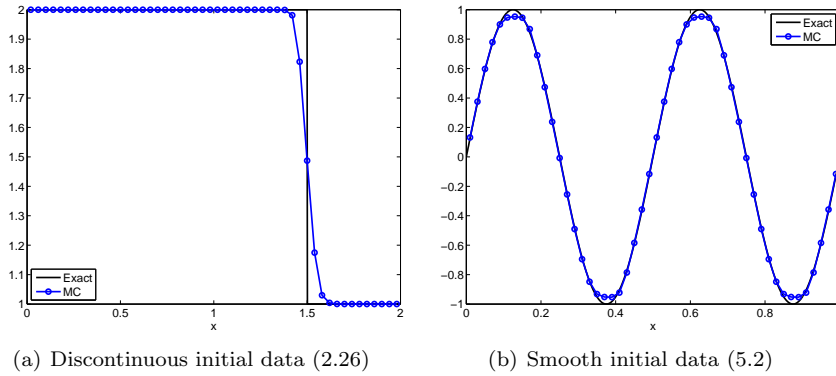


FIGURE 5.15. Linear advection equation computed to time  $t = 1$  50 cells using the MC limiter. [linAdv\_reconstr.m]

$$= aU_j^n + \frac{a}{2}(\Delta x - a\Delta t)\sigma_j^n.$$

Consequently, the second-order scheme (5.18) can be written in the finite volume form (4.13) with flux

$$(5.28) \quad F_{j+1/2}^n = aU_j^n + \frac{a}{2}(\Delta x - a\Delta t)\sigma_j^n.$$

A similar expression can be derived when the advection velocity  $a < 0$ . Defining  $\delta_{j+1/2}^n = \Delta x \sigma_j^n$ , we can rewrite (5.28) as

$$(5.29) \quad F_{j+1/2}^n = aU_j^n + \frac{a}{2} \left( 1 - \frac{a\Delta t}{\Delta x} \right) \delta_{j+1/2}^n.$$

Denote the jump of the solution at the interface  $x_{j+1/2}$  as

$$\llbracket U^n \rrbracket_{j+1/2} = U_{j+1}^n - U_j^n$$

and define

$$\theta_{j+1/2}^n = \frac{\llbracket U^n \rrbracket_{j-1/2}}{\llbracket U^n \rrbracket_{j+1/2}}.$$

The parameter  $\delta$  is an indication of the change in  $U$  around  $x_{j+1/2}$ . Therefore, we rewrite it in terms of cell interface jumps,

$$\delta_{j+1/2}^n = \phi(\theta_{j+1/2}^n) \llbracket U^n \rrbracket_{j+1/2}$$

for some function  $\phi(\theta)$ . The advantage of writing  $\delta$  in this *limited* form is immediate. If  $\phi \equiv 1$ , then the flux (5.29) immediately reduces to the Lax-Wendroff flux (5.9) for the linear advection equation. Hence, the second-order scheme can be thought as a version of the Lax-Wendroff scheme with the flux being limited by the limiter function  $\phi$ . In conclusion, a limiter on the slope of the linear function (5.13) can be written as a *flux limiter*.

### 5.5. Flux limiters and the TVD property.

We can realize slope limiters as flux limiters. Table 5.5 gives the flux limiter form of the slope limiters we have studied so far.

**Exercise 5.7.** For each flux limiter in Table 5.5, show that the resulting scheme is the same as when using the corresponding slope limiter in (5.18).

Method	Flux limiter function $\phi(\theta)$
Upwind	0
Lax-Wendroff	1
Beam-Warming	$\theta$
Minmod	$\min\text{mod}(1, \theta)$
Superbee	$\max(0, \min(1, 2\theta), \min(2, \theta))$
MC	$\max(0, \min(\frac{1+\theta}{2}, 2, 2\theta))$

TABLE 5.5. The most popular flux limiters.

Observe that the upwind, Lax-Wendroff and Beam-Warming flux limiters are linear (or linear affine), whereas the minmod, superbee and MC limiters are non-linear. It is essential to consider nonlinear limiters to obtain TVD second-order accurate schemes, as we will see in a moment.

We can rewrite the second-order scheme (5.18) in the incremental form

$$(5.30) \quad U_j^{n+1} = U_j^n + C_{j+1/2}^n \llbracket U^n \rrbracket_{j+1/2} - D_{j-1/2}^n \llbracket U^n \rrbracket_{j-1/2}$$

(consult (4.53)) with coefficients

$$\begin{aligned} C_{j+1/2}^n &= -\frac{1}{2}\lambda(1-\lambda)\phi\left(\theta_{j+1/2}^n\right), \\ D_{j-1/2}^n &= \lambda - \frac{1}{2}\lambda(1-\lambda)\phi\left(\theta_{j-1/2}^n\right), \end{aligned}$$

where  $\lambda = \frac{a\Delta t}{\Delta x}$ . However, realizing that  $\llbracket U^n \rrbracket_{j+1/2} = \frac{\llbracket U^n \rrbracket_{j-1/2}}{\theta_{j+1/2}^n}$ , we rewrite the coefficients in (5.30) in the more revealing form

$$(5.31) \quad \begin{aligned} C_{j+1/2}^n &= 0, \\ D_{j-1/2}^n &= \lambda + \frac{1}{2}\lambda(1-\lambda)\left(\frac{\phi(\theta_{j+1/2}^n)}{\theta_{j+1/2}^n} - \phi(\theta_{j-1/2}^n)\right). \end{aligned}$$

By Harten's Lemma 4.12, the scheme will be TVD if the coefficients  $C$  and  $D$  in (5.31) satisfy the criteria (4.55). Hence, we have to ensure that

$$(5.32) \quad 0 \leq D_{j-1/2}^n \leq 1.$$

We recall that  $\lambda \leq 1$  due to the CFL condition (2.18). Hence, a sufficient condition for (5.32) to hold is given by

$$(5.33) \quad \left| \frac{\phi(\theta_1)}{\theta_1} - \phi(\theta_2) \right| \leq 2 \quad \text{for } \theta_1, \theta_2 \in \mathbb{R}.$$

The set of solutions to (5.33) is rich. A particular class of solutions is given by

$$(5.34) \quad 0 \leq \phi(\theta) \leq \min\text{mod}(2, 2\theta) \quad \text{for } \theta \in \mathbb{R}.$$

It is straightforward to verify that (5.34) satisfies (5.33). The region defined by (5.34) is shown in Figure 5.16. Note that (5.34) requires that  $\phi \equiv 0$  whenever  $\theta < 0$ . This occurs precisely when there is a local extrema at  $x_j$ .

**Exercise 5.8.** *Show that any flux limiter that satisfies (5.34) reduces to first-order accuracy at extrema. In particular, show that when  $U_j^n$  has a local maximum or minimum in cell  $j$ , then  $p_j(x) \equiv U_j^n$ .*

**Exercise 5.9.** *Verify that the Lax-Wendroff method and the Beam-Warming method do not satisfy the TVD requirement (5.34), whereas the minmod, superbee and MC limiters do satisfy (5.34).*

## 5.6. High-resolution methods for nonlinear problems.

The limiter-based methods described above can be directly used for approximating the nonlinear equation (5.1), as the REA algorithm applies to any conservation law. However, it may be difficult to obtain explicit formulas like (5.28) in this case, as the flux integral (4.12) may not be possible to evaluate explicitly. Furthermore, an analogue of the exact solution (2.5) is not available in the nonlinear case, making an evaluation of (4.12) very complicated. Therefore, we use a related but slightly different approach to yield high-resolution methods for nonlinear equations.

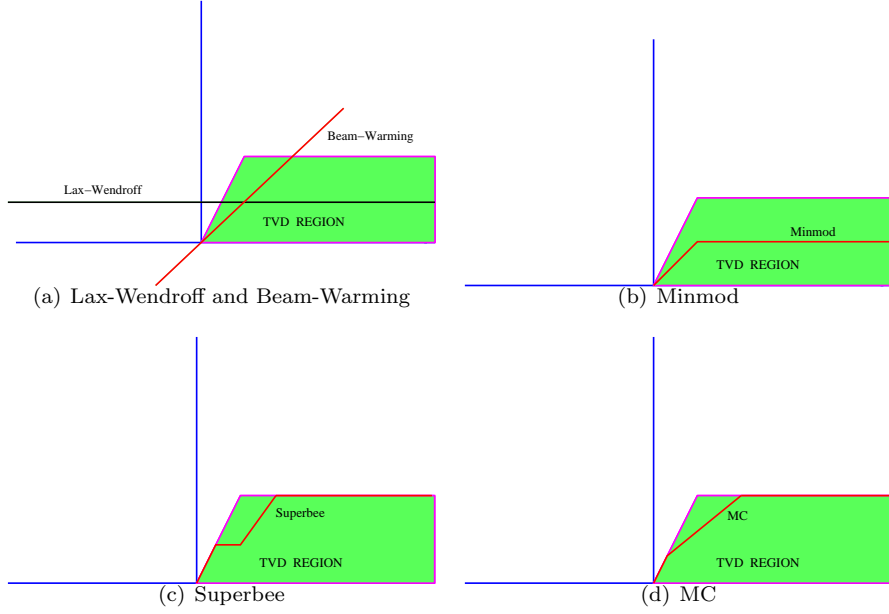


FIGURE 5.16. Sweby diagrams for different limiters.

**5.6.1. Semi-discrete formulation.** The starting point for obtaining high-resolution methods for the nonlinear conservation law (5.1) is still the reconstruction procedure, as described previously. Instead of employing the REA algorithm directly, we denote the cell average over  $\mathcal{C}_j$  as

$$U_j(t) = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} U(x, t) dx$$

and integrate the conservation law (5.1) over space to obtain

$$\begin{aligned} & \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} U_t + f(U)_x dx = 0 \\ \Rightarrow & \frac{d}{dt} U_j(t) + \frac{1}{\Delta x} \left( f(U(x_{j+1/2}^-, t)) - f(U(x_{j-1/2}^+, t)) \right) = 0. \end{aligned}$$

Here,  $U(x_{j+1/2}^-, t)$  denotes the left-limit of  $U$  at  $x_{j+1/2}$ ,

$$U(x_{j+1/2}^-, t) = \lim_{x \rightarrow x_{j+1/2}^-} U(x, t).$$

Letting  $F$  be an approximation of the above flux terms,

$$(5.35) \quad F_{j+1/2}^\pm(t) \approx f(U(x_{j+1/2}^\pm, t)),$$

we obtain the semi-discrete form of the finite volume scheme (4.13) as

$$(5.36) \quad \frac{d}{dt} U_j(t) + \frac{1}{\Delta x} \left( F_{j+1/2}^-(t) - F_{j-1/2}^+(t) \right) = 0.$$

The scheme (5.36) leads to a system of ODEs that must be integrated in time by a suitable time integration routine. Dropping the  $t$  in (5.43) for notational



convenience, we can use the following numerical flux function:

$$(5.37) \quad F_{j+1/2}^- = F_{j+1/2}^+ = F(U_j, U_{j+1}),$$

where  $F$  is an consistent two-point flux function like the Godunov (4.14), Rusanov (4.31) or Engquist-Osher (4.32) flux. If a standard forward Euler method is used to integrate the ODE system (5.36) in time, we obtain the standard first-order monotone finite volume scheme (4.13).

### 5.7. Second-order semi-discrete schemes.

The advantage of the semi-discrete formulation (5.36) lies in the fact that we can separately increase the order of spatial and temporal accuracy. The standard first-order scheme uses piecewise constants in space and a forward Euler routine for time integration. Based on our previous discussion, we know that one approach to increase the order of accuracy is to employ high-order reconstructions instead of the piecewise constant cell averages

$$(5.38) \quad U(x, t) = U_j(t) \quad \text{for } x_{j-1/2} \leq x < x_{j+1/2}.$$

The process of reconstructing suitable nonoscillatory (TVD) piecewise linear functions from cell averages was described in detail in the previous section. Given  $U_j$ , we can obtain the following linear function in the cell  $C_j$  (again dropping  $t$  for notational convenience):

$$(5.39) \quad p_j(x) = U_j + \sigma_j(x - x_j).$$

These linear functions are combined to form the piecewise linear function

$$(5.40) \quad p(x, t) = p_j(x) \quad \text{for } x_{j-1/2} \leq x < x_{j+1/2}.$$

In order to ensure that the reconstruction is TVD, i.e.,

$$\|p\|_{BV} \leq \|U^{\Delta x}\|_{BV} \quad \text{for all } t,$$

we need to choose the slope  $\sigma$  in (5.39) suitably. From our previous discussion, we have at least three choices that satisfy the above condition: the minmod limiter (5.23), the superbee limiter (5.26) and the MC limiter (5.27).

Last, we denote the reconstructed values at the cell interfaces as

$$(5.41) \quad U_j^+ = p_j(x_{j+1/2}), \quad U_j^- = p_j(x_{j-1/2}).$$

See Figure 5.17 for an illustration.

**5.7.1. The numerical flux.** The numerical flux in (5.36) is an approximation of the interface flux. If the data are represented as piecewise constant cell averages, then the two-point flux (5.37) suffices to define the numerical flux. Since we are representing the approximations as piecewise linear functions, we need to replace the cell averages in (5.37) with the relevant edge values (see Figure 5.17),

$$(5.42) \quad F_{j+1/2} = F(U_j^+, U_{j+1}^-),$$

where  $F$  is any consistent numerical flux. Since we require that the evolution be TVD, we will use monotone fluxes like the Godunov (4.14), Rusanov (4.31) and Engquist-Osher (4.32) flux. This completes the description of the second-order semi-discrete scheme (5.36).

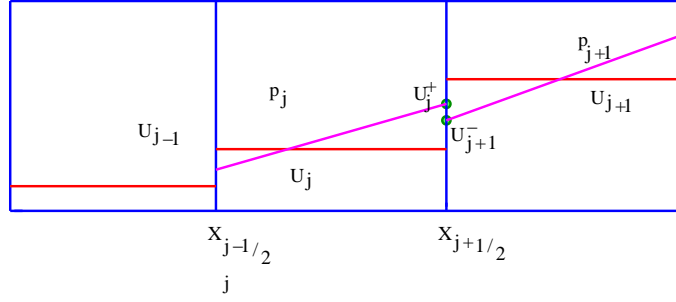


FIGURE 5.17. Second-order nonoscillatory reconstruction.

### 5.8. Time stepping

Denoting  $U(t)$  as the vector

$$U(t) = [\dots, U_{j-1}(t), U_j(t), U_{j+1}(t), \dots]$$

the finite volume scheme (5.36) can be rewritten as

$$(5.43) \quad \frac{d}{dt} U(t) = \mathcal{L}(U(t)).$$

Here, the operator  $\mathcal{L}$  acts pointwise on the vector  $U$  as

$$\mathcal{L}(U(t))_j := -\frac{1}{\Delta x} (F_{j+1/2}(t) - F_{j-1/2}(t)).$$

Thus, (5.43) is a system of ODEs that must be integrated in time. The simplest time integration routine is the forward Euler time integration,

$$(5.44) \quad U^{n+1} = U^n + \Delta t \mathcal{L}(U^n),$$

where  $U^n = U(t^n)$  is the vector of cell averages  $U$  at time  $t^n$ . The time step  $\Delta t$  should be determined by a suitable CFL condition like (4.9). By choosing a TVD reconstruction in (5.39) and a monotone flux in (5.42), we can ensure that the solution update in (5.44) is TVD, i.e.,

$$(5.45) \quad TV(U^{n+1}) \leq TV(U^n) \quad \text{for all } n.$$

However, the forward Euler method is only first-order accurate. Despite the second-order accuracy of the piecewise linear reconstruction (5.39), the first-order temporal accuracy leads to an overall first-order accuracy and negates the very purpose of using a piecewise linear reconstruction.

**5.8.1. Standard Runge-Kutta method.** The alternative is to employ high-order time stepping methods. These methods have been developed to a considerable extent; see standard textbooks like [4]. The standard high-order ODE methods are of the Runge-Kutta type. The well-established standard second-order Runge-Kutta method is of the form

$$(5.46) \quad U^{n+1} = U^n + \Delta t \mathcal{L} \left( U^n + \frac{\Delta t}{2} \mathcal{L}(U^n) \right).$$

This two-stage method is second-order accurate. However it fails to satisfy the TVD requirement (5.45) and may lead to oscillatory solutions [3]. Wishing to avoid such spurious oscillations, we need to search for Runge-Kutta methods that

preserve the TVD property. Such methods are termed *strong stability preserving* (SSP) Runge-Kutta methods. One second-order SSP Runge-Kutta method is

$$\begin{aligned}
 U^* &= U^n + \Delta t \mathcal{L}(U^n) \\
 U^{**} &= U^* + \Delta t \mathcal{L}(U^*) \\
 U^{n+1} &= \frac{1}{2}(U^n + U^{**})
 \end{aligned}
 \tag{5.47}$$

We have the following stability lemma for the second-order SSP Runge-Kutta method.

**Lemma 5.10.** *If the discrete differential operator  $\mathcal{L}$  is such that the forward Euler method (5.44) is TVD, then the Runge-Kutta method (5.47) is TVD.*

PROOF. We wish to prove the TVD property (5.45) for the Runge-Kutta method (5.47). Note that this method is a combination of two forward Euler stages. Therefore,

$$TV(U^*) \leq TV(U^n) \quad \text{and} \quad TV(U^{**}) \leq TV(U^*) \leq TV(U^n).$$

Thus,

$$TV(U^{n+1}) = TV\left(\frac{1}{2}(U^n + U^{**})\right) \leq \frac{1}{2}(TV(U^n) + TV(U^{**})) \leq TV(U^n).$$

□

### 5.9. High-resolution algorithm

With all the ingredients in place, we can state the algorithm for computing with a second-order scheme. Given cell averages  $U_j^n$  at time level  $t^n$ , we need to perform the following steps:

- Step 1 (Reconstruction): Given  $U_j$ , reconstruct the averages to obtain the piecewise linear function (5.39). Any nonoscillatory slope limiter like the minmod (5.23), superbee (5.26) or the MC (5.27) limiter can be used. Note that we only require the edge values  $U_j^\pm$  (5.41) in each cell.
- Step 2 (Flux evaluation): Given the edge values  $U_j^\pm$  in each cell, we plug these values into the numerical flux (5.42). In particular, monotone two-point fluxes like the Godunov (4.14), Engquist-Osher (4.32) and Rusanov (4.31) fluxes should be used.
- Step 3 (Time stepping): For second-order schemes, we use the second-order SSP Runge-Kutta method (5.47). As this method consists of two stages, steps 1 and 2 must be applied to each stage (e.g.,  $U^n$  and  $U^*$ ).

The time step  $\Delta t$  in (5.47) is determined by a CFL condition of the form (4.9). The second-order high resolution schemes are TVD as all the three ingredients are constructed to ensure this property.

The stencil of a second-order scheme consists of five points. This should be contrasted with the three point first-order schemes.

### 5.10. Numerical experiments

In this section, we present numerical experiments with both the linear transport equation (2.2) and Burgers' equation (3.3). To begin with, we consider the advection equation (2.2) with initial data (5.2) and periodic boundary conditions.

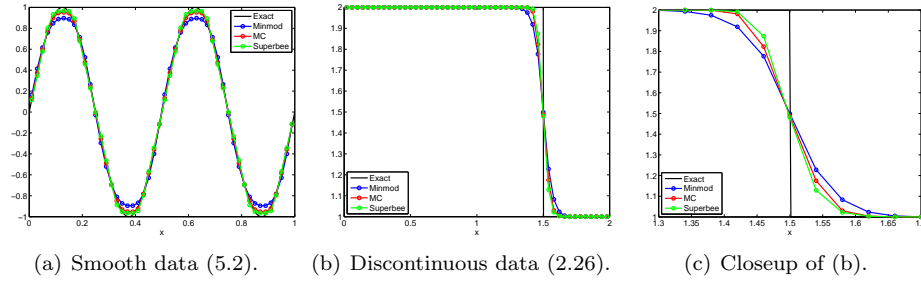


FIGURE 5.18. Linear advection equation (2.2) on a mesh of 50 cells, using various limiters.

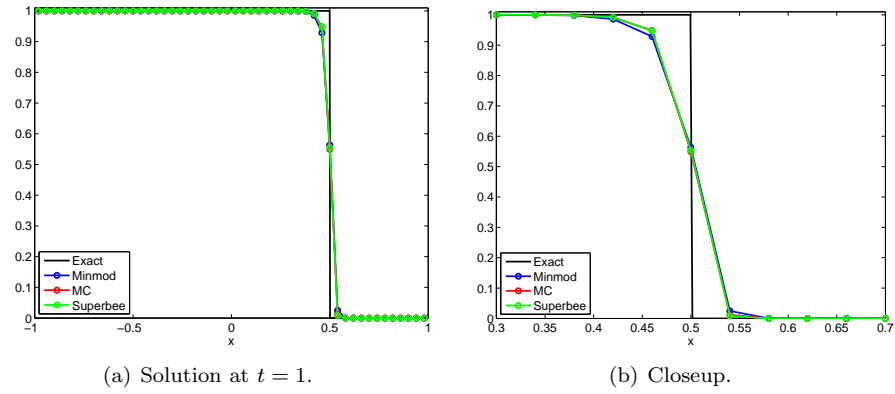


FIGURE 5.19. The Godunov scheme with different slope limiters on the problem (4.16). [burgers\_reconstr.m]

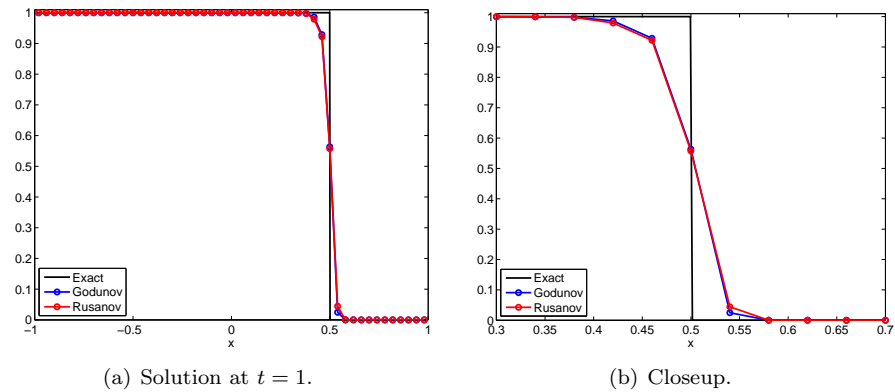


FIGURE 5.20. Comparison between the Godunov and Rusanov schemes with the minmod limiter on the problem (4.16). [burgers\_reconstr.m]

No. of cells	Godunov		Rusanov	
	Relative error	EOC	Relative error	EOC
20	2.86	–	3.51	–
40	1.48	0.951	1.66	1.08
80	0.74	0.998	0.839	0.988
160	0.37	1	0.42	0.997
320	0.185	1	0.21	1
640	0.0925	1	0.105	1
1280	0.0463	1	0.0525	1
2560	0.0231	1	0.0263	1

TABLE 5.6. Error and order of convergence for Burgers' equation computed to time  $t = 1$  with discontinuous initial conditions (4.16), using the minmod limiter. [burgers\_reconstr\_00C.m]

The second-order high-resolution scheme with all three choices of slope limiters are computed and the results are displayed in Figure 5.18(a). The results are very similar to those obtained with the second-order Lax-Wendroff scheme (5.9) (see Figure 5.3). This is to be expected, as the schemes basically only differ in their time integration routines. Analogous results hold for the discontinuous initial data (2.26) and are shown in Figure 5.18(b).

Next, we consider Burgers' equation with initial data (4.16) using the Godunov scheme. The results obtained with the high-resolution schemes are shown in Figures 5.19 and 5.20 – clearly there are very minor differences between the three choices of slope. The gain due to second-order accuracy in resolving the shock is not as much as the gain seen in the linear case.

In Figure 5.20, we display results comparing the Godunov flux and the Rusanov flux (4.31) using the minmod limiter. While the differences in results rendered by these fluxes is pronounced at first-order, the differences at second-order are considerably smaller. Error table 5.6 shows that the rate of convergence is exactly 1. This is to be expected, as the solution is not smooth, and so the truncation error is not well defined in this case.

Last, we consider the smooth initial value problem

$$U(x, 0) = \frac{1 + \sin(2\pi x)}{2} \quad \text{for } x \in [0, 1]$$

with periodic boundary conditions. The solution will develop a discontinuity at time  $t = -\frac{1}{\min U'(x, 0)} = \frac{1}{\pi} \approx 0.318$ . The solution computed with the Godunov scheme using the minmod limiter on a mesh of 50 cells is displayed in Figure 5.21. To assess the convergence rate of the scheme, we compute its experimental rate of convergence at  $t = 0.3$ , while the solution is still smooth. As seen in Table 5.7, the rate of convergence is now close to 2 for both the Godunov and Rusanov schemes.

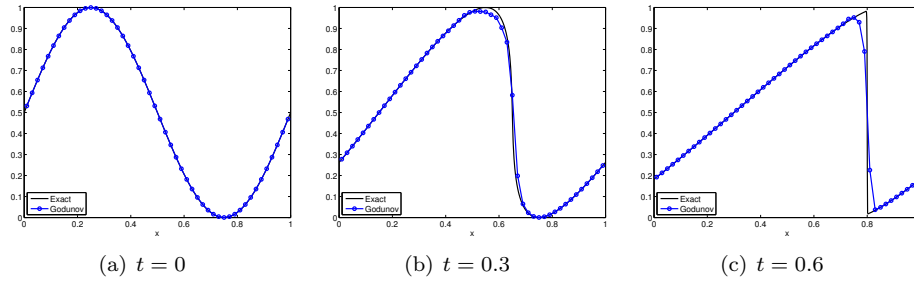


FIGURE 5.21. Smooth initial value problem developing into a discontinuous solution. [burgers\_sine\_reconstr.m]

No. of cells	Godunov		Rusanov	
	Relative error	EOC	Relative error	EOC
20	5.01	—	5.44	—
40	2.13	1.23	2.3	1.24
80	0.772	1.47	0.812	1.5
160	0.241	1.68	0.247	1.72
320	0.0712	1.76	0.0716	1.79
640	0.0214	1.73	0.0215	1.74
1280	0.00592	1.86	0.00592	1.86
2560	0.00168	1.82	0.00168	1.82

TABLE 5.7. Error and order of convergence for Burgers' equation for a smooth initial value problem. [burgers\_sine\_reconstr\_00C.m]

## CHAPTER 6

### Linear hyperbolic systems in one space dimension

In the last few chapters, we have treated the nonlinear scalar conservation law (3.4) and have designed robust high-order finite volume schemes to approximate its entropy solutions.

Despite their occurrence in many interesting models, scalar conservation laws are too simplistic to be used for modeling complex physical phenomena as these problems involve the interaction of many unknowns. These problems (see Chapter 1 for examples) are modeled by a nonlinear system of conservation laws:

$$(6.1) \quad \mathbf{U}_t + \mathbf{f}(\mathbf{U})_x = 0,$$

where  $\mathbf{U} = [U^1, U^2, \dots, U^m]^\top$  is the vector of unknowns and  $\mathbf{f} = [f^1, f^2, \dots, f^m]^\top$  is the flux vector. Note that (6.1) represents an  $m \times m$  system of conservation laws in one space dimension. Multi-dimensional systems will be considered later. The system (6.1) is supplemented with suitable initial and boundary conditions. When  $m = 1$ , the system reduces to the scalar conservation law (3.4).

To begin with, we consider the simplest case of (6.1) in the form a linear system,

$$(6.2) \quad \mathbf{U}_t + A\mathbf{U}_x = 0.$$

Here,  $A$  is an  $m \times m$  matrix with constant (in both space and time) entries. Linear systems arise for instance when *linearizing* (6.1) around a constant state  $\bar{\mathbf{U}}$ , which amounts to solving (6.2) with the constant matrix  $A = \frac{\partial \mathbf{f}}{\partial \mathbf{U}}(\bar{\mathbf{U}})$ .

A related linear system with variable coefficients takes the form

$$(6.3) \quad \mathbf{U}_t + (A(x, t)\mathbf{U})_x = 0.$$

It can be recast into the non-conservative form

$$(6.4) \quad \mathbf{U}_t + A(x, t)\mathbf{U}_x = -A_x(x, t)\mathbf{U}.$$

#### 6.1. Examples for linear systems

**6.1.1. The wave equation.** The simplest example of (6.2) is given by the one-dimensional wave equation,

$$(6.5) \quad u_{tt} - c^2 u_{xx} = 0.$$

This equation is also known as the string equation, as it models vibrations in media like strings and rods. Its derivation can be found in many text books; see for instance [7]. The wave equation can be written as a first-order system by defining auxiliary variables  $v = cu_x$  and  $w = -u_t$ . With this *change of variables*, it is easy to show that (6.5) transforms to

$$(6.6) \quad \begin{aligned} v_t + cw_x &= 0 \\ w_t + cv_x &= 0. \end{aligned}$$

Hence, the wave equation (6.6) is an example of the linear system (6.2) with

$$(6.7) \quad \mathbf{U} = \begin{bmatrix} v \\ w \end{bmatrix}, \quad A = \begin{bmatrix} 0 & c \\ c & 0 \end{bmatrix}.$$

**6.1.2. Maxwell's equations.** The dynamics of electromagnetic waves is modeled by the Maxwell equations of electrodynamics:

$$(6.8) \quad \begin{aligned} \mathbf{B}_t + \operatorname{curl}(\mathbf{E}) &= 0 \\ \mathbf{E}_t - c^2 \operatorname{curl}(\mathbf{B}) &= -\frac{\mathbf{j}}{\varepsilon_0} \\ \operatorname{div}(\mathbf{B}) &= 0 \\ \operatorname{div}(\mathbf{E}) &= \frac{\rho}{\varepsilon_0}. \end{aligned}$$

Here,  $\mathbf{B} = [B^1, B^2, B^3]^\top$  and  $\mathbf{E} = [E^1, E^2, E^3]^\top$  are the magnetic and electric fields, respectively. The charge and current density are denoted by  $\rho$  and  $\mathbf{j}$ , respectively. The speed of light is given by  $c^2 = \mu_0 \varepsilon_0$ , where  $\mu_0$  and  $\varepsilon_0$  are constants. In one space dimension, solving for electromagnetic waves propagating in the  $x$ -direction (with no current) reduces (6.8) to

$$(6.9) \quad \begin{aligned} B_t^2 + E_x^3 &= 0 \\ E_t^3 + c^2 B_x^2 &= 0. \end{aligned}$$

Thus, Maxwell's equations are equivalent to the wave equation (6.7) with  $\mathbf{U} = [cB^2, E^3]^\top$ , leading to another concrete application of the wave equation.

**6.1.3. Linearized Euler equations.** The Euler equations of gas dynamics (1.10) are a nonlinear system of conservation laws in several spatial dimensions. Let  $\rho$ ,  $u$  and  $p$  denote the density, velocity and pressure of the gas, respectively. Restricting the equations to one space dimension and linearizing (1.10) around a constant state  $\bar{\mathbf{U}} = [\bar{\rho}, \bar{u}, \bar{p}]^\top$ , we obtain the *linearized Euler equations*,

$$(6.10) \quad \begin{aligned} \rho_t + \bar{u}\rho_x + \bar{\rho}u_x &= 0 \\ u_t + \bar{u}u_x + \frac{1}{\bar{\rho}}p_x &= 0 \\ p_t + \gamma\bar{p}u_x + \bar{u}p_x &= 0. \end{aligned}$$

This system of equations can be written as a linear system (6.2) with

$$(6.11) \quad \mathbf{U} = \begin{bmatrix} \rho \\ u \\ p \end{bmatrix}, \quad A = \begin{bmatrix} \bar{u} & \bar{\rho} & 0 \\ 0 & \bar{u} & \frac{1}{\bar{\rho}} \\ 0 & \gamma\bar{p} & \bar{u} \end{bmatrix}.$$

The linearized Euler equations are a good approximation of the more complex Euler equations (1.10) when the solution only consists of small perturbations around constant states.

Given the above examples, the study of the linear system (6.2) is important in its own right. Furthermore, it serves as an example of the nonlinear system (6.1), and so the study of the linear system will be a starting point in the design of efficient numerical schemes for approximating (6.1).



### 6.2. Hyperbolicity and characteristic decomposition

We will consider the initial value problem associated with (6.2) by augmenting it with initial data

$$\mathbf{U}(x, 0) = \mathbf{U}_0(x).$$

One of the key notions underlying the behavior of the solutions of (6.2) is the concept of *hyperbolicity*. For scalar equations, the notion of hyperbolicity referred to the property of finite speed of propagation. Such a notion will hold true for a system.

**Definition 6.1** (Hyperbolic system). *The linear system (6.2) is hyperbolic if the matrix  $A$  is diagonalizable and has  $m$  real eigenvalues. Similarly, the equation (6.3) is hyperbolic if the matrix  $A(x, t)$  is diagonalizable and has  $m$  real eigenvalues for all  $x, t$ . Both systems are termed strictly hyperbolic if the eigenvalues are distinct.*

Thus, the concept of hyperbolicity for a system is tied to the eigenstructure of the underlying matrix. Why is the eigenstructure linked to finite speed of propagation? The answer lies in diagonalizability. A matrix  $A$  is diagonalizable if it has a complete set of eigenvectors, i.e., there exists  $m$  linearly independent vectors  $r_1, \dots, r_m \in \mathbb{R}^m$  and corresponding eigenvalues  $\lambda_1, \dots, \lambda_m \in \mathbb{R}$  such that

$$Ar_p = \lambda_p r_p.$$

We can then define the matrix of eigenvectors

$$(6.12) \quad R = [r_1 | r_2 | \dots | r_m]$$

and the diagonal matrix of eigenvalues

$$(6.13) \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$$

and use them to *diagonalize* the underlying matrix  $A$  – that is, write it in the form

$$(6.14) \quad A = R\Lambda R^{-1}.$$

The diagonalization of  $A$  leads us to the following decoupling:

$$\begin{aligned} \mathbf{U}_t + A\mathbf{U}_x &= 0 \\ \mathbf{U}_t + R\Lambda R^{-1}\mathbf{U}_x &= 0 \quad (\text{from (6.14)}) \\ \mathbf{U}_t + R\Lambda (R^{-1}\mathbf{U})_x &= 0 \quad (\text{constant coefficient}) \\ (R^{-1}\mathbf{U})_t + R^{-1}R\Lambda (R^{-1}\mathbf{U})_x &= 0 \quad (\text{multiplying both sides with } R^{-1}) \\ (R^{-1}\mathbf{U})_t + \Lambda (R^{-1}\mathbf{U})_x &= 0. \end{aligned}$$

Introducing the vector of characteristic variables

$$\mathbf{V} = R^{-1}\mathbf{U},$$

the above calculation leads to

$$(6.15) \quad \mathbf{V}_t + \Lambda \mathbf{V}_x = 0.$$

Thus, the hyperbolic linear system (6.2) is completely decoupled into a set of scalar linear transport equations, obtained by writing (6.15) componentwise:

$$(6.16) \quad V_t^p + \lambda_p V_x^p = 0 \quad \text{for } p = 1, \dots, m,$$

where  $V^p(x, t)$  is the  $p$ -th component of  $\mathbf{V}(x, t)$ . This decoupled set of transport equations can be solved explicitly by the method of characteristics (see Chapter 2) to obtain

$$(6.17) \quad V^p(x, t) = W_0^p(x - \lambda_p t),$$

where

$$\mathbf{V}_0(x) = R^{-1}\mathbf{U}_0(x)$$

and again  $W_0^p(x, t)$  is the  $p$ -th component of  $\mathbf{V}_0(x, t)$ . Therefore, we can explicitly write the initial value problem for a hyperbolic system (6.2) as

$$(6.18) \quad \mathbf{U}(x, t) = R\mathbf{V}(x, t).$$

### 6.3. Solutions of Riemann problems, waves

We illustrate the explicit solution (6.18) for the Riemann initial data

$$(6.19) \quad \mathbf{U}_0(x) = \begin{cases} \mathbf{U}_L & \text{if } x < 0 \\ \mathbf{U}_R & \text{if } x > 0. \end{cases}$$

The first step in finding an explicit solution is to determine  $\Lambda$  and  $R$  for the given system  $A$ . These are used to define  $\mathbf{V}_0 = R^{-1}\mathbf{U}_0$ , which will be of the form

$$\mathbf{V}_0(x) = \begin{cases} \mathbf{V}_L & \text{if } x < 0 \\ \mathbf{V}_R & \text{if } x > 0. \end{cases}$$

with  $\mathbf{V}_L = R^{-1}\mathbf{U}_L$  and  $\mathbf{V}_R = R^{-1}\mathbf{U}_R$ . Then we can solve (6.16) in terms of the transport equation (6.17) to obtain

$$V^p(x, t) = \begin{cases} V_L^p & \text{if } x < \lambda_p t \\ V_R^p & \text{if } x > \lambda_p t. \end{cases}$$

The solution  $\mathbf{U}$  is obtained by letting  $\mathbf{U} = R\mathbf{V}$ . This solution consists of  $m$  waves, one for each eigenvector. The  $p$ -th wave propagates with a speed of  $\lambda_p$ , and across this wave, there is a jump in only the  $p$ -th component of  $\mathbf{V}$ . This is seen clearly by decomposing the jump  $\mathbf{U}_R - \mathbf{U}_L$  in terms of eigenvectors:

$$(6.20) \quad \begin{aligned} \mathbf{U}_R - \mathbf{U}_L &= R(\mathbf{V}_R - \mathbf{V}_L) \\ &= \sum_{p=1}^m (V_R^p - V_L^p) r_p \\ &= \sum_{p=1}^m \alpha^p r_p, \end{aligned}$$

where  $\alpha^p := V_R^p - V_L^p$  is the *wave strength* of the  $p$ -th wave. Hence, the jump across the  $p$ -th wave is proportional to the eigenvector  $r_p$ . Thus, the solution of the Riemann problem for a linear system corresponds to decomposing the initial jump into  $m$  waves (with the corresponding eigenvalue being the wave speed), with the jump across the  $p$ -th wave being proportional to the corresponding eigenvector.

**6.3.1. Example: The wave equation.** We consider the linear wave equation (6.6). It is straightforward to find that the eigenvalues and eigenvectors of  $A$  are

$$(6.21) \quad \Lambda = \text{diag}(-c, c), \quad R = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Thus, the wave equation is strictly hyperbolic and diagonalizable, so the above construction of the explicit solution is applicable. Consider the Riemann problem (6.19) for the wave equation (6.6). Let  $c = 1$  for simplicity. The solution of the Riemann problem is then

$$(6.22) \quad \mathbf{U}(x, t) = \begin{cases} \mathbf{U}_L & \text{if } x < -t \\ \mathbf{U}_* & \text{if } -t < x < t \\ \mathbf{U}_R & \text{if } x > t, \end{cases}$$

where the mid-state  $\mathbf{U}_*$  is given by

$$\mathbf{U}_* = R \begin{bmatrix} \frac{v_R - u_R}{2} \\ \frac{v_L + u_L}{2} \end{bmatrix}.$$

Thus, it is straightforward to calculate solutions to the Riemann problem for the wave equation.

**6.3.2. Example: Linearized Euler equations.** The eigenstructure of the linearized Euler equations (6.10) can be explicit computed as

$$(6.23) \quad \Lambda = \text{diag}(\bar{u} - \bar{a}, \bar{u}, \bar{u} + \bar{a}), \quad R = \begin{bmatrix} -\frac{\bar{p}}{\bar{a}} & 1 & \frac{\bar{p}}{\bar{a}} \\ 1 & 0 & 1 \\ -\bar{\rho}\bar{a} & 0 & \bar{\rho}\bar{a} \end{bmatrix},$$

where

$$\bar{a} = \sqrt{\frac{\gamma \bar{p}}{\bar{\rho}}}$$

denotes the local sound speed. Hence, the linearized Euler equations are strictly hyperbolic. The two waves corresponding to  $r_1$  and  $r_3$  are termed the *acoustic* or *sound waves* and the wave corresponding to  $r_2$  is called a *contact wave*. The Riemann problem for the linearized Euler equations can be solved explicitly with the procedure outlined above.

## 6.4. Finite volume schemes

In this section, we construct finite volume schemes to approximate the linear system (6.2). For simplicity, we consider a uniform grid in both space and time with mesh size  $\Delta x$  and time step  $\Delta t$ . The mesh size and time step are related via a CFL type condition to be specified in the sequel. We denote the grid points as  $x_{j+1/2} = x_L + (j + 1/2)\Delta x$  and the time levels by  $t^n$ . The domain is divided into cells denoted by

$$\mathcal{C}_j = [x_{j+1/2}, x_{j-1/2}).$$

See Figure 4.1 for an illustration of the grid.

By now, we are familiar with the design procedure (see Chapter 4) and know that it consists of the following steps:

- (i) **Reconstruction:** If the solution  $\mathbf{U}$  is known at a certain time level  $t^n$ , we evaluate its cell averages  $\mathbf{U}_j^n$  as

$$\mathbf{U}_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}(x, t^n) dx.$$

- (ii) **Evolution:** The reconstruction procedure projects the solution into the space of piecewise constant functions and leads to a superposition of the Riemann problems

$$(6.24) \quad \begin{aligned} &\mathbf{U}_t + A\mathbf{U}_x = 0 \\ &\mathbf{U}(x_{j+1/2}, t^n) = \begin{cases} \mathbf{U}_j^n & \text{if } x_j < x < x_{j+1/2}, \\ \mathbf{U}_{j+1}^n & \text{if } x_{j+1/2} < x < x_{j+1} \end{cases} \end{aligned}$$

at each interface  $x_{j+1/2}$ . These Riemann problems can be solved exactly (as described in the previous section) or approximately (see the sequel).

We know that the fastest wave speed for each problem is bounded by the modulus of the maximum eigenvalue,

$$\lambda_{\max} = \max_{1 \leq p \leq m} |\lambda_p|,$$

where  $\lambda_p$  is the  $p$ -th eigenvalue of  $A$ . Hence, imposing a CFL condition

$$(6.25) \quad \lambda_{\max} \frac{\Delta t}{\Delta x} \leq \frac{1}{2}$$

prevents the waves from the neighboring problems to interact before time level  $t^{n+1}$ .

- (iii) **Averaging:** The solution of (6.24) is averaged over each cell to obtain  $\mathbf{U}_j^{n+1}$ . The cell average can be explicitly computed by integrating (6.24) over  $\mathcal{C}_j \times [t^n, t^{n+1})$ . This gives the finite volume scheme

$$(6.26) \quad \mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x} \left( \mathbf{F}_{j+1/2}^n - \mathbf{F}_{j-1/2}^n \right)$$

with

$$(6.27) \quad \mathbf{F}_{j+1/2}^n = \mathbf{F}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) := A\mathbf{U}_{j+1/2}(x_{j+1/2}, 0).$$

Here,  $\mathbf{U}_{j+1/2}$  is the solution of the (exact or approximate) solution of the Riemann problem (6.24). We remark that the solution is self-similar and its value at the interface is time independent.

The finite volume scheme (6.26) is similar to the scheme for scalar equations. The only difference lies in the more complicated structure of solutions of the Riemann problem (6.24). We need to complete (6.26) by computing the numerical flux explicitly.

**6.4.1. Godunov flux.** The Riemann problem (6.24) can be solved explicitly as described before. The resulting flux evaluation (6.27) defines the Godunov flux for the linear system (6.2). It turns out that we can derive a neat expression for the Godunov flux.

Let  $a^+ = \max(a, 0)$  and  $a^- = \min(a, 0)$ . Then  $a = a^+ + a^-$  and  $|a| = a^+ - a^-$  for any  $a \in \mathbb{R}$ . The explicit solution of the Riemann problem (6.24) can be computed

and the interface value  $\mathbf{U}_{j+1/2}(x_{j+1/2}, 0) = \mathbf{U}^*$  can be calculated (see (6.20)) by

$$(6.28) \quad \mathbf{U}^* = \mathbf{U}_j^n + \sum_{p: \lambda_p < 0} \alpha_{j+1/2}^p r_p,$$

where  $\alpha_{j+1/2}^p$  is the wave strength of the  $p$ -th wave, given by the  $p$ -th component of  $R^{-1}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n)$ . Since we are interested in calculating the interface flux in (6.27), we multiply both sides of (6.28) by  $A$  to obtain

$$(6.29) \quad \begin{aligned} A\mathbf{U}^* &= A\mathbf{U}_j^n + A \left( \sum_{p: \lambda_p < 0} \alpha_{j+1/2}^p r_p \right) \\ &= A\mathbf{U}_j^n + \sum_{p: \lambda_p < 0} \alpha_{j+1/2}^p A r_p \\ &= A\mathbf{U}_j^n + \sum_{p: \lambda_p < 0} \alpha_{j+1/2}^p \lambda_p r_p \quad (\text{as } r_p \text{ are eigenvectors of } A) \\ &= A\mathbf{U}_j^n + \sum_p \lambda_p^- \alpha_{j+1/2}^p r_p. \end{aligned}$$

Similarly, we have

$$\mathbf{U}^* = \mathbf{U}_{j+1}^n - \sum_{p: \lambda_p \geq 0} \alpha_{j+1/2}^p r_p.$$

Repeating the calculations of (6.29), we obtain

$$(6.30) \quad A\mathbf{U}^* = A\mathbf{U}_{j+1}^n - \sum_p \lambda_p^+ \alpha_{j+1/2}^p r_p.$$

Taking the average of (6.29) and (6.30) leads to

$$(6.31) \quad \begin{aligned} A\mathbf{U}^* &= \frac{1}{2} \left( A\mathbf{U}_j^n + A\mathbf{U}_{j+1}^n - \sum_p (\lambda_p^+ - \lambda_p^-) \alpha_{j+1/2}^p r_p \right) \\ &= \frac{1}{2} A (\mathbf{U}_j^n + \mathbf{U}_{j+1}^n) - \frac{1}{2} \sum_p |\lambda_p| \alpha_{j+1/2}^p r_p. \end{aligned}$$

Using the definition of the wave strength  $\alpha_{j+1/2}^p$  and defining

$$|\Lambda| = \text{diag}(|\lambda_1|, \dots, |\lambda_m|),$$

it is straightforward to rewrite (6.31) as

$$(6.32) \quad \mathbf{F}_{j+1/2}^n = A\mathbf{U}^* = \frac{1}{2} A (\mathbf{U}_j^n + \mathbf{U}_{j+1}^n) - \frac{1}{2} R |\Lambda| R^{-1} (\mathbf{U}_{j+1}^n - \mathbf{U}_j^n).$$

This is the explicit form of the Godunov flux for a linear system.

The Godunov scheme (6.26) with flux (6.32) for the linear system (6.2) should be compared with the upwind flux (2.17) for the linear transport equation (2.2). In particular, both fluxes are written in the central + diffusion form. The upwinding for a system is more complicated and involves characteristic decomposition. Note that the flux (6.32) reduces to the standard upwind flux for a scalar equation, when  $m = 1$ .

**6.4.2. Lax-Friedrichs and Rusanov flux.** The Godunov flux (6.32) requires a characteristic decomposition of the system. Explicit expressions for the eigenvalues and eigenvectors may not be available in some cases, and in others it might be computationally costly to evaluate them. Hence, there is scope for designing cheaper alternative numerical fluxes that are easy to implement. Approximate Riemann solvers (see Chapter 4) provide a recipe for designing such fluxes. The key idea is to replace the exact solution of the Riemann problem (6.24) with an approximate solution. The Lax-Friedrichs flux is based on a two-wave approximate Riemann solver (see Section 4.2.2). It is straightforward to generalize it to the case of a linear system: We replace the exact solution (consisting of  $m$  waves) with exactly two waves, one moving to the left and another to the right (see Figure 4.7). The wave speeds are  $-\frac{\Delta x}{2\Delta t}$  and  $\frac{\Delta x}{2\Delta t}$ , which is the maximum possible wave speed consistent with the CFL condition (6.25). The resulting flux is

$$(6.33) \quad \mathbf{F}_{j+1/2}^n = \mathbf{F}^{\text{LxF}}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) = \frac{1}{2}A(\mathbf{U}_j^n + \mathbf{U}_{j+1}^n) - \frac{\Delta x}{2\Delta t}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n).$$

(compare to (4.29)). This flux is trivial to implement as it does not require any characteristic information.

As in the scalar case, the Lax-Friedrichs scheme ((6.26) with (6.33)) is likely to be diffusive. A more reasonable choice of wave speeds in the approximate Riemann solver is the largest eigenvalue of  $A$ . (Note that the maximum wave speed in the Riemann problem (6.24) is always bounded by the maximum eigenvalue of  $A$ .) The resulting flux is the Rusanov flux

$$(6.34) \quad \mathbf{F}_{j+1/2}^n = \mathbf{F}^{\text{Rus}}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) = \frac{1}{2}A(\mathbf{U}_j^n + \mathbf{U}_{j+1}^n) - \frac{\lambda_{\max}}{2}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n)$$

(compare to (4.31)). The only characteristic information used in the Rusanov scheme is an estimate on the maximum eigenvalue of  $A$ .

## 6.5. Numerical experiments

In this section we present numerical experiments for linear systems (6.2) using the Godunov, Lax-Friedrichs and Rusanov schemes.

**6.5.1. Numerical experiment 1.** Consider the wave equation (6.6) with wave speed  $c = 1$ . Let the initial data be the Riemann problem

$$(6.35) \quad \begin{aligned} u_0(x) &= \begin{cases} 1 - \cos^2(2\pi(x - 0.25)) & \text{if } 0.25 \leq x \leq 0.75 \\ 0 & \text{otherwise} \end{cases} \\ v_0(x) &= 0. \end{aligned}$$

We compute on the periodic domain  $x \in [-1, 1]$  and divide the domain into 100 grid cells. In Figure 6.1 we present the computational results with the Godunov, Lax-Friedrichs and Rusanov schemes at time  $t = 0.5, 1$  and  $2$ . The initial profile for  $u$  first separates into two parts, and at time  $t = 1$  they meet at  $x = -0.5$  and merge together. This process is again repeated at time  $t = 2$ , at which time the solution is the same as at  $t = 0$ .

From Figure 6.1 it is clear that the Lax-Friedrichs scheme is more diffusive than the Godunov and Rusanov schemes. On the other hand, the Godunov and Rusanov schemes produce an identical profile. The reason for this is that in this case the Rusanov and Godunov schemes are identical.

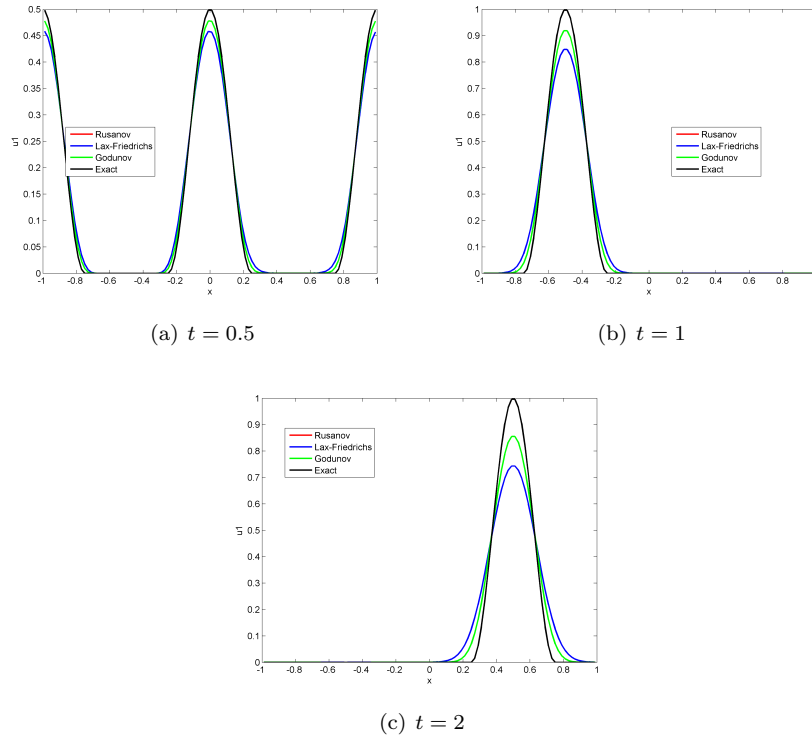


FIGURE 6.1. Wave equation (6.6) computed at time  $t = 0.5, 1$  and  $2$  with smooth initial conditions (6.35) and periodic boundary conditions on 100 cells using Rusanov, Lax-Friedrichs and Godunov fluxes. [waveBump.m]

**Exercise 6.2.** Show that the Godunov and Rusanov fluxes (6.32) and (6.34) are identical for the wave equation (6.7).

**6.5.2. Numerical experiment 2.** Consider the wave equation (6.6) with wave speed  $c = 1$  and initial data

$$(6.36) \quad \begin{aligned} u_0(x) &= \begin{cases} 1 & \text{if } 0.25 \leq x \leq 0.75 \\ 0 & \text{otherwise} \end{cases} \\ v_0(x) &= 0. \end{aligned}$$

We compute on the periodic domain  $x \in [-1, 1]$  and divide the domain into 100 grid cells. Results are presented in Figure 6.2. The results are again plotted at time  $t = 0.5, 1$  and  $2$ . The solution behaves similar to the smooth case. Again we see that the Lax-Friedrichs scheme is the most diffusive and that the Godunov and Rusanov schemes produce identical results.

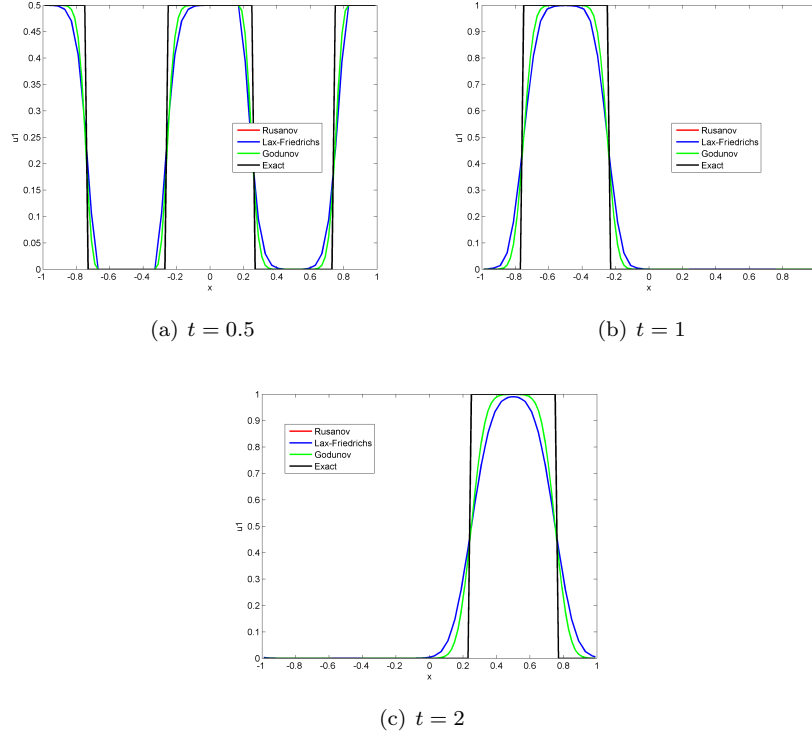


FIGURE 6.2. Wave equation (6.6) computed at time  $t = 0.5, 1$  and  $2$ , with discontinuous initial conditions (6.36) and periodic boundary conditions on 100 cells using Rusanov, Lax-Friedrichs and Godunov fluxes. `[waveDisc.m]`

**6.5.3. Numerical experiment 3.** Consider the linearized Euler equation (6.10) with parameters

$$\begin{bmatrix} \bar{\rho} \\ \bar{u} \\ \bar{p} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \gamma = \frac{7}{5}$$

and initial data

$$(6.37) \quad \begin{aligned} \rho_0(x) &= \begin{cases} 1 & \text{if } x < 0 \\ 0.2 & \text{if } x > 0 \end{cases} \\ u_0(x) &= 0 \\ p_0(x) &= \begin{cases} 1 & \text{if } x < 0 \\ 0.2 & \text{if } x > 0 \end{cases} \end{aligned}$$

in the computational domain  $[-1, 1]$  with Neumann (outflow) type boundary conditions. In Figure 6.3 we plot the density, velocity and pressure at time  $t = 0.5$ . The solution for density contain three waves, one for each eigenvalue. The fastest wave is moving with speed equal to the highest eigenvalue of the system. Note that the velocity and pressure contain only two waves.



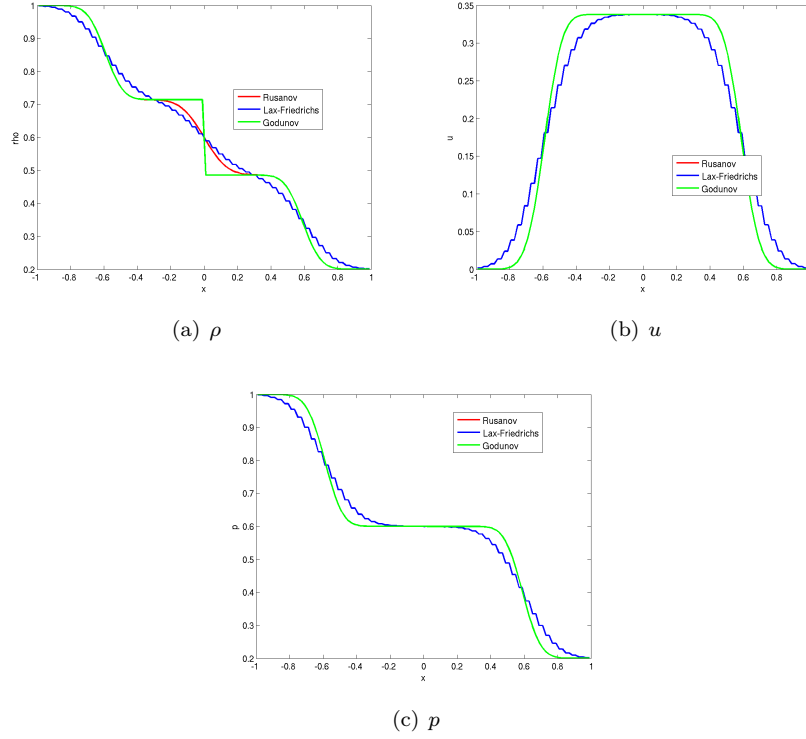


FIGURE 6.3. Linearized Euler equation (6.10) computed up to time  $t = 0.5$  with initial conditions (6.37) and outflow boundary conditions. [linEuler.m]

The most striking feature in Figure 6.3 is that the Godunov solver resolves the stationary wave in density *exactly*. This is a well-known feature of the Godunov scheme. In the velocity and pressure solutions, the Rusanov and Godunov schemes compute the exact same solutions. Of the three schemes, the Lax-Friedrichs scheme is by far the most inaccurate one.

### 6.6. High-order finite volume schemes

It is relatively straightforward to extend the second-order schemes of Chapter 5 to the linear system (6.2). Following Chapter 5, we use the semi-discrete form of the finite volume scheme (see (5.36)),

$$(6.38) \quad \frac{d}{dt} \mathbf{U}_j(t) + \frac{1}{\Delta x} (\mathbf{F}_{j+1/2}(t) - \mathbf{F}_{j-1/2}(t)),$$

where we denote the cell average as

$$\mathbf{U}_j(t) = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}(x, t) dx.$$

Given cell averages  $\mathbf{U}_j$ , we wish to obtain a linear reconstruction  $\mathbf{p}_j$  in each cell  $\mathcal{C}_j$  of the form

$$(6.39) \quad \mathbf{p}_j(x) = \mathbf{U}_j + \sigma_j(x - x_j).$$

These linear functions are combined to form the piecewise linear function

$$(6.40) \quad \mathbf{p}(x, t) = \mathbf{p}_j(x) \quad \text{for } x_{j-1/2} \leq x < x_{j+1/2}.$$

Contrary to the situation in Chapter 5, the slope  $\sigma_j$  is now a vector, and so the theory for scalar equations does not translate directly to the present situation. There are two basic approaches to constructing the slope  $\sigma_j$ .

**6.6.1. Reconstruction in primitive variables.** One approach to constructing  $\sigma_j$  is to reconstruct in primitive (conserved) variables  $U_j^p$  ( $p = 1, \dots, m$ ). This involves applying the reconstruction procedures of Chapter 5 to each component of  $\mathbf{U}_j$ , thus constructing  $\sigma_j$  component by component.

**6.6.2. Reconstruction in characteristic variables.** An alternative reconstruction involves using the characteristic variables. Given the cell averages  $\mathbf{U}_j$ , we define the corresponding characteristic average  $\mathbf{V}_j = R^{-1}\mathbf{U}_j$ . We can then perform a componentwise reconstruction in  $\mathbf{V}_j$ , as outlined in the previous section, to obtain a reconstruction

$$(6.41) \quad \mathbf{V}_j(x) = \mathbf{V}_j + \mu_j(x - x_j).$$

The primitive variables are then obtained from

$$(6.42) \quad \mathbf{p}_j(x) = R\mathbf{V}_j(x).$$

**6.6.3. The numerical flux.** As in the scalar case, the numerical flux  $\mathbf{F}$  in (6.38) is defined by

$$(6.43) \quad \mathbf{F}_{j+1/2} = \mathbf{F}(\mathbf{U}_j^+, \mathbf{U}_{j+1}^-),$$

where

$$(6.44) \quad \mathbf{U}_j^+ = \mathbf{p}_j(x_{j+1/2}), \quad \mathbf{U}_j^- = \mathbf{p}_j(x_{j-1/2}),$$

and  $\mathbf{F}$  is either the Godunov (6.32), Lax-Friedrichs (6.33) or Rusanov (6.34) numerical flux.

**6.6.4. Time stepping.** The time integration for (6.38) is performed with an SSP Runge-Kutta method, as described in Chapter 5. Denoting

$$\mathbf{U}(t) = [\dots, \mathbf{U}_{j-1}(t), \mathbf{U}_j(t), \mathbf{U}_{j+1}(t), \dots],$$

the finite volume scheme (6.38) can be rewritten as

$$(6.45) \quad \frac{d}{dt} \mathbf{U}(t) = \mathcal{L}(\mathbf{U}(t)),$$

where the operator  $\mathcal{L}$  acts pointwise on the vector  $\mathbf{U}$  as

$$\mathcal{L}(\mathbf{U}(t))_j := -\frac{1}{\Delta x} (\mathbf{F}_{j+1/2}(t) - \mathbf{F}_{j-1/2}(t)).$$

The second-order SSP Runge-Kutta method is then defined by

$$(6.46) \quad \begin{aligned} \mathbf{U}^* &= \mathbf{U}^n + \Delta t \mathcal{L}(\mathbf{U}^n) \\ \mathbf{U}^{**} &= \mathbf{U}^* + \Delta t \mathcal{L}(\mathbf{U}^*) \\ \mathbf{U}^{n+1} &= \frac{1}{2}(\mathbf{U}^n + \mathbf{U}^{**}). \end{aligned}$$

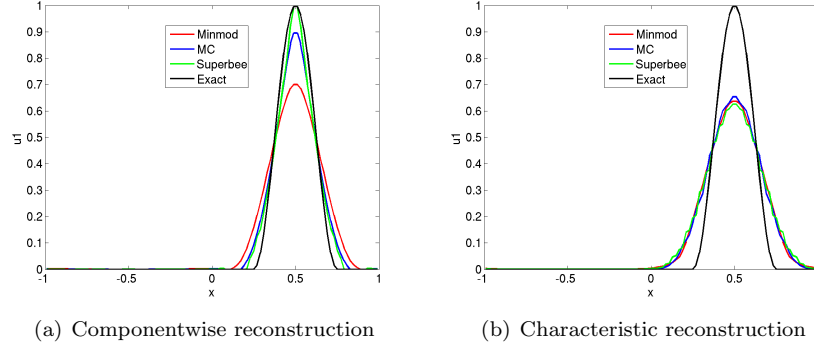


FIGURE 6.4. Wave equation (6.6) computed up to time  $t = 1$ , with smooth initial conditions (6.35) and periodic boundary conditions on a mesh of 100 grid cells. Computed with the Godunov flux using the minmod, superbee and MC limiters. [waveBump\_2nd.m]

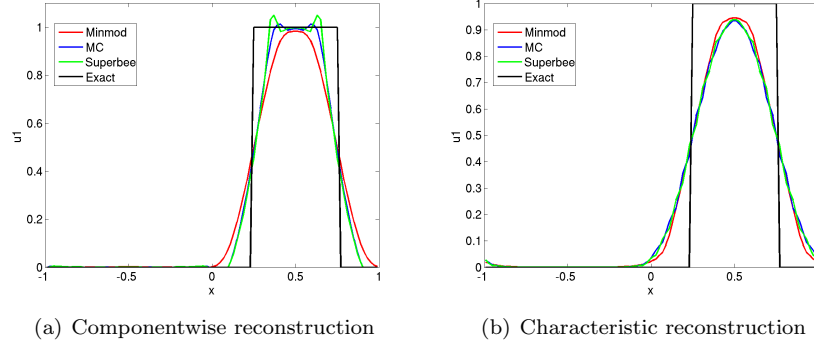


FIGURE 6.5. Wave equation (6.6) computed up to time  $t = 1$ , with discontinuous initial conditions (6.36) and periodic boundary conditions on a mesh of 100 grid cells. Computed with the Godunov flux using the minmod, superbee and MC limiters. [waveDisc\_2nd.m]

### 6.7. Numerical experiments

In this section we repeat the numerical experiments carried out in Section 6.5 using second order numerical scheme based on slope limiters. The computational results for the numerical experiment 1 are presented in Figure 6.4, where we compare the performance of the minmod, superbee and MC limiters. We note that both MC and superbee produce some oscillations. On the other hand, minmod is more diffusive compared to the other limiters. Similar conclusion can be drawn from the computational results of numerical experiment 2, presented in Figure 6.5.

In Figure 6.6 we present computational results for the linearized Euler equation using data from numerical experiment 3. Here, the solutions are much sharper than that of the first-order scheme. Similar to numerical experiments 1 and 2, we see

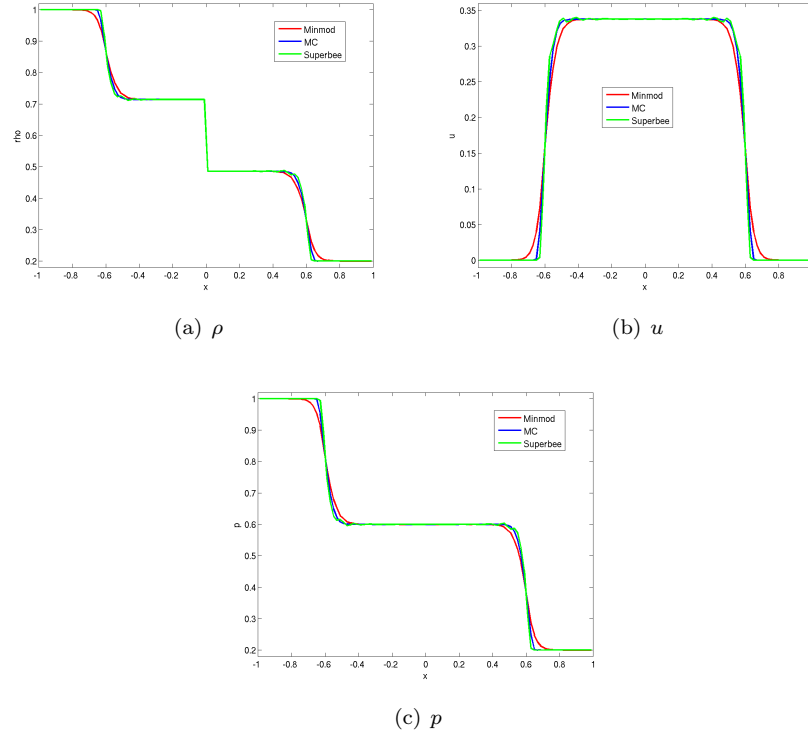


FIGURE 6.6. Linearized Euler equation (6.10) computed up to time  $t = 0.5$  with initial conditions (6.37) and outflow boundary conditions. Computed with the Godunov flux using the minmod, superbee and MC limiters. [linEuler\_2nd.m]

that superbee and MC can produce some oscillation, whereas minmod is slightly more diffusive. Note in particular that the stationary shock in density is perfectly resolved, as for the first-order Godunov scheme.

## APPENDIX A

### Results from real analysis

**Theorem A.1** (Gronwall's inequality). *Let  $\beta(t)$  be continuous and  $u(t)$  be differentiable on some interval  $[a, b]$ , and assume that*

$$u'(t) \leq \beta(t)u(t) \quad \forall t \in (a, b).$$

*Then*

$$u(t) \leq u(a) \exp \left( \int_a^t \beta(t) \right) \quad \forall t \in [a, b].$$



## Bibliography

1. E. Godlewski and P.A. Raviart, *Hyperbolic systems of conservation laws*, Ellipses, 1991.
2. S. K. Godunov, *A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations*, Math. Sbornik **47** (1959), 271–306.
3. Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Review **43** (2001), no. 1, 89–112.
4. Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner, *Solving ordinary differential equations*, 1987.
5. S N Kruzkov, *First order quasilinear equations in several independent variables*, Mathematics of the USSR-Sbornik **10** (1970), no. 2, 217–243.
6. L. D. Landau and E. M. Lifschitz, *Fluid mechanics*, Butterworth-Heinemann, 1987.
7. Aslak Tveito and Ragnar Winther, *Introduction to partial differential equations; a computational approach*, second ed., vol. 29, Springer-Verlag, 2009, 392 pages.