Notes on Godunov Methods

Hagala, R.; Hansteen, V; Mina, M.

1 Introduction

1.1 Motivation

We already learned about upwind schemes, which make sense for simple equations like the advection equation with constant velocity u. For more complex systems of equations, it is a lot less clear which information travels in which direction, and at what velocity. Furthermore, as we have seen in the past lectures, discontinuities in the solution of a differential equation are a big problem from the numerical point of view. The reason for this is that the *finite difference* schemes commonly applied to evaluate (spatial) derivatives are based on the Taylor expansion, which requires differentiability. An example is the forward difference derivative:

$$f'(x_i) \approx \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x}. \tag{1}$$

The local error when ignoring the rest of the Taylor series is related to the sum of the higher order derivatives of the function, $f''(x)\,\Delta x/2$ being the dominant error term. When a discontinuity is introduced – for instance where a shock front is forming – the local error terms for estimating the derivatives grow as the higher order derivatives diverge. A way to postpone (but not solve) this problem within the finite difference schemes would be to use higher order methods or higher resolution.

However, for systems of partial differential equations describing conservation laws (e.g. Euler equations) it is possible to construct a class of conservative numerical methods, which are able to capture and prevent shock formation in the numerical solution of the problem.

1.2 Definitions

1.2.1 Systems of First-Order Partial Differential Equations

A system of n first-order partial differential equations can be expressed on linear algebra form,

$$\mathbf{U}_t + \mathbf{A}\mathbf{U}_x + \mathbf{B} = 0. \tag{2}$$

Here $\mathbf{U} = (U_1, \cdots, U_n)^{\mathrm{T}}$ is a set of independent physical quantities in the system, $\mathbf{A}$ is an $n \times n$ matrix deciding how these quantities interact, and $\mathbf{B}$ is a vector of size n describing the source terms. Subscript t and x means an (element-wise) partial derivative with respect to time or space, $\mathbf{U}_x = \partial\mathbf{U}/\partial x$. Generally, both $\mathbf{A}$ and $\mathbf{B}$ could depend on any U_i.

1.2.2 Conservation Laws

A system describing a set of conserved physical quantities $\mathbf{U}$ (e.g. density, momentum) can in general be written

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = 0, \tag{3}$$

where $\mathbf{F}(\mathbf{U}) = (F_1(\mathbf{U}), \cdots, F_n(\mathbf{U}))^{\mathrm{T}}$ is called the *flux function*. Each F_i can be a function of all $(U_1, \cdots, U_n)^{\mathrm{T}}$.

1.2.3 Jacobian of the flux function

The Jacobian matrix is given by

$$\mathbf{A}(\mathbf{U}) = \frac{\partial \mathbf{F}}{\partial \mathbf{U}} = \begin{bmatrix} \dfrac{\partial F_1}{\partial U_1} & \cdots & \dfrac{\partial F_1}{\partial U_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial U_1} & \cdots & \dfrac{\partial F_n}{\partial U_n} \end{bmatrix}. \tag{4}$$

A (quasi-)linear system described by equation (3) can be put on matrix form (2) by using the jacobian matrix $\mathbf{A}$,

$$\mathbf{U}_t + \mathbf{A}(\mathbf{U})\,\mathbf{U}_x = 0. \tag{5}$$

1.2.4 Hyperbolic, strictly hyperbolic, and elliptic system

A system on the form (2) is *hyperbolic* at point (x,t) if all n eigenvalues of the matrix $\mathbf{A}$ are real and correspond to n linearly independent eigenvectors $\mathbf{K}^{(1)},\cdots,\mathbf{K}^{(n)}$. The system is *strictly hyperbolic* if all n eigenvalues are distinct. If *all* eigenvalues are imaginary, the system is *elliptic*.

Example: Gas Dynamics. Gas dynamics are described by the equations for conservation of mass and momentum. In 1D given by

$$\partial_t \rho + \partial_x (\rho u) = 0 \tag{6}$$

$$\partial_t (\rho u) + \partial_x \left(\rho u^2 + p\right) = 0 \tag{7}$$

From the above equations, one can directly see that we can write the system on conservative flux form (3) by using the set of conserved variables $\mathbf{U} = (\rho, \rho u)^{\mathrm{T}}$ and the flux function $\mathbf{F} = (\rho u, \rho u^2 + p)^{\mathrm{T}}$.

$$\begin{bmatrix} \rho \\ \rho u \end{bmatrix}_t = \begin{bmatrix} \rho u \\ \rho u^2 + p \end{bmatrix}_x \tag{8}$$

The jacobian of this system is given by

$$\mathbf{A} = \begin{bmatrix} \dfrac{\partial F_1}{\partial U_1} & \dfrac{\partial F_1}{\partial U_2} \\ \dfrac{\partial F_2}{\partial U_1} & \dfrac{\partial F_2}{\partial U_2} \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} \dfrac{\partial (\rho u)}{\partial (\rho)} & \dfrac{\partial (\rho u)}{\partial (\rho u)} \\ \dfrac{\partial (\rho u^2 + p)}{\partial (\rho)} & \dfrac{\partial (\rho u^2 + p)}{\partial (\rho u)} \end{bmatrix}. \tag{10}$$

Notice that $U_2 = \rho u$ is an independent variable, so $\partial(\rho u)/\partial\rho = 0$.

By assuming that $p = p(\rho)$, and realizing that $\mathbf{F} = (U_2, U_2^2/U_1 + p(\rho))^{\mathrm{T}}$, one finds

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -u^2 + \partial p/\partial\rho & 2u \end{bmatrix}. \tag{11}$$

Often one defines $c_s^2 \equiv \partial p/\partial\rho$, where c_s is the sound speed.

For an isothermal gas, we have $p = k\rho$ with some constant k. More generally, an adiabatic process has $p = k\rho^\gamma$, where γ is the adiabatic index. For an ideal monatomic gas, $\gamma = 5/3$, and for an ideal, diatomic gas (e.g. air), $\gamma = 1.4$.

The eigenvalues λ_i of this system can be found by solving

$$|\mathbf{A} - \lambda\mathbf{I}| = \det(\mathbf{A} - \lambda\mathbf{I}) = 0, \tag{12}$$

$$\Rightarrow \lambda^2 - 2u\lambda + u^2 - c_s^2 = 0. \tag{13}$$

With the solutions

$$\lambda_\pm = u \pm \frac{1}{2}\sqrt{4u^2 - 4u^2 + 4c_s^2} = u \pm c_s. \tag{14}$$

With a positive, real sound speed ($\partial p/\partial\rho > 0$), this is a strictly hyperbolic system.

1.3 The Riemann Problem

The Riemann problem is a specific Initial Value Problem (IVP) where the initial conditions are piecewise constant, with only one discontinuity (usually taken to be at the origin). In the linear case, the general Riemann problem is

$$\begin{cases} \mathbf{U}_t + \mathbf{A}\mathbf{U}_x = 0 \\ \mathbf{U}(x, t=0) = \begin{cases} \mathbf{U}_L & \text{for} \quad x < 0 \\ \mathbf{U}_R & \text{for} \quad x > 0 \end{cases} \end{cases} \tag{15}$$

The general solution $\mathbf{U}(x,t)$ in the strictly hyperbolic case is sketched in Figure 1. The n eigenvalues of $\mathbf{A}$ are ordered from lowest to highest, $\lambda_1, \lambda_2, \cdots, \lambda_n$. Each of them correspond to a characteristic wave travelling at velocity λ_i from the origin.

The state of the system to the left of the λ_1 wave will stay $\mathbf{U}_L$, since no information from the other state can travel faster than the frontier wave. Likewise, the state to the right of λ_n will stay $\mathbf{U}_R$. The wedges between the leftmost and rightmost wave will be called *star regions*. Here, the state $\mathbf{U}^*$ will be piecewise constant with discontinuities at each wavefront λ_i. $\mathbf{U}^*$ will be some linear combination of the decomposed right and left states.

To find the general solution, $\mathbf{U}$, let's first write the right and left states as expansions of the n linearly independent eigenvectors of $\mathbf{A}$, which we will denote by $\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(n)}$.

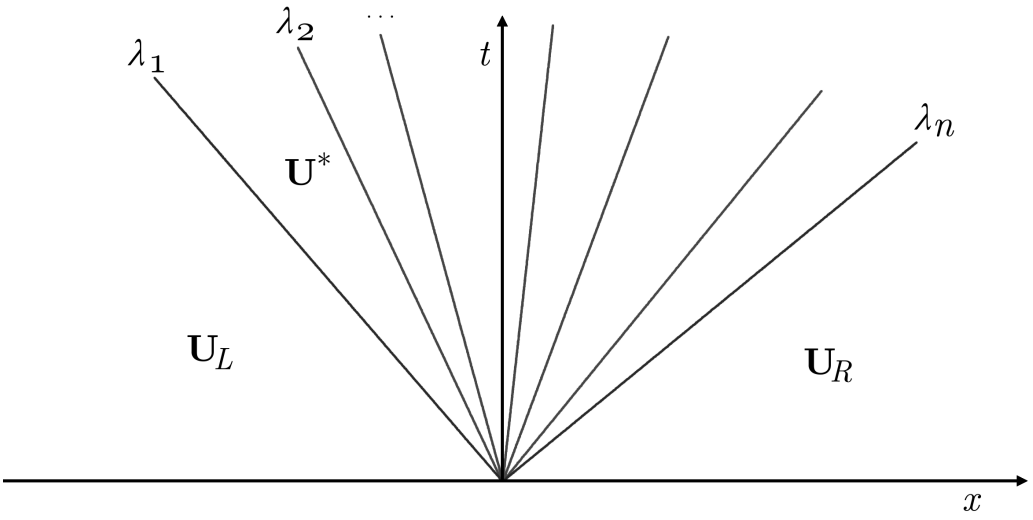

Figure 1: General solution to the strictly hyperbolic Riemann problem in n variables.

$$\mathbf{U}_L = \sum_{i=1}^{n} \alpha_i \mathbf{K}^{(i)}, \quad \mathbf{U}_R = \sum_{i=1}^{n} \beta_i \mathbf{K}^{(i)}. \tag{16}$$

The eigenvectors $\mathbf{K}^{(i)}$ describe the portion of the state information $\mathbf{U}$ that is carried along with the wave corresponding to the eigenvalue λ_i. We can combine the expansion (16) and our knowledge of the velocity of each characteristic wave to find the mixed state $\mathbf{U}^*$ in any region. At any point (x,t) where the solution is to be found, we only need to identify which characteristic lines are to the left, and which are to the right of that point. Specifically, all $\lambda_i < x/t$ are to the left of (x,t) – and have thus carried information from the *right* state to this point. All $\lambda_i > x/t$ are to the right of (x,t) and their corresponding information is carried from the *left* state. The general solution can in other words be written

$$\mathbf{U}(x,t) = \sum_{i:\lambda_i > x/t} \alpha_i \mathbf{K}^{(i)} + \sum_{i:\lambda_i < x/t} \beta_i \mathbf{K}^{(i)}. \tag{17}$$

Note that for a point where x/t is smaller than or larger than *all* λ_i, equation (17) is equivalent to the spectral decomposition of $\mathbf{U}_L$ or $\mathbf{U}_R$ re-

spectively (compare with equation 16). For a 2×2 system, there will only be three solutions at any time: a region where $\mathbf{U} = \mathbf{U}_L$, a region where $\mathbf{U} = \mathbf{U}_R$, and the star region where $\mathbf{U} = \mathbf{U}^*$.

Example: Riemann Problem for Linearised Gas Dynamics.
The equations for gas dynamics, (6) and (7), are *quasi-linear*, meaning that the matrix $\mathbf{A}$ is not constant, but depends on $\mathbf{U}$. A perturbative expansion of the variables $u = 0 + \delta u$ and $\rho = \rho_0 + \delta\rho$, while keeping only first order terms of δu and $\delta\rho$, gives the set of equations

$$\partial_t \rho + \rho_0 \partial_x u = 0, \tag{18}$$
$$\rho_0 \partial_t u + \partial_x p = 0, \tag{19}$$

where

$$\partial_x p = \frac{\partial p}{\partial \rho} \partial_x \rho \approx a^2 \partial_x \rho. \tag{20}$$

Here, a is constant, and equivalent to the sound speed when $\rho = \rho_0$.

On matrix form, this linearized system reads

$$\begin{bmatrix} \rho \\ u \end{bmatrix}_t + \begin{bmatrix} 0 & \rho_0 \\ a^2/\rho_0 & 0 \end{bmatrix} \begin{bmatrix} \rho \\ u \end{bmatrix}_x = 0. \tag{21}$$

With the initial conditions

$$\mathbf{U}_L = \begin{bmatrix} \rho_L \\ u_L \end{bmatrix}, \quad \mathbf{U}_R = \begin{bmatrix} \rho_R \\ u_R \end{bmatrix}. \tag{22}$$

Notice that the matrix $\mathbf{A}$ is constant in this case. This naturaly means that the characteristics λ_1 and λ_2 are the same for any Riemann problem with a given ρ_0 and a.

The eigenvalues of $\mathbf{A}$ are

$$\lambda_1 = -a\,, \quad \lambda_2 = a, \tag{23}$$

and a set of corresponding eigenvectors are

$$\mathbf{K}^{(1)} = \begin{bmatrix} \rho_0 \\ -a \end{bmatrix}, \quad \mathbf{K}^{(2)} = \begin{bmatrix} \rho_0 \\ a \end{bmatrix}. \tag{24}$$

The eigenvector decomposition (16) of the left state $\mathbf{U}_L$ is then

$$\mathbf{U}_L = \begin{bmatrix} \rho_L \\ u_L \end{bmatrix} = \alpha_1 \begin{bmatrix} \rho_0 \\ -a \end{bmatrix} + \alpha_2 \begin{bmatrix} \rho_0 \\ a \end{bmatrix}. \tag{25}$$

This system is a set of two equations which can be solved for the two unknown coefficients α_1 and α_2. The results are

$$\alpha_1 = \frac{a\rho_L - \rho_0 u_L}{2a\rho_0}, \quad \alpha_2 = \frac{a\rho_L + \rho_0 u_L}{2a\rho_0}, \tag{26}$$

and similarly, for the right hand state $\mathbf{U}_R$, we get

$$\beta_1 = \frac{a\rho_R - \rho_0 u_R}{2a\rho_0}, \quad \beta_2 = \frac{a\rho_R + \rho_0 u_R}{2a\rho_0}. \tag{27}$$

The non-trivial part of the general solution is $\mathbf{U}^*$, the star region wedged between the two characteristics λ_1 and λ_2. Equation (17) shows that the solution to the right of λ_1 and to the left of λ_2 is

$$\mathbf{U}^* = \begin{bmatrix} \rho^* \\ u^* \end{bmatrix} = \beta_1 \begin{bmatrix} \rho_0 \\ -a \end{bmatrix} + \alpha_2 \begin{bmatrix} \rho_0 \\ a \end{bmatrix}. \tag{28}$$

To find ρ^* and u^* expressed by u_L, ρ_L, u_R, ρ_R, ρ_0 and a, one can of course insert α_2 and β_1 found above.

2 Conservative methods

For this chapter we will always refer to the general system of conservation laws in one spatial dimension, expressed by equation (3). The treatment we perform in the next sections can be easily generalised in three spatial dimensions.

2.1 Integral form of conservation laws.

Each conservation law can always be expressed on an integral form. Starting from the system of conservation laws (3) and integrating the equations in time and space we obtain

$$\int_{x_1}^{x_2} \int_{t_1}^{t_2} \left[\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x\right] \mathrm{d}x\mathrm{d}t = 0. \tag{29}$$

For a generic function $w(a, b)$ of the independent variables a and b, the fundamental theorem of integral calculus states:

$$\int_{a_1}^{a_2} \frac{\partial w}{\partial a} \mathrm{d}a = w(a_2, b) - w(a_1, b). \tag{30}$$

Applying the the fundamental theorem of integral calculus on our system of conservation laws we obtain

$$\int_{x_1}^{x_2} \mathbf{U}(x,t_2)\mathrm{d}x = \int_{x_1}^{x_2} \mathbf{U}(x,t_1)\mathrm{d}x + \int_{t_1}^{t_2} \mathbf{F}(\mathbf{U}(x_1,t))\mathrm{d}t - \int_{t_1}^{t_2} \mathbf{F}(\mathbf{U}(x_2,t))\mathrm{d}t. \tag{31}$$

The equation (31) represents the integral form of a system of conservation laws and it is valid for any rectangular control volume $[x_1, x_2] \times [t_1, t_2]$.

2.2 Finite volume methods.

The Finite Volume Method (FVM) is a discretisation technique for partial differential equations, especially those that arise from physical conservation laws. FVM uses a volume integral formulation of the problem with a finite partitioning set of volumes to discretise the equations. Following the FVM's approach, the domain is divided into a number of control volumes or cells where the variable of interest is located at the centroid of the control volume, as shown in Figure 2. Then the equations are integrated over each control volume. In this way, the discretisation of the equations expresses the conservation principle for the variables inside the control volume. The most compelling feature of the FVM is that the resulting solution satisfies the conservation of quantities such as mass, momentum, energy, etc. This is exactly satisfied for any control volume as well as for the whole computational domain and for any number of control volumes. Even a coarse grid solution exhibits exact integral balances.

Starting from the integral form of our system of conservation laws (31), we discretise space and time in the following way:

$$\begin{cases} x & \rightarrow & i\Delta x \\ t & \rightarrow & n\Delta t \end{cases} \tag{32}$$

and we denote a generic control volume as $[x_{i-1/2}, x_{i+1/2}] \times [t_n, t_{n+1}]$. By replacing our generic control volume in the equation (31), we have:

$$\begin{aligned} \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x,t_{n+1})\mathrm{d}x = &\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x,t_n)\mathrm{d}x \\ &+ \int_{t_n}^{t_{n+1}} \mathbf{F}(\mathbf{U}(x_{i-1/2},t))\mathrm{d}t - \int_{t_n}^{t_{n+1}} \mathbf{F}(\mathbf{U}(x_{i+1/2},t))\mathrm{d}t \end{aligned} \tag{33}$$

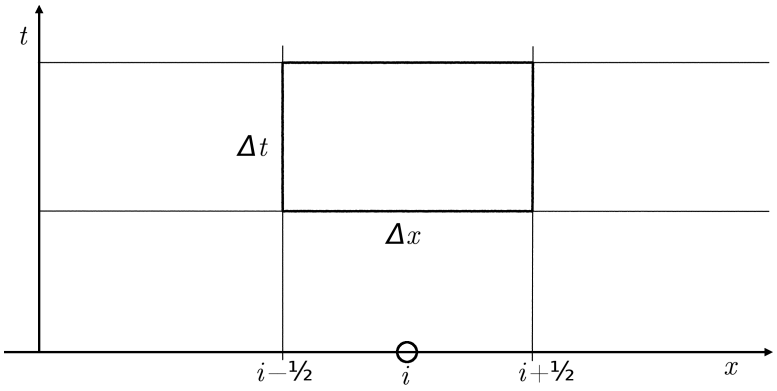

Figure 2: Illustration of a finite volume in 1D.

From the definition of integral average,

$$\bar{w} = \frac{1}{\Delta a}\int_{a}^{a+\Delta a} w(a)\mathrm{d}a, \tag{34}$$

we can rewrite the first term on right hand side of the equation (33) as

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x,t_n)\mathrm{d}x = \Delta x\bar{\mathbf{U}}. \tag{35}$$

The quantity $\bar{\mathbf{U}}$ contains a cell average of each conserved quantity and we assign it to each grid cell, obtaining our discretised quantities,

$$\bar{\mathbf{U}} = \mathbf{U}_i^n. \tag{36}$$

In the same way the cell averaged vector of conserved quantities at time t_{n+1} is given by the term on the left hand side of the equation (33) as follows:

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x,t_{n+1})\mathrm{d}x = \Delta x\mathbf{U}_i^{n+1}. \tag{37}$$

Furthermore, the last two terms on the right hand side of the equation (33) are related to the time averaged fluxes at the boundaries $x_{i\pm1/2}$ of the i^{th} cell. We introduce the following notation for the *numerical flux* $\mathbf{F}_{i\pm1/2}$:

$$\mathbf{F}_{i\pm1/2} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{F}(\mathbf{U}(x_{i\pm1/2},t))\mathrm{d}t. \tag{38}$$

The numerical flux is a (method dependent) numerical approximation of the physical flux. Putting everything together, our system of conservation laws reads:

$$\Delta x \mathbf{U}_i^{n+1} = \Delta x \mathbf{U}_i^n - \Delta t \big[\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}\big], \tag{39}$$

or equivalently:

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}\big[\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}\big]. \tag{40}$$

Because the flux entering a given cell is identical to that leaving the adjacent cell, the equation (40) represents a conservative scheme.

The numerical flux can be evaluated involving an arbitrary number of adjacent cells. A general numerical flux can be written as

$$\mathbf{F}_{i\pm1/2} = \mathbf{F}_{i\pm1/2}(\mathbf{U}_{i-l_L}, \ldots, \mathbf{U}_{i+l_R}), \tag{41}$$

where l_L and l_R are two non-negative integers. For any choice of l_L and l_R a corresponding conservative scheme results if the following consistency condition is satisfied:

$$\mathbf{F}_{i\pm1/2}(\mathbf{V}, \ldots, \mathbf{V}) = \mathbf{F}(\mathbf{V}) \tag{42}$$

This means that if the arguments used for the evaluation of the numerical fluxes are equal (e.g. when $\mathbf{U}$ represents a uniform distribution), then the numerical flux is the same as the physical flux.

3 Godunov's Method

When we deal with a conservative scheme in the form (40), we need to compute the fluxes at the interfaces of each cell in order to advance the solution from time t_n to time t_{n+1}. By exploiting the definition of cell average, we can construct at each time step a piece-wise constant distribution as shown in Figure 3. In this way we have a set of local Riemann problems at cell boundaries, each one is defined as:

$$\begin{cases} \mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = 0 \\ \mathbf{U}(x,0) = \begin{cases} \mathbf{U}_L & \text{for} \quad x_{i-1} < x < x_{i-1/2} \\ \mathbf{U}_R & \text{for} \quad x_{i-1/2} < x < x_i \end{cases} \end{cases} \tag{43}$$

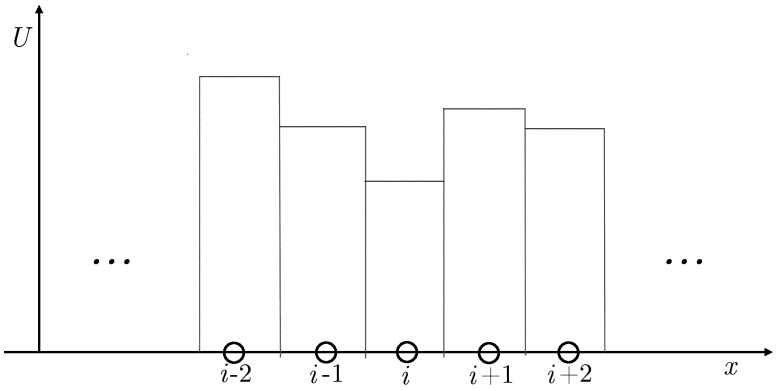

Figure 3: Piece-wise constant distribution used in Godunov methods.

As we can easily understand, in order to compute the fluxes at cell interfaces, we have to solve two Riemann problems for each cell. For example, if we are considering the i^{th} cell and we want compute the fluxes at the cell boundaries $x_{i\pm1/2}$, we have to solve the two Riemann problems $RP(\mathbf{U}_{i-1}, \mathbf{U}_i)$ and $RP(\mathbf{U}_i, \mathbf{U}_{i+1})$. Futhermore, since solutions of the two Riemann problems are constant in time along the $x_{i\pm1/2}$ axis, we do not need a time integral in evaluating the numerical flux, as required by (38), the CFL condition requires

$$\max(|\lambda_i|)\frac{\Delta t}{\Delta x} \leq 1, \tag{44}$$

where $\max(|\lambda_i|)$ is the absolute value of the highest eigenvalue of the Jacobian matrices (i.e. the highest wave speed of the entire system). If the condition (44) is satisfied, then no waves will travel across discontinuities of adjacent Riemann problems and no waves will interact with the solutions of other Riemann problems.

The idea described above was originally from Godunov. This method is called Godunov's scheme and it is a first order scheme. Higher order Godunov's schemes can be built, but for the purpose of these notes we will analyse only the first order scheme.

Example: Linear advection equation.. The linear advection equation describe the passive advection of a scalar field y at a constant velocity c:

$$y_t + f(y)_x = 0 \qquad \text{with} \qquad f(y) = cy \tag{45}$$

where $f(y)$ represent the flux of the scalar field y. Here we assume, without loss of generality, that the constant velocity c is positive. Since we are considering only one equation we have only one characteristic, which is defined by

$$x - ct = 0. \tag{46}$$

In order to find the flux at the boundary within a local Riemann problem, we do not need to do additional calculations, because the solution of each Riemann problem will always be:

$$y(x,t) = \begin{cases} y_L & \text{for} \quad x - ct < 0 \\ y_R & \text{for} \quad x - ct > 0 \end{cases} \tag{47}$$

By looking at Figure 4 we can see that since the velocity c is constant and positive, the flux at the boundary will always be given by

$$f_{i\pm 1/2} = cy_L \tag{48}$$

By considering a generic couple of Riemann problems, then we have:

$$\begin{cases} RP(y_{i-1}, y_i): & f_{i-1/2} = cy_L = cy_{i-1} \\ RP(y_i, y_{i+1}): & f_{i+1/2} = cy_L = cy_i \end{cases} \tag{49}$$

In this way, the Godunov method reduces to the scheme:

$$\begin{aligned} y_i^{n+1} &= y_i^n - \frac{\Delta t}{\Delta x}\left[f_{i+1/2} - f_{i-1/2}\right] \\ &= y_i^n - \frac{c\Delta t}{\Delta x}\left[y_i - y_{i-1}\right] \end{aligned} \tag{50}$$

The last expression of equation (50) corresponds to the upwind method discussed in previous lectures. The upwind scheme is one of the numerical method under the class of CIR methods.

Example: Non-linear gas dynamics: isothermal gas.. An isothermal gas is characterised by the equation of state:

$$p = k\rho, \tag{51}$$

where p is the pressure of the gas and k is the entropic function, which is constant during the evolution of an isothermal process.

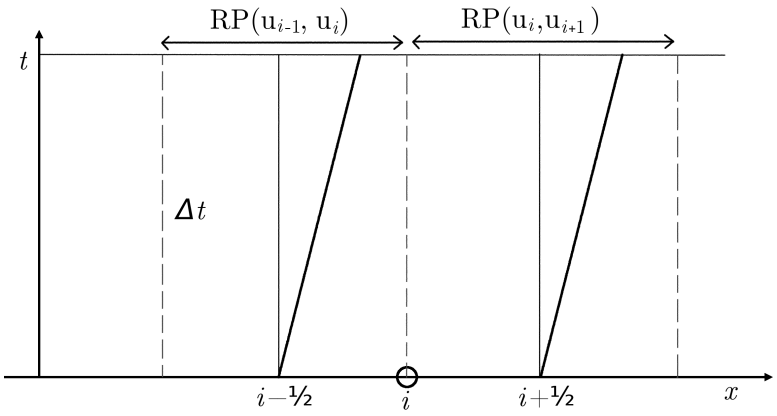

Figure 4: Solution if the Riemann problem for the advection equation with positive advection velocity.

The system of equations governing the gas dynamics, written in terms of conserved variables, reads:

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ &\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0. \end{aligned} \tag{52}$$

Since for an isothermal gas the speed of sound is constant and given by

$$c_s = \sqrt{\frac{\partial p}{\partial \rho}} = \sqrt{k}, \tag{53}$$

we can rewrite the system (52) as follows:

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ &\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + c_s^2 \rho) = 0. \end{aligned} \tag{54}$$

We can also write the system of conservation laws on matrix form (equation (3)) by defining the vectors:

$$\mathbf{U} = \begin{pmatrix} \rho \\ u\rho \end{pmatrix} \qquad \text{and} \qquad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u^2 + c_s^2 \rho \end{pmatrix} \tag{55}$$

As we have seen in the previous chapters, a system of conservation laws can be written using the Jacobian matrix of the system:

$$\mathbf{U}_t + \mathbf{A}(\mathbf{U})\mathbf{U}_x = 0 \qquad \text{with} \qquad \mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 \\ c_s^2 - u^2 & 2u \end{pmatrix}. \tag{56}$$

In this case the Jacobian matrix is not constant and to construct an exact Riemann solver can be tricky. For this reason we choose to employ an approximate Riemann solver: the Roe solver. Roe's approach to the problem of a non-linear Jacobian matrix consists into approximating the matrix $\mathbf{A}(\mathbf{U})$ with a constant matrix:

$$\mathbf{A}(\mathbf{U}) \to \tilde{\mathbf{A}}, \tag{57}$$

such that the flux of conserved quantities is given by:

$$\mathbf{F}(\mathbf{U}) \simeq \tilde{\mathbf{A}}\mathbf{U}. \tag{58}$$

The approximation of the Jacobian matrix must satisfy the following conditions:

- $\tilde{\mathbf{A}}$ has only real eigenvalues and a complete set of eigenvectors, in order to guarantee that the new linear system is truly hyperbolic and that the linear Riemann problem is solvable;
- $\tilde{\mathbf{A}}(\mathbf{U}_L, \mathbf{U}_R) \to \tilde{\mathbf{A}}(\mathbf{U})$ for $\mathbf{U}_L, \mathbf{U}_R \to \mathbf{U}$, in order to preserve the conservation properties of the numerical scheme;
- $\Delta\mathbf{F} = \tilde{\mathbf{A}}\Delta\mathbf{U}$, in order to preserve the jumps in the solution and to be consistent with the old system of conservation laws.

In order to find a matrix $\tilde{\mathbf{A}}$ satisfying all the previous conditions, we perform a change of variables:

$$\mathbf{U} \qquad \to \qquad \mathbf{Z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \frac{1}{\sqrt{\rho}}\mathbf{U} = \begin{pmatrix} \sqrt{\rho} \\ \sqrt{\rho}u \end{pmatrix}. \tag{59}$$

Then, in order to define a constant state at the discontinuity, we average the right and left values of the new set of variables:

$$\bar{\mathbf{Z}} = \frac{1}{2}\left(\mathbf{Z}_L + \mathbf{Z}_R\right) = \frac{1}{2}\begin{pmatrix} \sqrt{\rho_L} + \sqrt{\rho_R} \\ \sqrt{\rho_L}u_L + \sqrt{\rho_R}u_R \end{pmatrix}. \tag{60}$$

Now it is easy to check that

$$\Delta\mathbf{U} = \begin{pmatrix} 2\bar{z_1} & 0 \\ \bar{z_2} & \bar{z_1} \end{pmatrix} \Delta\mathbf{Z} = \tilde{\mathbf{B}}\Delta\mathbf{Z} \tag{61}$$

and

$$\Delta\mathbf{F} = \begin{pmatrix} \bar{z_2} & \bar{z_1} \\ 2c_s^2\bar{z_1} & 2\bar{z_2} \end{pmatrix} \Delta\mathbf{Z} = \tilde{\mathbf{C}}\Delta\mathbf{Z}. \tag{62}$$

The change of variables was performed in order to be able to express $\Delta\mathbf{U}$ and $\Delta\mathbf{F}$ in the form of some matrix times $\Delta\mathbf{Z}$, so that we can readily find the Roe matrix $\tilde{\mathbf{A}}$ as follows:

$$\begin{cases} \Delta\mathbf{U} = \tilde{\mathbf{B}}\Delta\mathbf{Z} \\ \Delta\mathbf{F} = \tilde{\mathbf{C}}\Delta\mathbf{Z} \end{cases} \qquad \Rightarrow \qquad \Delta\mathbf{F} = \tilde{\mathbf{C}}\tilde{\mathbf{B}}^{-1}\Delta\mathbf{U} = \tilde{\mathbf{A}}\Delta\mathbf{U} \tag{63}$$

Hence, for an isothermal gas, the Roe matrix $\tilde{\mathbf{A}}$ reads:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{C}}\tilde{\mathbf{B}}^{-1} = \begin{pmatrix} 0 & 1 \\ c_s^2 - \bar{z}_2^2/\bar{z}_1^2 & 2\bar{z}_2/\bar{z}_1 \end{pmatrix} \equiv \begin{pmatrix} 0 & 1 \\ c_s^2 - \bar{u}^2 & 2\bar{u} \end{pmatrix}, \tag{64}$$

where we have defined the Roe-averaged velocity by:

$$\bar{u} = \frac{\bar{z}_2}{\bar{z}_1} = \frac{\sqrt{\rho_L}u_L + \sqrt{\rho_R}u_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}. \tag{65}$$

We can see the Roe matrix $\tilde{\mathbf{A}}$ as the Jacobian matrix evaluated at the Roe-averaged velocity $\bar{u}$.

The spectral decomposition of the Roe matrix $\tilde{\mathbf{A}}$ reads:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{K}}^{-1} \tag{66}$$

where

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} \tilde{\lambda}_1 & 0 \\ 0 & \tilde{\lambda}_2 \end{pmatrix} \tag{67}$$

is the matrix containing the eigenvalues of $\tilde{\mathbf{A}}$ on its diagonal, and

$$\tilde{\mathbf{K}} = \begin{pmatrix} \vdots & \vdots \\ \tilde{\mathbf{K}}^{(1)} & \tilde{\mathbf{K}}^{(2)} \\ \vdots & \vdots \end{pmatrix} \tag{68}$$

is the matrix whose columns correspond to the eigenvectors of the matrix $\tilde{\mathbf{A}}$. For our case, the spectral decomposition leads to:

$$\tilde{\lambda}_1 = \bar{v} - c_s \qquad \tilde{\lambda}_2 = \bar{v} + c_s \tag{69}$$

$$\tilde{\mathbf{K}}^{(1)} = \begin{pmatrix} 1 \\ \bar{v} - c_s \end{pmatrix} \qquad \tilde{\mathbf{K}}^{(2)} = \begin{pmatrix} 1 \\ \bar{v} + c_s \end{pmatrix} \tag{70}$$

By solving the two linear systems:

$$\mathbf{U}_L = \tilde{\mathbf{K}}\mathbf{C}_L\,, \qquad \mathbf{U}_R = \tilde{\mathbf{K}}\mathbf{C}_R\,, \tag{71}$$

we can determine the coefficients of the eigenvector expansions for $\mathbf{U}_L$ and $\mathbf{U}_R$, respectively denoted as

$$\mathbf{C}_L = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}\,, \qquad \mathbf{C}_R = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}\,. \tag{72}$$

Once we have computed the coefficients of the eigenvector expansions, we can calculate the solution at the discontinuity as:

$$\tilde{\mathbf{U}} = \sum_{j:\tilde{\lambda}_j>0} \alpha_j \tilde{\mathbf{K}}^j + \sum_{j:\tilde{\lambda}_j<0} \beta_j \tilde{\mathbf{K}}^j \tag{73}$$

Then the flux at the discontinuity is given by:

$$\mathbf{F} = \tilde{\mathbf{A}}\tilde{\mathbf{U}} \tag{74}$$

Solving a generic couple of adjacent Riemann problems we end up with:

$$\begin{cases} RP(\mathbf{U}_{i-1}, \mathbf{U}_i): & \rightarrow & \mathbf{F}_{i-1/2} = \tilde{\mathbf{A}}_{i-1/2}\tilde{\mathbf{U}}_{i-1/2} \\ RP(\mathbf{U}_i, \mathbf{U}_{i+1}): & \rightarrow & \mathbf{F}_{i+1/2} = \tilde{\mathbf{A}}_{i+1/2}\tilde{\mathbf{U}}_{i+1/2} \end{cases} \tag{75}$$

and the solution of the system of conservation laws is updated through the scheme in equation (40):

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}\left[\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}\right].$$