Lectures on Hyperbolic Equations and Their Numerical Approximation

Eleuterio F. Toro

Abstract These introductory lecture notes on numerical methods for hyperbolic equations are suitable for advanced undergraduate and postgraduate students in mathematics and engineering disciplines. More advanced approaches exist and will be indicated as appropriate. The material is divided into four sections. Section 1 presents an overview of hyperbolic equations and also some basic concepts on numerical discretization techniques. Section 2 deals with a specific example, the system of non-linear shallow water equations; the equations are analysed and the Riemann problem is solved exactly in complete detail. In Sect. 3 we first present the Godunov method as applied to a generic hyperbolic system and then specialised to the shallow water system in one space dimension; approximate solution methods for the Riemann problem are also given. Finally, Sect. 4 gives a brief overview of the ADER approach to construct high-order numerical methods for hyperbolic equations, based on the first order Godunov method. Much of the material of these lectures has been taken from the author's text books (Toro, Riemann solvers and numerical methods for fluid dynamics. A practical introduction, 3rd edn. Springer, Berlin (2009) and Toro, Shock-capturing methods for free-surface shallow flows. Wiley, Chichester (2001)), where further reading material can be found. I also recommend the textbook by Godlewski and Raviart (Numerical approximation of hyperbolic systems of conservation laws. Springer, New York (1996)) and that by LeVeque (Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge (2002)).

E. F. Toro (✉)
University of Trento, Trento, Italy
e-mail: eleuterio.toro@unitn.it

A. Farina et al. (eds.), *Non-Newtonian Fluid Mechanics and Complex Flows*,
Lecture Notes in Mathematics 2212, https://doi.org/10.1007/978-3-319-74796-5_3

1 Hyperbolic Equations

Many problems in science and engineering (e.g. wave propagation and transport phenomena) are governed by advection-diffusion-reaction partial differential equations (PDEs). In the scalar case (a single equation) we may write

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = s(x,t,q(x,t)) + \partial_x(\alpha(x,t,q(x,t))\partial_x q(x,t)) \ , \tag{1}$$

where $q(x,t)$ is the *unknown*, called the *dependent variable*; $q(x,t)$ is a function of two *independent variables* x and t; $f(q)$ is a prescribed function of q called the *flux*, or *physical flux*; $s(x,t,q)$ is also a prescribed function, called the *source term*. The last term is called the diffusion term; $\alpha(x,t,q(x,t))$ is the *diffusion coefficient*. Equation (1) is parabolic due to the presence of the viscous term, a second-order term. In the rest of these lectures we shall be concerned exclusively with hyperbolic equations.

1.1 The Linear Advection Equation and Basic Concepts

A particular example of (1) is obtained by choosing

$$f(q) = \lambda q \ , \quad s(q) = 0 \ , \quad \alpha = 0 \ , \tag{2}$$

with λ a constant wave propagation speed, which leads to the *linear advection equation* (LAE)

$$\partial_t q + \lambda \partial_x q = 0 \ , \quad -\infty < x < \infty \ , \quad t > 0 \ . \tag{3}$$

Initial Value Problem (IVP) for the Linear Advection Equation We study the simplest case of (1), the linear advection equation, in which the spatial domain is infinite and an initial condition at the initial time $t = 0$ is prescribed, namely

$$\left.\begin{array}{ll} \text{PDE:} & \partial_t q + \lambda \partial_x q = 0 \ , \quad -\infty < x < \infty \ , \quad t > 0 \ , \\ \text{IC:} & q(x,0) = h(x) \ , \quad -\infty < x < \infty \ , \end{array}\right\} \tag{4}$$

where $h(x)$ is a prescribed function of distance x. Equation (4) defines a pure *initial-value problem* or *Cauchy problem*.

Characteristic Curves and the Solution *Characteristic curves*, or *characteristics*, are functions $x(t)$ in the x-t half-plane of independent variables satisfying the IVP

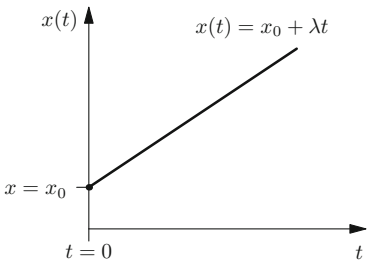

Fig. 1 Characteristic $x(t)$ in the t-x plane given by (6); x_0: foot of characteristic; positive characteristic speed λ

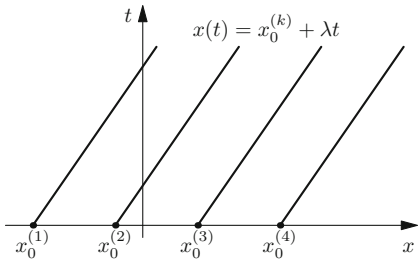

Fig. 2 Family of characteristic curves $x(t)$ in the x-t plane, for the case of positive characteristic speed λ. Compare with Fig. 1

for an ordinary differential equation (ODE), namely

$$\left.\begin{array}{ll} \text{ODE:} & \dfrac{dx}{dt} = \lambda \ , \quad t > 0 \ , \\ \text{IC:} & x(0) = x_0 \ , \end{array}\right\} \tag{5}$$

whose solution is immediate and reads

$$x = x_0 + \lambda t \ . \tag{6}$$

Figure 1 illustrates solution (6). In practice it is more common to represent characteristics in the x-t plane. The inclination of the characteristics depends on the characteristic speed λ, in fact on $1/\lambda$. In the linear case with constant coefficients, characteristics are all parallel to each other, as seen in Fig. 2.

Consider now the time-rate of change (or total derivative) of $q(x(t), t)$ along a characteristic curve $x = x(t)$

$$\frac{dq}{dt} = \frac{\partial q}{\partial t}\frac{dt}{dt} + \frac{\partial q}{\partial x}\frac{dx}{dt} \ . \tag{7}$$

But the curve $x(t)$ satisfies the ODE in (5). Then (7) becomes

$$\frac{dq}{dt} = \partial_t q + \lambda \partial_x q = 0 \ . \tag{8}$$

That is, $q(x,t)$ is constant along $x = x_0 + \lambda t$. Consequently, the PDE in (4) becomes an ODE, namely

$$\frac{dq}{dt} = 0 \text{ along the characteristic } x = x_0 + \lambda t \ .$$

This ODE states that $q(x,t)$ is constant along the characteristic. From the above observations, the value of $q(x,t)$ at a point (x,t) on the characteristic curve passing through (x,t) is equal to the value of q at the point x_0 called *the foot of the characteristic*. That is

$$q(x,t) = q(x_0,0) = h(x_0) \ . \tag{9}$$

But from (6)

$$x_0 = x - \lambda t$$

and therefore the solution of IVP (4) is

$$q(x,t) = h(x - \lambda t) \ , \tag{10}$$

which is the initial condition h in (4) evaluated at the position $x - \lambda t$. Figure 3 shows the three possible cases that can occur due to the value of the characteristic speed.

IVP Example Here we study in detail the following IVP

$$\begin{array}{l} \text{PDE: } \partial_t q + \lambda \partial_x q = 0 \ , \quad -\infty < x < \infty \ , \quad t > 0 \ , \\ \text{IC:} \quad q(x,0) = h(x) = \left\{ \begin{array}{ccc} 0 & \text{if} & x < -1 \ , \\ 1 - x^2 & \text{if} & -1 \le x \le 1 \ , \\ 0 & \text{if} & x > 1 \ . \end{array} \right. \end{array} \tag{11}$$

Solution According to formula (10) the solution of (11) is

$$q(x,t) = h(x - \lambda t) = \left\{ \begin{array}{ccc} 0 & \text{if} & x < -1 + \lambda t \ , \\ 1 - (x - \lambda t)^2 & \text{if} & -1 + \lambda t \le x \le 1 + \lambda t \ , \\ 0 & \text{if} & x > 1 + \lambda t \ . \end{array} \right. \tag{12}$$

Note that for a given speed λ and a chosen time t, the solution is simply a function of x, called a profile. See Fig. 4.

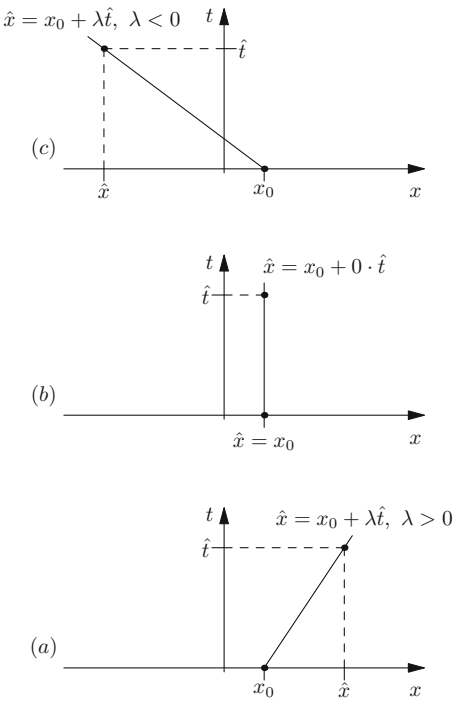

Fig. 3 The solution at point $(\hat{x}, \hat{t})$ is found by tracing the characteristic from $(\hat{x}, \hat{t})$ back to its foot x_0. There are three possibilities: (**a**) $\lambda > 0$, (**b**) $\lambda = 0$, (**c**) $\lambda < 0$

The Riemann Problem Riemann problem for the linear advection equation is the special IVP

$$\left.\begin{array}{l} \text{PDE: } \partial_t q + \lambda \partial_x q = 0\ , \quad -\infty < x < \infty\ , \quad t > 0\ , \\ \text{IC:} \quad q(x,0) = h(x) = \left\{ \begin{array}{l} q_L \text{ (constant) if } x < 0\ , \\ q_R \text{ (constant) if } x > 0\ , \end{array} \right. \end{array} \right\} \tag{13}$$

where q_L (left of 0) and q_R (right of 0) are constants.

Solution of the Riemann Problem From (10) it is obvious that the solution is

$$q(x,t) = h(x - \lambda t) = \left\{ \begin{array}{l} q_L \text{ if } x - \lambda t < 0 \leftrightarrow \frac{x}{t} < \lambda\ , \\ q_R \text{ if } x - \lambda t > 0 \leftrightarrow \frac{x}{t} > \lambda\ . \end{array} \right. \tag{14}$$

See Fig. 5.

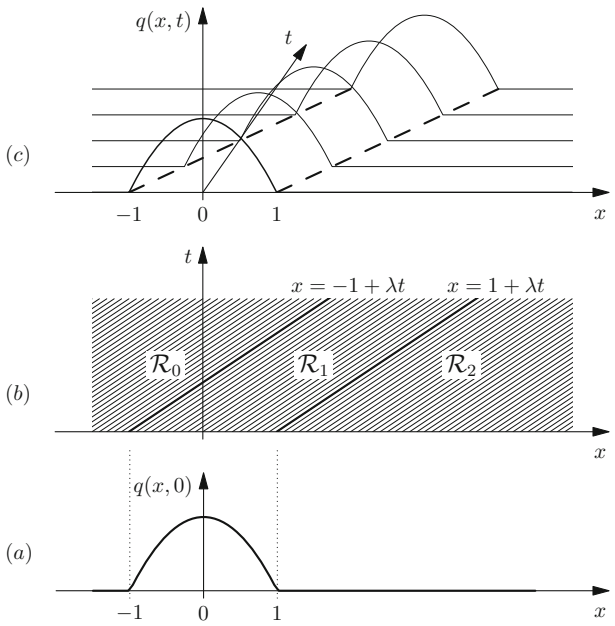

Fig. 4 Solution (12) of initial value problem (11). Frame (**a**) displays the initial condition $q(x, 0)$; frame (**b**) displays picture of characteristics in x-t space and frame (**c**) shows solution profiles $q(x, t_k)$ at different times t_k

1.2 Linear Systems

We now consider a general one-dimensional, time-dependent system of m linear hyperbolic equations with source terms, namely

$$\partial_t \mathbf{Q}(x,t) + \mathbf{A}\partial_x \mathbf{Q}(x,t) = \mathbf{S}(\mathbf{Q}(x,t)) \ . \tag{15}$$

Here $\mathbf{Q}$: unknowns, $\mathbf{A}$: matrix of coefficients (constant) and $\mathbf{S}(\mathbf{Q})$: source terms. These are given as follows

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ \dots \\ q_i \\ \dots \\ q_m \end{bmatrix} \ , \quad \mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ii} & \dots & a_{im} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mi} & \dots & a_{mm} \end{bmatrix} \ , \quad \mathbf{S}(\mathbf{Q}) = \begin{bmatrix} s_1 \\ \dots \\ s_i \\ \dots \\ s_m \end{bmatrix} \ . \tag{16}$$

Note that the linear advection equation (3) is a special case of (15).

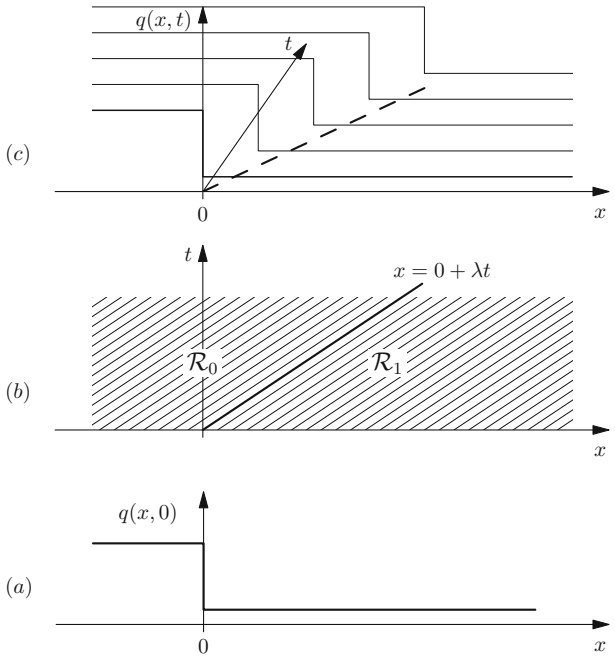

Fig. 5 Solution of Riemann problem (13). Frame (**a**) displays piece-wise constant initial condition $q(x, 0)$. Frame (**b**) displays picture of characteristics in x-t space. Frame (**c**) shows solution profiles $q(x, t_k)$ at different times t_k

Eigenvalues and Eigenvectors The **eigenvalues** of system (15) are the roots of the *characteristic polynomial*

$$P(\hat{\lambda}) \equiv Det(\mathbf{A} - \hat{\lambda}\mathbf{I}) = 0 \,. \tag{17}$$

Here $\mathbf{I}$: $m \times m$ unit matrix; $\hat{\lambda}$: a parameter; λ_i: eigenvalues, that is roots of (17), which if real numbers, are conventionally written in increasing order

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_i \leq \ldots \leq \lambda_{m-1} \leq \lambda_m \,. \tag{18}$$

A right eigenvector $\mathbf{R}_i$ of $\mathbf{A}$ corresponding to λ_i is column vector

$$\mathbf{R}_i = [r_{1i}\,, r_{2i}\,, \ldots\,, r_{ii}\,, \ldots\,, r_{mi}]^T\,, \tag{19}$$

such that

$$\mathbf{A}\mathbf{R}_i = \lambda_i \mathbf{R}_i \,. \tag{20}$$

The full set of m right eigenvectors corresponding to the eigenvalues (18) are

$$\mathbf{R}_1 \, , \mathbf{R}_2 \, , \ldots \, , \mathbf{R}_i \, , \ldots \, , \mathbf{R}_{m-1} \, , \mathbf{R}_m \, . \tag{21}$$

A left eigenvector $\mathbf{L}_i$ of $\mathbf{A}$ corresponding to λ_i is the row vector

$$\mathbf{L}_i = [l_{i1} \, , l_{i2} \, , \ldots \, , l_{ii} \, , \ldots \, , l_{im}] \, , \tag{22}$$

such that

$$\mathbf{L}_i \mathbf{A} = \lambda_i \mathbf{L}_i \, . \tag{23}$$

The m eigenvalues (18) generate corresponding m left eigenvectors

$$\mathbf{L}_1 \, , \mathbf{L}_2 \, , \ldots \, , \mathbf{L}_i \, , \ldots \, , \mathbf{L}_{m-1} \, , \mathbf{L}_m \, . \tag{24}$$

Hyperbolic System A system (15) is said to be hyperbolic if $\mathbf{A}$ has m real eigenvalues and a corresponding complete set of m linearly independent eigenvectors.

Note that for hyperbolicity, the eigenvalues are not required to be all distinct. What is important is that there is a *complete set of linearly independent eigenvectors, corresponding to the real eigenvalues*.

Strictly Hyperbolic System A hyperbolic system is said to be *strictly hyperbolic* if all eigenvalues of the system are distinct.

Weakly Hyperbolic System A system may have real but not distinct eigenvalues and still be hyperbolic if a *complete set* of linearly independent eigenvectors exists. However if all eigenvalues are real but no complete set of linearly independent eigenvectors exists then the system is called *weakly hyperbolic*, not to be mistaken with *non-strictly hyperbolic*.

Orthonormality of Eigenvectors The eigenvectors $\mathbf{L}_i$ and $\mathbf{R}_j$ are *orthonormal* if

$$\mathbf{L}_i \bullet \mathbf{R}_j = \begin{cases} 1 & \text{if } \; i = j \, , \\ 0 & \text{if } \; i \neq j \, . \end{cases} \tag{25}$$

Diagonalization and Characteristic Variables Consider $\mathbf{R} = [\mathbf{R}_1 \, , \ldots \, , \mathbf{R}_i \, , \ldots \, , \mathbf{R}_m]$: matrix whose columns are the right eigenvectors; Λ: diagonal matrix formed by eigenvalues. In full

$$\mathbf{R} = \begin{bmatrix} r_{11} & \ldots & r_{1i} & \ldots & r_{1m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{i1} & \ldots & r_{ii} & \ldots & r_{im} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{m1} & \ldots & r_{mi} & \ldots & r_{mm} \end{bmatrix} \, , \quad \Lambda = \begin{bmatrix} \lambda_1 & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \lambda_i & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots & \lambda_m \end{bmatrix} \, . \tag{26}$$

Proposition *If* $\mathbf{A}$ *is the coefficient matrix of a hyperbolic system (15) then*

$$\mathbf{A} = \mathbf{R}\Lambda\mathbf{R}^{-1} \text{ or } \Lambda = \mathbf{R}^{-1}\mathbf{A}\mathbf{R} \,. \tag{27}$$

In this case $\mathbf{A}$ is said to be *diagonalisable* and consequently system (15) is said to be *diagonalisable*. The proof is left as an exercise.

Characteristic Variables The existence of $\mathbf{R}^{-1}$ makes it possible to define the *characteristic variables* $\mathbf{C} = [c_1, c_2, \ldots, c_m]^T$ via

$$\mathbf{C} = \mathbf{R}^{-1}\mathbf{Q} \quad \leftrightarrow \quad \mathbf{Q} = \mathbf{R}\mathbf{C} \,. \tag{28}$$

Calculating the partial derivatives, recalling that the coefficient matrix is constant, we have

$$\partial_t\mathbf{Q} = \mathbf{R}\partial_t\mathbf{C} \,, \quad \partial_x\mathbf{Q} = \mathbf{R}\partial_x\mathbf{C}$$

and direct substitution of the these expressions into Eq. (15) gives

$$\mathbf{R}\partial_t\mathbf{C} + \mathbf{A}\mathbf{R}\partial_x\mathbf{C} = \mathbf{S} \,.$$

Multiplication of this equation from the left by $\mathbf{R}^{-1}$ and use of (27) gives

$$\partial_t\mathbf{C} + \Lambda\partial_x\mathbf{C} = \hat{\mathbf{S}} \,, \quad \hat{\mathbf{S}} = \mathbf{R}^{-1}\mathbf{S} \,. \tag{29}$$

This is called the *canonical form* or *characteristic form* of system (15). Assuming $\hat{\mathbf{S}} = \mathbf{0}$ and writing the equations in full, we have

$$\partial_t \begin{bmatrix} c_1 \\ \ldots \\ c_i \\ \ldots \\ c_m \end{bmatrix} + \begin{bmatrix} \lambda_1 & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \lambda_i & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots & \lambda_m \end{bmatrix} \partial_x \begin{bmatrix} c_1 \\ \ldots \\ c_i \\ \ldots \\ c_m \end{bmatrix} = \begin{bmatrix} 0 \\ \ldots \\ 0 \\ \ldots \\ 0 \end{bmatrix} \,. \tag{30}$$

Clearly, each equation i-th of this system is of the form

$$\partial_t c_i + \lambda_i\partial_x c_i = 0 \,, \quad i = 1, \ldots, m \tag{31}$$

and involves the *single unknown* $c_i(x, t)$, which is decoupled from the remaining variables. Moreover, this equations is identical to the linear advection equation (3), with characteristic speed λ_i.

We have m decoupled equations, each one defining a characteristic curve. Thus, at any chosen point $(\hat{x}, \hat{t})$ in the x-t half-plane there are m characteristic curves $x_i(t)$

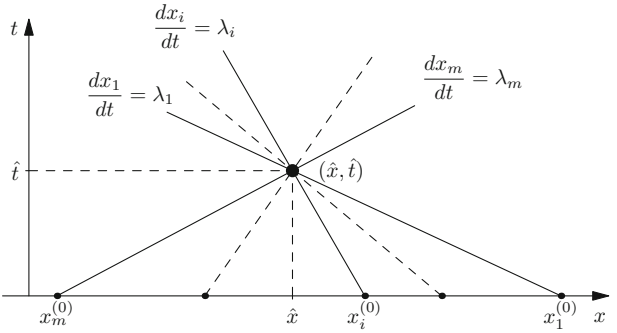

Fig. 6 The solution at a point $(\hat{x}, \hat{t})$ depends on the initial condition at the foot $x_i^{(0)}$ of each characteristic $x_i(t) = x_i^{(0)} + \lambda_i t$

passing through $(\hat{x}, \hat{t})$ and satisfying the m ODEs

$$\frac{dx_i}{dt} = \lambda_i \ , \quad \text{for } i = 1, \ldots, m \ , \tag{32}$$

as depicted in Fig. 6.

Remarks

- Each characteristic curve $x_i(t) = x_i^{(0)} + \lambda_i t$ intersects the x-axis at the point $x_i^{(0)}$, which is the *foot of the characteristic* passing through the point $(\hat{x}, \hat{t})$. The point $x_i^{(0)}$ is given as

$$x_i^{(0)} = \hat{x} - \lambda_i \hat{t} \ , \ \text{for } i = 1, 2, \ldots, m \ . \tag{33}$$

  See Fig. 6.
- Each Eq. (31) is just a linear advection equation whose solution at $(\hat{x}, \hat{t})$ is given by

$$c_i(\hat{x}, \hat{t}) = c_i^{(0)}(x_i^{(0)}) = c_i^{(0)}(\hat{x} - \lambda_i \hat{t}) \ , \ \text{for } i = 1, 2, \ldots, m \ , \tag{34}$$

  where $c_i^{(0)}(x)$ is the initial condition, at the initial time. The initial conditions for the characteristic variables are obtained from the transformation (28) applied to the initial condition $\mathbf{Q}(x, 0)$.
- Given the assumed order (18) of the distinct eigenvalues the following inequalities are satisfied.

$$x_m^{(0)} < x_{m-1}^{(0)} < \ldots < x_2^{(0)} < x_1^{(0)} \ . \tag{35}$$

Domain of Dependence The interval $[x_m^{(0)}, x_1^{(0)}]$ is called the domain of dependence of the point $(\hat{x}, \hat{t})$. See Fig. 6. The solution at $(\hat{x}, \hat{t})$ depends exclusively on initial data at points in the interval $[x_m^{(0)}, x_1^{(0)}]$. This is a distinguishing feature of hyperbolic equations. The initial data outside the domain of dependence can be changed in any manner we wish but this will not affect the solution at the point $(\hat{x}, \hat{t})$.

Proposition: The General Initial-Value Problem The solution of the general IVP for the linear homogeneous hyperbolic system

$$\left.\begin{array}{l} \text{PDEs: } \partial_t \mathbf{Q} + \mathbf{A}\partial_x \mathbf{Q} = \mathbf{0}\ , -\infty < x < \infty\ ,\ t > 0\ , \\ \text{IC:} \quad \mathbf{Q}(x,0) = \mathbf{Q}^{(0)}(x) \end{array}\right\} \tag{36}$$

is given by

$$\mathbf{Q}(x,t) = \sum_{i=1}^{m} c_i(x,t)\mathbf{R}_i\ . \tag{37}$$

The coefficient $c_i(x,t)$ of the right eigenvector $\mathbf{R}_i$ is a characteristic variable. The proof is left as an exercise.

Remarks

1. The function $c_i(x,t)$ is the coefficient of $\mathbf{R}_i$ in an *eigenvector expansion* of the solution vector $\mathbf{Q}(x,t)$.
2. Given a point (x,t) in the x-t plane, the solution $\mathbf{Q}(x,t)$ depends only on the initial data at the m points $x_0^{(i)} = x - \lambda_i t$. See Fig. 6.
3. These points are the intersections of the characteristics of speed λ_i with the x-axis.
4. Solution (37) represents superposition of m waves of unchanged shape $c_i^{(0)}(x)\mathbf{R}_i$ propagated with speed λ_i.

Proposition: The Riemann Problem Solution The solution of Riemann problem

$$\left.\begin{array}{l} \text{PDEs: } \partial_t \mathbf{Q} + \mathbf{A}\partial_x \mathbf{Q} = \mathbf{0}\ ,\ -\infty < x < \infty\ ,\ t > 0\ , \\ \text{IC:} \quad \mathbf{Q}(x,0) = \mathbf{Q}^{(0)}(x) = \begin{cases} \mathbf{Q}_\mathrm{L} \text{ if } x < 0\ , \\ \mathbf{Q}_\mathrm{R} \text{ if } x > 0\ , \end{cases} \end{array}\right\} \tag{38}$$

with $\mathbf{Q}_\mathrm{L}$ and $\mathbf{Q}_\mathrm{R}$ two constant vectors, is given by

$$\mathbf{Q}(x,t) = \sum_{i=1}^{I} c_{iR}\mathbf{R}_i + \sum_{i=I+1}^{m} c_{iL}\mathbf{R}_i\ , \tag{39}$$

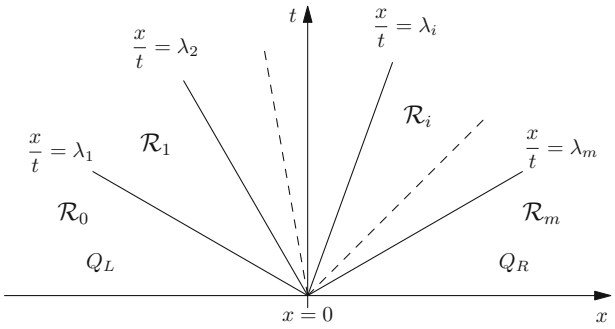

Fig. 7 Structure of the solution of the Riemann problem. There are m waves that divide the half x-t plane into $m + 1$ regions (wedges) $\mathscr{R}_i$, with $i = 0, 1, \ldots, m$

where

$$\sum_{i=1}^{m} c_{iL}\mathbf{R}_i = \mathbf{Q}_{\mathrm{L}} \ , \quad \sum_{i=1}^{m} c_{iR}\mathbf{R}_i = \mathbf{Q}_{\mathrm{R}} \tag{40}$$

and $I = I(x, t)$ is the maximum value of i for which $x - \lambda_i t > 0$. The proof is left as an exercise.

Remarks on the Solution of the Riemann Problem

1. The initial data consists of two constant vectors $\mathbf{Q}_{\mathrm{L}}$ and $\mathbf{Q}_{\mathrm{R}}$, separated by a discontinuity at $x = 0$.
2. This is a special case of IVP (36).
3. The structure of the solution of the Riemann problem (38) is depicted in Fig. 7, in the x-t plane.
4. The solution consists of a fan of m waves emanating from the origin, one wave for each eigenvalue λ_i. The speed of the wave i is the eigenvalue λ_i.
5. These m waves divide the x-t half plane into $m + 1$ constant regions

$$\mathscr{R}_i = \left\{ (x, t) / -\infty < x < \infty;\ t \geq 0;\ \lambda_i < \frac{x}{t} < \lambda_{i+1} \right\} \ , \tag{41}$$

for $i = 1, \ldots, m - 1$; $\mathscr{R}_0$ corresponds to the initial data $\mathbf{Q}_{\mathrm{L}}$ and $\mathscr{R}_m$ corresponds to the initial data $\mathbf{Q}_{\mathrm{R}}$. See Fig. 7.

Solving the Riemann problem means finding constant values for $\mathbf{Q}$ in regions $\mathscr{R}_i$ for $= 1, \ldots, m - 1$.

Corollary *The solution of the Riemann problem may be expressed as*

$$\mathbf{Q}(x, t) = \mathbf{Q}_{\mathrm{L}} + \sum_{i=1}^{I} \delta_i \mathbf{R}_i = \mathbf{Q}_{\mathrm{R}} - \sum_{i=I+1}^{m} \delta_i \mathbf{R}_i \ , \tag{42}$$

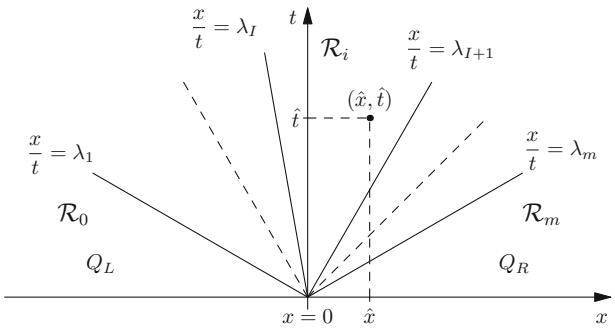

Fig. 8 The solution of the Riemann problem at a point $(\hat{x}, \hat{t})$ depends on the associated index $I = I(\hat{x}, \hat{t})$

where the coefficients $\Delta C = [\delta_1, \ldots, \delta_i, \ldots, \delta_m]^T$ *are the solution to the linear algebraic system*

$$\sum_{i=1}^{m} \delta_i \mathbf{R}_i = \Delta \mathbf{Q} \equiv \mathbf{Q}_R - \mathbf{Q}_L \, . \tag{43}$$

This form is more convenient. We only need to solve one linear system. The proof is left as an exercise. Figure 8 illustrates the solution at a point $(\hat{x}, \hat{t})$.

1.3 Non-linear Scalar Equations: Definitions and Examples

Consider the first-order PDE for the unknown function $q(x, t)$

$$\partial_t q + \partial_x f(q) = 0 \, . \tag{44}$$

This equation is called a **conservation law**, in which q is the **conserved variable**; $f(q)$ is the **flux function** or **physical flux**, a prescribed function of q. The equation is said to be written in *differential, conservative form*. One may express (44) in *quasi-linear form* as

$$\partial_t q + \lambda(q) \partial_x q = 0 \, , \qquad \lambda(q) = \frac{d}{dq} f(q) \equiv f'(q) \, . \tag{45}$$

Here $\lambda(q)$ is called **characteristic speed**.

Equations of the type (44) may be characterised by the behaviour of the flux $f(q)$ and its derivative, namely the characteristic speed $\lambda(q) = f'(q)$. There are three cases:

1. **Convex flux:** $\lambda(q)$ is a monotone *increasing* function of q, that is

$$\frac{d}{dq}\lambda(q) = \lambda'(q) = f''(q) > 0 , \quad \forall q . \tag{46}$$

2. **Concave flux:** $\lambda(q)$ is a monotone *decreasing* function of q, that is

$$\frac{d}{dq}\lambda(q) = \lambda'(q) = f''(q) < 0 , \quad \forall q . \tag{47}$$

3. **Non-convex, non-concave flux:** $\lambda(q)$ vanishes for some q, that is

$$\frac{d}{dq}\lambda(q) = \lambda'(q) = f''(q) = 0 , \quad \text{for some } q . \tag{48}$$

Example: The Inviscid Burgers' Equation

$$\left.\begin{aligned} &\partial_t q + \partial_x f(q) = 0 , \\ &f(q) = \frac{1}{2}q^2 , \\ &\lambda(q) = f'(q) = q , \quad \lambda'(q) = f''(q) = 1 > 0 , \quad \forall q . \end{aligned}\right\} \tag{49}$$

The flux is convex; the monotone increasing behaviour of $\lambda(q)$ means that larger values of q propagate faster than smaller values of q. This leads to wave distortion and shock formation. We note that the true Burgers equation is viscous, namely

$$\partial_t q + \partial_x f(q) = \alpha \partial_x^{(2)} q , \quad f(q) = \frac{1}{2}q^2 ,$$

where α is a viscosity (or diffusion) coefficient.

Example: A Traffic Flow Equation

$$\left.\begin{aligned} &\partial_t q + \partial_x f(q) = 0 , \\ &f(q) = u_{max}(1 - q/q_{max})q , \\ &\lambda(q) = f'(q) = u_{max}(1 - 2q/q_{max}) , \\ &\lambda'(q) = f''(q) = -2u_{max}/q_{max} < 0 , \quad \forall q . \end{aligned}\right\} \tag{50}$$

Here $u_{max} \geq 0$ and $q_{max} > 0$ are two constants, with $0 < q \leq q_{max}$. The flux is concave, larger values of q will propagate more slowly than smaller values of q, the opposite behaviour to that of Burgers' equation.

Solution Along Characteristics Consider the initial-value problem (or Cauchy problem)

$$\left.\begin{array}{ll} \text{PDE:} & \partial_t q + \partial_x f(q) = 0 \ , \\ \text{IC:} & q(x,0) = h(x) \ . \end{array}\right\} \tag{51}$$

As for LAE, solutions along characteristic curves $x = x(t)$, with

$$x = x_0 + \lambda(h(x_0))t \tag{52}$$

can be defined as

$$q(x,t) = h(x_0) = h(x - \lambda(h(x_0))t) \ . \tag{53}$$

Figure 9 depicts the situation.

Crossing Characteristics For non-linear equations, characteristics are no longer parallel, as in the linear case. Therefore, characteristic curves may cross, as illustrated in Fig. 10.

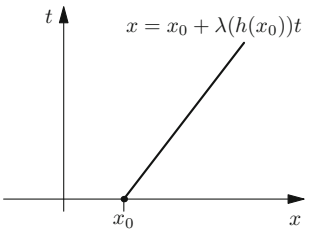

Fig. 9 Characteristic curve $x(t) = x_0 + \lambda(h(x_0))\ t$ emanating from x_0: foot of the characteristic

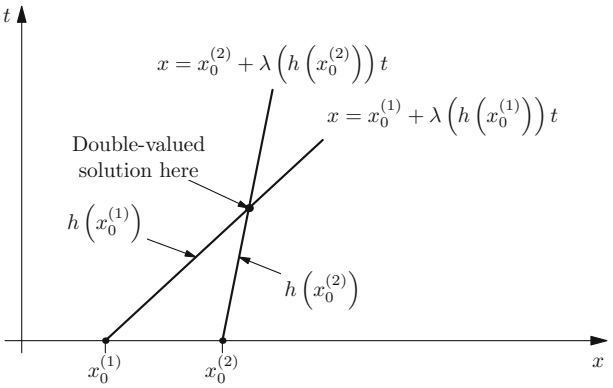

Fig. 10 Characteristics from $x_0^{(1)}$ and $x_0^{(2)}$ carry different initial values $h(x_0^{(1)})$ and $h(x_0^{(2)})$, leading to multi-valued solutions

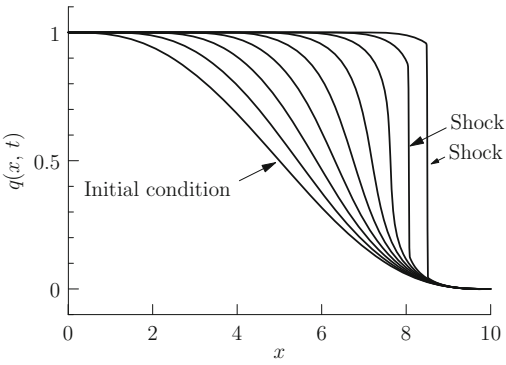

Fig. 11 Shock wave formation from smooth initial condition at time $t = 0$. Burgers' equation solved numerically with the first-order Godunov method on a very fine mesh

Shock Formation: A Numerical Example For non-linear equations, even if the initial data is continuous, discontinuities may develop in time. This is illustrated in Fig. 11 below, where a sequence of profiles corresponding to an increasing sequence of time values is shown, starting from $t = 0$, the initial condition.

The phenomenon of shock formation in non-linear equations calls for the extension of the definition of solution. To this end the equations are reformulated in terms of integral relations that no longer require continuity of the solution.

Integral Forms of the Equation Consider the general case written in differential conservative form

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = s(q(x,t)) \ . \tag{54}$$

This equation includes a source term and is thus called a **balance law**. If $s(q(x,t)) = 0$ then the equation is a conservation law.

Here we study **integral forms**, to accommodate discontinuous solutions. We shall also derive a condition to be satisfied at discontinuities. To this end we consider a **control volume** V in the x-t plane, depicted in Fig. 12, defined as

$$V = [x_L, x_R] \times [t_1, t_2] \ . \tag{55}$$

We integrate Eq. (54) in space and time in the control volume V

$$\int_{x_L}^{x_R} \int_{t_1}^{t_2} [\partial_t q(x,t) + \partial_x f(q(x,t))] \, dxdt = \int_{x_L}^{x_R} \int_{t_1}^{t_2} s(q(x,t)) \, dxdt \ . \tag{56}$$

On rearranging the space and time integrals we obtain

$$\left.\begin{aligned} \int_{x_L}^{x_R} \left[\int_{t_1}^{t_2} \partial_t q(x,t) dt \right] dx = -\int_{t_1}^{t_2} \left[\int_{x_L}^{x_R} \partial_x f(q(x,t)) dx \right] dt \\ + \int_{x_L}^{x_R} \int_{t_1}^{t_2} s(q(x,t)) dxdt \ . \end{aligned}\right\} \tag{57}$$

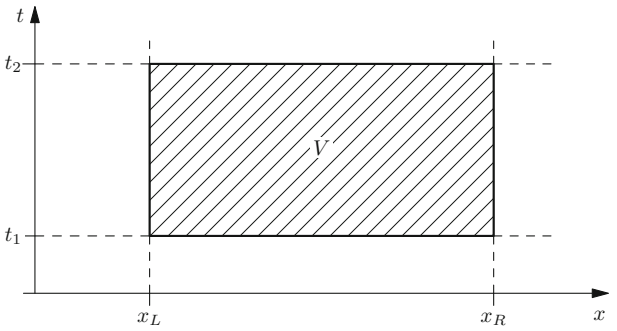

Fig. 12 Control volume $V = [x_L, x_R] \times [t_1, t_2]$ in x-t space. Equations will be integrated exactly on this volume to derive integral forms of the conservation laws

Exact space-time integration gives the **integral form** of the balance law (54), namely

$$\left.\begin{aligned}\int_{x_L}^{x_R} q(x,t_2)dx = \int_{x_L}^{x_R} q(x,t_1)dx - \left[\int_{t_1}^{t_2} f(q(x_R,t))dt - \int_{t_1}^{t_2} f(q(x_L,t))dt\right] \\ + \int_{x_L}^{x_R}\int_{t_1}^{t_2} s(q(x,t))dxdt\ .\end{aligned}\right\} \tag{58}$$

In the absence of the source term, the integral form states that *the amount of* $q(x,t)$ *in the interval* $[x_L, x_R]$ *at time* $t = t_2$ *is equal to the amount of* $q(x,t)$ *in the interval* $[x_L, x_R]$ *at time* $t = t_1$ *plus a difference of time integrals of the fluxes at the extreme points*. In the presence of a source term this statement is modified appropriately.

It is also convenient to obtain an averaged version of (58), namely

$$\left.\begin{aligned}\frac{1}{\Delta x}\int_{x_L}^{x_R} q(x,t_2)dx = \frac{1}{\Delta x}\int_{x_L}^{x_R} q(x,t_1)dx \\ -\frac{\Delta t}{\Delta x}\left[\frac{1}{\Delta t}\int_{t_1}^{t_2} f(q(x_R,t))dt - \frac{1}{\Delta t}\int_{t_1}^{t_2} f(q(x_L,t))dt\right] \\ +\frac{\Delta t}{\Delta x \Delta t}\int_{x_L}^{x_R}\int_{t_1}^{t_2} s(q(x,t))dxdt\ .\end{aligned}\right\} \tag{59}$$

The Finite Volume Formula The integral expression (59) can be written as

$$q^{new} = q^{old} - \frac{\Delta t}{\Delta x}\left[f_{right} - f_{left}\right] + \Delta t\ s_{vol}\ , \tag{60}$$

which is exact, with the following definitions

$$\left.\begin{array}{rcl} q^{new} & = & \frac{1}{\Delta x}\int_{x_L}^{x_R} q(x,t_2)dx\ , \\ q^{old} & = & \frac{1}{\Delta x}\int_{x_L}^{x_R} q(x,t_1)dx\ , \\ f_{right} & = & \frac{1}{\Delta t}\int_{t_1}^{t_2} f(q(x_R,t))dt\ , \\ f_{left} & = & \frac{1}{\Delta t}\int_{t_1}^{t_2} f(q(x_L,t))dt\ , \\ s_{vol} & = & \frac{1}{\Delta x \Delta t}\int_{x_L}^{x_R}\int_{t_1}^{t_2} s(q(x,t))dxdt\ . \end{array}\right\} \tag{61}$$

Numerical methods called finite volume methods, use the *finite volume formula* (60) to compute approximate solutions in which q^{old} is a known average of the solution at the previous time level and the remaining terms on the right hand side of (60) are found by appropriate approximations of the integrals in (61). The computational parameters Δt and Δx must be prescribed to complete the scheme to compute q^{new}.

Generalised Solutions and Rankine-Hugoniot Conditions A generalised (or weak) solution of the conservation law (54) is a function $q(x,t)$ that satisfies the integral form (58). Weak solutions admit discontinuities (shocks), which satisfy the *Rankine-Hugoniot jump condition.*

Proposition: Rankine-Hugoniot Condition A discontinuity of a weak solution of the conservation law (54), no source term, satisfies the Rankine-Hugoniot jump condition across it, namely

$$f(q(s_R,t)) - f(q(s_L,t)) = [q(s_R,t) - q(s_R,t)]\, s\ , \tag{62}$$

where $q(s_L,t)$ and $q(s_R,t)$ are limiting values from left and right of the discontinuity; $f(q(s_R,t))$ and $f(q(s_L,t))$ are the corresponding flux values and s is the speed of the discontinuity. For the proof see [1].

Summarising in order to admit discontinuous solutions one needs to formulate the equations in integral form and enforce the Rankine-Hugoniot condition across discontinuities, while in smooth parts of the solution one may formulate equations in differential form.

Example: Burgers's Equation Assume a shock wave of speed s with states q_L and q_R. The Rankine-Hugoniot condition gives

$$f(q_R) - f(q_L) = \frac{1}{2}q_R^2 - \frac{1}{2}q_L^2 = s(q_R - q_L)\ ,$$

from which the shock speed is given by

$$s = \frac{1}{2}(q_L + q_R) \ . \tag{63}$$

This is a very special case. The shock speed is a simple arithmetic average of the characteristic speeds either side of the shock.

A Non-uniqueness Example The enlarged set of solutions of the integral formulation includes smooth (classical) and discontinuous solutions. However, now the set is too large, it contains spurious, non-physical solutions. Hence this requires an admissibility criterion to discard *unphysical shocks*. To illustrate the question of non-uniqueness we consider the following example:

$$\left.\begin{array}{l} \text{PDE}: \partial_t q + \partial_x f(q) = 0 \ , \ f(q) = \frac{1}{2}q^2 \ , \\ \text{IC}: \quad q(x,0) = h(x) = \left\{ \begin{array}{l} q_L = 0 \text{ if } x < 0 \ , \\ q_R = 1 \text{ if } x > 0 \ . \end{array} \right. \end{array} \right\} \tag{64}$$

Solution 1: Rarefaction Wave One solution of the problem is the rarefaction wave (smooth)

$$q(x,t) = \left\{ \begin{array}{ll} q_{\text{L}} = 0 & \text{if} \quad x/t < 0 \ , \\ x/t & \text{if} \ \ 0 \le x/t \le 1 \ , \\ q_R = 1 & \text{if} \quad x/t > 1 \ . \end{array} \right. \tag{65}$$

Figure 13 illustrates solution and the corresponding picture of characteristics.

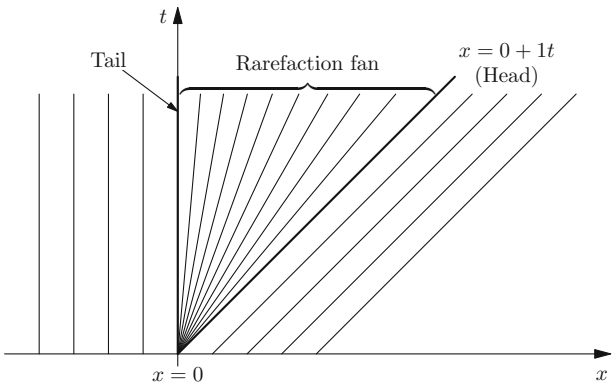

Fig. 13 Illustration of the rarefaction solution (65) to initial-value problem (64)

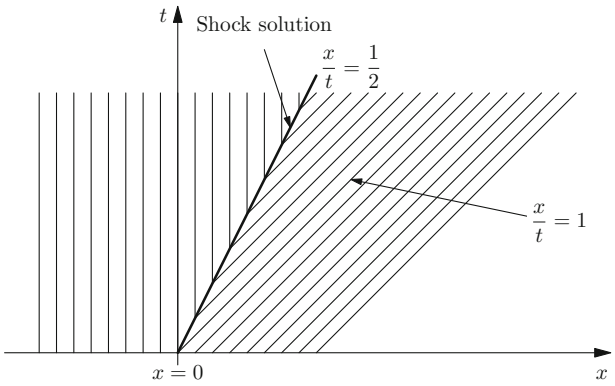

Fig. 14 Illustration of the shock solution (66) to problem (64). Characteristics diverge from the shock path

Solution 2: Shock Wave Another, discontinuous, solution (shock) is given as

$$q(x,t) = \begin{cases} 0 & \text{if } x/t < s = 1/2 \,, \\ 1 & \text{if } x/t > s = 1/2 \,. \end{cases} \tag{66}$$

Figure 14 shows the shock solution to problem (64). Note that characteristics diverge from the shock and the solution is therefore non-admissible. So the initial value problem (64) has at least two solutions.

Admissible Shocks: The Lax Entropy Condition The proposed solution (66) is not accepted as a *physical* solution. *Rarefaction shocks* are excluded. Admissible discontinuities are those arising from *compression*. This *compressibility* condition is ensured by the *Lax entropy condition:*

$$\lambda(q_L) > s > \lambda(q_R) \,. \tag{67}$$

s: shock speed, $\lambda(q_L)$ and $\lambda(q_R)$ are characteristic speeds. Note that characteristics *run into the shock*, which is *compressed* by the characteristics, see Fig. 15.

The Riemann Problem for Burgers's Equation The problem is defined as

$$\left.\begin{aligned} &\text{PDE}: \partial_t q + \partial_x f(q) = 0 \,, \quad f(q) = \tfrac{1}{2}q^2 \,, \\ &\text{IC}: \quad q(x,0) = \begin{cases} q_L & \text{if } x < 0 \,, \\ q_R & \text{if } x > 0 \,. \end{cases} \end{aligned}\right\} \tag{68}$$

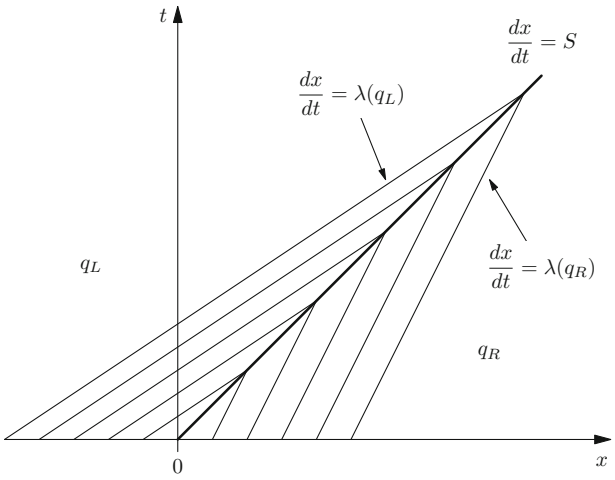

Fig. 15 Picture of characteristics for an entropy-satisfying shock. Characteristic curves run into the shock path

The solution is given by the following two cases, shock if $q_L > q_R$ and rarefaction otherwise:

$$\left.\begin{aligned} q(x,t) &= \left\{\begin{array}{ll} q_L & \text{if } x - st < 0 \\ q_R & \text{if } x - st > 0 \end{array}\right\} \text{ if } q_L > q_R \,, \\ s &= \tfrac{1}{2}(q_L + q_R) \\ q(x,t) &= \left\{\begin{array}{ll} q_L & \text{if } \dfrac{x}{t} \le q_L \\ \dfrac{x}{t} & \text{if } q_L < \dfrac{x}{t} < q_R \\ q_R & \text{if } \dfrac{x}{t} \ge q_R \end{array}\right\} \text{ if } q_L \le q_R \,. \end{aligned}\right\} \tag{69}$$

Figure 16 illustrates the solution structure for the two cases. The bottom frame shows the shock case while the top frame shows the rarefaction case.

First-Order Non-linear Systems To end this introductory section we state that the general setting is that of non-linear systems of m hyperbolic balance laws in three space dimensions, which written in differential conservative form read

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) + \partial_y \mathbf{G}(\mathbf{Q}) + \partial_z \mathbf{H}(\mathbf{Q}) = \mathbf{S}(\mathbf{Q}) \,, \tag{70}$$

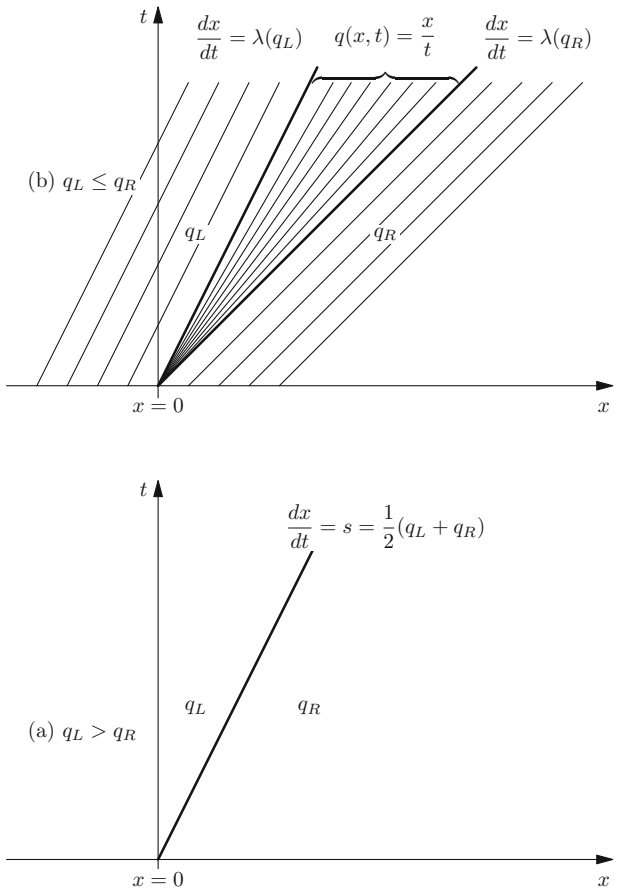

Fig. 16 Solution of the Riemann problem for the Burgers equation. Frame (**a**): shock wave if $q_L > q_R$. Frame (**b**): rarefaction wave if $q_L \leq q_R$

where

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ \cdots \\ q_m \end{bmatrix} \; ; \; \mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{bmatrix} \; ; \; \mathbf{G} = \begin{bmatrix} g_1 \\ g_2 \\ \cdots \\ g_m \end{bmatrix} \; ; \; \mathbf{H} = \begin{bmatrix} h_1 \\ h_2 \\ \cdots \\ h_m \end{bmatrix} \; ; \; \mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \cdots \\ s_m \end{bmatrix} . \tag{71}$$

Here the independent variables are: x, y, z and t. $\mathbf{Q}(x, y, z, t)$ is the vector of dependent variables, called *conserved variables*; $\mathbf{F}(\mathbf{Q})$ is the flux vector in the x-direction; $\mathbf{G}(\mathbf{Q})$ is the flux vector in the y-direction and $\mathbf{H}(\mathbf{Q})$ is the flux vector in

the z-direction; $\mathbf{S}(\mathbf{Q})$ is the vector of source terms. Fluxes and sources are prescribed functions of $\mathbf{Q}(x, y, z, t)$.

In this chapter we deal exclusively with the one-dimensional case (1D). For the more general case see for example [1–3] and [4].

1.4 Numerical Approximation of Hyperbolic Equations

Here we introduce some basic concepts on numerical discretization methods for hyperbolic equations, all based on the simplest equation. To this end we first consider the initial-boundary value problem (IBVP) for the linear advection equation

$$\left.\begin{array}{ll} \text{PDE:} & \partial_t q + \lambda \partial_x q = 0 \ , \quad x \in [a, b] \ , \ t > 0 \ , \\ \text{IC:} & q(x, 0) = h(x) \ , \quad x \in [a, b] \ , \ t = 0 \ , \\ \text{BCs:} & q(a, t) = b_L(t) \ ; \quad q(b, t) = b_R(t) \ , \ t \geq 0 \ . \end{array}\right\} \tag{72}$$

Here $[a, b]$ defines the spatial domain; $h(x)$ is the initial condition (IC) at the initial time $t = 0$, a prescribed function of x; $b_L(t)$ and $b_R(t)$ are prescribed functions of time and define boundary conditions (BCs) at $x = a$ (left) and at $x = b$ (right).

Finite Difference Discretisation One approach to solve problem (72) is by the method of finite differences, which requires the following steps:

1. *Partition of the spatial domain* $[a, b]$ into $M + 2$ equidistant points

$$x_i = a + i\Delta x \ , \quad i = 0, \ldots, M + 1 \ , \quad \Delta x = \frac{b - a}{M + 1} \ , \tag{73}$$

   where M is a chosen positive integer. See Fig. 17. There are M *interior points*: $x_1, x_2, \ldots, x_M$; and two *boundary points*: $x_0 = a$ and $x_{M+1} = b$.
2. *Partition of the temporal domain* $[0, T_{out}]$ into a set of time points, or time levels,

$$t_n = n\Delta t \ , \quad n = 0 \ , \ldots, N_{out} \ , \ldots \ . \tag{74}$$

   See Fig. 17. Here $t_0 = 0$: initial time; $T_{out} = \Delta t N_{out}$; Δt: timestep. We assume a fixed relationship between Δt and Δx of the form

$$\Delta x = \Delta t \times K \ , \quad K > 0 : \text{constant} \ . \tag{75}$$

The spatial mesh parameter Δx is chosen through the choice of M, that is, the number of interior points. There are no particular constraints in choosing M. The choice of the time step Δt is constrained by accuracy or stability considerations [4].

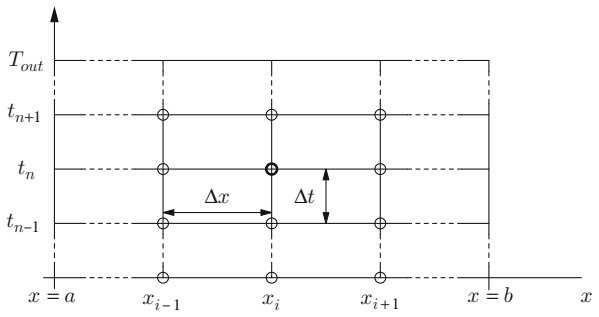

Fig. 17 Finite difference mesh defining a discrete set of points (x_i, t_n) resulting from partitions of the spatial x and temporal t domains

Discrete Values The continuous domain $[a, b] \times [0, \infty)$ has been replaced by a mesh made up of a finite number of points (x_i, t_n). We now need to replace the continuous distribution of the function $q(x, t)$ by a finite number of discrete values $q(x_i, t_n)$ associated with these points. Then in order to solve the differential equation in this discrete setting we also need to represent in discrete form the partial derivatives $\partial_t q(x, t)$ and $\partial_x q(x, t)$ in (72). Here we do so by finite difference approximations. In this manner the partial differential equation is represented by a *difference equation*, an expression that relates approximate discrete values of the solution at neighbouring points. The *differential operator* is replaced by a *numerical operator*, as we shall see.

Consider the generic point (x_i, t_n) of the mesh, as shown in Fig. 17. We seek an approximation to $q(x_i, t_n)$ and this will be denoted by q_i^n, that is

$$q_i^n \approx q(x_i, t_n) \ . \tag{76}$$

The temporal partial derivative $\partial_t q(x, t)$ can be approximated in a variety of ways, such as

$$\partial_t q(x_i, t_n) = \begin{cases} \dfrac{q(x_i, t_{n+1}) - q(x_i, t_n)}{\Delta t} + \mathscr{O}(\Delta t) \ , & \text{Forward} \ , \\ \dfrac{q(x_i, t_n) - q(x_i, t_{n-1})}{\Delta t} + \mathscr{O}(\Delta t) \ , & \text{Backward} \ , \\ \dfrac{q(x_i, t_{n+1}) - q(x_i, t_{n-1})}{2\Delta t} + \mathscr{O}(\Delta t^2) \ , & \text{Centred} \ . \end{cases} \tag{77}$$

Analogously, for the spatial partial derivative $\partial_x q(x, t)$ in (72) at the point (x_i, t_n) we write

$$\partial_x q(x_i, t_n) = \begin{cases} \dfrac{q(x_{i+1}, t_n) - q(x_i, t_n)}{\Delta x} + \mathscr{O}(\Delta x) \ , & \text{Forward} \ , \\ \dfrac{q(x_i, t_n) - q(x_{i-1}, t_n)}{\Delta x} + \mathscr{O}(\Delta x) \ , & \text{Backward} \ , \\ \dfrac{q(x_{i+1}, t_n) - q(x_{i-1}, t_n)}{2\Delta x} + \mathscr{O}(\Delta x^2) \ , & \text{Centred} \ . \end{cases} \tag{78}$$

Now, various combinations of these finite difference approximations will lead to various well-known methods.

The Method of Godunov: Finite Difference Version This method uses the following approximations to partial derivatives

$$\left.\begin{aligned} \partial_t q(x_i, t_n) &= \frac{q(x_i, t_{n+1}) - q(x_i, t_n)}{\Delta t} + \mathscr{O}(\Delta t) \ , \\ \partial_x q(x_i, t_n) &= \begin{cases} \dfrac{q(x_i, t_n) - q(x_{i-1}, t_n)}{\Delta x} + \mathscr{O}(\Delta x) \text{ if } \lambda > 0 \ , \\ \dfrac{q(x_{i+1}, t_n) - q(x_i, t_n)}{\Delta x} + \mathscr{O}(\Delta x) \text{ if } \lambda < 0 \ . \end{cases} \end{aligned}\right\} \tag{79}$$

Remarks

1. The time derivative is approximated by a *forward-in-time* formula.
2. The space derivative is approximated by a one-sided, **upwind**, space derivative discretisation, according to the sign of the wave propagation speed.
3. For linear equations the method was first proposed Courant, Isaacson and Rees (1952).
4. Godunov [5] extended the upwind method in **conservation form** to solve non-linear systems of hyperbolic equations, see Sect. 3.

The differential operator in (72) is

$$L_e(q) \equiv \partial_t q(x, t) + \lambda \partial_x q(x, t) = 0 \ , \tag{80}$$

which when applied to the point (x_i, t_n) of the mesh, for $\lambda > 0$, becomes

$$\left.\begin{aligned} L_e(q(x_i, t_n)) &= \partial_t q(x_i, t_n) + \lambda \partial_x q(x_i, t_n) \\ &= \frac{q(x_i, t_{n+1}) - q(x_i, t_n)}{\Delta t} + \mathscr{O}(\Delta t) \\ &\quad + \lambda [\frac{q(x_i, t_n) - q(x_{i-1}, t_n)}{\Delta x}] + \mathscr{O}(\Delta x) \\ &= 0 \ . \end{aligned}\right\} \tag{81}$$

Suppressing $\mathscr{O}(\Delta t) + \mathscr{O}(\Delta x)$ and replacing $q(x_i, t_n)$ by q_i^n gives

$$\frac{q_i^{n+1} - q_i^n}{\Delta t} + \lambda \left(\frac{q_i^n - q_{i-1}^n}{\Delta x} \right) = 0 \ .$$

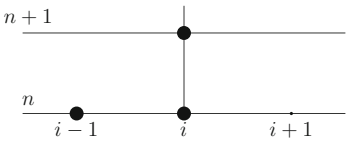

Fig. 18 Stencil for Godunov's method for positive characteristic speed λ. Note the one-sided (upwind) character of the stencil

Solving for q_i^{n+1} we obtain the numerical scheme

$$q_i^{n+1} = q_i^n - \frac{\lambda \Delta t}{\Delta x}\left(q_i^n - q_{i-1}^n\right) \ . \tag{82}$$

The **Courant-Friedrichs-Lewy number**, or the CFL number, or simply **Courant number** is defined as

$$c = \frac{\lambda \Delta t}{\Delta x} = \frac{\lambda}{\Delta x/\Delta t} \ . \tag{83}$$

This is a dimensionless quantity, it is the ratio of the speed λ in the PDE in (72) and the *mesh speed* $\Delta x/\Delta t$. Then the Godunov upwind scheme becomes

$$q_i^{n+1} = q_i^n - c\left(q_i^n - q_{i-1}^n\right) \ . \tag{84}$$

Figure 18 displays the stencil of scheme (84), which is the set of points of the mesh that contribute to the scheme

The FTCS method (Forward-in-Time Centred-in-Space) results from the following approximations to the partial derivatives

$$\left.\begin{aligned} \partial_t q(x_i, t_n) &= \frac{q(x_i, t_{n+1}) - q(x_i, t_n)}{\Delta t} + \mathscr{O}(\Delta t) \ , \\ \partial_x q(x_i, t_n) &= \frac{q(x_{i+1}, t_n) - q(x_{i-1}, t_n)}{2\Delta x} + \mathscr{O}(\Delta x^2) \ . \end{aligned}\right\} \tag{85}$$

Substituting of these into the PDE, suppressing error terms and replacing exact values by approximate values, yields

$$\frac{q_i^{n+1} - q_{i-1}^n}{\Delta t} + \lambda\left(\frac{q_{i+1}^n - q_{i-1}^n}{2\Delta x}\right) = 0 \ . \tag{86}$$

Solving for q_i^{n+1} we obtain the FTCS numerical scheme

$$q_i^{n+1} = q_i^n - \frac{1}{2}c(q_{i+1}^n - q_{i-1}^n) \ . \tag{87}$$

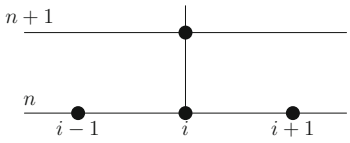

Fig. 19 Stencil for the FTCS method. Note the symmetric character of the stencil

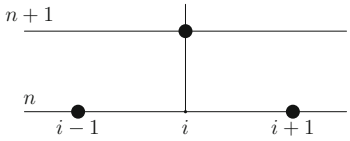

Fig. 20 Stencil for the Lax-Friedrichs method. Note the symmetry of the stencil and the missing point (x_i, t_n)

Figure 19 shows the stencil. Unfortunately, FTCS is useless; it is *unconditionally unstable*. FTCS uses the same approximation to the time derivative as the Godunov method, but the spatial derivative is approximated via a centred, second-order accurate, discretization. Naively, one would have expected a better method than Godunov's method. There are two ways to rescue FTCS. One modification results in the explicit Lax-Friedrichs scheme. The other way is to resort to an implicit version.

The Lax-Friedrichs method results from replacing q_i^n in the approximation to the time derivative of FTCS by a mean value, that is

$$q_i^n \longrightarrow \frac{1}{2}(q_{i-1}^n + q_{i+1}^n) \, .$$

Then

$$\frac{q_i^{n+1} - \frac{1}{2}(q_{i-1}^n + q_{i+1}^n)}{\Delta t} + \lambda \left(\frac{q_{i+1}^n - q_{i-1}^n}{2\Delta x} \right) = 0 \, , \tag{88}$$

yielding the Lax-Friedrichs scheme

$$q_i^{n+1} = \frac{1}{2}(1+c) q_{i-1}^n + \frac{1}{2}(1-c) q_{i+1}^n \, , \tag{89}$$

whose stencil is shown in Fig. 20.

The Lax-Wendroff Method The construction of this method follows a different approach, via the following steps:

1. The solution at (x_i, t_{n+1}) is expressed as a Taylor series in time

$$q(x_i, t_{n+1}) = q(x_i, t_n) + \Delta t \partial_t q(x_i, t_n) + \frac{1}{2} \Delta t^2 \partial_t^{(2)} q(x_i, t_n) + \mathscr{O}(\Delta t^3) \, . \tag{90}$$

2. By means of the Cauchy-Kowalewskaya method (or Lax-Wendroff procedure, as is sometimes called) one uses the PDE in (72) to replace time derivatives by

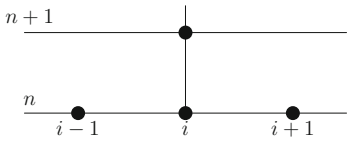

Fig. 21 Stencil for the Lax-Wendroff method. Note the symmetry of the stencil

space derivatives

$$\partial_t q(x,t) = -\lambda \partial_x q(x,t) \ , \quad \partial_t^{(2)} q(x,t) = \lambda^2 \partial_x^{(2)} q(x,t) \ . \tag{91}$$

In fact, for any order k, one can prove

$$\partial_t^{(k)} q(x,t) = (-\lambda)^k \partial_x^{(k)} q(x,t) \ . \tag{92}$$

3. By substituting (91) into (90) one obtains

$$q(x_i, t_{n+1}) = q(x_i, t_n) - \Delta t \lambda \partial_x q(x_i, t_n) + \frac{1}{2} \Delta t^2 \lambda^2 \partial_x^{(2)} q(x_i, t_n) + \mathscr{O}(\Delta t^3) \tag{93}$$

4. The spatial derivatives are approximated by centred finite differences

$$\left.\begin{aligned} \partial_x q(x_i, t_n) &= \frac{q(x_{i+1}, t_n) - q(x_{i-1}, t_n)}{2\Delta x} + \mathscr{O}(\Delta x^2) \ , \\ \partial_x^{(2)} q(x_i, t_n) &= \frac{q(x_{i+1}, t_n) - 2q(x_i, t_n) + q(x_{i-1}, t_n)}{\Delta x^2} + \mathscr{O}(\Delta x^2) \ . \end{aligned}\right\} \tag{94}$$

5. Finally, by substituting (94) into (93), neglecting truncation errors and replacing exact values $q(x_i, t_n)$ by q_i^n one obtains the Lax-Wendroff scheme

$$q_i^{n+1} = \frac{1}{2} c(1+c) q_{i-1}^n + (1 - c^2) q_i^n - \frac{1}{2} c(1-c) q_{i+1}^n \ , \tag{95}$$

whose stencil is shown in Fig. 21.

General Form of a Scheme and Examples All explicit schemes studied so far can be written in the general form

$$q_i^{n+1} = H(q_{i-l}^n, \ldots, q_i^n, \ldots, q_{i+r}^n) \ , \tag{96}$$

with l, r two non-negative integers and $H(\ldots)$ a real-valued function of $l + r + 1$ arguments and

$$q_i^n \approx q(x_i, t_n) \ , \quad q_i^n \to 0 \text{ as } |i| \to \infty \tag{97}$$

is a point-wise value that approximates the true solution $q(x, t)$ at the mesh point (x_i, t_n), with $x_i = i\Delta x$, $t_n = n\Delta t$.

Example: The Godunov Finite Difference Method When the Godunov scheme is written as in (96), we have

$$\left.\begin{array}{l} \text{For } \lambda > 0 \; H = cq_{i-1}^n + (1-c)q_i^n \; , \\ \text{For } \lambda < 0 \; H = (1+c)q_i^n - cq_{i+1}^n \; . \end{array}\right\} \tag{98}$$

Linear Schemes Linear schemes are a special class of schemes (96) for the linear advection equation in (72), of the form

$$q_i^{n+1} = \sum_{k=-l}^{k=r} b_k q_{i+k}^n \; , \tag{99}$$

in which the coefficients b_k are constant, that is, they do not depend on the solution.

Consider now two examples.

1. For the Godunov finite difference method we have two cases: For $\lambda > 0$ $l = 1$, $r = 0$, $b_{-1} = c$ and $b_0 = 1 - c$. For $\lambda < 0$ we have $l = 0$, $r = 1$, $b_0 = 1 + c$, $b_1 = -c$.
2. For the Lax-Wendroff method we have $l = 1$, $r = 1$, $b_{-1} = \frac{1}{2}(1 + c)c$, $b_0 = 1 - c^2$, $b_1 = -\frac{1}{2}(1 - c)c$.

Monotone Schemes A numerical scheme of the form (96) is called monotone if H satisfies

$$\frac{\partial}{\partial q_k^n} H(q_{i-l}^n, q_{i-l+1}^n, \ldots, q_i^n, \ldots, q_{i+r}^n) \geq 0 \; , \quad i - l \leq k \leq i + r \; . \tag{100}$$

Remark a linear scheme is monotone if and only if all its coefficients are non-negative. This follows from the definitions of linear schemes and monotonicity.

A Shortcut to Accuracy Through the Accuracy Lemma A linear scheme of the form (99) is p-th order accurate in space and time ($p \geq 0$) in the sense of local truncation error, if and only if

$$\sum_{k=-l}^{r} k^\eta b_k = (-c)^\eta \; , \quad \eta = 0, 1, \ldots, p \; , \quad c : \text{Courant Number.} \tag{101}$$

For notational convenience we introduce $0^0 = 1$.

Proof For the proof and extensions to two and three dimensions see [1].

Example: The Godunov Upwind Finite Difference Method For $\lambda > 0$ the scheme is

$$q_i^{n+1} = H(q_{i-l}^n, q_i^n) = cq_{i-l}^n + (1-c)q_i^n \ . \tag{102}$$

$l = 1, r = 0, b_{-1} = c, b_0 = 1 - c$. Then we need to verify identity (101) for all possible non-negative integer values of η.

$$\eta = 0 \ : \qquad (-1)^0 \times c + 0^0 \times (1-c) = c + 1 - c = 1 = (-c)^0 \ .$$

This merely says that the sum of the coefficients of the scheme is unity.

$$\eta = 1 \ : \qquad (-1)^1 \times c + 0^1 \times (1-c) = -c = (-c)^1 \ .$$

The Godunov scheme is first-order accurate. But just for fun we try:

$$\eta = 2 \ : \qquad (-1)^2 \times c + 0^2 \times (1-c) = c \neq (-c)^2 \ .$$

Thus the Godunov scheme is **not** second-order accurate, except for the trivial cases $c = 0$ and $c = 1$.

Godunov's Theorem [5] There are no monotone, linear schemes (99) for the linear advection equation with constant λ, of accuracy two or higher.

Proof It is sufficient to prove that there is no second order, linear, monotone method for LAE. Proceed by contradiction and assume there is a second order, linear, monotone method for LAE. From the accuracy lemma we must have:

$$s_\eta = \sum_{k=-l}^{r} k^\eta b_k = \left\{ \begin{array}{l} s_0 = 1 \ , \eta = 0 \ , \\ s_1 = -c \ , \eta = 1 \ , \\ s_2 = c^2 \ , \eta = 2 \ . \end{array} \right. \tag{103}$$

But, in particular, from (103) plus some algebraic manipulations one obtains

$$\left. \begin{array}{rl} s_2 = & \displaystyle\sum_{k=-l}^{r} k^2 b_k \\ = & \displaystyle\sum_{k=-l}^{r} (k+c)^2 b_k - 2c \sum_{k=-l}^{r} k b_k - c^2 \sum_{k=-l}^{r} b_k \\ = & \left[\displaystyle\sum_{k=-l}^{r} (k+c)^2 b_k \right] - 2cs_1 - c^2 s_0 \ . \end{array} \right\} \tag{104}$$

Use of (103) into (104) gives

$$c^2 = \left[\sum_{k=-l}^{r} (k+c)^2 b_k \right] + c^2 \ . \tag{105}$$

This implies a contradiction; for a monotone scheme all coefficients b_k are non-negative but not simultaneously zero. Thus Godunov's theorem is true □.

Consequences of Godunov's Theorem From the theorem we have that linear monotone schemes are at most first-order accurate. But first-order methods are too inaccurate to be of practical use and therefore one must search for other classes of schemes. This is down to finding ways of circumventing Godunov's theorem.The key to this lies on the assumption of linear schemes. Thus a necessary condition for a numerical scheme to be oscillation-free (without new extrema) and of high-order of accuracy (for smooth solutions) is to be non-linear. In simple terms: *Schemes must be non-linear, even when applied to linear equations.*

Recall that schemes can be expressed in **the general form** (96). In what follows we introduce other forms.

The conservative form is a particular class of schemes for hyperbolic equations and can be written in the form

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x} \left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}} \right) \ , \tag{106}$$

where $f_{i+\frac{1}{2}}$ is the *numerical flux*. See definition (60).

The Viscous Form of a Scheme This requires a function $d_{i+\frac{1}{2}}$ of $2k$ variables

$$d_{i+\frac{1}{2}} = d_{i+\frac{1}{2}}(q_{i-k+1}^n, q_{i-k+1}^n, \ldots, q_i^n, \ldots, q_{i+k}^n) \ , \tag{107}$$

such that a three-point scheme can be written as

$$q_i^{n+1} = q_i^n - \frac{1}{2}\frac{\Delta t}{\Delta x}[f(q_{i+1}^n) - f(q_{i-1}^n)] + \frac{1}{2}(d_{i+\frac{1}{2}} \Delta q_{i+\frac{1}{2}} - d_{i-\frac{1}{2}} \Delta q_{i-\frac{1}{2}}) \ . \tag{108}$$

The function $d_{i+\frac{1}{2}}$ is called the *coefficient of numerical viscosity*.

Viscous Form of a Three-Point Linear Scheme We study the viscous form a three-point linear scheme of the form

$$q_i^{n+1} = b_{-1} q_{i-1}^n + b_0 q_i^n + b_1 q_{i+1}^n \ . \tag{109}$$

The coefficients b_{-1}, b_0 and b_1 are constant. Assume the scheme to be at least first-order. Then from the accuracy lemma, see (101), we have

$$b_{-1} + b_0 + b_1 = 1 \ , \quad b_{-1} - b_1 = c \ . \tag{110}$$

System (110) gives a one-parameter family of solutions. From the first equation we introduce $d = b_{-1} + b_1 = 1 - b_0$ and thus

$$b_{-1} = \frac{1}{2}(d + c) \ , \quad b_0 = 1 - d \ , \quad b_1 = \frac{1}{2}(d - c) \ . \tag{111}$$

Now in terms of d scheme (109) becomes

$$q_i^{n+1} = q_i^n - \frac{1}{2}c(q_{i+1}^n - q_{i-1}^n) + \frac{1}{2}d(q_{i+1}^n - 2q_i^n + q_{i-1}^n) \ . \tag{112}$$

This is the *viscous form* of scheme (109) and d is the *coefficient of numerical viscosity* of the scheme.

Remarks on the Viscous Form

1. Particular values of d give particular schemes, as we shall see.
2. The stability condition becomes

$$c^2 \leq d \leq 1 \ . \tag{113}$$

3. The monotonicity condition is

$$c \leq d \leq 1 \ . \tag{114}$$

4. A truncation error analysis gives *coefficient of numerical viscosity*

$$\alpha_{visc} = \frac{1}{2}\Delta x \lambda \left(\frac{d - c^2}{c} \right) \ . \tag{115}$$

Thus effectively the coefficient d measures the truncation error of the scheme.

Proposition *The Godunov upwind scheme for the linear advection equation is the monotone scheme with the smallest truncation error. The proof is left as an exercise.*

Well-known schemes are obtained by an appropriate choice d. The following four choices for d give four well-known numerical schemes:

$$d = \begin{cases} 1 & \rightarrow \text{ Lax-Friedrichs} \ , \\ \frac{1}{2}(1 + c^2) & \rightarrow \text{ FORCE} \ , \\ |c| & \rightarrow \text{ Godunov upwind} \ , \\ c^2 & \rightarrow \text{ Lax-Wendroff} \ . \end{cases} \tag{116}$$

Figure 22 shows the coefficient of numerical viscosity for all four schemes above. The region of monotone methods is contained in the dark triangular region. Schemes outside this region are not monotone. Of the monotone methods the least accurate

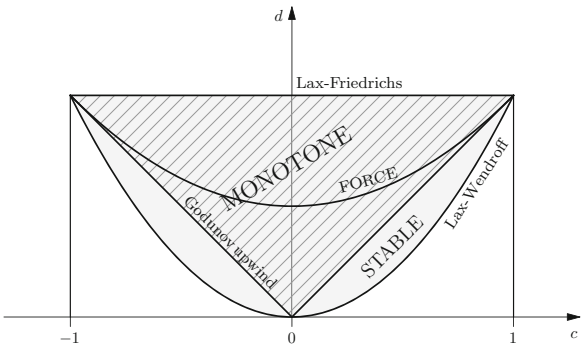

Fig. 22 Coefficient of numerical viscosity d for four schemes as functions of the Courant number c. Monote schemes lie inside the triangular region defined by the Godunov method (bottom) and the Lax-Friedrichs scheme (top)

method is the Lax-Friedrichs method and the most accurate method is the Godunov method. The FORCE method [6] is seen to lie in between these two methods. The Law-Wendroff method is the most accurate scheme of them all but it is not monotone. Stable schemes lie above the Lax-Wendroff method.

Sample Numerical Results Figures 23 and 24 show numerical results for the linear advection equation (symbols) compared to the exact solution (line) for the Lax-Friedrichs, Godunov and the Lax-Wendroff methods. Figure 23 shows the case of a smooth solution, while Fig. 24 shows the case of a discontinuous solution. For the smooth case of Fig. 23 we see that the Lax-Friedrichs method is the least accurate, just look at the peak value (unity); this is followed by the Godunov method, with Lax-Wendroff displaying the most accurate result. However, even for this smooth test problem, the Lax-Wendroff method shows *spurious oscillations* (overshoots and undershoots), mainly behind the wave. In fact the numerical solution has some negative values, which would be unphysical if $q(x, t)$ represented a concentration variable, for example.

Figure 24 shows results for the discontinuous case. Again the least accurate method is the Lax-Friedrichs method; note also the pairing of numerical values, which is a typical feature of this method. The Godunov method is a little bit more accurate but still far from representing well the square wave with its two discontinuities. The Lax-Wendroff method shows less spreading of the discontinuities (numerical diffusion) and its peak value is closer to the exact value; however the spurious oscillations, with negative values, make this method unsuitable for computing discontinuous solutions.

Note that the Lax-Friedrichs and Godunov methods do not show over and undershoots; this is due to the fact that these schemes are monotone. This property will prove useful when computing solutions to general hyperbolic systems. However, monotone methods are at most first-order accurate and thus they need to be extended to higher order of accuracy, by circumventing the Godunov theorem via the construction of non-linear methods. This subject will be addressed in Sect. 4.

Further Reading For further reading we recommend the following books [1–4].

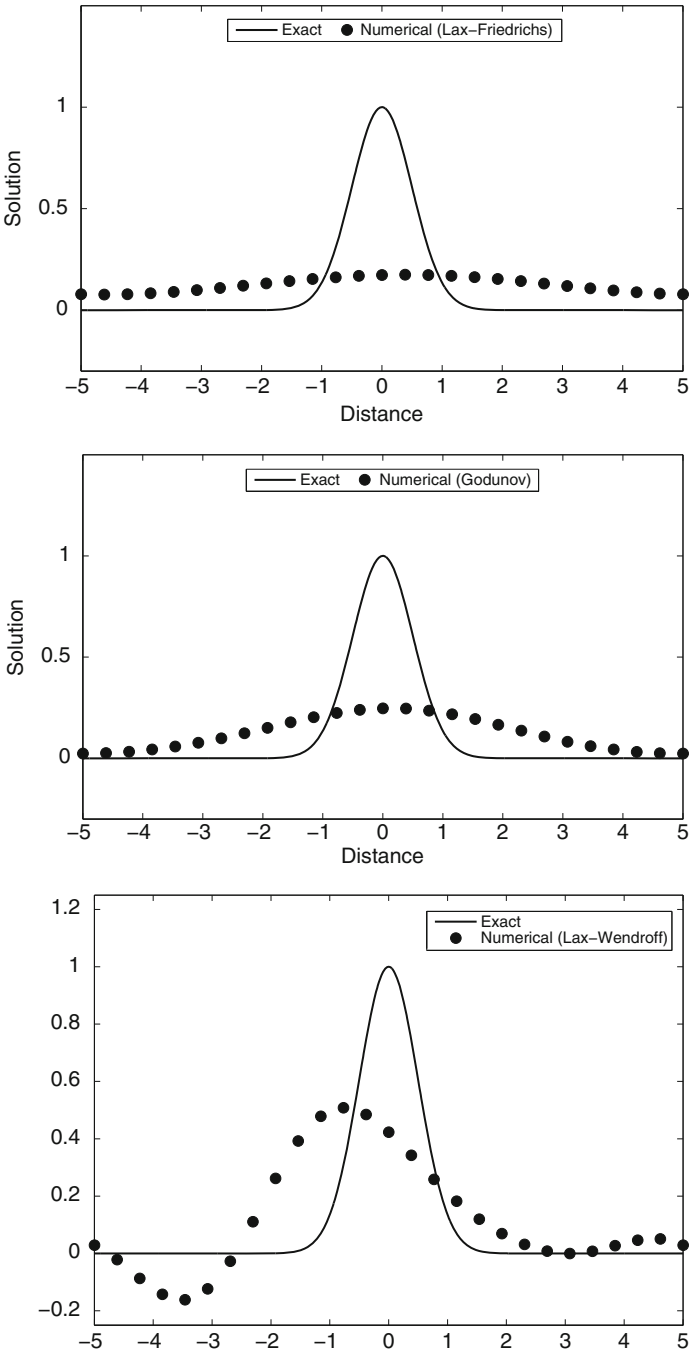

Fig. 23 Test 1 for smooth solution. Results at time $t = 100$ from the Lax-Friedrichs, Godunov and Lax-Wendroff methods. Mesh used M = 25 and Courant number CFL = 0.9

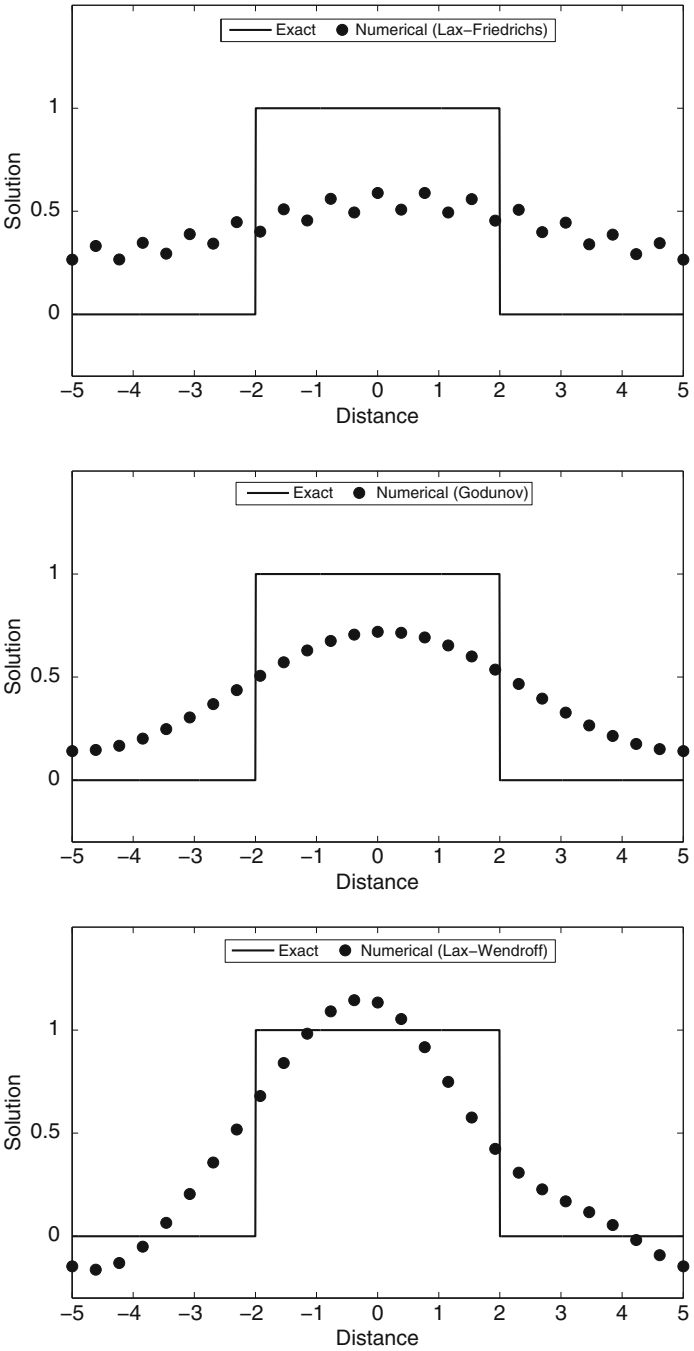

Fig. 24 Test 2 for discontinuous solution. Results at time $t = 100$ from the Lax-Friedrichs, Godunov and Lax-Wendroff methods. Mesh used M = 25 and Courant number CFL = 0.9

2 The Shallow Water Equations and the Riemann Problem

In this section we study a particular non-linear hyperbolic system of practical interest, namely the shallow water equations. We first establish the governing equations and some of their properties and then solve exactly the corresponding Riemann problem. For further reading see [2].

2.1 Equations, Properties and Wave Relations

The equation for conservation of mass reads

$$\partial_t h + \partial_x (hu) = 0 \,, \tag{117}$$

where $h(x,t)$ is water depth and $u(x,t)$ is the particle velocity. The equation for conservation of momentum reads

$$\partial_t (hu) + \partial_x (hu^2 + \frac{1}{2} g h^2) = 0 \,, \tag{118}$$

where g is the acceleration due to gravity. Recall that the *celerity* is defined as

$$a = \sqrt{gh} \,, \tag{119}$$

which is analogous to the *speed of sound* in a gas. In certain applications it is of interest to consider an additional PDE

$$\partial_t \psi + u \partial_x \psi = 0 \,. \tag{120}$$

$\psi(x,t)$ is transported with $u(x,t)$ and is often called a *passive scalar*. If we assume that solutions are smooth, then from (117) and (120) we obtain the conservation equation

$$\partial_t (h\psi) + \partial_x (h\psi u) = 0 \,. \tag{121}$$

Now the three equations of interest are (117), (118) and (121). These can be written in conservation form as

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{0} \,, \tag{122}$$

with

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} h \\ hu \\ h\psi \end{bmatrix} \,, \quad \mathbf{F}(\mathbf{Q}) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} hu \\ hu^2 + \frac{1}{2} g h^2 \\ h\psi u \end{bmatrix} \,. \tag{123}$$

Here $\mathbf{Q}$ is called the vector of conserved variables and $\mathbf{F}(\mathbf{Q})$ if the physical flux vector.

Quasi-linear Form and Eigenvalues Equation (122) can be written in quasi-linear form as follows

$$\partial_t \mathbf{Q} + \mathbf{A}(\mathbf{Q})\partial_x \mathbf{Q} = \mathbf{0}\,, \tag{124}$$

where $\mathbf{A}(\mathbf{Q})$ is the Jacobian matrix given as

$$\mathbf{A}(\mathbf{Q}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial q_1} & \dfrac{\partial f_1}{\partial q_2} & \dfrac{\partial f_1}{\partial q_3} \\ \dfrac{\partial f_2}{\partial q_1} & \dfrac{\partial f_2}{\partial q_2} & \dfrac{\partial f_2}{\partial q_3} \\ \dfrac{\partial f_3}{\partial q_1} & \dfrac{\partial f_3}{\partial q_2} & \dfrac{\partial f_3}{\partial q_3} \end{bmatrix}\,. \tag{125}$$

From (123) we have

$$\mathbf{F}(\mathbf{Q}) = \begin{bmatrix} f_1(q_1, q_2, q_3) \\ f_2(q_1, q_2, q_3) \\ f_3(q_1, q_2, q_3) \end{bmatrix} = \begin{bmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ h\psi u \end{bmatrix} = \begin{bmatrix} q_2 \\ \dfrac{q_2^2}{q_1} + \dfrac{1}{2}gq_1^2 \\ \dfrac{q_2 q_3}{q_1} \end{bmatrix}\,. \tag{126}$$

Note that each component f_k of the flux vector has been expressed in terms of the components q_j of the vector of conserved variables. This is necessary before proceeding to calculate the partial derivatives. Calculating now the partial derivatives in (125) and then using the physical variables u, a and ψ we may write the Jacobian matrix as

$$\mathbf{A}(\mathbf{Q}) = \begin{bmatrix} 0 & 1 & 0 \\ a^2 - u^2 & 2u & 0 \\ -u\psi & \psi & u \end{bmatrix}\,. \tag{127}$$

The eigenvalues are the roots of the *characteristic polynomial*

$$P(\hat{\lambda}) = Det(\mathbf{A} - \hat{\lambda}\mathbf{I}) = 0\,, \tag{128}$$

where $\mathbf{I}$ is the identity matrix and $\hat{\lambda}$ is a parameter. It is easily verified that

$$P(\hat{\lambda}) = (u - \hat{\lambda})[\hat{\lambda}(2u - \hat{\lambda}) + a^2 - u^2] = 0\,, \tag{129}$$

a cubic equation, for which three real solutions exist, and therefore the system has three real eigenvalues, namely

$$\lambda_1 = u - a \ , \quad \lambda_2 = u \ , \quad \lambda_3 = u + a \ . \tag{130}$$

Note that all three roots are distinct if $a \neq 0$.

Right Eigenvectors A right eigenvector $\mathbf{R}$ corresponding to $\hat{\lambda}$ satisfies

$$\mathbf{A}\mathbf{R} = \hat{\lambda}\mathbf{R} \ . \tag{131}$$

For a generic $\mathbf{R} = [r_1, r_2, r_3]^T$ we have

$$\left.\begin{array}{rl} r_2 & = \hat{\lambda} r_1 \ , \\ (a^2 - u^2) r_1 + 2 u r_2 & = \hat{\lambda} r_2 \ , \\ -u\psi r_1 + \psi r_2 + u r_3 & = \hat{\lambda} r_3 \ . \end{array}\right\} \tag{132}$$

To find $\mathbf{R}_i$ corresponding to λ_i we substitute λ_i into (132) and solve the resulting system for r_1, r_2 and r_3 in terms of free parameters α_i. The result is

$$\mathbf{R}_1 = \alpha_1 \begin{bmatrix} 1 \\ u - a \\ \psi \end{bmatrix} \ , \quad \mathbf{R}_2 = \alpha_2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \ , \quad \mathbf{R}_3 = \alpha_3 \begin{bmatrix} 1 \\ u + a \\ \psi \end{bmatrix} \ , \tag{133}$$

where α_1, α_2 and α_3 are arbitrary *scaling factors* which can be chosen as desired.

Left Eigenvectors To compute a *left eigenvector* $\mathbf{L} = [l_1, l_2, l_3]$ corresponding to an eigenvalue $\hat{\lambda}$, we solve the system of algebraic equations

$$\mathbf{L}\mathbf{A} = \hat{\lambda}\mathbf{L} \ . \tag{134}$$

The *left eigenvectors* of $\mathbf{A}$ are given by

$$\left.\begin{array}{l} \mathbf{L}_1 = \beta_1 \left[-(u + a) \ , \ 1 \ , \ 0 \right] \ , \\ \mathbf{L}_2 = \beta_2 \left[-\psi \ , \ 0 \ , \ 1 \right] \ , \\ \mathbf{L}_3 = \beta_3 \left[-(u - a) \ , \ 1 \ , \ 0 \right] \ , \end{array}\right\} \tag{135}$$

where the coefficients β_1, β_2, β_3 are arbitrary *scaling factors*.

Bi-orthonormality of Left and Right Eigenvectors The reader can easily verify that the right and left eigenvectors (133), (135) of the Jacobian matrix $\mathbf{A}$ are

bi-orthonormal, that is they satisfy the relations

$$\mathbf{L}_i \cdot \mathbf{R}_j = \begin{cases} 1 \text{ if } i = j \,, \\ \\ 0 \text{ if } i \neq j \,, \end{cases} \tag{136}$$

if the scaling factors are chosen thus

$$\beta_1 = \frac{1}{2a\alpha_1} \,, \quad \beta_2 = \frac{1}{\alpha_2} \,, \quad \beta_3 = -\frac{1}{2a\alpha_3} \,. \tag{137}$$

Nature of Characteristic Fields First recall that a λ_i-characteristic field is said to be *linearly degenerate* if

$$\nabla\lambda_i(\mathbf{Q}) \cdot \mathbf{R}_i(\mathbf{Q}) = 0 \,, \quad \forall \mathbf{Q} \in \Re^m \tag{138}$$

$$\nabla\lambda_i(\mathbf{Q}) = \left[\frac{\partial}{\partial q_1}\lambda_i \,, \; \frac{\partial}{\partial q_2}\lambda_i \,, \; \ldots \,, \; \frac{\partial}{\partial q_m}\lambda_i \right]^{\mathrm{T}} \,. \tag{139}$$

Now we show that *the λ_2-characteristic field is linearly degenerate*.

$$\lambda_2(\mathbf{Q}) = u = \frac{hu}{h} = \frac{q_2}{q_1}$$

$$\nabla\lambda_2(\mathbf{Q}) = \left[\frac{\partial}{\partial q_1}\lambda_2 \,, \; \frac{\partial}{\partial q_2}\lambda_2 \,, \; \frac{\partial}{\partial q_3}\lambda_2 \right]^{\mathrm{T}} = \left[-\frac{u}{h} \,, \; \frac{1}{h} \,, \; 0 \right]^{\mathrm{T}} \,.$$

Then

$$\nabla\lambda_2(\mathbf{Q}) \cdot \mathbf{R}_2(\mathbf{Q}) = 0 \tag{140}$$

for $\mathbf{Q} \in \Re^3$ and thus the λ_2-characteristic field is *linearly degenerate*.

The λ_1- and λ_3-characteristic fields are genuinely nonlinear. First recall that a λ_i-characteristic field is said to be *genuinely non-linear* if

$$\nabla\lambda_i(\mathbf{Q}) \cdot \mathbf{R}_i(\mathbf{Q}) \neq 0 \,, \quad \forall \mathbf{Q} \in \Re^m \,. \tag{141}$$

Simple calculations give

$$\nabla\lambda_1(\mathbf{Q}) \cdot \mathbf{R}_1(\mathbf{Q}) = -\tfrac{3}{2a} \neq 0 \;\text{ and }\; \nabla\lambda_3(\mathbf{Q}) \cdot \mathbf{R}_3(\mathbf{Q}) = \tfrac{3}{2a} \neq 0 \,. \tag{142}$$

Therefore the $\lambda_1(\mathbf{Q})$ and $\lambda_3(\mathbf{Q})$ characteristic fields are *genuinely non-linear*, if $a \neq 0$.

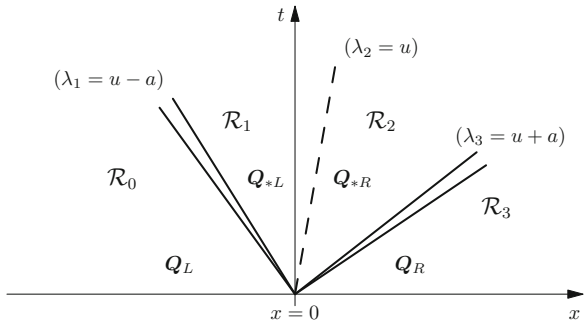

Fig. 25 Structure of the solution of the Riemann problem for the augmented 1D shallow water equations

2.2 The Riemann Problem

The Riemann problem for the shallow water equation (122) is the initial value problem

$$\left.\begin{array}{l} \text{PDEs: } \partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{0}\,, \quad -\infty < x < \infty\,,\ t > 0\,, \\ \text{ICs: } \quad \mathbf{Q}(x,0) = \begin{cases} \mathbf{Q}_L & \text{if } x < 0\,, \\ \mathbf{Q}_R & \text{if } x > 0\,. \end{cases} \end{array}\right\} \tag{143}$$

The vector of conservative variables $\mathbf{Q}$ and the vector of fluxes $\mathbf{F}(\mathbf{Q})$ are given in (123). $\mathbf{Q}_L$ and $\mathbf{Q}_R$ are two constant vectors that define the initial conditions of the problem.

Structure of the Solution of the Riemann Problem
The structure of the solution in the $x-t$ plane is shown in Fig. 25. Note that there are three wave families separating four constant regions. The outer waves are non-linear and are associated with shocks or rarefactions. The middle wave is linear (called contact discontinuity). The solution in regions $\mathscr{R}_0$ and $\mathscr{R}_3$ is known, corresponding to the initial data on the left and right respectively. The solution in regions $\mathscr{R}_1$ and $\mathscr{R}_2$ (Star Region) is unknown. The full problem of solving the Riemann problem is divided into two subproblems: Problem 1: *The Star Problem* and Problem 2: *The Complete Solution*. We start with the *The Star Problem* for which we first establish some conventional wave relations.

2.2.1 Wave Relations

Rarefactions and Generalized Riemann Invariants Generalized Riemann Invariants (GRIs) are relations that apply **across** the wave structure of simple waves in $x - t$ space. For a system of m equations consider the $\lambda_j(\mathbf{Q})$-characteristic

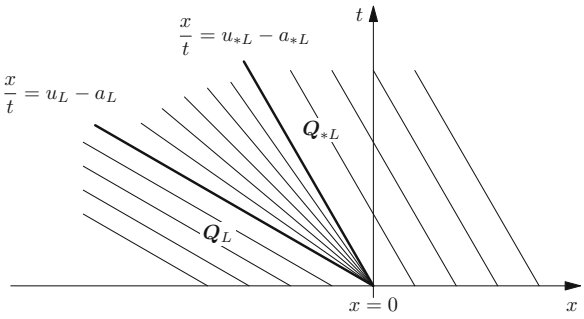

Fig. 26 Left rarefaction wave connecting states $\mathbf{Q}_L$ and $\mathbf{Q}_{*L}$. The characteristic line $x/t = u_L - a_L$ defines the **head** and the characteristic line $x/t = u_{*L} - a_{*L}$ defines the **tail**

field and the corresponding right eigenvector

$$\mathbf{R}_j = \left[r_{1j}, r_{2j}, \cdots, r_{mj}\right]^{\mathrm{T}} \ . \tag{144}$$

The GRIs apply *across* the wave structure and lead to $m - 1$ ODEs in phase space:

$$\frac{dq_1}{r_{1j}} = \frac{dq_2}{r_{2j}} = \frac{dq_3}{r_{3j}} = \cdots = \frac{dq_m}{r_{mj}} \ . \tag{145}$$

Equation (145) relate ratios of dq_i to r_{ij} and we emphasize that the ratios are to be interpreted as meaning proportionality, that is

$$dq_i \propto r_{ij} \ . \tag{146}$$

If $r_{ij} = 0$ then $dq_i = 0$ and therefore q_i does not change across the respective wave. We now apply these wave relations to study a particular class of waves.

Left Rarefaction Wave Assume a left rarefaction wave connecting $\mathbf{Q}_L$ (left) and $\mathbf{Q}_{*L}$ (right). See Fig. 26. The rarefaction wave occupies a wedge $\mathscr{R}_L$ defined as

$$\mathscr{R}_L = \left\{(x,t)/ \quad u_L - a_L \le \frac{x}{t} \le u_{*L} - a_{*L}\right\} \ , \tag{147}$$

where the characteristic line $x/t = u_L - a_L$ defines the **head** and the characteristic line $x/t = u_{*L} - a_{*L}$ defines the **tail**. $\lambda_1(\mathbf{Q})$ increases monotonically across the wave from head to tail. Application of GRIs across the λ_1-wave with $\mathbf{Q} = [h, hu, h\psi]^T$ and $\mathbf{R}_1 = [1, u - a, \psi]^T$ gives

$$\frac{dh}{1} = \frac{d(hu)}{u - a} = \frac{d(h\psi)}{\psi} \ . \tag{148}$$

From the first and third ratios $d\psi = 0$ and so across the λ_1 wave

$$\psi : \text{constant} \ . \tag{149}$$

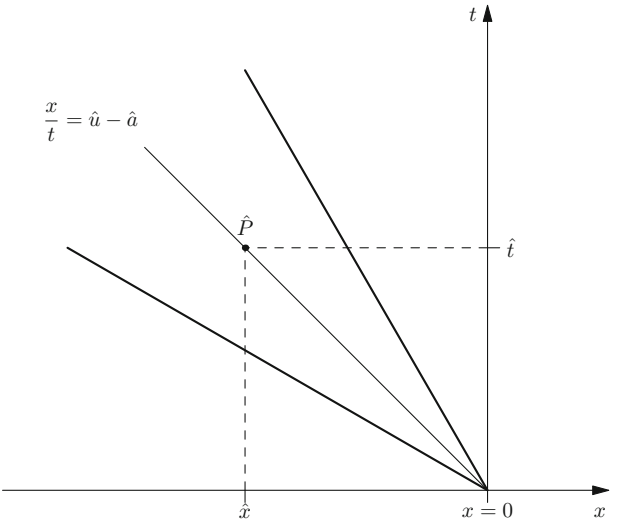

Fig. 27 Point $\hat{P} = (\hat{x}, \hat{t})$ inside left rarefaction wave. We seek the solution for the celerity a and the particle velocity u at the point $\hat{P}$ in terms of its prescribed coordinates $\hat{x}, \hat{t}$

Analogously, from first and second ratios, along with integration in phase space we obtain

$$u + 2a = \text{constant} \, . \tag{150}$$

From here we establish

$$u_{*L} + 2a_{*L} = u_L + 2a_L \, , \tag{151}$$

which we also express as

$$u_{*L} = u_L - f_L \, ; \quad f_L = 2(a_{*L} - a_L) \, . \tag{152}$$

Solution Inside a Rarefaction Consider a left rarefaction wave and a point inside the wave. See Fig. 27. The point inside the rarefaction wave is $\hat{P} = (\hat{x}, \hat{t}) \in \mathscr{R}_L$. Consider now a characteristic line through $\hat{P} = (\hat{x}, \hat{t})$ and the origin $(0, 0)$, of slope (known)

$$\frac{\hat{x}}{\hat{t}} = \hat{u} - \hat{a} \, . \tag{153}$$

The unknowns of the problem are $\hat{u} = u(\hat{x}, \hat{t})$ and $\hat{a} = a(\hat{x}, \hat{t})$, noting that h follows from a. Application of the left Riemann invariant (150) to connect the point $\hat{P}$ to the

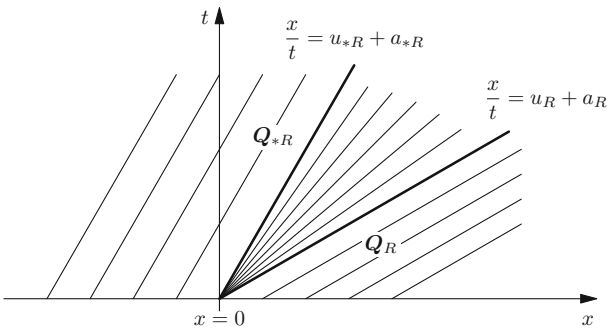

Fig. 28 Right rarefaction wave connecting states $\mathbf{Q}_{*R}$ and $\mathbf{Q}_R$. The characteristic line $x/t = u_R + a_R$ defines the head while $x/t = u_{*R} + a_{*R}$ defines the tail

left initial condition gives

$$\hat{u} + 2\hat{a} = u_L + 2a_L \, . \tag{154}$$

Equations (153) and (154) are two equations for the two unknowns $\hat{a}$ and $\hat{u}$, whose solution is

$$\hat{a}_L = a(\hat{x}, \hat{t}) = \frac{1}{3}(u_L + 2a_L - \frac{\hat{x}}{\hat{t}}) \, , \quad \hat{u}_L = u(\hat{x}, \hat{t}) = \frac{1}{3}(u_L + 2a_L + \frac{2\hat{x}}{\hat{t}}) \, . \tag{155}$$

Right Rarefaction Wave Assume a right rarefaction wave, as depicted in Fig. 28, connecting the constant states $\mathbf{Q}_{*R}$ (left) and $\mathbf{Q}_R$ (right). The wave occupies a wedge $\mathscr{R}_R$

$$\mathscr{R}_R = \left\{ (x,t)/ \quad u_{*R} + a_{*R} \leq \frac{x}{t} \leq u_R + a_R \right\} \, . \tag{156}$$

$\lambda_3(\mathbf{Q})$ is monotone. The right generalized Riemann invariant gives

$$u - 2a = \text{constant} \, , \quad \psi \, : \text{constant} \, . \tag{157}$$

From here we obtain

$$u_{*R} - 2a_{*R} = u_R - 2a_R \, , \tag{158}$$

which we also express as

$$u_{*R} = u_R + f_R \, ; \quad f_R = 2(a_{*R} - a_R) \, . \tag{159}$$

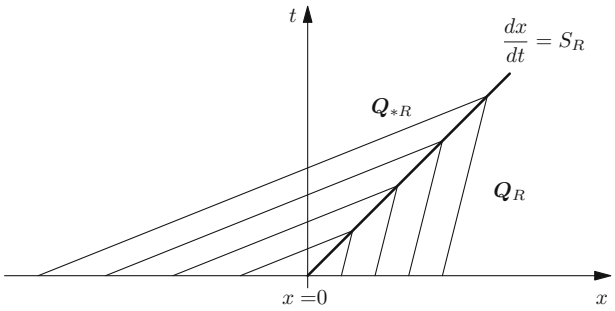

Fig. 29 Right-facing shock wave of speed S_R connecting constant states $\mathbf{Q}_R$ (ahead) and $\mathbf{Q}_{*R}$ (behind)

The solution at $\hat{P} = (\hat{x}, \hat{t}) \in \mathscr{R}_R$ inside the right rarefaction wave can easily be found to be

$$\hat{a}_R = \frac{1}{3}(-u_R + 2a_R + \frac{\hat{x}}{\hat{t}}) \ , \ \hat{u}_R = \frac{1}{3}(u_R - 2a_R + \frac{2\hat{x}}{\hat{t}}) \ . \tag{160}$$

Right-Facing Shock Wave Consider an isolated right-facing shock wave of speed S_R associated with the λ_3-characteristic field, as depicted in Fig. 29. For system (122), across the shock, the **Rankine-Hugoniot Conditions** apply and thus we have

$$S_R(\mathbf{Q}_R - \mathbf{Q}_{*R}) = \mathbf{F}(\mathbf{Q}_R) - \mathbf{F}(\mathbf{Q}_{*R}) \ . \tag{161}$$

In addition, the shock must also satisfy the Lax entropy condition

$$\lambda_3(\mathbf{Q}_{*R}) > S_R > \lambda_3(\mathbf{Q}_R) \ . \tag{162}$$

Characteristics run into the shock path, as illustrated in Fig. 29. Now we apply the transformation

$$\hat{u}_{*R} = u_{*R} - S_R \ , \quad \hat{u}_R = u_R - S_R \ , \tag{163}$$

which is illustrated in Fig. 30. In the new frame the shock propagation speed is 0 and the vectors of conserved variables and fluxes ahead of the shock are

$$\hat{\mathbf{Q}}_R = \begin{bmatrix} h_R \\ h_R\hat{u}_R \\ h_R\psi_R \end{bmatrix} \ , \quad \hat{\mathbf{F}}_R = \begin{bmatrix} h_R\hat{u}_R \\ h_R\hat{u}_R^2 + \frac{1}{2}gh_R^2 \\ h_R\hat{u}_R\psi_R \end{bmatrix} \ , \tag{164}$$

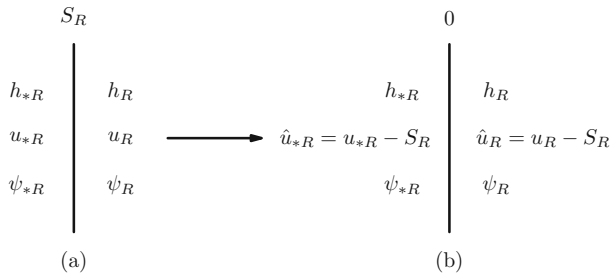

Fig. 30 Right shock wave in two frames of reference. Frame (**a**) is the original frame of reference and frame (**b**) is the moving frame of references in which the shock is stationary

while those behind the shock are

$$\hat{\mathbf{Q}}_{*R} = \begin{bmatrix} h_{*R} \\ h_{*R}\hat{u}_{*R} \\ h_{*R}\psi_{*R} \end{bmatrix} , \quad \hat{\mathbf{F}}_{*R} = \begin{bmatrix} h_{*R}\hat{u}_{*R} \\ h_{*R}\hat{u}_{*R}^2 + \frac{1}{2}gh_{*R}^2 \\ h_{*R}\hat{u}_{*R}\psi_{*R} \end{bmatrix} . \tag{165}$$

The Rankine-Hugoniot conditions in the moving frame are

$$\mathbf{F}(\hat{\mathbf{Q}}_{*R}) - \mathbf{F}(\hat{\mathbf{Q}}_R) = \mathbf{0} \times (\hat{\mathbf{Q}}_{*R} - \hat{\mathbf{Q}}_R) , \tag{166}$$

which give

$$\mathbf{F}(\hat{\mathbf{Q}}_{*R}) = \mathbf{F}(\hat{\mathbf{Q}}_R) .$$

The above flux equality written in full gives

$$\left.\begin{array}{rcl} h_{*R}\hat{u}_{*R} & = & h_R\hat{u}_R , \\ h_{*R}\hat{u}_{*R}^2 + \frac{1}{2}gh_{*R}^2 & = & h_R\hat{u}_R^2 + \frac{1}{2}gh_R^2 , \\ h_{*R}\hat{u}_{*R}\psi_{*R} & = & h_R\hat{u}_R\psi_R . \end{array}\right\} \tag{167}$$

The first equation in (167) says that the mass flux is constant across the shock, that is

$$-M_R \equiv h_{*R}\hat{u}_{*R} = h_R\hat{u}_R . \tag{168}$$

Using this into the third of Eq. (167) gives

$$\psi_{*R} = \psi_R . \tag{169}$$

That is, ψ is constant across the shock wave. Thus we only need to work with the first two equations in (167); the second one gives

$$(h_{*R}\hat{u}_{*R})\hat{u}_{*R} - (h_R\hat{u}_R)\hat{u}_R = \frac{1}{2}g(h_R^2 - h_{*R}^2) . \tag{170}$$

Use of (168) into (170) gives

$$M_R = \frac{\frac{1}{2}g(h_R^2 - h_{*R}^2)}{\hat{u}_R - \hat{u}_{*R}} \ . \tag{171}$$

But from (168) we write

$$\hat{u}_{*R} = -\frac{M_R}{h_{*R}} \ , \quad \hat{u}_R = -\frac{M_R}{h_R} \ . \tag{172}$$

Use of (172) into (171) followed by some manipulations yields

$$M_R = \sqrt{\frac{1}{2}gh_Rh_{*R}(h_R + h_{*R})} \ . \tag{173}$$

From (163)

$$u_{*R} = u_R + (\hat{u}_{*R} - \hat{u}_R) \ . \tag{174}$$

Inserting (172) into (174) followed by some algebraic manipulations gives

$$u_{*R} = u_R + f_R \ ; \quad f_R = (h_{*R} - h_R)\sqrt{\frac{1}{2}g\frac{(h_{*R} + h_R)}{h_Rh_{*R}}} \ . \tag{175}$$

From (163) we have

$$S_R = u_R - \hat{u}_R \ . \tag{176}$$

Use of (172) into (176) followed by manipulations gives

$$S_R = u_R + q_Ra_R \ , \quad q_R = \sqrt{\frac{1}{2}\frac{(h_R + h_{*R})h_{*R}}{h_R^2}} \ . \tag{177}$$

This expression relates the shock speed to the unknown depth h_{*R} behind the shock. Note that for the limiting case $h_{*R}/h_R = 1$ the shock speed coincides with the characteristic speed, that is $S_R = u + a$, as expected.

Left-Facing Shock Wave For a left-facing shock of speed S_L associated with the eigenvalue $\lambda_1 = u - a$ the analysis is similar to that of a right shock. See Fig. 31. First we define the transformation

$$\hat{u}_{*L} = u_{*L} - S_L \ ; \quad \hat{u}_L = u_L - S_L \ . \tag{178}$$

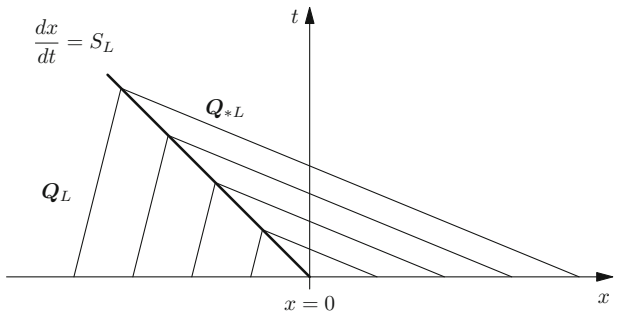

Fig. 31 Left-facing shock wave of speed S_L connecting states $\mathbf{Q}_L$ (ahead) and $\mathbf{Q}_{*L}$ (behind)

Then the Rankine-Hugoniot conditions give

$$\left.\begin{array}{rcl} h_{*L}\hat{u}_{*L} & = & h_L\hat{u}_L\,, \\ h_{*L}\hat{u}_{*L}^2 + \frac{1}{2}gh_{*L}^2 & = & h_L\hat{u}_L^2 + \frac{1}{2}gh_L^2\,, \\ h_{*L}\hat{u}_{*L}\psi_{*L} & = & h_L\hat{u}_L\psi_L\,. \end{array}\right\} \tag{179}$$

The first of Eq. (179) says that the mass flux

$$M_L \equiv h_{*L}\hat{u}_{*L} = h_L\hat{u}_L \tag{180}$$

is constant across the shock wave. Using this condition into the third of Eq. (179) gives

$$\psi_{*L} = \psi_L\,. \tag{181}$$

In other words the passive scalar ψ is constant across the right shock. Analogous manipulations to those for a right-facing shock yield

$$M_L = \sqrt{\frac{1}{2}gh_Lh_{*L}(h_L + h_{*L})} \tag{182}$$

and

$$u_{*L} = u_L - f_L\,; \quad f_L = (h_{*L} - h_L)\sqrt{\frac{1}{2}g\frac{(h_{*L} + h_L)}{h_Lh_{*L}}}\,. \tag{183}$$

This relates u_{*L} to h_{*L} via the function f_L. Also, from (178)

$$S_L = u_L - \hat{u}_L\,. \tag{184}$$

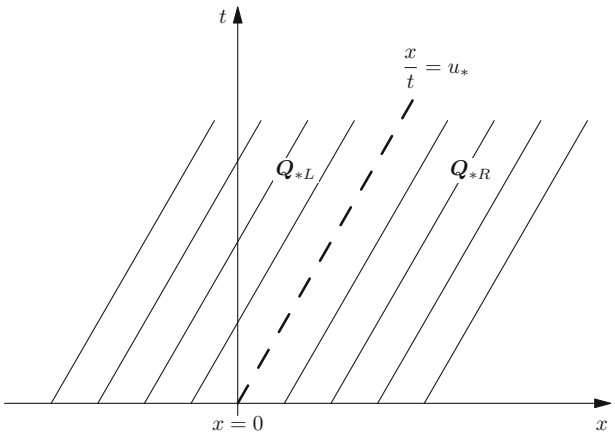

Fig. 32 Contact wave associated with the linearly degenerate field λ_2, connecting states $\mathbf{Q}_{*L}$ and $\mathbf{Q}_{*R}$. Characteristics either side of the wave are parallel to the wave, just as in the linear advection equation

Use of (180) into (184) followed by manipulations gives

$$S_L = u_L - q_L a_L \; ; \quad q_L = \sqrt{\frac{1}{2}\frac{(h_L + h_{*L})h_{*L}}{h_L^2}} \; . \tag{185}$$

This expression relates S_L to h_{*L}. Again, in the limiting case $h_{*L}/h_L = 1$ we have $S_L = u - a$.

Contact Discontinuity Wave An isolated contact discontinuity connecting the (constant) states $\mathbf{Q}_{*L}$ and $\mathbf{Q}_{*R}$ associated with the λ_2-characteristic field is depicted in Fig. 32. The wave is a single discontinuity travelling with speed u_* and characteristics either side of the discontinuity run parallel to it, namely

$$\lambda_2(\mathbf{Q}_{*L}) = u_* = \lambda_2(\mathbf{Q}_{*R}) \; . \tag{186}$$

An eigenvector analysis provides the sought jump conditions across the contact discontinuity. The right eigenvector corresponding to λ_2 is $\mathbf{R}_2 = [0, 0, 1]^T$, from which we have

$$\left.\begin{array}{l} u_{*L} = u_{*R} = u_* \; , \\ h_{*L} = h_{*R} = h_* \; , \\ \psi_{*L} \neq \psi_{*R} \; . \end{array}\right\} \tag{187}$$

Exercise Show that the above solution for the contact discontinuity wave satisfies the Rankine-Hugoniot conditions across the wave.

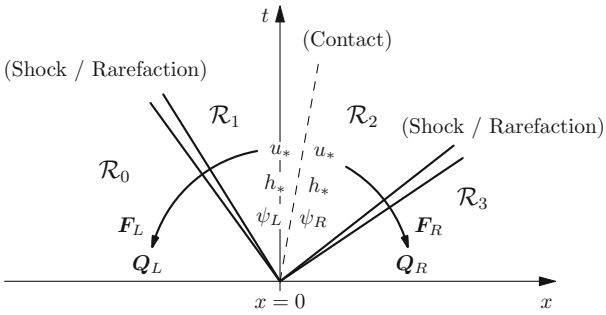

Fig. 33 Structure of the solution of the Riemann problem for the augmented shallow water equations

2.2.2 Solution of Problem 1: The Star Problem

Figure 33 depicts the structure of the solution of the Riemann problem in the $x-t$ plane. The left and right waves can be shocks or rarefactions. The velocity and depth are constant in the *Star Region*; ψ is also constant in $\mathscr{R}_1 \cup \mathscr{R}_0$ and in $\mathscr{R}_2 \cup \mathscr{R}_3$ but with a discontinuous jump across the contact wave. To find the velocity u_* and the depth h_* we first assemble together all the wave relations derived for each elementary wave in isolation. Note that the velocity u_* is connected to $\mathbf{Q}_L$ via a function f_L and that the velocity u_* is connected to $\mathbf{Q}_R$ via a function f_R; the functions f_L and f_R depend on the unknown depth h_*, the *wave type* (shock or rarefaction) and, parametrically, on the initial conditions $\mathbf{Q}_L$ and $\mathbf{Q}_R$, that is

$$f_L = f_L(h_*, w_L; \mathbf{Q}_L) \; ; \quad f_R = f_R(h_*, w_R; \mathbf{Q}_R) \; . \tag{188}$$

Here w_L and w_R denote logical variables that identify the wave type; w_K denotes either a shock or a rarefaction, for $K = L$ and $K = R$. The complete solution procedure for the *Star Problem* is then summarised in the following proposition.

Proposition *The solution h_* for the Riemann problem (143) is the root of*

$$f(h) \equiv f_L(h, w_L; h_L) + f_R(h, w_R; h_R) + \Delta u = 0 \; , \quad \Delta u \equiv u_R - u_L \; , \tag{189}$$

$$f_L(h, w_L; h_L) = \begin{cases} 2(\sqrt{gh} - \sqrt{gh_L}) & \text{if } \; h \le h_L \; (w_L\text{: rarefaction}) \; , \\ (h - h_L)\sqrt{\dfrac{1}{2} g \dfrac{(h + h_L)}{h h_L}} & \text{if } \; h > h_L \; (w_L\text{: shock}) \; , \end{cases} \tag{190}$$

$$f_R(h, w_R; h_R) = \begin{cases} 2(\sqrt{gh} - \sqrt{gh_R}) & \text{if } h \le h_R \ (w_R\text{: rarefaction}) , \\ (h - h_R)\sqrt{\dfrac{1}{2} g \dfrac{(h + h_R)}{h h_R}} & \text{if } h > h_R \ (w_R\text{: shock}) , \end{cases} \tag{191}$$

Once the depth h_* has been found the solution for the velocity u_* follows as

$$u_* = \tfrac{1}{2}(u_L + u_R) + \tfrac{1}{2}[f_R(h_*, w_R; h_R) - f_L(h_*, w_L; h_L)] . \tag{192}$$

Sketch of the Proof First note that the particle velocity u_* and depth h_* are constant across the contact discontinuity according to (187). In fact u_* and h_* are constant in the entire *Star region*. Then, the function f_L is used to relate u_* to the left initial condition $\mathbf{Q}_L$ across the left wave. In case the left wave is a shock we have the relation (183) and if it is a rarefaction we use (152). Analogously, the function f_R is used to relate u_* to the right initial condition $\mathbf{Q}_R$ across the right wave. If the right wave is a shock we have the relation (175) and if it is a rarefaction we use (159). As $u_* = u_{*L} = u_{*R}$, see (187), we can eliminate u_* resulting in Eq. (189). Then the particle velocity could be written in terms of the function f_L, for both the shock and rarefaction cases. See (183) and (152). So we could compute u_* directly from f_L once h_* is known. Alternatively, we could compute u_* directly from f_R using (175) or (159). Solution (192) results from a mean of the two possible solutions. This concludes the proof.

Iterative Solution for h_* We need to solve the algebraic non-linear equation (189) for the unknown h_* in the *Star Region*. To my knowledge, there is no general close-form solution available to this equation and therefore we must solve it *numerically* through an iteration procedure. To perform this task there are several methods available, one choice being the Newton-Raphson method

$$h^{(k+1)} = h^{(k)} - \frac{f(h^{(k)})}{f'(h^{(k)})} , \tag{193}$$

for $k = 0, 1, \ldots, K$. Here $f'(h)$ denotes the derivative of f with respect to h. The iteration (193) is stopped whenever the change in h is smaller than a prescribed small positive tolerance TOL, that is when

$$\Delta h \equiv \frac{|h^{(k+1)} - h^{(l)}|}{(h^{(k+1)} + h^{(l)})/2} < TOL . \tag{194}$$

Usually one takes $TOL = 10^{-6}$. Having formulated and solved numerically the Eq. (189) for h_*, the solution for u_* follows directly from (192).

The Two-Rarefaction Case and Guess Value The iterative procedure (193) requires a guess value $h^{(0)}$ to start the iteration. To this end we use a two-rarefaction

type approximation, as we now describe. Assume a-priori that the two non-linear waves associated with the eigenvalues λ_1 and λ_3 are both rarefaction waves. See Fig. 33. Then the functions f_L and f_R in (190), (191) respectively are those corresponding to rarefaction waves. Then (189) becomes

$$f(h) \equiv 2(a - a_L) + 2(a - a_R) + u_R - u_L = 0 \,, \tag{195}$$

which has exact solution, called the **Two-Rarefaction Solution**. For the celerity a one has

$$a_{TR} = \frac{1}{2}(a_L + a_R) - \frac{1}{4}(u_R - u_L) \,. \tag{196}$$

From the definition of celerity we obtain

$$h^{(0)} = \frac{a_{TR}^2}{g} \,, \tag{197}$$

which is used as a starting value in the iteration procedure (193).

2.2.3 Solution of Problem 2: The Complete Solution

Now we put together all the components of the solution so as to be able to compute the solution $\mathbf{Q}(x, t)$ for any given point (x, t) in the x-t half plane, $-\infty < x < \infty$ and $t \geq 0$. See Fig. 34. We call this task the solution sampling procedure in which we assume that the depth h_* and velocity u_* in the *Star Region* are already known. The solution $\mathbf{Q}(x, t)$ is sought at a specified time $\hat{t}$ for any x in a finite interval $[x_l, x_r]$ containing the full wave system, as depicted in Fig. 34. Then $\mathbf{Q}(x, \hat{t})$ is a function of x alone and gives a profile at time $\hat{t}$. To sample the solution we make use of the

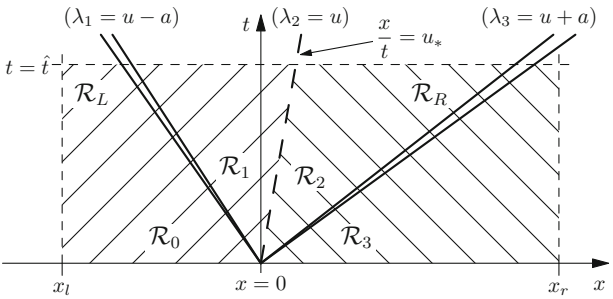

Fig. 34 Sampling the solution through the complete wave structure, at a chosen time $\hat{t}$

contact discontinuity to divide the full domain into the two subregions

$$\mathscr{R}_L = \left\{(x,t)/ \quad \frac{x}{t} \leq u_*\right\} , \quad \mathscr{R}_R = \left\{(x,t)/ \quad u_* < \frac{x}{t}\right\} . \tag{198}$$

To perform the sampling we represent the solution in terms of the vector of physical variables $\mathbf{W} = [h, u, \psi]^T$ and make use of the similarity variable

$$\xi = x/\hat{t} \tag{199}$$

to locate the sampling point and assign the corresponding solution $\mathbf{W}(\xi)$. Note that ξ has dimensions of *velocity*. There are two cases.

- **Sampling point lies to the left of the contact.** The solution $\mathbf{W}(\xi)$ for $(x, \hat{t}) \in \mathscr{R}_L$ depends on the wave type. There are two possibilities:
  Left shock. If the left wave is a shock of speed S_L, then $\mathscr{R}_L$ is again subdivided into two subregions and the solution is

$$\mathbf{W}(\xi) \equiv \begin{cases} \mathbf{W}_{*L} = [h_*, u_*, \psi_L]^T & \text{if } S_L \leq \xi \leq u_* , \\ \mathbf{W}_L = [h_L, u_L, \psi_L]^T & \text{if } \quad \xi < S_L , \end{cases} \tag{200}$$

  where the shock speed S_L is given by (185).

  Left Rarefaction If the left wave is a rarefaction then $\mathscr{R}_L$ is subdivided into three subregions and the solution is

$$\mathbf{W}(\xi) = \begin{cases} \mathbf{W}_L = [h_L, u_L, \psi_L]^T & \text{if } \quad \xi \leq u_L - a_L , \\ \mathbf{W}_{Lfan} = \left[\hat{h}_L, \hat{u}_L, \psi_L\right]^T & \text{if } u_L - a_L \leq \xi \leq u_* - a_* , \\ \mathbf{W}_{*L} = [h_*, u_*, \psi_L]^T & \text{if } \quad u_* - a_* \leq \xi \leq u_* , \end{cases} \tag{201}$$

  where $\hat{h}_L$ and $\hat{u}_L$ inside the left rarefaction are obtained from (155).

- **Sampling point lies to the right of the contact.** The solution $\mathbf{W}(\xi)$ for $(x, \hat{t}) \in \mathscr{R}_R$ depends on the type of the left wave present. Again there are two possibilities.
  Right shock. If the right wave is a shock of speed S_R, then $\mathscr{R}_R$ is again subdivided into two subregions and the solution is

$$\mathbf{W}(\xi) \equiv \begin{cases} \mathbf{W}_{*R} = [h_*, u_*, \psi_R]^T & \text{if } u_* \leq \xi \leq S_R , \\ \mathbf{W}_R = [h_R, u_R, \psi_R]^T & \text{if } \quad \xi > S_R , \end{cases} \tag{202}$$

  where the shock speed S_R is given by (177).

Table 1 Initial conditions for two Riemann problems for the shallow water equations

Test	x_0	T_{out}	h_L	u_L	h_R	u_R
1	10.0	7.0	1.0	2.5	0.1	0.0
2	25.0	2.5	1.0	10.0	1.0	-10.0

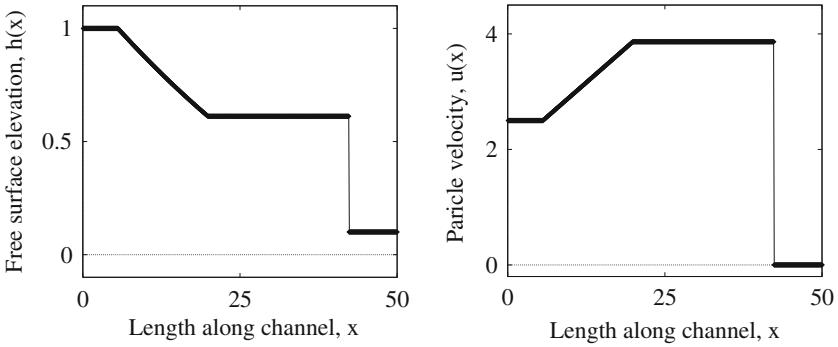

Fig. 35 Test 1. Solution profiles for h and u at the output time $T_{out} = 7.0s$. The solution consists of a left rarefaction and a right shock

Right Rarefaction If the right wave is a rarefaction then $\mathscr{R}_R$ is subdivided into three subregions and the solution is

$$\mathbf{W}(\xi) = \begin{cases} \mathbf{W}_R &= [h_R, u_R, \psi_R]^T \quad \text{if} \quad \xi > u_R + a_R \,, \\ \mathbf{W}_{Rfan} &= \left[\hat{h}_R, \hat{u}_R, \psi_R\right]^T \quad \text{if} \quad u_{*R} + a_* \le \xi \le u_R + a_R \,, \\ \mathbf{W}_{*R} &= [h_*, u_*, \psi_R]^T \quad \text{if} \quad u_* \le \xi \le u_{*R} + a_* \,, \end{cases} \tag{203}$$

where $\hat{h}_R$ and $\hat{u}_R$ inside the right rarefaction are derived from (160).

Test Problems Here we solve two specific Riemann problems for the shallow water equations in a finite channel of length 50 m. Table 1 gives the initial conditions and computational details. Column 2 gives the position of the initial discontinuity and column 3 gives the output time. The remaining columns give the initial conditions for depth h and velocity u. Note that in these examples we have not considered the equation for a passive scalar. Figures 35 and 36 show profiles for tests 1 and 2 respectively.

2.3 Concluding Remarks

In this section we have introduced the 1D shallow water equations augmented by a passive scalar, and studied its salient mathematical properties. We have solved *exactly* the corresponding Riemann problem, whose solution can be used to construct Godunov-type finite volume numerical methods and discontinuous Galerkin

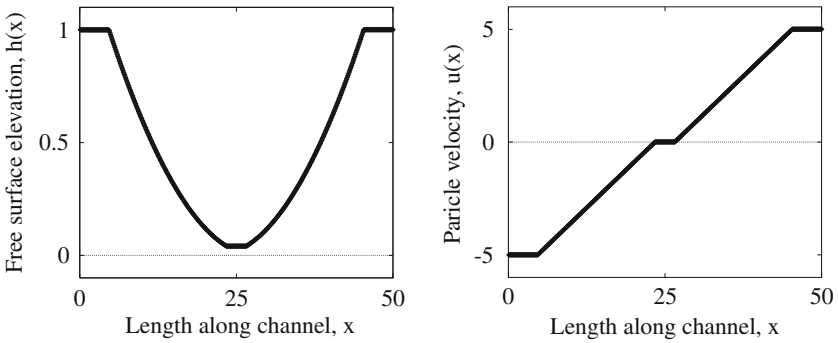

Fig. 36 Test 2. Solution profiles for h and u at the output time $T_{out} = 2.5s$. The solution consists of two rarefaction waves

finite element methods. Moreover, this exact solution can be used to construct approximate solutions (approximate Riemann solvers) to be used in numerical methods. Note also that the exact solution can be used to assess the correctness and accuracy of numerical computations intended for solving the shallow water equations.

Further reading material is found in [2] and [1] and in references therein.

3 Godunov's Method for the Shallow Water Equations

We study the Godunov method [5] as applied to a general non-linear hyperbolic system, and in particular as applied to the 1D shallow water equations. We consider two approaches for computing the Godunov flux: the first requires the calculation of *the Godunov state*, that is the state along the t-axis in the solution of the Riemann problem, see Sect. 2. Then, the numerical flux is simply the *physical flux* function evaluated at this Godunov state. In the second approach one calculates a numerical flux directly.

General Initial-Boundary Value Problem (IBVP) First we apply the Godunov's method to a generic nonlinear hyperbolic system. Consider the IBVP for any non-linear hyperbolic system

$$\left.\begin{array}{ll} \text{PDEs:} & \partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{0} \ , \quad x \in [a,b] \ , \quad t > 0 \ , \\ \text{ICs:} & \mathbf{Q}(x,0) = \mathbf{Q}^{(0)}(x) \ , \quad x \in [a,b] \ , \\ \text{BCs:} & \mathbf{Q}(a,t) = \mathbf{B}_L(t) \ , \quad \mathbf{Q}(b,t) = \mathbf{B}_R(t) \ , \quad t \geq 0 \ . \end{array}\right\} \tag{204}$$

$\mathbf{Q}(x,t)$ is the vector of conserved variables; $\mathbf{F}(\mathbf{Q})$ is the flux function, or *physical flux*; $\mathbf{Q}^{(0)}(x)$ is the initial condition; $\mathbf{B}_L(t)$ and $\mathbf{B}_R(t)$ are the boundary conditions on the left and right boundaries respectively, two prescribed functions of time.

3.1 The Finite Volume Method

Unlike the finite difference method introduced in Sect. 1, the finite volume discretisation of the domain considers a partition of the entire $x-t$ domain into space-time *finite volumes*, as in Fig. 12 of Sect. 1. In the numerical context these finite volumes are denoted as $V_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t_n, t_{n+1}]$. Figure 37c shows three consecutive finite volumes. Here $\Delta t = t_{n+1} - t_n$ denotes the time step and $\Delta = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ denotes the cell spatial size, or mesh size; $x_{i+\frac{1}{2}}$ denotes the volume interface. With this notation, the exact integration of the equations in the generic finite volume V_i gives the finite volume formula

$$\mathbf{Q}_i^{n+1} = \mathbf{Q}_i^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{i+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}\right) , \tag{205}$$

with

$$\mathbf{Q}_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{Q}(x, t_n)dx \tag{206}$$

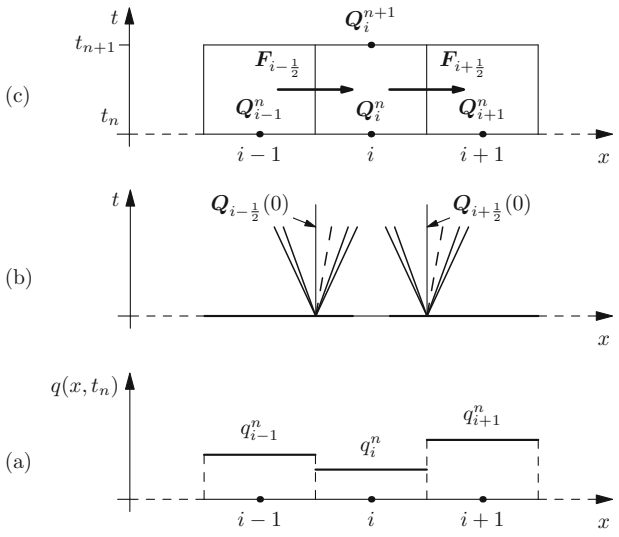

Fig. 37 Godunov's method for a hyperbolic system: (**a**) integral averages for one component q of the vector $\mathbf{Q}$ in each interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ at time t_n give piece-wise constant data; (**b**) structure of solutions of Riemann problems at intercell boundaries determined by piece wise constant data; (**c**) finite volume formula to update averages using numerical fluxes

and

$$\mathbf{F}_{i+\frac{1}{2}} \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{F}(\mathbf{Q}(x_{i+\frac{1}{2}}, t))dt \ . \tag{207}$$

See Eqs. (60) and (61) in Sect. 1. Formula (205) serves to update approximations to spatial integral averages (206) using numerical fluxes that are approximations to time integral averages (207) at the cell interface $x_{i+\frac{1}{2}}$. See Fig. 37.

3.1.1 The Godunov Flux

To define the finite volume scheme (205) we prescribe suitable approximations to the integral (207) to obtain the *numerical flux* $\mathbf{F}_{i+\frac{1}{2}}$. The Godunov upwind numerical flux $\mathbf{F}_{i+\frac{1}{2}}$ is computed from (207), making use of the solution $\mathbf{Q}_{i+\frac{1}{2}}(x/t)$ of the *local* Riemann problem

$$\left.\begin{array}{l} \text{PDEs: } \partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{0} \ , \\ \\ \text{ICs:} \quad \mathbf{Q}(x,0) = \left\{ \begin{array}{ll} \mathbf{Q}_L \equiv \mathbf{Q}_i^n & \text{if } \ x < 0 \ , \\ \\ \mathbf{Q}_R \equiv \mathbf{Q}_{i+1}^n & \text{if } \ x > 0 \ . \end{array} \right. \end{array} \right\} \tag{208}$$

The Godunov flux is computed from (207) and becomes

$$\mathbf{F}_{i+\frac{1}{2}} = \mathbf{F}(\mathbf{Q}_{i+\frac{1}{2}}(0)) \ . \tag{209}$$

$\mathbf{Q}_{i+\frac{1}{2}}(0)$ is called the *Godunov state* and results from $\mathbf{Q}_{i+\frac{1}{2}}(x/t)$ evaluated at the interface $x/t = 0$. Note that for convenience, at each interface $x_{i+\frac{1}{2}}$ and time level t_n we use *local coordinates* through a change from global to local coordinates as follows:

$$\left.\begin{array}{ll} \bar{x} = x - x_{i+\frac{1}{2}} & , \bar{t} = t - t^n \ , \\ x \in [x_i, x_{i+1}] & , \ t \in [t^n, t^{n+1}] \ , \\ \bar{x} \in [-\frac{\Delta x}{2}, \frac{\Delta x}{2}] & , \bar{t} \in [0, \Delta t] \ . \end{array} \right\} \tag{210}$$

We then use (x, t) to mean the local coordinates $(\bar{x}, \bar{t})$. See Fig. 38. In what follows we specialise Godunov's method to the shallow water equations.

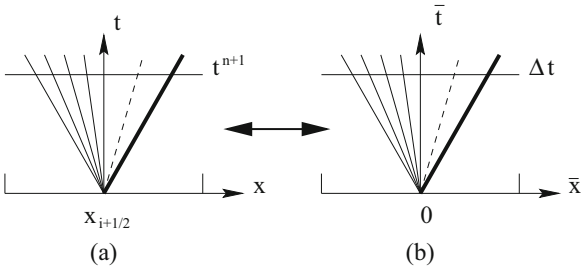

Fig. 38 Correspondence between the global (**a**) and local (**b**) frames of reference for the solution of the Riemann problem

3.1.2 Godunov Flux with the Exact Riemann Solver

One first solves the **Star problem**; the solution for h_* and u_* in the *Star Region* is found. To this end one solves the non-linear equation (using Newton-Raphson method, for example)

$$f(h) \equiv f_L(h) + f_R(h) + \Delta u = 0 \,, \quad \Delta u \equiv u_R - u_L \,. \tag{211}$$

All details on the Riemann problem are given in Sect. 2. Once the water depth $h = h_*$ has been found the velocity u_* is calculated as

$$u_* = \tfrac{1}{2}(u_L + u_R) + \tfrac{1}{2}[f_R(h_*) - f_L(h_*)] \,. \tag{212}$$

Note that not always the *Godunov state* needed for flux evaluation corresponds to the **Star State**, for which it is necessary to go through a *sampling procedure* to find the Godunov state $\mathbf{Q}_{i+\frac{1}{2}}(0)$ for flux evaluation, see Sect. 2.
If a **passive scalar** is present in the equations, then one simply chooses

$$\psi(x,t) = \begin{cases} \psi_L \ \text{ if } \ \dfrac{x}{t} < u_* \,, \\ \\ \psi_R \ \text{ if } \ \dfrac{x}{t} > u_* \,. \end{cases} \tag{213}$$

This completes the description of Godunov's flux as applied to the augmented 1D shallow water equations, using the exact Riemann solver.

In practice one resorts to approximate solution methods to find a Godunov-type flux. We next describe several approaches but before doing so we address some issues that emerge when having to choose an approximate Riemann solver. We first recall that the Godunov method is the most accurate monotone numerical method, as shown for the scalar linear case in Sect. 1. For systems, first recall that the concept of monotone method does not exist, it is only a scalar concept. Then, on the accuracy question, one knows that this depends crucially on the particular *Riemann solver* used. The exact solver is the best but at the cost of (i) complexity and (ii) computational expense. Computational expense however has to be seen in light of

efficiency, that is, in relation to the error. The computational expense of the exact Riemann solver is not excessive for systems such as for blood flow, shallow water and ideal gas dynamics. Still, approximate Riemann solvers can and are used for these systems but great care is required in choosing the appropriate approximation. The following remarks are in order:

1. Good approximate Riemann solvers are required to be:
   - **Complete:** their **wave model** contains all characteristic fields of the exact Riemann problem.
   - **Non-linear.** Linearised Riemann solvers have various defects and are thus to be avoided whenever possible.
2. The simplest Riemann solver is the Rusanov solver, as we shall see its wave model contains just one wave.
3. At the bottom of the hierarchy of numerical fluxes are **centred methods**, such as the Lax-Friedrichs and FORCE fluxes.
4. Centred, or symmetric, methods may be the simplest but not the most efficient, as we shall see later.

3.2 A Simple Linearised Riemann Solver

As an academic example here, we study a linearised Riemann solver, even if in practice, such solvers are to be avoided. We look for approximations to h_* and u_* in the *Star Region*. First we re-write the governing equations in terms of primitive, or physical, variables h, u and ψ.

$$\partial_t \mathbf{P} + \mathbf{M}(\mathbf{P})\partial_x \mathbf{P} = \mathbf{0} \,, \tag{214}$$

with

$$\mathbf{P} = \begin{bmatrix} h \\ u \\ \psi \end{bmatrix} \,, \quad \mathbf{M}(\mathbf{P}) = \begin{bmatrix} u & h & 0 \\ g & u & 0 \\ 0 & 0 & u \end{bmatrix} \,. \tag{215}$$

Denote the initial conditions for the Riemann problem as

$$\mathbf{P}_L = \begin{bmatrix} h_L \\ u_L \\ \psi_L \end{bmatrix} \,, \quad \mathbf{P}_R = \begin{bmatrix} h_R \\ u_R \\ \psi_R \end{bmatrix} \,. \tag{216}$$

Now assume $\mathbf{P}_L$ is close to $\mathbf{P}_R$ and linearise system (214) about

$$\tilde{h} = \frac{1}{2}(h_L + h_R) \,, \quad \tilde{u} = \frac{1}{2}(u_L + u_R) \tag{217}$$

so that the nonlinear system (214) becomes the linear system

$$\partial_t \mathbf{P} + \hat{\mathbf{M}} \partial_x \mathbf{P} = \mathbf{0} \ , \tag{218}$$

with constant coefficient matrix

$$\hat{\mathbf{M}} = \begin{bmatrix} \hat{u} & \hat{h} & 0 \\ g & \hat{u} & 0 \\ 0 & 0 & \hat{u} \end{bmatrix} \ . \tag{219}$$

The linear Riemann problem for (218) with initial conditions (216) is solved *exactly* by using standard methods for hyperbolic linear systems, see Sect. 1, to obtain

$$\left.\begin{aligned} h_* &= \frac{1}{2}(h_L + h_R) - \frac{1}{2}(u_R - u_L)/\bar{C} \ , \\ u_* &= \frac{1}{2}(u_L + u_R) - \frac{1}{2}(h_R - h_L)\bar{C} \ , \\ \bar{C} &= \sqrt{\frac{2g}{h_L + h_R}} \ . \end{aligned}\right\} \tag{220}$$

Remarks About the Linearised Solution

1. The solution for $\psi(x, y)$ is as given by (213), though u_* is an approximation.
2. A sampling procedure to find $\mathbf{Q}_{i+\frac{1}{2}}(0)$ for evaluating the numerical flux is required.
3. This Riemann solver is very simple but not robust enough.
4. It fails for strong rarefactions, near the vacuum state.
5. It fails for trans-critical (or sonic) flow, leading to entropy violating shocks (or rarefaction shocks).
6. This Riemann solver is **complete** but **linear**.
7. In general, linearised Riemann solvers are not recommended for practical use.

3.3 A Two-Rarefaction Riemann Solver

Starting from the exact Riemann solver, by directly assuming that both non-linear waves are rarefactions, constancy of Riemann invariants leads to

$$u_* + 2c_* = u_L + 2c_L \ , \quad u_* - 2c_* = u_R - 2c_R \ . \tag{221}$$

There follows that

$$\left.\begin{aligned} u_* &= \tfrac{1}{2}(u_L + u_R) - (c_R - c_L) \ , \\ c_* &= \tfrac{1}{2}(c_L + c_R) - \tfrac{1}{4}(u_R - u_L) \ , \\ h_* &= \tfrac{1}{g}(c_*)^2 \ . \end{aligned}\right\} \tag{222}$$

The solution for $\psi(x, y)$ is as given by (213), though u_* is an approximation. The sampling procedure to find $\mathbf{Q}_{i+\frac{1}{2}}(0)$ for evaluating the numerical flux is required; this is the same as for the exact Riemann solver. This Riemann solver is very simple, **complete** and **non-linear**; in practice it is also shown to be very robust.

3.4 The Harten-Lax-van Leer (HLL) Riemann Solver

We want to solve the Riemann problem (208) approximately with the aim of finding directly a numerical flux of the form

$$\mathbf{F}_0 = \frac{1}{T}\int_0^T \mathbf{F}(\mathbf{Q}(0, t))dt \tag{223}$$

for an arbitrary time $T > 0$ and where $\mathbf{Q}(0, t)$ is an approximate solution of the Riemann problem along the t-axis (the Godunov state); see Fig. 39. Here we construct a numerical flux following the HLL approach proposed by Harten, Lax and van Leer [7]. We first establish some useful relations obtained by applying the integral form of the conservation laws on appropriately chosen control volumes.

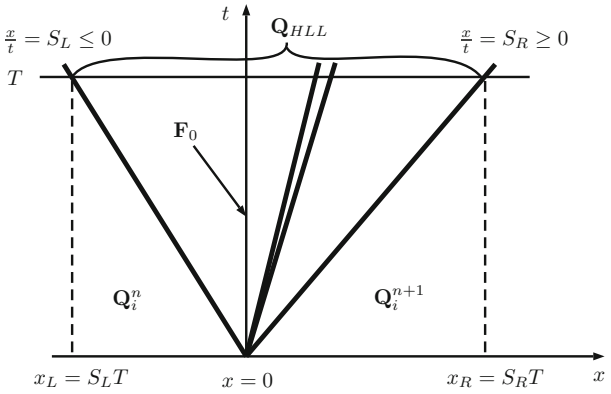

Fig. 39 Wave pattern used for the derivation of the HLL flux for a subcritical, or subsonic, wave pattern, $S_L \leq 0$ and $S_R \geq 0$

Consider the control volume $[x_L, 0] \times [0, T]$ in the space-time configuration of Fig. 39. Assume the fastest signals perturbing the constant initial states $\mathbf{Q}_L \equiv \mathbf{Q}_i^n$ and $\mathbf{Q}_R \equiv \mathbf{Q}_{i+1}^n$ emerging from the Riemann problem solution are S_L (for left travelling signals) and S_R (for right travelling signals). Assume the wave configuration is subsonic, that is $S_L \leq 0$ and $S_R \geq 0$. Then, for an arbitrary time $T > 0$ we define the distances

$$x_L = TS_L \ , \quad x_R = TS_R \ . \tag{224}$$

Applying the integral form of the conservation laws (204) in the control volume $[x_L, 0] \times [0, T]$ we obtain

$$\int_{x_L}^{0} \mathbf{Q}(x, T)dx = \int_{x_L}^{0} \mathbf{Q}(x, 0)dx + \int_{0}^{T} \mathbf{F}(\mathbf{Q}(x_L, t))dt - \int_{0}^{T} \mathbf{F}(\mathbf{Q}(0, t))dt \ . \tag{225}$$

Evaluation of the first and second terms on the right hand side gives

$$\int_{x_L}^{0} \mathbf{Q}(x, 0)dx = -S_L T\mathbf{Q}_L \ ; \quad \int_{0}^{T} \mathbf{F}(\mathbf{Q}(x_L, t))dt = T\mathbf{F}(\mathbf{Q}_L) \ . \tag{226}$$

Inserting these into (225) and dividing through by T gives

$$\mathbf{F}_0 = \frac{1}{T}\int_{0}^{T} \mathbf{F}(\mathbf{Q}(0, t))dt = -S_L\mathbf{Q}_L + \mathbf{F}(\mathbf{Q}_L) - \frac{1}{T}\int_{x_L}^{0} \mathbf{Q}(x, T))dx \ . \tag{227}$$

To define $\mathbf{F}_0$ approximately it is sufficient to find an approximation to the integral on the right hand side of (227). This is accomplished by finding an approximate state $\mathbf{Q}(x, T)$ by adopting an approach analogous to the Lax-Wendroff method, or to the Godunov centred method; see [1]. Applying the integral form (225) of the conservation laws (204) in the control volume $[x_L, x_R] \times [0, T]$, see Fig. 39, we obtain

$$\int_{x_L}^{x_R} \mathbf{Q}(x, T)dx = \int_{x_L}^{x_R} \mathbf{Q}(x, 0)dx + \int_{0}^{T} \mathbf{F}(\mathbf{Q}(x_L, t))dt - \int_{0}^{T} \mathbf{F}(\mathbf{Q}(x_R, t))dt \ . \tag{228}$$

The first term on the right hand side gives

$$\int_{x_L}^{x_R} \mathbf{Q}(x, 0)dx = -S_L T\mathbf{Q}_L + S_R T\mathbf{Q}_R \ . \tag{229}$$

Substitution of this expression into (228) and evaluation of the integrals gives

$$\int_{x_L}^{x_R} \mathbf{Q}(x, T)dx = T[S_R\mathbf{Q}_R - S_L\mathbf{Q}_L + \mathbf{F}(\mathbf{Q}_L) - \mathbf{F}(\mathbf{Q}_R)] \ . \tag{230}$$

On division through by $x_R - x_L = T(S_R - S_L)$ we obtain the averaged state

$$\mathbf{Q}^{HLL} = \frac{1}{(x_R - x_L)} \int_{x_L}^{x_R} \mathbf{Q}(x, T)dx = \frac{S_R\mathbf{Q}_R - S_L\mathbf{Q}_L + \mathbf{F}(\mathbf{Q}_L) - \mathbf{F}(\mathbf{Q}_R)}{S_R - S_L} . \tag{231}$$

We now use the state $\mathbf{Q}^{HLL}$ to evaluate the integral on the right hand side of (227). The resulting intercell flux is

$$\mathbf{F}_0 = \frac{S_R\mathbf{F}(\mathbf{Q}_L) - S_L\mathbf{F}(\mathbf{Q}_R) + S_LS_R(\mathbf{Q}_R - \mathbf{Q}_L)}{S_R - S_L} . \tag{232}$$

The HLL Flux Finally the HLL flux for the approximate Godunov method is

$$\mathbf{F}^{HLL}_{i+\frac{1}{2}} = \begin{cases} \mathbf{F}_L & \text{if} \quad 0 \le S_L , \\ \dfrac{S_R\mathbf{F}(\mathbf{Q}_L) - S_L\mathbf{F}(\mathbf{Q}_R) + S_LS_R(\mathbf{Q}_R - \mathbf{Q}_L)}{S_R - S_L} & \text{if} \quad S_L \le 0 \le S_R , \\ \mathbf{F}_R & \text{if} \quad 0 \ge S_R . \end{cases} \tag{233}$$

To complete the HLL scheme it is necessary to find estimates for S_L and S_R. In [2], the following estimates are suggested

$$S_L = u_L - q_Lc_L , \quad S_R = u_R + q_Rc_R . \tag{234}$$

Here q_K $(K = L, R)$ are obtained according to the type of non-linear waves present

$$q_K = \begin{cases} \sqrt{\dfrac{1}{2}\dfrac{(\bar{h}_* + h_K)\bar{h}_*}{h_K^2}} & \text{if} \quad \bar{h}_* > h_K , \\ 1 & \text{if} \quad \bar{h}_* \le h_K . \end{cases} \tag{235}$$

The scheme is complete by defining $\bar{h}_* \approx h_*$, the depth in the Star Region. Here we suggest to use the simple but robust estimate from (222). Given wave speed estimates S_L and S_R, HLL is most easily implemented noting also that HLL is a **non-linear** Riemann solver and entropy satisfying. HLL is **complete** but only for systems of two equations. For larger systems HLL is **incomplete**.

HLL Rusanov and Lax-Friedrichs Schemes Well-known methods can be derived from HLL, as a special cases. For example, **The Rusanov flux** [8] can be derived

from HLL by assuming $S^+ = S_R$ and $S_L = -S^+$. Then we obtain

$$\mathbf{F}^{Rus}_{i+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}(\mathbf{Q}_L) + \mathbf{F}(\mathbf{Q}_R)\right] - \frac{1}{2}S^+(\mathbf{Q}_R - \mathbf{Q}_L)\ . \tag{236}$$

The Rusanov scheme is the simplest upwind method, it has a 1-wave model and is non-linear. But obviously the Rusanov method is incomplete for any system of equations.

The well-known Lax-Friedrichs method can also be derived from HLL and more specifically from Rusanov by choosing $S^+ = \frac{\Delta x}{\Delta t}$, producing the Lax-Friedrichs flux

$$\mathbf{F}^{LF}_{i+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}(\mathbf{Q}_L) + \mathbf{F}(\mathbf{Q}_R)\right] - \frac{1}{2}\frac{\Delta x}{\Delta t}(\mathbf{Q}_R - \mathbf{Q}_L)\ . \tag{237}$$

The wave model of the Lax-Friedrichs method has zero waves. It is the most diffusive (most inaccurate) stable method for hyperbolic equations. I would not recommend its use for practical computations.

3.5 The HLLC Riemann Solver

HLL ignores intermediate waves in systems of three or more equations, leading to excessive numerical dissipation for these waves. A possible improvement, called HLLC, was first proposed by Toro and collaborators in 1992 [9]; see also [10] and [11]. The HLLC approximate Riemann solver is a modification of HLL that accounts for intermediate waves in the solution of the Riemann problem. See Fig. 40.

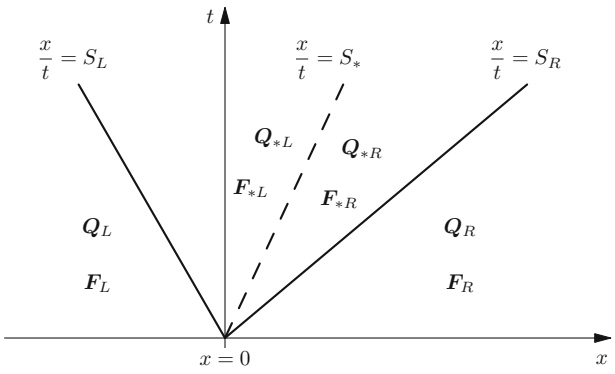

Fig. 40 Assumed wave pattern for the HLLC Riemann solver. The Star Region contains two sub-regions separated by the intermediate wave

Consider the wave pattern depicted in Fig. 40, where an intermediate wave of speed S_* is present. Application of the integral form of the conservation laws in $[x_L, 0] \times [0, T]$ and in $[0, x_R] \times [0, T]$ yield

$$\mathbf{F}_{*L} = \mathbf{F}_L + S_L(\mathbf{Q}_{*L} - \mathbf{Q}_L) \ , \quad \mathbf{F}_{*R} = \mathbf{F}_R + S_R(\mathbf{Q}_{*R} - \mathbf{Q}_R) \ . \tag{238}$$

Here there are two vector equations for four unknown vectors $\mathbf{Q}_{*L}$, $\mathbf{Q}_{*R}$, $\mathbf{F}_{*L}$ and $\mathbf{R}_{*R}$. To solve this overdetermined algebraic system we make the following additional assumptions

$$h_{*L} = h_{*R} = h_* \ , \quad u_{*L} = u_{*R} = u_* = S_* \ . \tag{239}$$

As a matter of fact the above assumptions are true for the exact Riemann solver, as seen in Sect. 2. From the first component of the first vector equation in (238) we write

$$h_* u_* = h_L u_L + S_L(h_* - h_L) \ . \tag{240}$$

From the first component of the second vector equation in (238) we write

$$h_* u_* = h_R u_R + S_R(h_* - h_R) \ . \tag{241}$$

From (240) and (241) we write

$$h_* = \frac{h_R(u_R - S_R)}{u_* - S_R} = \frac{h_L(u_L - S_L)}{u_* - S_L} \ . \tag{242}$$

From here we obtain

$$u_* = S_* = \frac{S_L h_R(u_R - S_R) - S_R h_L(u_L - S_L)}{h_R(u_R - S_R) - h_L(u_L - S_L)} \ . \tag{243}$$

If S_L and S_R are prescribed, then h_* is known from (242)–(243). Then the vectors $\mathbf{Q}_{*L}$ and $\mathbf{Q}_{*R}$ in (238) are given as

$$\mathbf{Q}_{*L} = h_* \begin{bmatrix} 1 \\ S_* \\ \psi_L \end{bmatrix} \ , \quad \mathbf{Q}_{*R} = h_* \begin{bmatrix} 1 \\ S_* \\ \psi_R \end{bmatrix} \ . \tag{244}$$

Now the vectors $\mathbf{F}_{*L}$ and $\mathbf{F}_{*R}$ in (238) are determined and finally the HLLC flux is given as

$$\mathbf{F}^{HLLC}_{i+\frac{1}{2}} = \begin{cases} \mathbf{F}_L & \text{if} \quad 0 \le S_L \ , \\ \mathbf{F}_{*L} = \mathbf{F}_L + S_L(\mathbf{Q}_{*L} - \mathbf{Q}_L) & \text{if} \quad S_L \le 0 \le S_* \ , \\ \mathbf{F}_{*R} = \mathbf{F}_R + S_R(\mathbf{Q}_{*R} - \mathbf{Q}_R) & \text{if} \quad S_* \le 0 \le S_R \ , \\ \mathbf{F}_R & \text{if} \quad 0 \ge S_R \ , \end{cases} \tag{245}$$

where the states $\mathbf{Q}_{*L}$, $\mathbf{Q}_{*R}$ are given by (244). The wave speed estimates for S_L and S_R are as for the HLL flux, see (234), and for S_* we use (243).

The use of HLLC instead of HLL for a system including the passive scalar ψ makes a dramatic difference to the resolution of the contact wave. This is particularly evident for long-time evolution problems.

3.6 The Dumbser-Osher-Toro Riemann Solver: DOT

Here we present a modification of the Osher-Solomon Riemann solver [12] that makes the approach much more practical and applicable to any hyperbolic system for which the complete eigenstructure is known, either analytically of numerically. The resulting scheme is non-linear and complete. The modification is due to Dumbser and Toro [13, 14].

3.6.1 Definitions and Notation

Consider a general $m \times m$ hyperbolic system

$$\partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{0} \,, \tag{246}$$

with conserved variables and flux vectors respectively denoted as

$$\mathbf{Q} = [q_1, q_2, \ldots, q_m]^T \,, \quad \mathbf{F} = [f_1, f_2, \ldots, f_m]^T \,. \tag{247}$$

The real eigenvalues are $\lambda_i(\mathbf{Q})$ and the corresponding right eigenvectors are $\mathbf{R}_i(\mathbf{Q})$, for $i = 1, 2, \ldots, m$. Here we consider Godunov-type finite volume schemes to solve (246)

$$\mathbf{Q}_i^{n+1} = \mathbf{Q}_i^n - \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}) \,, \tag{248}$$

where $\mathbf{F}_{i+\frac{1}{2}}$ is the numerical flux found by solving the Riemann problem for (246) with initial condition

$$\mathbf{Q}(x, 0) = \begin{cases} \mathbf{Q}_0 & \text{if } x < 0 \,, \\ \mathbf{Q}_1 & \text{if } x > 0 \,. \end{cases} \tag{249}$$

Recall that hyperbolicity of system (246) is equivalent to saying that the Jacobian matrix $\mathbf{A}(\mathbf{Q})$ of the flux $\mathbf{F}(\mathbf{Q})$ is diagonalizable, that is

$$\mathbf{A}(\mathbf{Q}) = \mathbf{R}(\mathbf{Q}) \Lambda(\mathbf{Q}) \mathbf{R}^{-1}(\mathbf{Q}) \,, \tag{250}$$

where $\mathbf{R}(\mathbf{Q})$ is the matrix formed by the right eigenvectors $\mathbf{R}_i(\mathbf{Q})$, $\mathbf{R}^{-1}(\mathbf{Q})$ is its inverse and $\Lambda(\mathbf{Q})$ is the diagonal matrix whose diagonal entries are the eigenvalues $\lambda_i(\mathbf{Q})$.

We introduce the definitions

$$\lambda_i^+(\mathbf{Q}) = \max(\lambda_i(\mathbf{Q}), 0) \ , \quad \lambda_i^-(\mathbf{Q}) = \min(\lambda_i(\mathbf{Q}), 0) \tag{251}$$

and consider the associated diagonal matrices $\Lambda^+(\mathbf{Q})$, $\Lambda^-(\mathbf{Q})$ and $|\Lambda^-(\mathbf{Q})|$, whose diagonal entries are $\lambda_i^+(\mathbf{Q})$, $\lambda_i^-(\mathbf{Q})$ and $|\lambda_i(\mathbf{Q})|$ respectively. Note that

$$|\lambda_i(\mathbf{Q})| = \lambda_i^+(\mathbf{Q}) - \lambda_i^-(\mathbf{Q}) \tag{252}$$

and hence

$$|\Lambda(\mathbf{Q})| = \Lambda^+(\mathbf{Q}) - \Lambda^-(\mathbf{Q}) \ . \tag{253}$$

Then we introduce

$$|\mathbf{A}(\mathbf{Q})| = \mathbf{R}(\mathbf{Q})|\Lambda(\mathbf{Q})|\mathbf{R}^{-1}(\mathbf{Q}) \ . \tag{254}$$

Osher and Solomon [12] defined the numerical flux as

$$\mathbf{F}_{i+\frac{1}{2}} = \frac{1}{2}(\mathbf{F}(\mathbf{Q}_0) + \mathbf{F}(\mathbf{Q}_1)) - \frac{1}{2}\int_{\mathbf{Q}_0}^{\mathbf{Q}_1} |\mathbf{A}(\mathbf{Q})| d\mathbf{Q} \ . \tag{255}$$

This requires the evaluation of an integral in phase space, which depends on the chosen integration path joining $\mathbf{Q}_0$ to $\mathbf{Q}_1$. Originally, Osher and Solomon proposed two ways of choosing integration paths *so as to make the actual integration tractable*, (a) the P-ordering and (b) the O-ordering. However, the analytical calculations to be performed are still too involved for general hyperbolic systems. Full details of the original Osher-Solomon Riemann solver are found in Chapter 12 of Toro [1].

3.6.2 The DOT Riemann Solver

Dumbser and Toro [13, 14] made two simple but effective suggestions: (i) choose any path, without considerations regarding computational tractability of the scheme; (ii) evaluate matrices by numerical integration in phase space. The simplest path to evaluate the integral in (255) is the *canonical path*

$$\psi(s; \mathbf{Q}_0, \mathbf{Q}_1) = \mathbf{Q}_0 + s(\mathbf{Q}_1 - \mathbf{Q}_0) \ , \quad s \in [0, 1] \ . \tag{256}$$

Obviously, other choices are available. Then, under a change of variables we obtain

$$\mathbf{F}_{i+\frac{1}{2}} = \frac{1}{2}(\mathbf{F}(\mathbf{Q}_0)+\mathbf{F}(\mathbf{Q}_1))-\frac{1}{2}\left(\int_0^1 |\mathbf{A}(\psi(s;\mathbf{Q}_0,\mathbf{Q}_1))|\,ds\right)(Q_1-Q_0)\ . \tag{257}$$

Finally, the integral in (257) is computed *numerically* along the path ψ using a Gauss type quadrature rule with G points s_j and associated weights ω_j in the unit interval $I = [0, 1]$. We obtain

$$\mathbf{F}_{i+\frac{1}{2}} = \frac{1}{2}(\mathbf{F}(\mathbf{Q}_0)+\mathbf{F}(\mathbf{Q}_1)) - \frac{1}{2}\left(\sum_{j=1}^{G}\omega_j\left|\mathbf{A}(\psi(s_j;\mathbf{Q}_0,\mathbf{Q}_1))\right|\right)(Q_1-Q_0)\,. \tag{258}$$

Note that $\left|\mathbf{A}(\psi(s_j;\mathbf{Q}_0,\mathbf{Q}_1))\right|$ must be decomposed as in (254) for each s_j.

Remarks on the DOT Scheme

1. The complete eigenstructure of the system is needed and is used at each integration point in (258).
2. The scheme is **non-linear and complete**, as it contains all characteristic fields of the exact problem.
3. The scheme is very general. The original version of Osher and Solomon was restricted to very simple hyperbolic systems.
4. The new DOT scheme also applies to non-conservative hyperbolic systems.

3.6.3 Sample Numerical Results, Accuracy and Efficiency

The purpose here is to show some numerical results for a wider range of equations than those studied in these lecture notes. We first show some selected numerical results for the Euler equations of gas dynamics; see [1] for background. Then we also address the crucial issues of accuracy and efficiency of Riemann solvers; this is done in terms of the blood flow equations [15].

Numerical Results for the Euler Equations Figure 41 shows computations from a linearised Riemann solver not studied here, namely the Roe Riemann solver [16]. Results shown are for the Euler equations [1]. The left frame shows results from the original Roe scheme without an *entropy fix*; an entropy violating shock (or rarefaction shock) is computed, which is not physically admissible. The right frame shows results from a modified Roe scheme through a so-called *entropy fix*; now the results look correct and also accurate, recalling that the corresponding Godunov method is only first order accurate.

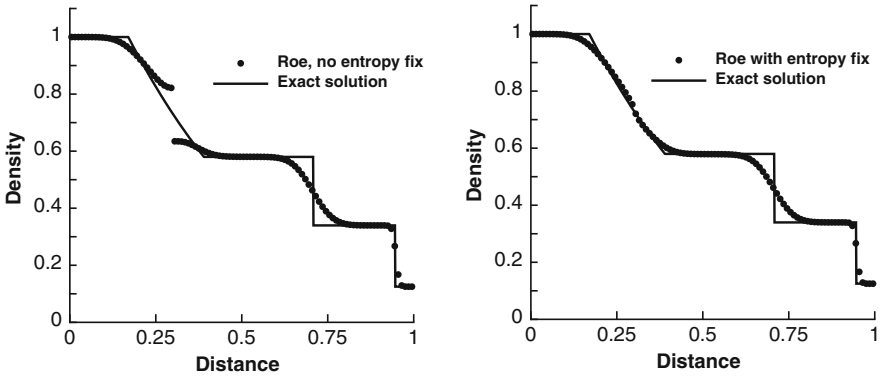

Fig. 41 Sonic flow test problem for the Euler equations taken from [1]. Comparison between numerical (symbol) and exact (line) solutions. Left: linearised Roe solver without entropy fix. Right: linearised Roe solver with entropy fix

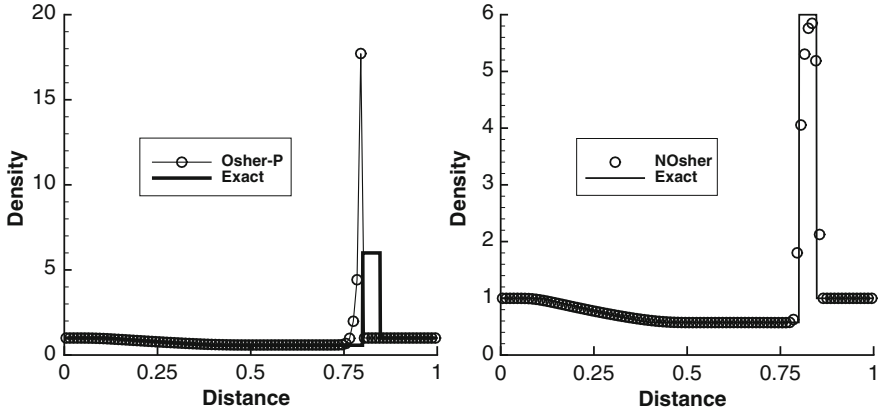

Fig. 42 Test problem for the Euler equations taken from [1]. Comparison between numerical (symbol) and exact (line) solutions. Left: Original Osher-Solomon scheme. Right: new DOT scheme

In Fig. 42 we show results for another test problem for the Euler equations taken from [1]. Comparison between numerical (symbol) and exact (line) solutions is shown. The left frame shows results from the original Osher-Solomon scheme [12], which as seen in the figure, are completely wrong. The right frame shows results from the new DOT scheme [13, 14]; these results are very accurate, especially for the narrow region between the contact discontinuity and the shock wave.

Figure 43 shows results for a special test problem for the Euler equations, also taken from [1]. The test problem consists of a single, isolated stationary contact discontinuity. The left frame shows numerical results from the HLL scheme [7] and the FORCE scheme [6]. Both numerical methods show large errors due to numerical diffusion of the intermediate wave in the Riemann problem for the Euler equations.

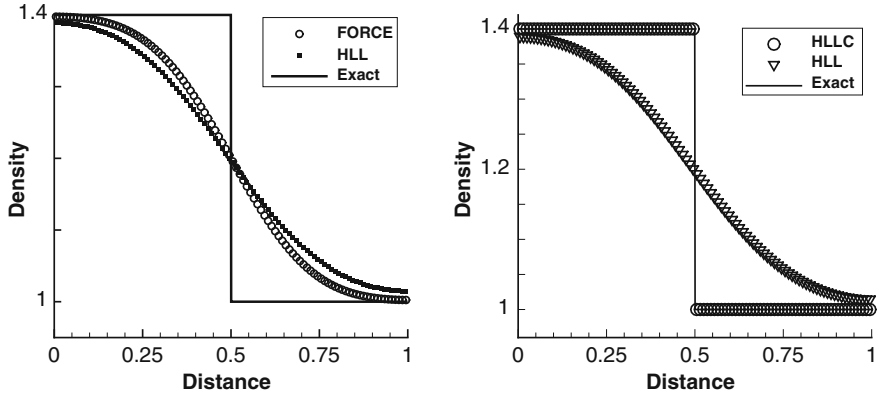

Fig. 43 Test problem for the Euler equations containing a single, isolated stationary contact discontinuity; taken from [1]. Left: FORCE and HLL versus the exact solution. Right: HLL and HLLC versus the exact solution

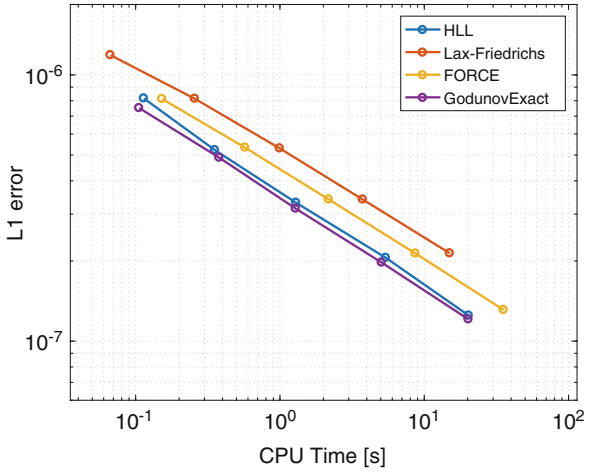

Fig. 44 Efficiency test for the blood flow equations. The test is a Riemann problem containing two rarefaction waves. Error against CPU time. Results are shown for the Godunov method used in conjunction with Exact Riemann solver, the HLL Riemann solver, FORCE and the Lax-Friedrichs flux (Courtesy of PhD student Christian Contarino, University of Trento, Italy)

The right frame shows results from HLL [7] and from HLLC [11], noting that the latter reproduces the exact solution.

Accuracy and Efficiency We have already mentioned the question of efficiency, which relates error (or accuracy) to computational cost. Figure 44 shows results from an efficiency test for the 1D blood flow equations [15], where error is measured against CPU time. Comparison is made for the Godunov method with four Riemann solvers: the exact Riemann solver, HLL, FORCE and Lax-Friedrichs.

What the results of Fig. 44 show is that for this test problem with smooth solution the Godunov method is the most efficient method. Compared to the Lax-Friedrichs method, it is about five times more efficient. To see this, imagine a horizontal line through the last point of the Lax-Friedrichs curve with the smallest error and look for its intersection with the exact Riemann solver curve; these two intersection points give two respective CPU times.

3.6.4 Concluding Remarks

The Godunov method for the augmented one-dimensional shallow water equations has been introduced. The Godunov scheme works with the exact and with approximate Riemann solvers. Examples of approximate Riemann solvers have also been presented, along with some selected numerical results for the Euler equations and for the blood flow equations, not studied here. The first-order Godunov schemes studied in this section can be extended to high order of accuracy following a variety of procedures available in the literature. In the next section we present the ADER approach to construct high-order numerical methods.

4 High Order Methods: The ADER Approach

In this section we present one approach, the ADER approach, to construct high-order accurate extensions of the first-order methods presented previously.

4.1 Overview

We are interested in time-dependent partial differential equations of the form

$$\left.\begin{array}{c} \partial_t \mathbf{Q}(\mathbf{x},t) + \mathbf{A}(\mathbf{Q}(\mathbf{x},t)) = \mathbf{S}(\mathbf{Q}(\mathbf{x},t)) + \mathbf{D}(\mathbf{Q}(\mathbf{x},t)) \ , \\ \\ \mathbf{x} \in \Omega \ , \quad t > 0 \ , \quad \text{ICs} \ , \quad \text{BCs} \ , \end{array}\right\} \tag{259}$$

along with appropriate initial and boundary conditions. Here $\mathbf{Q}(\mathbf{x},t)$ is the vector of unknowns; $\mathbf{A}(\mathbf{Q}(\mathbf{x},t))$ is an advection differential operator in 1D, 2D or 3D; $\mathbf{D}(\mathbf{Q}(\mathbf{x},t))$ is a dissipative operator in 1D, 2D or 3D and $\mathbf{S}(\mathbf{Q}(\mathbf{x},t))$ is a source term vector, a prescribed function of the unknowns.

The ADER approach was first presented by Toro and collaborators [17] for linear hyperbolic systems in 1D, 2D and 3D on structured meshes; schemes of upto 10th order of accuracy in space and time were constructed and implemented. The ADER schemes were further developed in [18] and [19] for non-linear systems;

in [20] ADER was formulated, in a unified manner, in both the finite volume and the discontinuous Galerkin finite element frameworks. For an introduction to ADER see Chapters 19 and 20 of [1] and the many references therein, up to 2009. Distinguishing features of the ADER approach include:

1. Accuracy is arbitrary in both **space and time**.
2. Schemes are non-linear schemes, in the sense of Godunov; computed shock waves and other discontinuities have none or controlled spurious oscillations.
3. Schemes are suitable for general geometries in multiple space dimensions, treated with both structured or unstructured meshes.
4. Schemes work in both the finite volume and the discontinuous Galerkin finite element frameworks.
5. Schemes are applicable to conservative and non-conservative hyperbolic systems.

Why is High Accuracy Important? Because of Efficiency Figure 45 shows computational results for an acoustic problem modelled by the linearised two dimensional Euler equations solved by ADER schemes taken from [21]. The

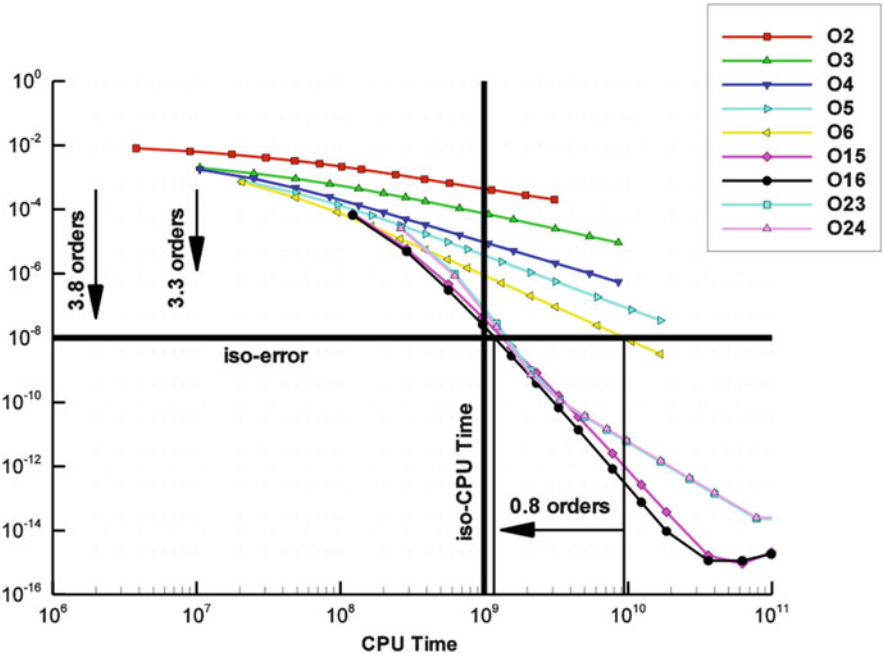

Fig. 45 Efficiency plot: error against CPU cost for nine high-order ADER schemes, from the 2nd order to the 24th order of accuracy. For a chosen fixed error there corresponds a horizontal line (e.g. black horizontal line); its intersection with the various curves gives corresponding times, which give the cost of the corresponding scheme to compute the solution with that error. Taken from [21]

original paper reports computations of orders of accuracy from 1 to 24 in space and time. Figure 45 displays some selected results from 2nd to 16th order.

4.2 ADER in the Finite Volume Framework

Consider the general system of hyperbolic equations with source terms (hyperbolic balance laws) in one space dimension

$$\partial_t \mathbf{Q}(x,t) + \mathbf{F}(\mathbf{Q}(x,t)) = \mathbf{S}(\mathbf{Q}(x,t)) \ . \tag{260}$$

Exact integration of (260) in the control volume $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [0, \Delta t]$ gives a finite volume like formula

$$\hat{\mathbf{Q}}_i^{n+1} = \hat{\mathbf{Q}}_i^n - \frac{\Delta t}{\Delta x}(\hat{\mathbf{F}}_{i+\frac{1}{2}} - \hat{\mathbf{F}}_{i-\frac{1}{2}}) + \Delta t \hat{\mathbf{S}}_i \ , \tag{261}$$

where

$$\left.\begin{aligned} \hat{\mathbf{Q}}_i^n &= \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{Q}(x,t^n)dx \ , \\ \hat{\mathbf{F}}_{i+\frac{1}{2}} &= \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{Q}(x_{i+\frac{1}{2}},t))dt \ , \\ \hat{\mathbf{S}}_i &= \frac{1}{\Delta t \Delta x}\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{Q}(x,t))dxdt \ . \end{aligned}\right\} \tag{262}$$

Relation (261) with definitions (262) is exact and motivates an approximate formula, namely

$$\mathbf{Q}_i^{n+1} = \mathbf{Q}_i^n - \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}) + \Delta t \mathbf{S}_i \ . \tag{263}$$

See Sect. 1. Equation (263) defines a one-step, fully discrete finite volume numerical scheme with **numerical flux**

$$\mathbf{F}_{i+\frac{1}{2}} \approx \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}(\mathbf{Q}_{i+\frac{1}{2}}(\tau))d\tau \tag{264}$$

and **numerical source**

$$\mathbf{S}_i \approx \frac{1}{\Delta t \Delta x}\int_0^{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{Q}_i(x,\tau))dxd\tau \ . \tag{265}$$

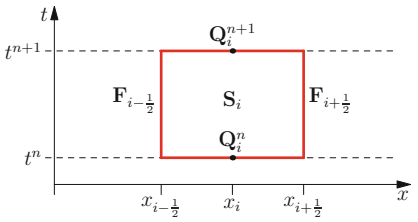

Fig. 46 Illustration of the finite volume scheme (263) to solve the system of hyperbolic equation (260) with source terms. The scheme requires numerical fluxes at interfaces and the numerical source within the control volume

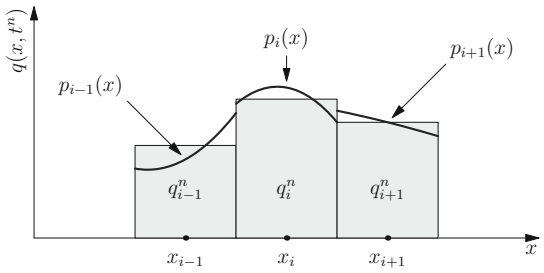

Fig. 47 Illustration of the reconstruction procedure for one variable $q(x, t)$ in one space dimension on a regular mesh. From the set of (constant) integral averages $\{q_i^n\}$ one obtains an interpolant $p_i(x)$ satisfying a conservation property and a non-linear property to circumvent Godunov's theorem, using for example criteria such as TVD, ENO and WENO. Note that at each interface one now has reconstructed data that defines a generalised Riemann problem

Figure 46 illustrates scheme (263) to solve (260). The finite volume ADER scheme (263) aims at computing approximations (264) and (265) as accurately as possible.

4.3 Ingredients of ADER

The ADER method to solve (260) is based on the finite volume formula (263) and requires the accurate evaluation of integrals (264) for the intercell numerical flux and (265) for the numerical source. In order to achieve this, the following steps are required.

1. **Reconstruction:** high-order *non-linear* spatial reconstruction, once per time step, using any of the methodologies available, such as TVD, ENO and WENO. Figure 47 illustrates the reconstruction process. For background on reconstruction techniques see for example [1, 4, 22, 23] and [24].

2. **Generalised Riemann problem (GRP) and numerical flux.** At each interface one must solve a Riemann problem with piece-wise smooth data, not piece-wise constant, as in the conventional case. This GRP may also include the source terms in case these are present in the equations.
3. **Numerical source.** This is an additional term in the case in which the equations include source term.

4.4 Generalized Riemann Problem

Starting from reconstructed data, at each interface one defines the following initial value problem, called the generalized Riemann problem, or GRP

$$\left.\begin{array}{l} \text{PDEs: } \partial_t \mathbf{Q} + \partial_x \mathbf{F}(\mathbf{Q}) = \mathbf{S}(\mathbf{Q}) \ , \quad x \in (-\infty, \infty) \ , \quad t > 0 \ , \\ \\ \text{ICs:} \quad \mathbf{Q}(x,0) = \begin{cases} \mathbf{Q}_L(x) & \text{if } \ x < 0 \ , \\ \\ \mathbf{Q}_R(x) & \text{if } \ x > 0 \ . \end{cases} \end{array}\right\} \tag{266}$$

In the GRP (266) the governing equations include source terms and the initial conditions are piece-wise smooth (e.g. polynomials of any degree). This Riemann problem also generalises the case in which the data is piece-wise linear, which is associated with the second-order GRP scheme of Ben-Artzi and Falcovitz [25].

Figure 48 illustrates the classical Riemann problem (left) and the generalised Riemann problem (right). Figure 49 shows an example of a generalised Riemann problem for the Euler equations of gas dynamics. There are so far several published methods for solving the generalised Riemann problem for hyperbolic systems. The first practical solver for non-linear hyperbolic systems with source terms is due to Toro and Titarev [18]. This solver is suitable for non-stiff source terms. Other solvers include [26–30]. An important development was that in [27] in which the proposed solver can deal with stiff source terms, reconciling in this way, stiffness and high-order of accuracy.

4.5 Numerical Examples

Here we show some sample numerical results, first for the 1D linear advection equation and then for the 2D Euler equations of gas dynamics with the ideal equation of state.

Figure 50 shows computed (symbols) and the exact solution (line) for linear advection equation using a mesh of $M = 50$ cells, a Courant number coefficient $C_{cfl} = 0.95$ at the output time $t_{out} = 1000\pi$. The top frame displays results from a

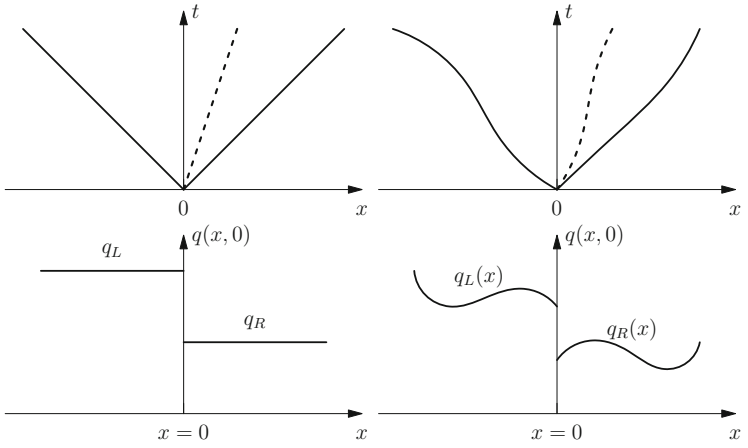

Fig. 48 Classical Riemann problem (left) and generalised Riemann problem (right). Bottom frames depict the initial conditions (for a single variable) and top frames depict the structure of the solution of the initial value problem in the $x - t$ plane

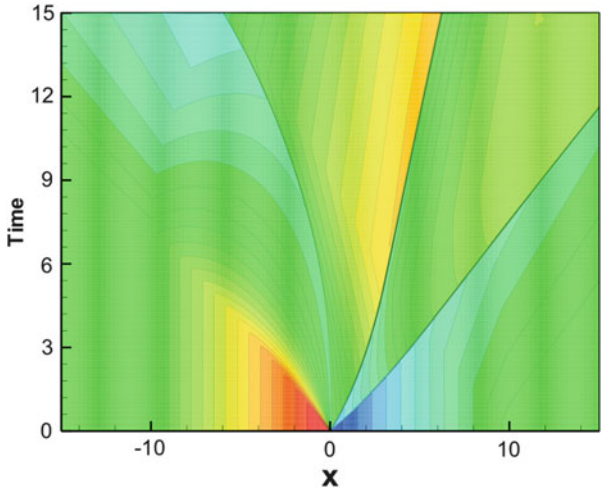

Fig. 49 Structure of the solution of a generalized Riemann problem for the Euler equations. Characteristics are curved in the $x - t$ plane (Courtesy of Dr VA Titarev)

second-order TVD method used in conjunction with the MINBEE limiter [1]. The bottom frame shows results from the 5th-order ADER scheme (5th order in space and time) with WENO (non-linear) reconstruction. The results speak by themselves. The second order TVD method is unable to resolve the wave packet and there is not even a hint of waves; the profile is virtually flat. The killer here is the long evolution time, $t_{out} = 1000\pi$. Long time evolution problems expose the limitations of low order methods. The fifth order method is just perfect.

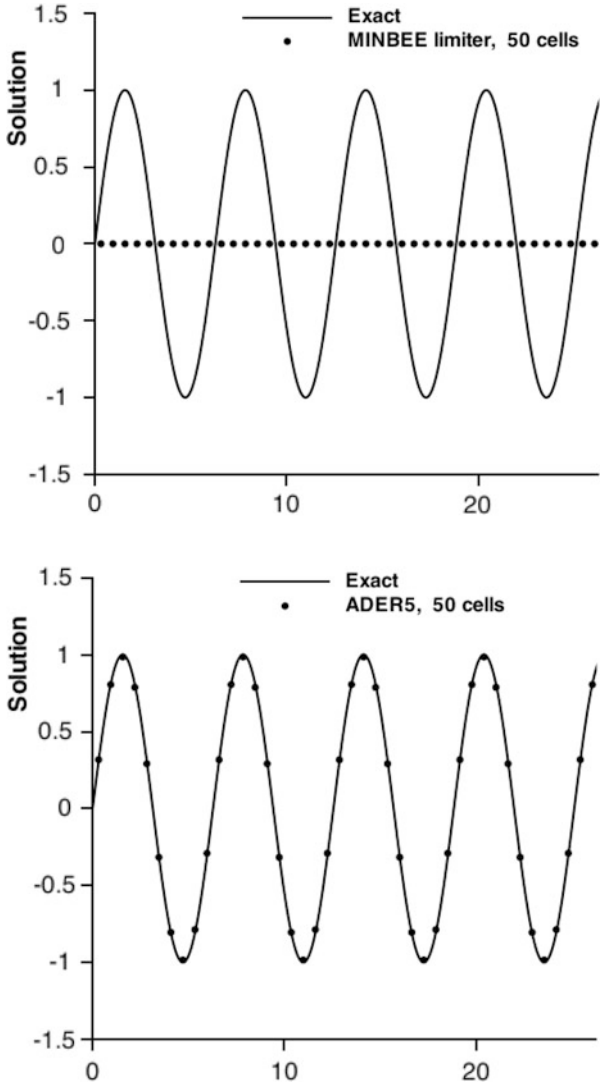

Fig. 50 Computed (symbols) and exact solution (line) of linear advection equation using a mesh of $M = 50$ cells, Courant number coefficient $C_{cfl} = 0.95$ and output time $t_{out} = 1000\pi$. Top frame displays results from second-order TVD method with the MINBEE limiter. Bottom frame shows results from the 5th-order ADER scheme with WENO reconstruction

Figure 51 shows computed results for the two-dimensional Euler equations of gas dynamics with the ideal gas equation of state. This test problem is well known in the gas dynamics community. The domain is a rectangular region with a solid fixed triangle in its interior (white object). The top and bottom boundaries are reflecting fixed walls, while the left and right boundaries are transmissive. The initial condition

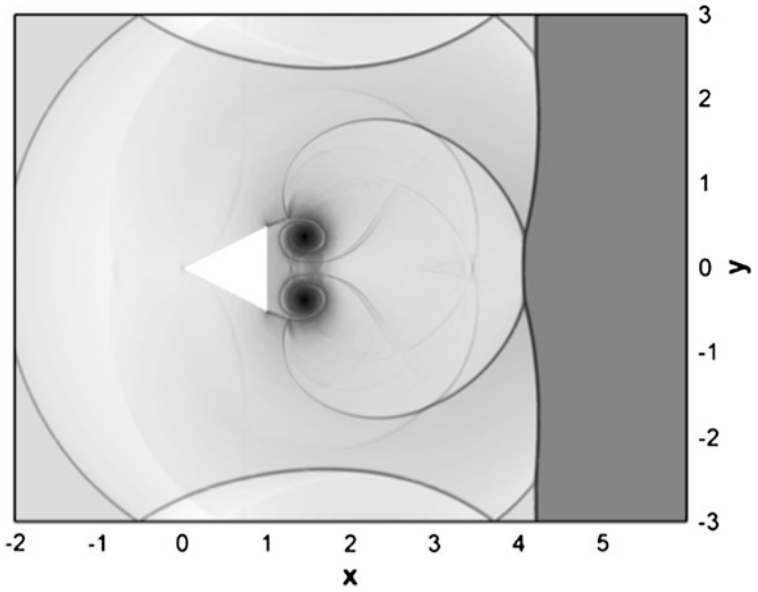

Fig. 51 Shock wave impinging on stationary triangular body. Numerical solution of the Euler equations of gas dynamics on a triangular mesh using a fourth order ADER method (Courtesy of Prof. M Dumbser, University of Trento, Italy)

is an isolated shock wave of Mach number 1.3 positioned between the left boundary and the triangle. The evolution of this initial condition gives rise to a complex pattern of waves propagating and interacting. There are experimental visualization results for this problem. The ADER solution represents those experiments well. In addition to the dominant shock waves everywhere there are also regions of smooth flow and many low amplitude waves; these are the flow features that are difficult to capture with low order methods, they are simply wiped out, just as seen for the linear advection example of Fig. 50.

4.6 Concluding Remarks

In this last section we have given a very brief introduction to one approach to construct high-order numerical methods for hyperbolic equations, namely the ADER approach. This is a fully discrete approach that requires a spatial reconstruction procedure and the solution of the generalised Riemann problem. There are indeed alternative methods to achieve high order of accuracy. Prominent examples are the ENO and WENO semidiscrete approaches pioneered by Shu and collaborators [22–24].

Accumulated experience over the last few years has shown that high-order methods are much more efficient than low order methods if small errors are sought, that is if accurate solutions are sought. By efficiency we mean that given an error deemed acceptable, then high order methods attain that error much more efficiently on a coarse mesh than low order methods on a fine mesh. This is illustrated in Fig. 45.

The issues of accurate solutions and efficiency are becoming increasingly important given the growing trend to use mathematical models (PDEs) to understand the physics they embody. Only very accurate solutions of the PDEs will achieve this and also reveal limitations of the mathematical models (the governing equations and their parameters). Very long time evolution simulations, as in wave propagation problems for long distances, require the use of high order methods, as illustrated in Fig. 50.

References

1. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction, 3rd edn. Springer, Berlin (2009). ISBN 978-3-540-25202-3. http://link.springer.com/book/10.1007%2Fb79761
2. Toro, E.F.: Shock-Capturing Methods for Free-Surface Shallow Flows. Wiley, Chichester (2001)
3. Godlewski, E., Raviart, P.A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws. Springer, New York (1996)
4. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge University Press, Cambridge (2002)
5. Godunov, S.K.: A Finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics. Math. Sb. **47**, 357–393 (1959)
6. Toro, E.F., Billett, S.J.: Centred TVD schemes for hyperbolic conservation laws. IMA J. Numer. Anal. **20**, 47–79 (2000)
7. Harten, A., Lax, P.D., van Leer, B.: On spstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**(1), 35–36 (1983)
8. Rusanov, V.V.: Calculation of interaction of non-steady shock waves with obstacles. J. Comput. Math. Phys. USSR **1**, 267–279 (1961)
9. Toro, E.F., Spruce, M., Speares, W.: Restoration of the contact surface in the HLL Riemann solver. Technical Report, Department of Aerospace Science, College of Aeronautics, Cranfield Institute of Technology. CoA-9204 (1992)
10. Toro, E.F., Spruce, M., Speares, W.: Restoration of the contact surface in the HLL Riemann solver. Shock Waves **4**, 25–34 (1994)
11. Toro, E.F., Chakraborty, A.: Development of an approximate Riemann solver for the steady supersonic Euler equations. Aeronaut. J. **98**, 325–339 (1994)
12. Osher, S., Solomon, F.: Upwind difference schemes for hyperbolic conservation laws. Math. Comput. **38**(158), 339–374 (1982)
13. Dumbser, M., Toro, E.F.: A simple extension of the Osher Riemann solver to general non-conservative hyperbolic systems. J. Sci. Comput. **48**, 70–88 (2011)
14. Dumbser, M., Toro, E.F.: On universal Osher-type schemes for general nonlinear hyperbolic conservation laws. Commun. Comput. Phys. **10**, 635–671 (2011)
15. Toro, E.F.: Brain venous haemodynamics, neurological diseases and mathematical modelling. A review. Appl. Math. Comput. **272**, 542–579 (2016)

16. Roe, P.L.: Approximate Riemann solvers, parameter vectors and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
17. Toro, E.F., Millington, R.C., Nejad, L.A.M.: Towards very high-order Godunov schemes. In: Toro, E.F. (ed.) Godunov Methods: Theory and Applications. Edited Review. Conference in Honour of Godunov SK, vol. 1, pp. 897–902. Kluwer Academic/Plenum Publishers, New York (2001)
18. Toro, E.F., Titarev, V.A.: Solution of the generalised Riemann problem for advection-reaction equations. Proc. R. Soc. London, Ser. A **458**, 271–281 (2002)
19. Titarev, V.A., Toro, E.F.: ADER: arbitrary high order Godunov approach. J. Sci. Comput. **17**, 609–618 (2002)
20. Dumbser, M., Balsara, D., Toro, E.F., Munz, C.D.: A unified framework for the construction of one-step finite-volume and discontinuous Galerkin schemes on unstructured meshes. J. Comput. Phys. **227**, 8209–8253 (2008)
21. Dumbser, M., Schwartzkopff, T., Munz, C.D.: Arbitrary high order finite volume schemes for linear wave propagation. In: Computational Science and High Performance Computing II. Notes on Numerical Fluid Mechanics and Multidisciplinary Design Book Series (NNFM), vol. 91, pp. 129–144. Springer, Berlin (2006)
22. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
23. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes II. J. Comput. Phys. **83**, 32–78 (1989)
24. Jiang, G.S., Shu, C.W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
25. Ben-Artzi, M., Falcovitz, J.: A second order Godunov-type scheme for compressible fluid dynamics. J. Comput. Phys. **55**, 1–32 (1984)
26. Castro, C.E., Toro, E.F.: Solvers for the high-order Riemann problem for hyperbolic balance laws. J. Comput. Phys. **227**, 2481–2513 (2008)
27. Dumbser, M., Enaux, C., Toro, E.F.: Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. J. Comput. Phys. **227**, 3971–4001 (2008)
28. Toro, E.F., Montecinos, G.I.: Implicit, semi-analytical solution of the generalized Riemann problem for stiff hyperbolic balance laws. J. Comput. Phys. **303**, 146–172 (2015)
29. Götz, C.R., Iske, A.: Approximate solutions of generalized Riemann problems for nonlinear systems of hyperbolic conservation laws. Math. Comput. **85**, 35–62 (2016)
30. Götz, C.R., Dumbser, M.: A novel solver for the generalized Riemann problem based on a simplified LeFloch-Raviart expansion and a local space-time discontinuous Galerkin formulation. J. Sci. Comput. **69**(2), 805–840 (2016)