Eight Lectures on Discretisation Principles and Hyperbolic Conservation Laws

P.W. Hemker

Version: December 13, 1999

Contents

Books used:
[52, 20, 19, 31, 58, 60, 26, 16, 42, 22, 23, 43, 41, 49, 34, 37, 8, 11, 51]

1

Introduction to PDEs

1.1 Classification

In these lectures we study numerical methods for partial differential equations (PDEs). In the later part, in particular, we will concentrate on so called elliptic equations. These equations are used in many areas of application, as e.g. mechanics, electromagnetics, etc..

To get insight into the quantitative behaviour of the solutions of the PDEs, and to determine the solutions with some accuracy, in most cases analytic methods fail and we have to rely on numerical methods. The numerical techniques to compute (approximations to) the solution are often split in two parts. First, the PDE (the solution of which is a continuous function or set of functions) is transformed into a set of linear or nonlinear algebraic equations, usually with a large number of unknowns. This transformation is called the discretisation process. The second step is the solution of the large algebraic system of equations, in order to determine the unknowns that describe (approximately) the solution of the PDE.

In the present chapter different kinds of differential equations are classified, and some elementary properties of the PDEs are explained. In Chapter 2 we will give a survey of the methods that are used for discretisation, for different kinds of PDEs. In Chapter 3 we study the Finite Element Method for the discretisation of elliptic equations.

1.1.1 Ordinary and partial differential equations

An *ordinary differential equation* (ODE) implicitly describes a function depending on a single variable and the ODE expresses a relation between the solution and one or more of its derivatives. The *order of the differential equation* is the order of the highest derivative in the equation. Beside the ODE, usually one or more additional (initial) conditions are needed to determine the unknown function uniquely.

Example 1.1.1
An example of a first order ODE is

$$u(x) = u'(x), \quad u(0) = 1.$$

An example of a second order ODE is

$$u(x) = -c^2\, u''(x), \quad u(0) = 0, u'(0) = 1.$$

In many interesting cases a function depends on more independent variables and a relation is given for the function and its partial derivatives. This relation describes a *partial differential equation* (PDE). In many practical problems the independent variables represent the time and space coordinates (t, x, y, z). To determine the unknown function, again additional relations are required: initial and/or boundary conditions.

Example 1.1.2
Let $u(t, x, y, z)$ denote the mass fraction of a chemical species, m, in a medium with density $\rho(t, x, y, z)$. In the presence of a velocity field $\mathbf{v}(t, x, y, z)$ the conservation of the species m is expressed by

$$\frac{\partial}{\partial t}(\rho u) + \text{div } (\rho u \mathbf{v}) = s, \tag{1.1}$$

where $s(t, x, y, z)$ is a possible source [1].

If diffusion also plays a role, then the *diffusive flux* $\mathbf{J}_d$ is described by Fick's law of diffusion [2]

$$\mathbf{J}_d = -d \, \text{grad } u,$$

where d is the diffusion coefficient, and the complete equation describing the behaviour of $u(t, x, y, z)$ reads

$$\frac{\partial}{\partial t}(\rho u) + \text{div } (\rho \mathbf{v} u) + \text{div } \mathbf{J}_d = s. \tag{1.2}$$

In this equation we distinguish between the diffusive flux, $\mathbf{J}_d$, and the *convective flux*, $\mathbf{J}_c = \rho u \mathbf{v}$.

The *principle part* of a differential equation is the part (the set of terms) that contain the highest derivatives. E.g. the principle part of (1.1) is $(\rho u)_t +$ div $(\rho u \mathbf{v})$ and the principle part of (1.2) is $(\rho u)_t -$ div $(d \, \text{grad } u)$. PDEs can have quite diverse properties. A few of such typical forms are given in

---

[1] The *divergence* of the vector field $\mathbf{v}$ is defined by div $\mathbf{v} = \nabla \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$. An important relation for the divergence is *Gauss' theorem*

$$\int_\Omega \text{div } \mathbf{u} \, d\Omega = \oint_\Gamma \mathbf{u} \cdot \mathbf{n} \, d\Gamma,$$

where $\Omega \subset \mathbb{R}^3$ is a bounded volume with surface Γ and $\mathbf{n}$ is the outward pointing normal on Γ.

[2] The *gradient gradient* of a scalar function $\phi(x, y, z)$ is the vector field defined by

$$\text{grad } \phi = \nabla \phi = \begin{pmatrix} \partial \phi / \partial x \\ \partial \phi / \partial y \\ \partial \phi / \partial z \end{pmatrix}.$$

the next example. Most typical properties are related with the shape of the principle part of the equation.

Example 1.1.3

- With $\Omega \subset \mathbb{R}^2$ an open domain, the *Poisson equation* is

$$u_{xx} + u_{yy} = f(x,y), \tag{1.3}$$

  where $f(x,y)$ is a prescribed function. In the special case of $f \equiv 0$ equation (1.3) is called *Laplace's equation*. The solution of these equations can be determined if the value of $u(x,y)$ is given at the boundary of Ω. Notice that the equation (in its three-dimensional form) is obtained from (1.2) in the case of a time-independent solution with $\mathbf{v} = 0$ and $s = 0$.

- The *diffusion equation*

$$u_t = d\,u_{xx}, \tag{1.4}$$

  describes the one-dimensional unsteady case of simple diffusion, without a velocity field or a source term.

- A different kind of PDE is the *wave equation*

$$u_{tt} = a^2 u_{xx}, \tag{1.5}$$

  where a is a positive real number. It is simple to see that a solution of this equation is given by

$$u(x,y) = \phi(at - x) + \psi(at + x), \tag{1.6}$$

  where ϕ and ψ are arbitrary (twice differentiable) functions.

All the above mentioned examples are differential equations that are linear in the dependent variable u. Although most theory available is for such linear partial differential equations, many important equations are nonlinear. E.g. the equation (1.1) or (1.2) would be nonlinear if the velocity field $\mathbf{v}$ was dependent on u or the source s was a nonlinear function of u. An important subclass of the nonlinear differential equations are the *quasi-linear* equations, i.e. equations in which the highest derivatives appear linearly.

In general, the dependent variable u in the PDE is a function of time, t, and the three space coordinates, (x, y, z). In numerical methods we have to choose how to represent the function u, and often we select values of the independent variables at which the values of u will be calculated. The fewer the number of degrees of freedom in the approximation of $u(t, x, y, z)$, the less computational effort is needed to determine an approximation of $u(t, x, y, z)$. Fortunately, not all problems require all independent variables.

By symmetry considerations (i.e. by a judicious choice of the coordinate system) the number of space dimensions can often be reduced to two or one.

If a solution is sought that is not time-dependent the solution is called steady and the time-variable can be neglected. The corresponding equations are the *steady equations*.

To understand the behaviour of the solution it is important to notice that in some equations we can distinguish a direction of "flow of information". For the solution (1.6) of equation (1.5) we see that the information contained in the part ϕ of the solution flows forward (i.e. in the positive x-direction) with speed a, whereas the information in the part ψ flows backward as time proceeds. Also, in equation (1.1) the flow of information follows the vector field $\mathbf{v}$. However, no such direction is found in (1.3). For time-dependent PDEs such as (1.2) or (1.4) we can always distinguish a unique direction "forward in time".

The motivation for the discussion of "direction of flow of information" is that, if such a direction can be identified, the behaviour of the solution can be better understood. The knowledge also may be used to reduce the computer storage and/or the computer time needed to find the numerical solution. E.g. in a time-dependent problem as (1.1), for which the solution has to be computed over a long time-interval $[t_0, t_2]$, the solution at time $t = t_1$, with $t_0 < t_1 < t_2$, contains all information to determine the solution for $t > t_1$. Having computed the solution for $t \leq t_1$, it may be efficient -from the point of view of computer resources- to disregard all or most information that was obtained about the solution for $t < t_1$ (or $t < t_1 - \tau$, for some small $\tau > 0$) before the solution on $(t_1, t_2]$ is determined.

The notion of direction "forward in time" shows also that in time-dependent problems all or part of the side-conditions will be given as "initial conditions", that determine the solution (i.e. the state of the physical system) at a later stage.

1.1.2 Elliptic, parabolic or hyperbolic equations

In the previous section we noticed that different PDEs may show a different character. This character is mainly determined by the highest derivatives that appear in the equation. To characterise some typical classes of PDEs, we consider the general linear second order differential equation in two space dimensions:

$$\begin{aligned} Lu := \quad & a(x,y)u_{xx} + 2\,b(x,y)u_{xy} + c(x,y)u_{yy} \\ & + 2\,d(x,y)u_x + 2\,e(x,y)u_y + f(x,y)u = g(x,y). \end{aligned} \tag{1.7}$$

L is a linear differential operator of order two. With this differential operator we associate the *characteristic polynomial*

$$\begin{aligned} P_{xy} = \quad & a(x,y)\xi^2 + 2\,b(x,y)\xi\eta + c(x,y)\eta^2 + \\ & + 2\,d(x,y)\xi + 2\,e(x,y)\eta + f(x,y). \end{aligned} \tag{1.8}$$

The second order differential equation (1.7) is called *elliptic* at (x,y) if $a(x,y)c(x,y) - b^2(x,y) > 0$; it is *hyperbolic* at (x,y) if $a(x,y)c(x,y) - b^2(x,y) < 0$, and it is *parabolic* at (x,y) if $a(x,y)c(x,y) - b^2(x,y) = 0$. Thus, this terminology is related with the geometric interpretation of the characteristic polynomial. The differential equation is called elliptic, parabolic or hyperbolic (without reference to a specific point) if it satisfies the corresponding (in)equalities for all $(x,y) \in \Omega$, where Ω is the domain on which the differential equation is defined. We see that the Poisson equation, the diffusion equation and the wave equation are elliptic, parabolic and hyperbolic, respectively.

For an equation of more variables $(x_1, x_2, ..., x_n)$, the general second order equation reads

$$Lu := \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) u_{x_i x_j} + \sum_{i=1}^{n} a_i(\mathbf{x}) u_{x_i} + a(\mathbf{x}) u. \tag{1.9}$$

Now, for the *principle part* of the equation, the *characteristic polynomial* in $\xi_1, \xi_2, ..., \xi_n$ is

$$P_{\mathbf{x}}(\xi_1, ..., \xi_n) = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\, \xi_i \xi_j. \tag{1.10}$$

The operator is elliptic if the matrix $(a_{ij}(\mathbf{x}))$ is positive or negative definite; it is hyperbolic if $(a_{ij}(\mathbf{x}))$ has one negative (positive) eigenvalue and $n-1$ positive (negative) eigenvalues; it is parabolic if one eigenvalue vanishes and the other $n-1$ have the same sign. It is clear that in more dimensions not all linear second order PDEs can be classified as elliptic, parabolic or hyperbolic.

The characterisation of the different PDEs is important for distinguishing the kind of additional conditions that is needed to specify a (unique) solution for the equation and for understanding the "flow of information" that may appear in its solution. We show this for the three linear examples given earlier in this section.

The hyperbolic equation

For the wave equation (1.5) with initial conditions $u(0,x) = f(x)$, $u_t(0,x) = g(x)$, the general solution for $t \geq 0$ reads

$$u(t,x) = \frac{1}{2} f(x+at) + \frac{1}{2} f(x-at) + \frac{1}{2a} \int_{x-at}^{x+at} g(\xi)\, d\xi. \tag{1.11}$$

This means that the solution at (t,x) depends only on the initial solution on the interval $[x-at, x+at]$.

The parabolic equation

For the linear diffusion equation (1.4), with initial condition $u(0,x) = f(x)$, with $x \in \mathbb{R}$, the solution reads

$$u(t,x) = \frac{1}{2\sqrt{\pi d t}} \int_{-\infty}^{+\infty} f(\xi)\, e^{-(\xi - x)^2/(4dt)}\, d\xi. \tag{1.12}$$

This implies that the solution at $x \in \mathbb{R}$ for any $t > 0$ depends on all of the solution $u(0,x)$ over the interval $-\infty < x < +\infty$.

The elliptic equation

For Laplace's equation, written in polar coordinates (r, ϕ), it is known from the theory of harmonic functions that the solution $u(x,y) = \tilde{u}(r,\phi)$ in the interior of a circle with radius R, is given by

$$\tilde{u}(r,\phi) = \frac{1}{2\pi} \oint_0^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \phi) + r^2}\, \tilde{u}(R,\theta)\, d\theta. \tag{1.13}$$

This shows that the behaviour of the solution inside a circle, $r = R$, is completely determined by *all* the values of the solution at the boundary of that circle. Generally, the function of an elliptic second order equation is determined by its values at the boundary of its domain of definition.

1.1.3 Conservation laws

A slight extension of equation (1.1) is the system of equations

$$\frac{\partial}{\partial t}\mathbf{u}(t,\mathbf{x},\mathbf{v}) + \operatorname{div}\, \mathbf{J}(t,\mathbf{x},\mathbf{v})) = \mathbf{s}(t,\mathbf{x},\mathbf{v}), \tag{1.14}$$

where $\mathbf{v}(t,\mathbf{x})$ is the solution sought, $\mathbf{u}(t,\mathbf{x},\mathbf{v})$ is a state vector and $\mathbf{J}(t,\mathbf{x},\mathbf{v})$ is the flux of the state variables[3]; $\mathbf{s}(t,\mathbf{x},\mathbf{v})$ is a possible source term.

Equation(1.14) is called a system of conservation laws because the equations describe the conservation of the quantities $\mathbf{u}$. This is seen by considering an arbitrary volume Ω and by integration of (1.14) over Ω. Using Gauss' theorem, we see that the increase of the amount of $\mathbf{u}$ inside the volume Ω should be caused either by inflow over its boundary or by a source described in $\mathbf{s}(t,\mathbf{x},\mathbf{v})$.

In natural coordinates, a physical system can often be described with $\mathbf{u} = \mathbf{u}(\mathbf{v})$, $\mathbf{J} = \mathbf{J}(\mathbf{v})$, $\mathbf{s} = \mathbf{s}(\mathbf{v})$. Then, the dependence on $(t,\mathbf{x})$ may be caused by a coordinate transformation.

---

[3] Because we consider a system of equations, now each component of $\mathbf{J}$ is a vector.

Example 1.1.4 The Euler equations.
The *Euler equations* of gas dynamics describe the flow of an inviscid non-heat-conducting compressible fluid (a gas). They represent the conservation of mass, momentum and energy. With $\rho(t,\mathbf{x})$ density, $\mathbf{v}(t,\mathbf{x})$ velocity, $e(t,\mathbf{x})$ specific energy (temperature) and $p(t,\mathbf{x})$ the presure of the gas, these conservation laws read respectively

$$\frac{\partial}{\partial t}(\rho) + \text{div }(\rho\mathbf{v}) = 0, \tag{1.15}$$

$$\frac{\partial}{\partial t}(\rho v_i) + \text{div }(\rho v_i\mathbf{v}) = -\frac{\partial p}{\partial x_i}, \quad i = 1,2,3,$$

$$\frac{\partial}{\partial t}(\rho e) + \text{div }(\rho e\mathbf{v}) = -\text{div }(p\mathbf{v}),$$

and $p = p(\rho, e, \mathbf{v})$. For a perfect gas we have the state equation

$$p = (\gamma - 1)\rho(e - \frac{1}{2}|\mathbf{v}|^2). \tag{1.16}$$

With suitable boundary conditions the variables $\rho, \mathbf{v}, e$ and p can be solved from (1.15) and (1.16). These variables play the role of the vector $\mathbf{v}$ in (1.14). Notice that they are different from the conserved quantities $\mathbf{u}$.

1.2 Hyperbolic equations

The second order equation (1.5) allows decomposition into a system of two first order equations. Similar to the factorisation of the characteristic polynomial $P(\tau,\xi) = \tau^2 - a^2\xi^2 = (\tau + a\xi)(\tau - a\xi)$ of (1.5), the first order operators are found to be $\frac{\partial}{\partial t} + a\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}$. The decomposition first can be written[4] as a system

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t} = -a\frac{\partial p}{\partial x}, \\ \frac{\partial p}{\partial t} = -a\frac{\partial u}{\partial x}. \end{array} \right. \tag{1.17}$$

If we make the change of dependent variables $v = u + p$ and $w = u - p$, this equation (1.17) uncouples and transforms into

$$\left\{ \begin{array}{l} \frac{\partial v}{\partial t} + a\frac{\partial v}{\partial x} = 0, \\ \frac{\partial w}{\partial t} - a\frac{\partial w}{\partial x} = 0. \end{array} \right. \tag{1.18}$$

The solution of the system reads $v(t,x) = v(x - at)$, $w(t,x) = w(x + at)$. These equations show that the quantities $u + p$ (respectively $u - p$) are constant along the line $x - at =$ constant $(x + at =$ constant). These

[4] The minus sign is by convention only.

lines are called the *characteristics* of the differential equation (or of the differential operator).

A typical first order "one-way" wave equation in $\mathbb{R}^d$, $d = 1, 2, 3$, is

$$\frac{\partial u}{\partial t} - \mathbf{v} \cdot \operatorname{grad} u = 0. \tag{1.19}$$

It shares with (1.18) the property of having a characteristic direction, $\mathbf{v}$, along which the solution is constant. Therefore also the first order equations are called hyperbolic PDE. To illustrate further the importance of the characteristics, we consider the inhomogeneous equation

$$\frac{\partial u}{\partial t} - \mathbf{v} \operatorname{grad} u = s(u). \tag{1.20}$$

We assume that the vector field $\mathbf{v}(\mathbf{x})$ is sufficiently smooth so that it determines a complete family of characteristics in Ω. Then we will show in Section 1.2.1 that the solution of this PDE is uniquely determined as soon as *one* value of the solution is given for each characteristic. Along the characteristic the solution is now determined by a simple ODE.

1.2.1 Characteristics

We consider a general first order (quasi-)linear DE in more dimensions. We write it down for dimension $d = 3$; for the one-, two- or more- dimensional case the treatment is completely analogous. We consider the equation

$$Pp + Qq + Rr = S\,, \tag{1.21}$$

where $p = u_x$, $q = u_y$, $r = u_z$, and P, Q, R, S are functions of (x, y, z, u). I.e. (x, y, z) are the independent variables and u is the dependent variable.

Let $u(x, y, z)$ be the solution of (1.21), then, by definition of p, q, and r, (1.21) is satisfied as well as

$$du = pdx + qdy + rdz.$$

Or, if A is a point (x, y, z, u) on a solution, then at A we find

$$\begin{array}{lcl} (p, q, r, -1) & \perp & (P, Q, R, S), \\ (p, q, r, -1) & \perp & (dx, dy, dz, du), \\ (p, q, r, -1) & \perp & (\dot{x}, \dot{y}, \dot{z}, \dot{u}), \end{array}$$

with $\dot{} = \partial/\partial s$, and $A(s)$ is an (arbitrary) curve in a solution surface. Apparently, the vector field $(P, Q, R, S)(x, y, z, u)$ is tangential to the solution in each point of the solution. Streamlines in this vector field are characteristic solutions. The projections of these characteristic solutions on the (x, y, z)-space are characteristics (characteristic direction lines). For a solution along a characteristic solution line (parametrised by s) we have

$$(P, Q, R, S) \parallel (\dot{x}, \dot{y}, \dot{z}, \dot{u}), \tag{1.22}$$

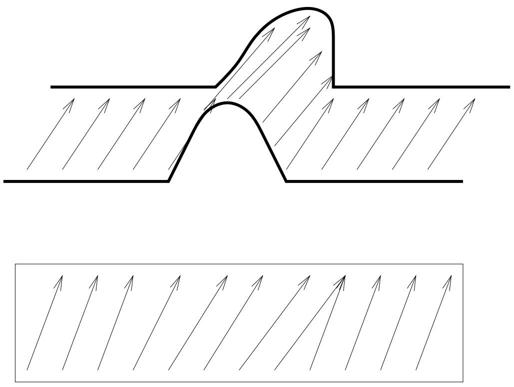


FIGURE 1.1. The generation of a shock for Burgers' equation

i.e. the two vectors are linearly dependent. It follows that the characteristic direction can be computed from

$$\frac{P}{dx} = \frac{Q}{dy} = \frac{R}{dz} = \frac{S}{du}. \tag{1.23}$$

Further, it follows that, if a value of (x, y, z, u) is given in one point of a characteristic, then the characteristic and the solution on it can be found by the solution of the simple ODE (1.23).

If the solution is prescribed on a suitable $(d-1)$-dimensional manifold (suitable means that the manifold intersects all characteristics once) then the solution is apparently uniquely determined.

Example 1.2.1 Inviscid Burgers' equation.
inviscid Burgers' equation A well-known model problem is the inviscid Burgers' equation

$$u_t + (\frac{1}{2}u^2)_x = 0\,.$$

(See also example (1.2.2)). We notice that

$$\frac{1}{dt} = \frac{u}{dx} = \frac{0}{du}\,.$$

The characteristic satisfies $dx/dt = u$ and $du/dt = 0$, so that the solution is constant along the characteristics, and the characteristics are straight lines. If the initial solution is increasing in some area, the characteristics will intersect and no smooth solution will further be possible. This will be considered further in the next section.

We may study the above arguments more generally in d-dimensional space. Let a solution be prescribed along a $(d-1)$-dimensional manifold

C. Under what conditions is a solution determined in a layer in the neighbourhood of C? We easily extend the question to the case of systems of equations. For simplicity we take $d = 3$. Let

$$\begin{cases} a_1 u_x + b_1 u_y + c_1 u_z + d_1 v_x + e_1 v_y + f_1 v_z + \cdots = g_1 , \\ a_2 u_x + b_2 u_y + c_2 u_z + d_2 v_x + e_2 v_y + f_2 v_z + \cdots = g_2 , \\ \cdots , \end{cases} \tag{1.24}$$

or, in short,

$$\sum_{\substack{j=1,2,3 \\ i=1,2,\cdots,m}} a_{kij} \frac{\partial u_i}{\partial x_j} = b_k , \quad k = 1, \cdots , m \, . \tag{1.25}$$

We introduce a function $\phi(x_1, x_2, x_3)$ such that a surface Φ is determined by $\phi(x_1, x_2, x_3) = 0$. (The surface Φ is our $(d-1)$-dimensional manifold, with d=3.) Moreover, we introduce a new set of independent variables (ϕ_1, ϕ_2, ϕ_3) such that the surface is determined by $\phi_1 =$ constant; e.g. $\phi_1 = 0$. (I.e. we introduce local coordinates around the surface Φ.) Now we have

$$\frac{\partial u_i}{\partial x_j} = \sum_{l=1,2,3} \frac{\partial u_i}{\partial \phi_l} \frac{\partial \phi_l}{\partial x_j} \cdot$$

Equation (1.25) becomes

$$\sum_{\substack{l,j=1,2,3 \\ i=1,2,\cdots,m}} a_{kij} \frac{\partial u_i}{\partial \phi_l} \frac{\partial \phi_l}{\partial x_j} = b_k , \quad k = 1, \cdots , m \, , \tag{1.26}$$

and we get

$$\sum_{i=1,\cdots,m} \left[\sum_{j=1,2,3} a_{kij} \frac{\partial \phi_1}{\partial x_j} \right] \frac{\partial u_i}{\partial \phi_1} = b_k - \sum_{\substack{l=2,3 \\ j=1,2,3 \\ i=1,2,\cdots,m}} a_{kij} \frac{\partial u_i}{\partial \phi_l} \frac{\partial \phi_l}{\partial x_j} . \tag{1.27}$$

This is a linear $m \times m$-system for the computation of $\partial u_i / \partial \phi_1$, i.e. the normal component of u_i on the surface Φ. For the computation we need (i) the coefficients a_{kij}; (ii) the right-hand-sides b_k; (iii) $\partial u_i / \partial \phi_l$, $l = 2, 3$, on the surface $\phi_1 = 0$.

If the solution $u(x, y, z)$ of the system of PDEs is known in the surface Φ, then we can also find the normal derivatives (and hence the analytic solution in a layer around Φ), provided that

$$\det \left(\sum_{j=1,2,3} a_{kij} \frac{\partial \phi_1}{\partial x_j} \right) \neq 0. \tag{1.28}$$

If, on the contrary, det $= 0$, then -in general- no smooth solution exists in the neighbourhood of Φ.

1.2.2 Discontinuous solutions

For simplicity, in this section we first consider the one-dimensional case. We have seen that the solution of

$$u_t(t,x) + a\, u_x(t,x) = 0, \tag{1.29}$$

with $a \in \mathbb{R}$, and the initial condition

$$u(0,x) = u_0(x)\,, \quad x \in \mathbb{R}, \tag{1.30}$$

is $u(t,x) = u_0(x - at)$. Also, with $a(x)$ a continuous function, the behaviour of $u(t,x)$ is easily understood for $t > 0$. The problem gets considerably more interesting if the problem is quasi-linear, i.e. if we consider

$$u_t(t,x) + a(u)\, u_x(t,x) = 0, \tag{1.31}$$

where $a(u)$ is a continuous function depending on the solution. Together with the initial condition (1.30), it now turns out that not in all cases a smooth solution exists for all $t > 0$, even if $u_0(x)$ is a smooth function. The reason is that characteristic lines for the equation (1.31) may intersect. As we have seen, a classical solution of (1.31) should be constant on a characteristic line, and different characteristics can carry different values. At points where characteristics intersect a shock-line is formed and a discontinuity appears. The characteristic line ends. But this cannot be described by the differential equation.

There are essentially two ways out of this difficulty. The first approach is to generalise the concept of a solution. The class of possible solutions is extended from the class of differentiable functions to the class of integrable functions, and the equation is considered in a weak sense, i.e. it is written in an integral form. (We see this form later in (1.36).) However, the weak solutions turn out to be non-unique, (for a given set of initial data), and it remains to characterise the "physically relevant" weak solutions.

Before we consider this question, we present the second approach. Instead of (1.31) we now consider the slightly perturbed equation

$$u_t(t,x) + a(u)\, u_x(t,x) = \epsilon u_{xx}, \quad 0 < \epsilon \ll 1. \tag{1.32}$$

This corresponds with the transition from (1.1) to (1.2) and it means that we add a small diffusion term to the original equation without diffusion. Under reasonable conditions for the initial function $u_0(x)$ the elliptic equation (1.32) always has a unique smooth solution $u^\epsilon(t,x)$ for $\epsilon > 0$, and a (non-smooth) limit function $u^0(t,x) = \lim_{\epsilon \to 0} u^\epsilon(t,x)$ exists.

Now we return to the first approach, where we obtain weak solutions of (1.31). As the functions $u^\epsilon(t,x)$ converge to such a weak solution, then that solution is called the *physically relevant* weak solution of (1.31).

The integral or weak form[5] of the conservation law in divergence form[6]

$$u_t + \text{div } \mathbf{J}(u) = 0, \tag{1.34}$$

is found by integration over an area $R = [t_0, t_1] \times \Omega$. The integral form reads

$$\int_{[t_0,t_1]} \int_\Omega u_t + \text{div } \mathbf{J}(u)\, d\mathbf{x}\, dt = 0, \tag{1.35}$$

which can also be written as

$$\int_\Omega u(t_1)\, d\Omega - \int_\Omega u(t_0)\, d\Omega = -\int_{[t_0,t_1]} \oint_{\partial\Omega} \mathbf{n} \cdot \mathbf{J}(u)\, d\Gamma\, dt. \tag{1.36}$$

The integral form holds also if there is a $(d-1)$-dimensional manifold on which $u(t,x)$ is discontinuous. Thus, (1.35) is a generalisation of (1.34).

For $d = 2$ we immediately see (cf. Figure ???) that at a discontinuity D we have

$$\left(\frac{dx}{dt}\right)_D [u] = [\mathbf{J}_\perp(u)], \tag{1.37}$$

where $[u]$ denotes the jump of u over the discontinuity. Denoting the propagation speed of the discontinuity by s, we see[7]

$$s = \frac{\mathbf{J}_\perp(\mathbf{u}_R) - \mathbf{J}_\perp(\mathbf{u}_L)}{u_R - u_L}. \tag{1.38}$$

---

[5]The general principle of *weak solution* for a differential equation $N(u) = 0$ on $\mathbb{R}^d$ is actually to take an arbitrary $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$, i.e. an arbitrary smooth function with compact support, and to consider the integral

$$\int_0^\infty \int_{\mathbb{R}^d} \phi(t, \mathbf{x})\, N(u) d\mathbf{x}\, dt = 0.$$

In our case, with $N(u) = u_t + \text{div } \mathbf{J}(u)$, on the half space $t \geq 0$, after partial integration this comes down to

$$-\int_0^\infty \int_{\mathbb{R}^d} u\phi_t + \mathbf{J}(u) \cdot \text{grad } \phi\ d\mathbf{x}\, dt = -\int_{\mathbb{R}^d} \phi(0, \mathbf{x})\, u_0(\mathbf{x})\ d\mathbf{x} \tag{1.33}$$

for all $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$. We see that here the requirement of a differentiable u has disappeared from the formulation of the equation. By taking a sequence of functions $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$, that have the characteristic function on $[t_0, t_1] \times \Omega$ as its limit, it can be shown that (1.33) and (1.36) are equivalent.

[6]Notice that (1.19) is written as (1.34) with $\mathbf{J}(u) = -\mathbf{v}u$, with div $\mathbf{v} = 0$, because $\text{div}(-\mathbf{v}u) = -u\nabla_j(v_j u) = -u\nabla_j v_j - v_j \nabla_j u = -\mathbf{v}\text{grad}\, u$.

[7]In more dimensions we get

$$\mathbf{J}_\perp(\mathbf{u}_R) - \mathbf{J}_\perp(\mathbf{u}_L) = s(\mathbf{u}_R - \mathbf{u}_L)$$

where $\mathbf{J}_\perp$ denotes the flux vector component in the direction perpendicular to the discontinuity.

This expression for the shock speed is called the *Rankine-Hugoniot* relation.

Apparently two types of discontinuities may appear in first order hyperbolic equations: linear and nonlinear ones. The first type is present in the initial condition and is carried along a characteristic. It is convected with a speed $\mathbf{v}$ that does not essentially depend on the solution. The second type may appear even if the initial solution is smooth, and the speed essentially depends the solution.

Example 1.2.2 Burgers' equation.
The appearance of a nonlinear discontinuity is often shown for Burgers' equation. In one space dimension it reads

$$u_t + uu_x = \epsilon u_{xx}. \tag{1.39}$$

If the r.h.s. is neglected, i.e. if we take $\epsilon = 0$, in a positive solution the top of a wave moves faster than the bottom. Hence, after a sufficiently long period this would lead to a multiple valued function (if that would be possible). Because such solutions are not allowed, a discontinuity appears of which the speed s is determined by

$$\mathbf{J}(u) = \frac{u^2}{2}, \tag{1.40}$$

and it follows that the discontinuity moves with a speed $s = (u_{\text{left}} + u_{\text{right}})/2$.

If we do allow $\epsilon > 0$, then the r.h.s. will always have a significant value in those regions the solution is steep and the dissipation will prevent the discontinuity to appear[8].

[8] For some initial conditions the shape of the dissipated shock is easy to determine for Burgers' equation. With

$$u(0,x) = \begin{cases} u_{-\infty} = q > 0, & x < 0, \\ u_{\infty} = -q < 0, & x > 0, \end{cases} \tag{1.41}$$

the steady solution of (1.39) is

$$u(\infty, x) = -q \tanh\left(\frac{qx}{2\epsilon}\right). \tag{1.42}$$

Notice that the initial condition is already a weak solution of the reduced Burgers' equation ((1.39) with $\epsilon = 0$). A non-trivial steady solution does not appear for $u_{-\infty} < 0 < u_{\infty}$. In that case, after a first dissipative phase, the nonlinear convection will level out all differences in u.

The dissipated shock moves at a speed $a = (u_L + u_R)/2$, and its shape is given by

$$u(t,x) = \frac{u_L + u_R}{2} - \frac{u_L - u_R}{2} \tanh\left(\frac{(u_L + u_R)(x - at)}{2\epsilon}\right).$$

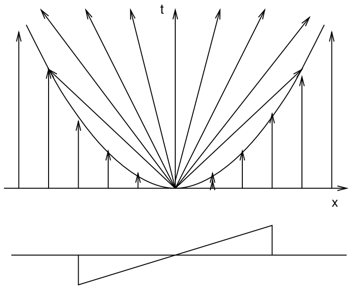


FIGURE 1.2. A curved shock solution for Burgers' equation

Example 1.2.3 *A discontinuous solution.*
As an example (from [37]), the following wave is also a solution to the inviscid *Burgers' equation*:

$$u(t,x) = \begin{cases} x/t, & -\sqrt{t} < x < \sqrt{t}, \\ 0, & \text{otherwise.} \end{cases} \tag{1.43}$$

This solution (see Figure 1.2) has two shocks, propagating with speeds $\pm 1/2\sqrt{t}$. The right shock has left- and right states $u_L = 1/\sqrt{t}$ and $u_R = 0$, so the Rankine-Hugoniot condition is satisfied. The other shock behaves similarly.

We can use the relation (1.38) to solve explicitly some initial value problems that are not classically solvable. However, now it appears that these solutions are not always uniquely determined. We take again the inviscid Burgers' equation as an example.

Example 1.2.4 *A unique solution.*
We take Burgers' equation with initial function

$$u_0(x) = \begin{cases} 1, & x < 0, \\ 1 - x, & 0 \le x \le 1, \\ 0, & x > 1. \end{cases} \tag{1.44}$$

The solution we find reads (see Figure (1.3))

$$u(t,x) = \begin{cases} 1, & x < t, \text{ or} \\ & x < 1 + (t-1)/2, \\ (1-x)/(1-t), & t \le x \le 1, \\ 0, & x > 1, \text{ or} \\ & x > 1 + (t-1)/2. \end{cases} \tag{1.45}$$

Example 1.2.5 *A non-unique solution.*
We take Burgers' equation with initial function

$$u_0(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases} \tag{1.46}$$

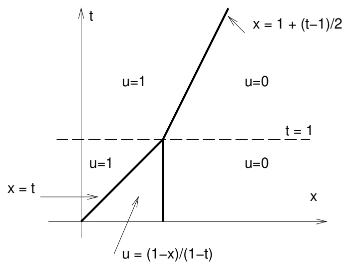


FIGURE 1.3. A continuous solution developing a discontinuity

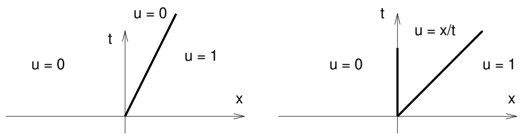


FIGURE 1.4. A discontinuous and a continuous solution solution

The solutions we find are (see Figure (1.4)) are a family of discontinuous functions for $a \in (0,1)$

$$u(t,x) = \begin{cases} 0, & x < at/2, \\ a, & at/2 < x < (1+a)t/2, \\ 1, & (1+a)t/2 < x. \end{cases} \tag{1.47}$$

and a continuous function

$$u(t,x) = \begin{cases} 0, & x < 0, \\ x/t, & 0 < x < t, \\ 1, & x > t. \end{cases} \tag{1.48}$$

Example 1.2.6 *Another non-unique solution.*
We may extend the previous example to obtain a more complex non-unique solution. We again take Burgers' equation, now with initial function

$$u_0(x) = \begin{cases} -1, & x < 0, \\ 1, & x > 0. \end{cases} \tag{1.49}$$

The solution we find reads (see Figure (1.5)), with some $a \in (0,1)$,

$$u_a(t,x) = \begin{cases} -1, & 2x < -(1+a)t, \\ -a, & -(1+a)t < 2x < 0, \\ +a, & 0 < 2x < (1+a)t, \\ +1, & (1+a)t < 2x. \end{cases} \tag{1.50}$$

To pick out the "physically relevant solution" we have to find the solution $u(t,x) = \lim_{\epsilon \to 0} u^\epsilon(t,x)$, with u^ϵ the solution of

$$u_t + f(u)_x = \epsilon u_{xx} .$$

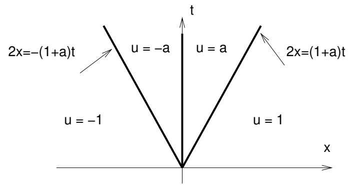


FIGURE 1.5. Another discontinuous solution

First, however, we formulate a condition that guarantees unicity of the discontinuous solution. Such conditions are generally called *entropy conditions*. Later we shall see a relationship between these entropy conditions and the physical relevance of a solution. In the case of a single equation $u_t + f(u)_x = 0$ with $f'' > 0$, one can show that there exists a unique solution that satisfies the "entropy" condition [*Oleinik*] (1.51),

$$\frac{u(t, x+h) - u(x,t)}{h} \leq \frac{E}{t}, \quad h > 0, \ t > 0, \tag{1.51}$$

where E is independent[9]of x, t, and h. This condition implies that if we fix $t > 0$, and we let x go from $-\infty$ to $+\infty$, then we can only jump down (in one direction) over the discontinuity.

This entropy condition of *Oleinik* guarantees uniqueness of the weak solution. Condition (1.51) says that, at discontinuities, we have $u_L > u_R$. From the Rankine-Hugoniot condition we have

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R} = f'(\zeta)\,,$$

for some $\zeta \in [u_R, u_L]$. We had assumed $f'' > 0$ (i.e. f is convex) and hence we finally find that Oleiniks entropy condition implies

$$f'(u_L) > s > f'(u_R)\,. \tag{1.52}$$

Sometimes this inequality is also called the *entropy condition*. The inequality (1.52) means that the characteristic lines on both sides of the discontinuity run into the discontinuity. The entropy condition (1.51) of Oleinik has not been extended to systems of conservation laws. However (1.52) has been extended by Lax [35].

Example 1.2.7 *Traffic flow (shocks and rarefaction waves).*
(This example is taken from [37].) Let $0 \leq \rho(x,t) \leq \rho_m$ denote car density on a road, and $u(x,t)$ car speed. Because cars don't disappear we have

$$\rho_t + (\rho u)_x = 0\,.$$

[9]E depends on $\|u_0\|_\infty$, $\min\{f''(u)\}$ and $\max\{f'(u)\}$. We do not go into details. The interested reader is referred to Chapter 16 of [47].

Let u be given as $u = u(\rho)$, e.g. $u(\rho) = u_m(1 - \rho/\rho_m)$. Combining both equations, gives

$$\rho_t + f(\rho)_x = 0\,.$$

with $f(\rho) = \rho u_m(1 - \rho/\rho_m)$. We easily derive the characteristic speed

$$f'(\rho) = u_m(1 - 2\rho/\rho_m)\,,$$

and the shock speed for a jump in the car density from ρ_L to ρ_R:

$$s = \frac{f(\rho_L) - f(\rho_R)}{\rho_L - \rho_R} = u_m(1 - (\rho_L + \rho_R)/\rho_m)\,. \tag{1.53}$$

The entropy condition implies that a shock must satisfy $f'(\rho_L) > f'(\rho_R)$, which implies $\rho_L < \rho_R$. Now, with initial data

$$\rho(x,0) = \begin{cases} \rho_L, & x < 0, \\ \rho_R, & x > 0, \end{cases}$$

several situations can be distinguished. With $0 < \rho_L < \rho_R < \rho_m$ a shock wave travels with speed (1.53). This shock speed can move in the positive or in the negative direction. With $0 < \rho_R < \rho_L < \rho_m$ we rather have a "rarefaction wave'.

2

Discretisation Principles

In many cases it is important to obtain quantitative data about physical, technical or other real-life systems. One may obtain these through measurements. However, if we have sufficiently accurate mathematical models for a system, we can also try to derive quantitative data from these models. Models for technical problems often take the form of a PDE or a system of PDEs, and although much analytical theory is available for PDEs, most of these equations do not allow an explicit solution in closed form. Therefore numerical methods are used to get insight in their quantitative behaviour.

In practice it is clear that there are many restrictions in gathering reliable data from real-life measurements. In the numerical modelling of these systems, limitations of accuracy are caused by possible inaccuracies in the mathematical model, because the model is often a simplification of reality, and also because there is always a limitation of computing resources.

Numerical mathematics takes the mathematical equations as a starting point and strives for an efficient and accurate approximation of the the quantitative data.

2.1 Discrete representation of the solution

In many problems from practice, the solution of a PDE is a steady or time-dependent (vector-) function in d space variables ($d = 1, 2, 3$). Thus the solution is an element of an infinitely-dimensional space, and -in general-it cannot be described by a finite number of real numbers. Because computational capacity is always finite, the problem has to be discretised: i.e. the solution has to be approximated by a function that *can* be represented by a finite set of numbers.

Example 2.1.1
Let $d = 1$, and let the solution of the equation be a simple continuous function $u(x)$, $x \in [a, b] = \Omega$. A few possible ways to represent this function are:
i) a table of equidistant function values

$$u_i \approx u(x_i), \quad i = 0, 1, 2, \cdots, N, \quad x_i = a + i(b - a)/N;$$

ii) a table of non-equidistant function values

$$u_i \approx u(x_i), \quad i = 0, 1, 2, \cdots, N, \quad a \leq x_0 \leq x_1 \leq \cdots \leq \cdots \leq x_N \leq b;$$

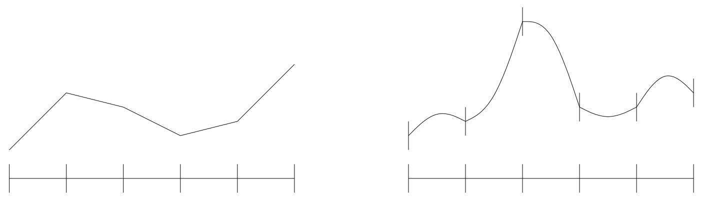


FIGURE 2.1. Piecewise polynomial approximations

iii) a table of function values and values of derivatives of that function (either equidistant or non-equidistant)

$$u_{i,k} = \left(\frac{d}{dx}\right)^k u(x_i); \quad k = 0, 1; \quad i = 0, 1, 2, \cdots, N.$$

In all these cases the solution at the intermediate values of x can be approximated by interpolation. This can be done e.g. by a choosing a polynomial, a spline or a piecewise polynomial as the interpolating function.

Another way to approximate the function is by choosing -a priory- a finite-dimensional function space, selecting a basis in this space, and determining the coefficients of the approximation with respect to that basis. Then we have to select a family of functions $\{\phi_i(x) | i = 1, \cdots, N\}$, which forms the basis for the approximating function space, and we compute coefficients $\{a_i | i = 1, \cdots, N\}$ so that the function

$$\phi(x) = \sum_{i=1}^{N} a_i \phi_i(x) \tag{2.1}$$

approximates the function that is to be represented.

Possible choices for $\{\phi_i\}$ are, e.g. (i) the polynomials $\phi_i(x) = x^i$; (ii) goniometric functions $\{\phi_i(x)\} = \{1, \sin x, \cos x, \sin(2x), \cos(2x), \cdots\}$; or (iii) piecewise polynomials on a partition of the interval $[a, b]$, e.g. piecewise linear or piecewise parabolic functions.

To obtain a more accurate approximation of the function $u(x)$, we need -in general- a denser set of function values or more coefficients to compute. If a partition of nodal points is made, or a partition in subintervals, the largest distance between two neighbouring points, or the size of the largest interval in the partition is usually denoted by a parameter h. This parameter is called the mesh-size. For a smaller h the approximation gets (in general) (i) more accurate, and (ii) more laborious (time-consuming) ("more expensive") to compute.

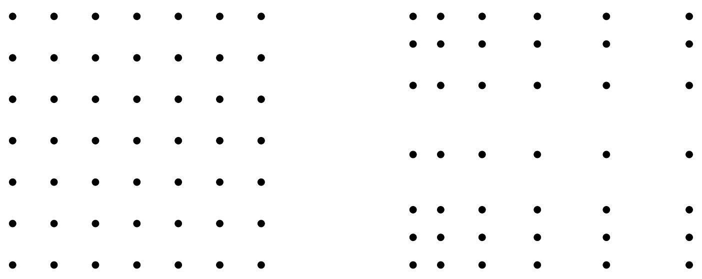


FIGURE 2.2. Sets of nodal points in rectangular regions

We notice that the classification of approximation methods does not yield disjoint classes of methods. In the representation (2.1) the coefficients may coincide with function values or values of derivatives. Example (iii) can also be seen as an interpolation method for a set of function values, given at the sub-interval endpoints. If there is a set of nodal point $\{x_i\}$ and a set of basis function $\{\phi_i\}$, such that they satisfy the relation

$$\phi_i(x_j) = \delta_{i,j},$$

where $\delta_{i,j}$ denotes the Kronecker delta, the same method can be considered as a pointwise approximation and as a functional approximation as well.

2.1.1 Discretisation of the domain, solution and equation

In the transformation of the PDE into a finite set of equations, (i.e. in the discretisation process) it is convenient to distinguish a number of stages:

i) the discretisation of Ω, the domain of definition of the PDE;

ii) the discretisation of u, the solution (i.e. the selection of the representation for the solution); and

iii) the discretisation of the equation.

Before we describe the techniques how to discretise the equations, we first describe different methods for the discretisation of the domain and of the solution.

To find the discrete representation of the solution, one usually selects in the domain where the PDE is defined, a (regular or irregular) *mesh* of points, or a *partitioning* of the domain in triangles or quadrilaterals (See Figure 2.2 or 2.3).

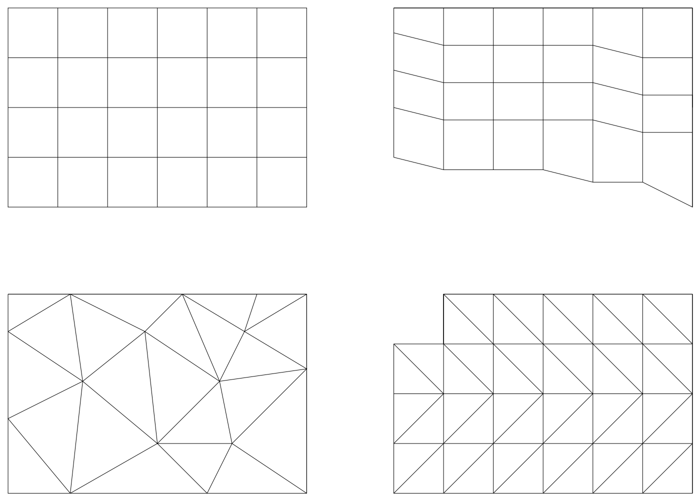


FIGURE 2.3. Partition of two-dimensional domains

2.1.2 Finite difference approximation

For the finite difference method (FDM), the domain Ω is represented by a finite subset of points $\{\mathbf{x}_i\} = \Omega_h \subset \Omega$. These points are the so called "nodal points" of the grid. This grid is almost always arranged in (uniform or non-uniform) manner, see Figure 2.2.

In the finite difference method (FDM) the function u is represented by a set of function values u_i that approximate $u(\mathbf{x}_i)$. The discrete equations are obtained by replacing the differentials in the PDE by finite differences in the discrete equations.

2.1.3 Finite element approximation

In the finite element method (FEM) the domain Ω is partioned in a a finite set of elements $\{\Omega_i\}$, so that $\Omega_i \cap \Omega_j = \emptyset$, for $i \neq j$, and $\cup \overline{\Omega_i} = \overline{\Omega}$. Usually one takes for Ω_i triangles or quadrangles. Then the function is approximated by

$$u_h = \sum a_i \phi_i(x)\,,$$

where ϕ_i are functions that are polynomials on each Ω_i (i.e. *piecewise polynomials*). Usually the functions ϕ_i are continuous polynomials of a low degree. Further they are constructed so that their support extends only over a small number of elements.

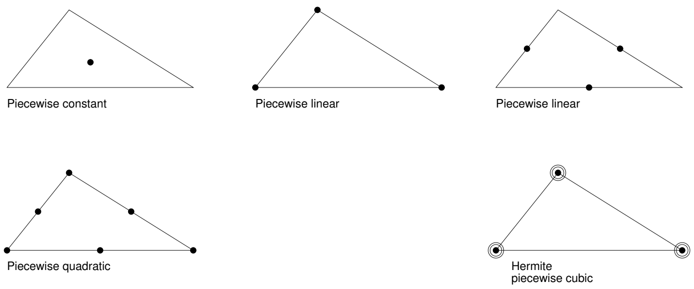


FIGURE 2.4. Examples of nodal points in a triangular domain

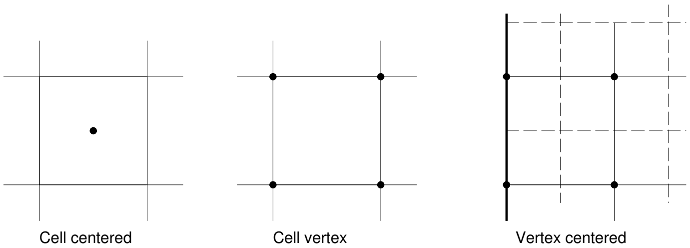


FIGURE 2.5. Examples of nodal points in a quadrangular domain

2.1.4 Finite volume approximation

In this case the domain Ω is also partitioned in a finite set of volumes (mostly triangles or quadrangles) so that $\Omega_i \cap \Omega_j = \emptyset$, for $i \neq j$, and $\cup\overline{\Omega_i} = \overline{\Omega}$. Now in A *cell centered* (or *block centered*) finite volume method (FVM), the function u is approximated by

$$u_h = \sum a_j \phi_j(x)\,, \tag{2.2}$$

where the (possibly discontinuous) functions $\phi_i(x)$ are defined only on a single volume. or on the boundaries of the volumes.

If there is a single coefficient a_i associated with each volume (so that a_i may represent $u(x)$ for $x \in \Omega_i$), then the grid is called a *block centered* or *cell centered.* If a (single) coefficient is associated with each vertex, then the method is called *cell vertex* (see figure 2.5). If the discretisation is constructed on cells (dashed lined in Figure 2.5) that are located around vertices of a grid (solid lines), then the mesh is called *vertex centered.*

2.1.5 Spectral methods

The function u is again approximated by

$$u_h = \sum a_i \phi_i(x)$$

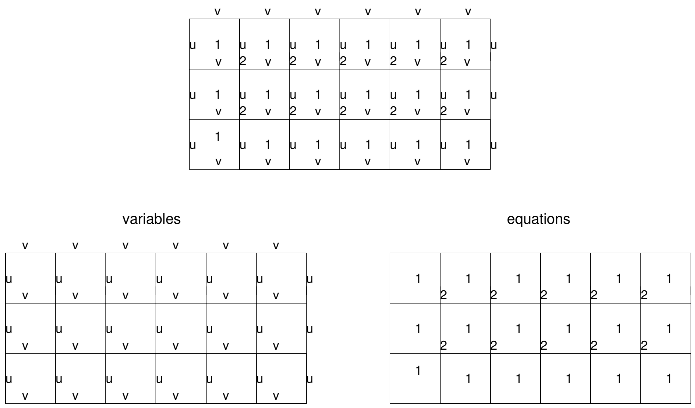


FIGURE 2.6. A staggered grid for the Cauchy-Riemann equations

where the functions $\phi_i(x)$ are defined on all of Ω (they have a support that extends over more than only a small part of Ω) then the methods are called global or spectral methods. (We see that the different types of methods are not always clearly to distinguish.)

2.1.6 Staggered grid

For systems of equations all methods mentioned above are easy to generalise. All functions can simply be replaced by vector functions. However, it is also possible to approximate the different (dependent) variables by different representations. One way is to approximate the different components of the solution on different point sets. This is the case e.g. with "staggered grids". We explain this by two examples.

Example 2.1.2
We consider the *Cauchy Riemann* equations

$$\left.\begin{array}{ll} (i) & u_x + v_y = f_1 \\ (ii) & u_y - v_x = f_2 \end{array}\right\} \quad \text{on} \quad \Omega \subset \mathbb{R}^2\,, \tag{2.3}$$

Ω is a rectangular domain, and Ω is partitioned in a number of equal subrectangles. Approximations are computed for u and v. The values of u are calculated for the midpoints of the vertical edges, the values of v for the midpoints of the horizontal edges. The discrete equations corresponding with equation (2.3(i)) are related with the midpoints of the cells. The discrete equations corresponding with equation (2.3(ii)) are constructed at the cell vertices.

To understand the character of these equations, we write the Cauchy-

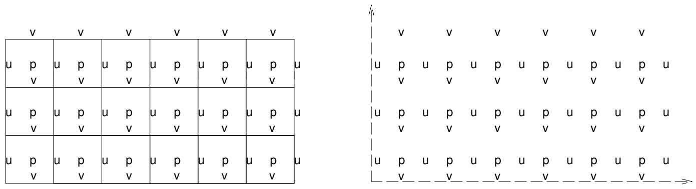


FIGURE 2.7. A staggered grid for Stokes equations

Riemann equations as

$$L\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & -\frac{\partial}{\partial x} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{2.4}$$

We see that

$$\det(L) = -\left(\frac{\partial}{\partial x}\right)^2 - \left(\frac{\partial}{\partial y}\right)^2 = -\Delta\,.$$

This shows that the system is elliptic and second order. It is also directly seen by elimination that [1] $\Delta u = \text{div } \mathbf{f}$, or $\Delta v = -\text{curl } \mathbf{f}$.

The system being second order elliptic, it needs *one* boundary condition along all the boundary. We take

$$u.n_x + v.n_y = \mathbf{v}.\mathbf{n} = g(x,y)\,.$$

In our discrete system, this condition fixes the values of u and v on the boundary. Now the number of unknowns is 27 (viz. 3×5 for u, and 2×6 for v), and the number of equations is 28 (3×6 for equation (i), and 2×5 for equation (ii)). Hence, the system seems to be overdetermined. However, it is easily seen that the boundary function $g(x,y)$ should satisfy a condition. From $u_x + v_y = \text{div } \mathbf{v} = f_1$, it follows that

$$\int_\Omega u_x + v_y \, d\Omega = \int_\Omega \text{div } \mathbf{v}\, d\Omega = \oint_{\partial\Omega} \mathbf{v}.\mathbf{n}\, d\Gamma = \oint_{\partial\Omega} g(x,y)\, d\Gamma\,,$$

i.e. the *compatibility condition*

$$\int_\Omega f_1(x,y)\,\Omega = \oint_{\partial\Omega} g(x,y)\, d\Gamma\,.$$

---

[1] In three dimensions

$$\text{curl}\,\mathbf{v} = \begin{pmatrix} \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1} \\ \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \end{pmatrix},$$

in two dimensions $\text{curl}\,\mathbf{v} = \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}$.

Similarly it follows from adding all discrete equations $(u_R - u_L)/h + (u_T - u_B)/h = f_1$ for each cell, that -in total-

$$h\sum_k g_k = h\sum_i u_{R,i} - h\sum_i u_{L,i} + h\sum_i v_{T,i} - h\sum_i v_{B,i} = h^2\sum_i f_{1,i}, \quad (2.5)$$

i.e.

$$\sum_{k\in\text{boundary}} g_k = h \sum_{i\in\text{interior}} f_{1,i} \quad (2.6)$$

Here $f_{1,i}$ denotes the value of $f_1(x,y)$ in the i-th cell in the interior of Ω, and g_k denotes the value of $g(x,y)$ in the k-th boundary edge element. If the boundary values satisfy condition (2.6), one (linear) dependence relation exists between the 28 equations and we get a solvable system of equations.

Example 2.1.3 *Stationary Stokes equations*
Analogous to the staggered grid for the Cauchy-Riemann equations, we can construct one for the stationary *Stokes equations*,

$$\begin{cases} \eta\Delta u - p_x = 0, \\ \eta\Delta v - p_y = 0, \\ u_x + v_y = 0. \end{cases}$$

This can be written as a system with linear differential operator L,

$$L\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} \eta\Delta & 0 & \nabla_1 \\ 0 & \eta\Delta & \nabla_2 \\ \nabla_1 & \nabla_2 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The character of the system of equations is determined by the determinant of the principle part of the differential operator

$$\det L = \eta\Delta\cdot\nabla_1^2 + \eta\Delta\cdot\nabla_2^2 + 0 = \eta\Delta^2\,.$$

This is a 4-th order elliptic operator, for which we need two conditions over all of the boundary. The operator Δ is represented by the finite difference stencil $\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}$, the operator ∇_1 by $\begin{bmatrix} -1 & 1 \end{bmatrix}$ and ∇_2 by $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

2.2 Techniques for discretisation of PDEs

2.2.1 Finite difference methods

For this method, we have already seen examples. The discrete equation is constructed by replacing differential operators by difference operators. The construction of the difference operators is as follows: in the neighbourhood

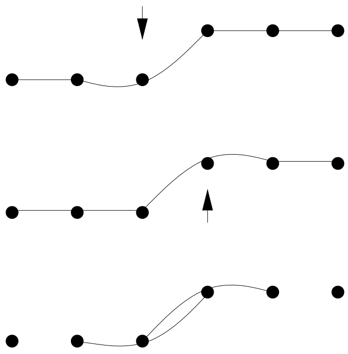


FIGURE 2.8. Difference approximation and polynomial interpretation

of a point $\mathbf{x}\in\Omega$ the approximation of $u(x)$ is replaced by a truncated Taylor series:

$$u(\mathbf{x}) = a_0 + a_{10}x + a_{01}y + a_{11}x^2 + \cdots .$$

With the values u_{ij} in the nodal points x_{ij} around $\mathbf{x}$, the coefficients a are expressed in these function values u_{ij}. These coefficients a_{pq} are associated with the derivatives of the approximation in $\mathbf{x}$. Now, in the PDE at this point the differentials are replaced by difference approximations. In this way, for each unknown u_{ij} an equation is set up at the point x_{ij}.

For a regular grid, in the interior points of the domain the same difference approximation can be used. Different difference approximations appear near the boundary of the domain, or in parts of the domain where the nodal points are not regularly placed.

In the neighbourhood of $\mathbf{x}_{ij}$, the approximation of $u(\mathbf{x})$ is, thus, considered as a piecewise polynomial (of a low degree). That does *not* mean that the complete function u is considered as a polynomial. The interpretation of the polynomial representation depends on the nodal point where the interpretation is made.

The (linear or nonlinear) algebraic system of equations that results from the discretisation process has an important property. By the difference approximations the value of the unknown function at a particular point is only coupled with the values at neighbouring points. In this way a "sparse system" of algebraic equations appears. The corresponding matrix (of the linearisation) only contains a few non-zero elements in each row; most entries in the matrix vanish. Whereas the systems that appear as a result of discretisations can become very large, they mostly have this nice property, by which the solution can often be found in a relatively efficient[2] manner. As this property is shared by most discretisations of PDEs, special methods

[2]Compared with a full system of equations of the same dimensions.

are sought (and found) for the efficient solution of such sparse systems of equations.

2.2.2 Variational method

Variational method for difference methods

An important class of PDEs, the symmetric elliptic equations, can be written as a minimisation problem. This gives another possibility to construct the discrete equations. We show this by means of the Poisson equation

$$\left.\begin{array}{ll} \Delta u = f, & \text{on } \Omega, \\ u = 0, & \text{on } \Gamma\,. \end{array}\right\}$$

This equation can be written as the minimisation problem: find a continuous function u such that $u = 0$ on Γ, and

$$J(u) = \int_\Omega \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + 2fu\; d\Omega$$

exists and is minimal. (Exists: $u_x^2 + u_y^2$ and fu should be integrable functions.)

Now we can set up a discrete analogue for this minimisation problem [8]. On a regular square grid (i.e. Ω a rectangle, and square meshes with a meshsize h) we can write: find $u_h = \{u_{ij}\}$ such that $u_{ij} = 0$ on Γ, and

$$J_h(u_h) = \sum_{ij} \left(\frac{u_{i+1,j} - u_{i,j}}{h}\right)^2 + \left(\frac{u_{i,j+1} - u_{i,j}}{h}\right)^2 + 2f_{ij}u_{i,j}$$

is minimal. The minimisation of this quadratic functional in the unknowns $\{u_{ij}\}$ leads to linear system in $\{u_{ij}\}$. This system also yields a difference equation for each internal point of the grid.

The variational method for coefficients

For problems that can be written as minimisation problems, we can use the *variational method* also if we seek an approximation in the form

$$u_h(x,y) = \sum_i a_i \phi_i(x,y)\,.$$

We take the same example (2.2.2), so that with $\phi_i(x,y) = 0$ on Γ. Now we compute the coefficients a_i such that

$$\int_\Omega \left\{\sum_i a_i \frac{\partial}{\partial x}\phi_i(x,y)\right\}^2 + \left\{\sum_i a_i \frac{\partial}{\partial y}\phi_i(x,y)\right\}^2 + 2f(x,y)\sum_i a_i\phi_i(x,y)\, d\Omega$$

attains its minimum. This minimisation problem also leads to a linear system, now in the unknown coefficients (a_i). The same technique is applicable on irregular grids.

2.2.3 Weighted residual methods

A quite general discretisation technique is based on the following consideration. Let the PDE be given by

$$\left.\begin{aligned} \mathcal{N}_\Omega(u) = s_\Omega, &\quad \text{on } \Omega, \\ \mathcal{N}_\Gamma(u) = s_\Gamma, &\quad \text{on } \Gamma, \end{aligned}\right\} \tag{2.7}$$

then, for any function w we have

$$\int_\Omega w_\Omega(\mathcal{N}_\Omega(u) - s_\Omega)\, d\Omega + \oint_\Gamma w_\Gamma(\mathcal{N}_\Gamma(u) - s_\Gamma)\, d\Gamma \;=\; 0. \tag{2.8}$$

Or, in an even more general notation, let the equation be $\mathcal{N}(u) = s$ and let $s, \mathcal{N}(u) \in V$, with V a Banach space, then for an arbitrary linear functional $l_w \in V'$ the equation[3]

$$l_w\left(\mathcal{N}(u) - s\right) = 0$$

is satisfied. Now, let S be a set of functions in which the solution u is sought: $\mathcal{N} : S \to V$, then the PDE can be formulated as: find $u \in S$ such that

$$l_w(\mathcal{N}(u) - s) = 0, \quad \text{for all } l_w \in V'.$$

In a weighted residual method an n-dimensional subspace $S_h \subset S$ is selected and an n-dimensional subspace $V_h^D \subset V'$. The discretisation now reads: find $u_h \in S_h$ such that

$$l_w(\mathcal{N}(u_h) - s) = 0 \;\text{ for all } l_w \in V_h^D.$$

In this way, the discretisation yields an $n \times n$-system of equations.

Of course, there are many different applications of the weighted residual principle. We will treat the most important.

2.2.4 Collocation methods

Collocation methods are weighted residual methods with the functionals l_w in the space V_h^D that is spanned by

$$\{l_{x_i} \,|\, x_i \in \Omega;\; i = 1, \cdots, N\}\;.$$

Here l_{x_i} is defined by

$$l_{x_i}(f) = f(x_i), \text{ for } f \in V.$$

This means that the discrete equations are

$$\mathcal{N}(u_h)(x_i) = s(x_i), \text{ for } i = 1, \cdots, N.$$

This means that the approximate solution u_h should satisfy the original PDE precisely in n given points, where $n = \dim(S_h)$.

[3]The dual space of a Banach space V (denoted by V') is the family of all bounded linear functionals defined on V, furnished with the obvious addition and scalar multiplication.

2.2.5 Galerkin methods

When the functionals l are defined by integrals with integrable weight functions, we obtain "Galerkin"-methods. Typically $l_w \in V_h^D$ is defined by

$$l_w(f) = \int_\Omega w(x)\, f(x)\, d\Omega\,.$$

The space V_h^D is now defined by n linearly independent functions $w_i(x)$.

If we choose $\{w_i(x)\}_{i=1,2,\cdots,N}$, such that $V_h := \text{span}\,(w_i) = S_h$ then the methods are called *Bubnov-Galerkin* methods. An obvious choice is the $w_i = \phi_i$. Then the discrete equations read

$$\int_\Omega \phi_j\, \mathcal{N}(\sum a_i \phi_i)\, d\Omega = \int_\Omega \phi_j\, s\, d\Omega.$$

If we take $V_h := \text{span}\,(w_i) \neq S_h$, the method is called a *Petrov-Galerkin* method.

Almost all usual Finite Elements methods are of the (Bubnov-) Galerkin type. When applied to symmetric positive definite elliptic PDEs, these methods can also be regarded as variational methods.

2.2.6 Box methods = Finite Volume methods

Box methods are a special kind of Galerkin methods. Here the domain Ω is partitioned in a number of volumes $\{\Omega_i\}_{i=1,\cdots,N}$ I.e. $\Omega_i \cap \Omega_j = \emptyset$, if $i \neq j$, and $\cup \overline{\Omega_i} = \overline{\Omega}$. The weight functions are now the characteristic functions on these Ω_i,

$$\left.\begin{array}{l} w_i(x) = 1, \quad \text{if } x \in \Omega_i, \\ w_i(x) = 0, \quad \text{if } x \notin \Omega_i, \end{array}\right\}$$

and the discrete equations read

$$\int_{\Omega_i} \mathcal{N}\left(\sum a_i \phi_i\right)\, d\Omega_i = \int_{\Omega_i} s\, d\Omega_i.$$

The volumes Ω_i are called "boxes", "cells" or "control volumes".

Box methods are particularly popular for the discretisation of equations in divergence form

$$\text{div}\ \mathbf{J}(u) = s(u), \quad \text{on } \Omega.$$

Then the discrete equation gets the form

$$\oint_{\partial\Omega_i} \mathbf{J}(u_h) \cdot \mathbf{n}\, d\Gamma_i = \int_{\Omega_i} s(u_h)\, d\Omega_i, \quad \text{for} \quad i = 1, \cdots, N. \tag{2.9}$$

If the computation of $\mathbf{J}(u_h)$ at an interface between two cells is made consistently for both neighbouring cells, then the discretisation satisfies a discrete conservation law. This means that for an arbitrary union of cells $G = \cup \Omega_i$ hold:

$$\int_G \text{div}\ \mathbf{J}(u_h)\, d\Omega = \int_G s(u_h)\, d\Omega\,.$$

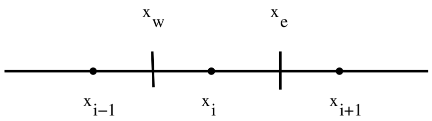


FIGURE 2.9. Nodal points and boxes on the real line

2.3 Examples

2.3.1 THE DIFFUSION EQUATION

To show the principle of discretisation, in this section we discretise a steady one-dimensional diffusion equation, with a source-term s and a variable diffusion coefficient k,

$$-\frac{d}{dx}\left(k(x)\,\frac{d}{dx}u\right) = s(x,u), \quad \text{on} \;\; \Omega \subset \mathbb{R}, \tag{2.10}$$

together with a boundary condition (e.g. $u = 0$, on $\partial\Omega$). This may be seen as the description of heat-conduction through a bar, where $u(x)$ denotes the temperature, $k(x)$ the heat diffusion coefficient, and $s(x,u)$ the heat generation per unit length, by some source.

We divide Ω in (unequal) boxes Ω_i and we choose for each Ω_i a nodal point x_i (see Figure 2.9).

We may choose the box-walls halfway the nodal points, or we may put the nodal points in the midpoint of the boxes. (Such choice is not relevant at the moment.)

The discrete approximation is represented by the function values at the nodal points $u_i \approx u(x_i)$. Integration of the equation over Ω_i yields

$$\int_{x_w}^{x_e} s(x,u)dx = -\int_{x_w}^{x_e} (ku_x)_x dx = -\, k(x)u_x(x)|_{x_w}^{x_e}\,. \tag{2.11}$$

To set up the discrete equations we need an approximation of $u(x)$ over Ω_i, and approximations $u_x(x_e)$ and $u_x(x_w)$. Starting from $u_i \approx u(x_i)$ we can approximate u, e.g. by
i) taking a constant value u_i on Ω_i;
ii) taking a piecewise linear approximation between the nodal points.
With (i) we cannot approximate the r.h.s. of (2.11), but with (ii) we can. Then we obtain

$$k(x_e)\frac{u_{i+1}-u_i}{x_{i+1}-x_i} - k(x_w)\frac{u_i-u_{i-1}}{x_i-x_{i-1}} = -\int_{x_w}^{x_e} s(x,u_h)\,dx = -\overline{s_i}(u_h)\,(x_e - x_w)\,, \tag{2.12}$$

where $\overline{s_i}$ is the mean value of $s(x,u)$ over the volume Ω_i.

2.3.2 The source term and solution of the system

If s is independent of u, or if s depends linearly on u, then (2.12) is a linear system of equations that can be solved immediately (as soon as the values of the endpoints u_0 and u_n are given) if we take $s_i(u_h) = s(u_i)$.

On the other hand, if $s = s(u)$ is a nonlinear function of u, then (i) $s_i(u_h)$ should be approximated, and (ii) the equation has to be solved iteratively. The source term $\overline{s_i}$ can **either** be approximated by a constant $\overline{s_i} := s_i(u_h^{(n)})$, where the previous approximation $u_h^{(n)}$ is used directly to obtain a (hopefully better) approximation $u_h^{(n+1)}$ by (2.12), **or** by *linearisation*, i.e. taking a linear approximation of s as $\overline{s_i} = s_i^C(u_h^{(n)}) + s_i^L(u_h^{(n)})(u_i^{(n+1)} - u_i^{(n)}$. In this case we have to adapt (2.12) slightly.

The difference between (i) and (ii) is that in (i) $\overline{s_i}$ depends completely on the previous approximation, whereas in (ii) linear dependence of s on u is taken into account; s_i^C and s_i^L are constants, that have to be determined for each nodal point. This makes that the linear system to solve is changed:

$$\begin{aligned} \tfrac{k(x_e)}{x_{i+1}-x_i} u_{i+1} - \left(\tfrac{k(x_e)}{x_{i+1}-x_i} + \tfrac{k(x_w)}{x_i - x_{i-1}} - (x_e - x_w) s_i^L(u_h^{(n)}) \right) u_i + \\ + \tfrac{k(x_w)}{x_i - x_{i-1}} u_{i-1} = -(x_e - x_w) s_i^C(u_h^{(n)}) \end{aligned} \tag{2.13}$$

where s_i^L and s_i^C are such that $s_i^L(u_h)u_i + s_i^C(u_h) = \overline{s_i}(u_h)$. Here, the process reads: first make an initial estimate for u_h, then determine $s_i^C(u_h)$, and $s_i^L(u_h)$, and solve the linear system (2.13). This process is repeated until convergence is reached.

The final solution of the two processes (i) and (ii) is the same if

$$s_i(u_h) = s_i^C(u_h) + s_i^L(u_h)\, u_i\,.$$

An advantage of the alternative (ii) is that, by a judicious choice of s_i^L we may expect that the iterative process (ii) converges faster than process (i). The processes are identical if we take $s_i^L = 0$ for all i. However, if s is a linear function of u, and if we know $s^L(u) = ds/du$, then process (ii) may converge in one step.

We see that by (2.12) or (2.13) u_i is only coupled with u_{i-1} nd u_{i+1}, the values in the neighbouring cells. Hence the linear system is tridiagonal. This band structure of the matrix makes that the linear system can be solved very efficiently.

In the construction of (2.12) we have used the simplest possible representation for u_h that still allows an approximation of (ku_x). Other representations are possible as well. Apparently, it is not necessary to be consistent in the choice for u_h, (i.e. in the assumptions for the approximation of u) when the different parts in the differential operator are discretised. E.g. we could take $\overline{s_i} = s(u_i)$ piecewise constant, whereas a piecewise linear approximation was used to determine u_x.

2.3.3 The convection equation

Now we consider the problem

$$\frac{\partial}{\partial x}(v(x)u(x)) = s(u,x), \tag{2.14}$$

where v(x) is a given function. As a boundary condition we use e.g. $u = 0$ at the left boundary of the interval. Integration, analogous to (2.11) gives

$$\int_{x_w}^{x_e} s(u,x)dx = \int_{x_w}^{x_e} (vu)_x dx = v(x)u(x)|_{x_w}^{x_e}. \tag{2.15}$$

We take the same grid as in (2.3.1) and, again, a piecewise linear approximation for u_h. This yields the discrete equation

$$v(x_e)\frac{u_{i+1}+u_i}{2} - v(x_w)\frac{u_i+u_{i-1}}{2} = \overline{s_i}(u_h)\,(x_e - x_w)\,. \tag{2.16}$$

For a constant $v(x) = v$, this leads to

$$\frac{v}{2}(u_{i+1} - u_{i-1}) = \overline{s_i}(u_h)\,(x_e - x_w)\,. \tag{2.17}$$

Here we see that an important problem appears. In the simplest case of a constant v and s, we see that in equation (2.16) u_{i+1} is coupled with u_{i-1} only. The approximations of u at the even nodes are coupled to each other, and also the values at the odd nodes, but there is no coupling between the odd and the even nodes. If the problem is defined for $0 \le i \le N$, and the boundary condition is given at the inflow boundary x_0 ($i = 0$), then there is no boundary condition to determine the values at the odd nodes. Because the linear system is singular there is a family of possible solutions. The homogeneous equation allows nontrivial solutions. This is also called instability: the discrete solution is not bounded by the rhs of the discrete equation.

Notice that the technique for the discretisation of the equations in the interior domain cannot be used at the outflow boundary. There a different discretisation method should be used, because no value u_n is available. In fact it is this particular discretisation at the outflow boundary that determines the behaviour of the discrete solution at the odd points. We consider this completely intolerable.

2.4 Techniques for time-discretisation

2.4.1 Time-space discretisation or semi-discretisation

We have seen that (because the time-direction is unique) it makes sense to treat the time-dependence of a problem different from the space-dependence.

The discrete solution is determined in a number of subsequent time-steps. Usually one uses the same time-steps for all nodal points in a grid. Thus, in a 2-dimensional finite difference discretisation, the solutions are represented by $\{u_{ij}^n\}$, where (i,j) denotes the nodal point x_{ij} and n the time-coordinate t_n.

By the special character of the time-variable (the future depends on the past, and not the past on the future) the values $\{u_{ij}^n\}$ are computed on basis of the known values $\{u_{ij}^k\}$, $k = n-1, n-2, \cdots$.

Another possibility for discretisation of time-dependent PDEs is to defer the time-discretisation and to apply semi-discretisation: the solution of the equations is represented by $\{u_{ij}(t)\}$. The value of the solution at the point x_{ij} (or another coefficient used for the space-discretisation) is considered as a function in time. Then, the semi-discretised PDE leads to a very large system of ODEs for the unknown functions $\{u_{ij}(t)\}$, as e.g. in

$$\frac{d}{dt} u_{ij}(t) = F_{ij}(\{u_{kl}(t)\}_{kl}), \quad \forall i, j.$$

For the solution of these equations any suitable technique for the solution of ODEs can be used.

2.4.2 FDM for a linear hyperbolic problem

In this section we will show and analyse a number of characteristic problems that are encountered in the treatment of time-dependent problems. We do this by means of the simple convection equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0. \tag{2.18}$$

As a grid we simply take a set of equidistant nodal point $\{x_i\}$. Halfway the nodal points we situate the cell interfaces of the cell Ω_i at $x_{i\pm\frac{1}{2}}$. Integration of (2.18) over Ω_i yields

$$\frac{\partial}{\partial t} \int_{\Omega_i} u \, dx + f(u(x)) \Big|_{x=x_{i-\frac{1}{2}}}^{x=x_{i+\frac{1}{2}}} = 0\,. \tag{2.19}$$

We represent the discrete approximation by $\{u_i\}$, with u_i the mean value of u over the cell Ω_i. Then we arrive at the equation

$$(x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}) \frac{\partial}{\partial t} u_i(t) + f_{x_{i+\frac{1}{2}}} - f_{x_{i-\frac{1}{2}}} = 0\,. \tag{2.20}$$

Here, $f_{x_{i\pm\frac{1}{2}}}$ denotes an approximation for $f(u(x_{i\pm\frac{1}{2}}))$.

For the space discretisation we use *upwind, downwind* and *central* flux computations. To construct a system of equations for the variables $\{u_i\}$ we still have to decide how we relate $f(u(x_{i\pm\frac{1}{2}}))$ to the values of $\{u_i\}$. As in section 2.3.3, one possible choice is

$$f_{x_{i+\frac{1}{2}}} = \frac{1}{2}(f(u_i) + f(u_{i+1})),$$

the central approximation. Other choices are e.g.

$$f_{x_{i+\frac{1}{2}}} = f(u_i) \quad \text{or} \quad f_{x_{i+\frac{1}{2}}} = f(u_{i+1}),$$

respectively the backward and the forward discretisation[4].

For convenience we will combine the three choices in one formula, with a parameter w. After some analysis we make a choice for w. We write

$$f_{i+\frac{1}{2}} = wf(u_i) + (1-w)f(u_{i+1}), \quad i = 1, 2, 3, \cdots. \tag{2.21}$$

Now we have completed our semi-discretisation:

$$\frac{\partial}{\partial t}u_i(t) = \frac{w}{\Delta x}f(u_{i-1}) + \frac{1-2w}{\Delta x}f(u_i) - \frac{1-w}{\Delta x}f(u_{i+1}), \tag{2.22}$$

where $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$. We write this equation briefly as

$$\frac{\partial}{\partial t}u_h(t) = g_h(u_h(t)). \tag{2.23}$$

We realise the time discretisation of (2.23) simply by time-stepping with steps Δt. This can be done e.g. by the *forward Euler method*:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = g_h(u_h(t^n)). \tag{2.24}$$

In this case the new values $u_h(t^{n+1})$ are expressed explicitly in the old values $u_h(t^n)$.

An alternative possibility is the *backward Euler method*:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = g_h(u_h(t^{n+1})), \tag{2.25}$$

or also the *Cranck-Nicolson method*

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = (\ g_h(u_h(t^n)) + g_h(u_h(t^{n+1}))\ )/2. \tag{2.26}$$

These two methods are "implicit" time-discretisation methods. Here we have to solve a system of equations to compute the new values $u_h(t^{n+1})$.

---

[4]With $\partial f/\partial u > 0$, the approximation $f_{x_{i+\frac{1}{2}}} = f(u_i)$ is called upwind approximation, because the flow of information is from left to right and for the approximation of $f_{x_{i+\frac{1}{2}}}$ the upstream value u_i is used. The other approximation is called downstream approximation.

We summarise the three possibilities, again, in one parametrised formula:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = (1-\theta)g_h(u_h(t^n)) + \theta g_h(u_h(t^{n+1}))\,. \qquad (2.27)$$

The fully discretised system reads:

$$\begin{aligned} u_i^{n+1} - u_i^n = \;& \tfrac{\Delta t}{\Delta x}\theta && \{w f_{i-1}^{n+1} + (1-2w) f_i^{n+1} - (1-w) f_{i+1}^{n+1}\} + \\ & \tfrac{\Delta t}{\Delta x}(1-\theta) && \{w f_{i-1}^{n} + (1-2w) f_i^{n} - (1-w) f_{i+1}^{n}\}, \end{aligned} \qquad (2.28)$$

where f_j^k represents $f(u_j^k)$.

In general, to compute u_i^{n+1} from u_i^n, we have to solve a system of nonlinear equations. Because it is difficult to analyse nonlinear equations we simplify the problem by linearising it. We set $\frac{df}{du} = f'$ constant, and we combine a few parameters in a single: $\lambda = f'\Delta t/\Delta x$. Now equation (2.28) becomes

$$\begin{aligned} u_i^{n+1} - u_i^n = \;& \lambda\theta && \{w u_{i-1}^{n+1} + (1-2w) u_i^{n+1} - (1-w) u_{i+1}^{n+1}\} + \\ & \lambda(1-\theta) && \{w u_{i-1}^{n} + (1-2w) u_i^{n} - (1-w) u_{i+1}^{n}\}\,. \end{aligned} \qquad (2.29)$$

Having simplified the equation so far, it is simple to solve it also analytically. We will use this knowledge about the analytic solution to check if our numerical techniques yield reliable results.

The equation now is $u_t + f'u_x = 0$, and the solution is $\psi(x - f't)$ for an arbitrary initial solution $\psi(x) = u(0,x)$. In time the solution doesn't really change, it doesn't grow or shrink, it only moves with the velocity f'. With the different possible choices for the numerical method, we can see which properties of the true solution are inherited by the numerical counterparts.

In order to study the behaviour of the solution of the discrete system, we consider the behaviour in time of the approximate solution in the domain $\Omega = (-\infty, +\infty)$, for an initial function $u(t_0, x) = e^{j\omega x}$. We use the notation $j := \sqrt{-1}$ in order not to interfere with the index i previously used. (We refer to a textbook in Fourier analysis to understand the relevance of this choice.) The discrete initial function, then, is

$$u_i^0 = e^{j\omega h i}.$$

Using equation (2.29) we see immediately that the solution at a later stage is given by

$$u_i^n = \gamma^n\, e^{j\omega h i}, \qquad (2.30)$$

where γ is a complex number that satisfies

$$\begin{aligned} \gamma - 1 = \;& \lambda\theta && \{w\gamma e^{-j\omega h} + (1-2w)\gamma - (1-w)\gamma e^{+j\omega h}\} + \\ & \lambda(1-\theta) && \{w e^{-j\omega h} + (1-2w) - (1-w) e^{+j\omega h}\}\,. \end{aligned} \qquad (2.31)$$

This can also be written as

$$\gamma = \frac{1 + \lambda(1-\theta)\{(1-2w) + w e^{-j\omega h} - (1-w) e^{+j\omega h}\}}{1 - \lambda\theta\{(1-2w) + w e^{-j\omega h} - (1-w) e^{+j\omega h}\}} \qquad (2.32)$$

$$= \frac{1+\lambda(1-\theta)\{(1-2w)(1-\cos(\omega h)) - j\sin(\omega h)\}}{1-\lambda\theta\{(1-2w)(1-\cos(\omega h)) - j\sin(\omega h)\}};$$

γ is the *amplification factor* for the time-integration: $|u_i^n| = |\gamma|^n$. The solutions increases for $t \to \infty$ if $|\gamma| > 1$, and it decreases for $|\gamma| < 1$. We see that

$$|\gamma|^2 = \frac{\{1+\lambda(1-\theta)(1-2w)(1-\cos(\omega h))\}^2 + \lambda^2(1-\theta)^2\sin^2(\omega h)}{\{1-\lambda\theta(1-2w)(1-\cos(\omega h))\}^2 + \lambda^2\theta^2\sin^2(\omega h)}. \tag{2.33}$$

Now we want to study the effect of the parameter w, that was introduced in the space discretisation, and of θ that was introduced in the time discretisation. Further we can see what is the effect of the parameter λ.

- If $w = \frac{1}{2}$ (i.e. for central discretisation in space) equation(2.33) reduces to
  $$|\gamma|^2 = \frac{1+\lambda^2(1-\theta)^2\sin^2(\omega h)}{1+\lambda^2\theta^2\sin^2(\omega h)}, \tag{2.34}$$
  and we find
  $$|\gamma| < 1 \Leftrightarrow (1-\theta)^2 \le \theta^2 \Leftrightarrow \theta \ge \frac{1}{2}.$$
  Depending on θ we can draw the following conclusion:
  i) $\theta = \frac{1}{2} \Leftrightarrow |\gamma| = 1$. This implies neutral stability (Crank-Nicolson).
  ii) $\theta = 0 \Leftrightarrow |\gamma| > 1$. This means unconditional instability (forward Euler).
  iii) $\theta = 1 \Leftrightarrow |\gamma| < 1$. This means unconditional stability (backward Euler).

- If $w = 1$ (i.e. for "backward" discretisation in space)
  stability $\Leftrightarrow |\gamma| \le 1 \Leftrightarrow \lambda(\lambda(1-2\theta) - 1) \le 0$. Depending on the sign of f' we find:

  - $f' > 0 \Leftrightarrow \lambda > 0 \Leftrightarrow$ stability for: $\lambda(1-2\theta) \le 1$.
    This implies stability for $\theta \ge \frac{1}{2}$. If $\theta < \frac{1}{2}$ stability may depend on λ; stability is obtained for $\lambda \le \frac{1}{1-2\theta}$ and we obtain stability only for λ (i.e. the time-step) small enough.
  - $f' < 0 \Leftrightarrow \lambda < 0 \Leftrightarrow$ stability for: $\lambda(1-2\theta) \ge 1$.
    This implies instability for $\theta \in [0, \frac{1}{2}]$. If $\theta > \frac{1}{2}$ then stability is guaranteed for $|\lambda| \ge \frac{1}{2\theta - 1}$ and we obtain stability only for λ large (!) enough.

- If $w = 0$ (i.e. for "forward" discretisation in space)
  stability $\Leftrightarrow |\gamma| \le 1 \Leftrightarrow \lambda(\lambda(1-2\theta) + 1) \le 0$

  - $f' > 0 \Leftrightarrow \lambda > 0 \Leftrightarrow$ stability for $\lambda(1-2\theta) \le 1$.
    This implies instability for $\theta \in [0, \frac{1}{2}]$; if $\theta > \frac{1}{2}$ then get stability for $|\lambda| \ge \frac{1}{2\theta-1}$, i.e. only for λ large(!) enough.

– $f' < 0 \Leftrightarrow \lambda < 0$; this implies stability for $\theta \geq \frac{1}{2}$; if $\theta \in [0, \frac{1}{2}]$ we get stability for $|\lambda| \leq \frac{1}{1-2\theta}$, i.e. we obtain stability for λ small enough.

We recognise a symmetry If $f' > 0$ then $w = 1$ corresponds with upwind discretisation and $w = 0$ with downwind; with $f' < 0$ it is the other way. We notice that the upwind discretisation is unconditionally stable when $\theta \geq \frac{1}{2}$. For $\theta < \frac{1}{2}$ it is stable if the time-step is small enough. The downwind discretisation is instable for all reasonable cases.

2.4.3 The equivalent differential equation

In this section we study the local discretisation error (i.e. the amount to which the true solution -at the nodal points- does not satisfy the discrete equation [5]) for (2.28). We consider again the linear problem

$$\frac{\partial}{\partial t}u + f'\frac{\partial}{\partial x}u = 0, \tag{2.35}$$

and its discretisation equation (2.28). We can write (2.29) also as

$$\begin{aligned} u_i^{n+1} - u_i^n = \; & -\lambda\theta\{\tfrac{1}{2}(u_{i+1}^{n+1} - u_{i-1}^{n+1}) + (\tfrac{1}{2} - w)(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1})\} + \\ & -\lambda(1-\theta)\{\tfrac{1}{2}(u_{i+1}^n - u_{i-1}^n) + (\tfrac{1}{2} - w)(u_{i+1}^n - 2u_i^n + u_{i-1}^n)\}. \end{aligned} \tag{2.36}$$

In order to compute the local truncation error, we expand the function $u(x,t)$ in a Taylor series around the point $x = x_i$, and $t_\star = t_n + \theta\Delta t$. We

---

[5]For the differential operator $L : X \to Y$ and its discrete operator $L_h : X_h \to Y_h$, and the restriction operators $R_h : X \to X_h$ and $\overline{R}_h : Y \to Y_h$, we remember the definitions for *global discretisation error*

$$|u(x_i) - u_i| = |(R_h u)_i - u_i|,$$

and for the *local discretisation error*

$$|(L_h R_h u - \overline{R}_h L u)_i|.$$

denote $u_\star := u(x_i, t_\star)$. We obtain

$$
\begin{array}{ll}
(i) & u_i^n = u_\star - \theta\Delta t \left(\frac{\partial u}{\partial t}\right)_\star + \frac{1}{2}\theta^2\Delta t^2 \left(\frac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3);\\
(ii) & u_i^{n+1} = u_\star + (1-\theta)\Delta t \left(\frac{\partial u}{\partial t}\right)_\star + \frac{1}{2}(1-\theta)^2\Delta t^2 \left(\frac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3);\\
(iii) & u_{i-1} = u_i - \Delta x \left(\frac{\partial u}{\partial x}\right)_i + \frac{1}{2}\Delta x^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
(iv) & u_{i+1} = u_i + \Delta x \left(\frac{\partial u}{\partial x}\right)_i + \frac{1}{2}\Delta x^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
(v) & \frac{1}{2}(u_{i+1} - u_{i-1}) = +\Delta x \left(\frac{\partial u}{\partial x}\right)_i + \mathcal{O}(\Delta x^3);\\
(vi) & (u_{i+1} - 2u_i + u_{i-1}) = \Delta x^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
(vii) & \left(\frac{\partial u}{\partial x}\right)_i^n = \left(\frac{\partial u}{\partial x}\right)_\star - \theta\,\Delta t \left(\frac{\partial^2 u}{\partial x \partial t}\right)_\star + \mathcal{O}(\Delta t^2);\\
(viii) & \left(\frac{\partial u}{\partial x}\right)_i^{n+1} = \left(\frac{\partial u}{\partial x}\right)_\star + (1-\theta)\,\Delta t \left(\frac{\partial^2 u}{\partial x \partial t}\right)_\star + \mathcal{O}(\Delta t^2);\\
(ix) & (1-\theta)\left(\frac{\partial u}{\partial x}\right)_i^n + \theta\left(\frac{\partial u}{\partial x}\right)_i^{n+1} = \left(\frac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2);\\
(x) & (1-\theta)\left(\frac{\partial^2 u}{\partial x^2}\right)_i^n + \theta\left(\frac{\partial^2 u}{\partial x^2}\right)_i^{n+1} = \left(\frac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta t^2).
\end{array}
\tag{2.37}
$$

Substitution of this in (2.36) yields for the left-hand-side

$$
\begin{array}{l}
u_i^{n+1} - u_i^n\\
= \Delta t \left(\frac{\partial u}{\partial t}\right)_\star + \frac{1}{2}(1-2\theta)\Delta t^2 \left(\frac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3) =
\end{array}
$$

and for the right-hand-side

$$
\begin{array}{l}
-\lambda\theta \left\{\Delta x \left(\frac{\partial u}{\partial x}\right)_i^{n+1} + (\frac{1}{2} - w)\Delta x^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_i^{n+1} + \mathcal{O}(\Delta x^3)\right\}\\
-\lambda(1-\theta) \left\{\Delta x \left(\frac{\partial u}{\partial x}\right)_i^{n} + (\frac{1}{2} - w)\Delta x^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_i^{n} + \mathcal{O}(\Delta x^3)\right\}\\
= -\lambda \left\{\Delta x \left[\left(\frac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2)\right] + (\frac{1}{2} - w)\Delta x^2 \left[\left(\frac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta t^2)\right]\right\} + \mathcal{O}(\Delta x^3)\,.
\end{array}
$$

Making use of the parameter $\lambda = f'\Delta t/\Delta x$ we get the expression for the truncation error (in equation (2.36) $\tau_\star(u) = \text{rhs} - \text{lhs}$)

$$
\begin{array}{l}
\dfrac{\tau_\star(u)}{\Delta t} = -\left(\dfrac{\partial u}{\partial t}\right)_\star - (\dfrac{1}{2} - \theta)\Delta t \left(\dfrac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^2)\\
\quad - f' \left\{\left(\dfrac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2) + (\dfrac{1}{2} - w)\Delta x \left(\dfrac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta x^2)\right\}\\
= (\theta - \dfrac{1}{2})\Delta t \left(\dfrac{\partial^2 u}{\partial t^2}\right)_\star + f'(w - \dfrac{1}{2})\Delta x \left(\dfrac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta t^2)\,.
\end{array}
$$

This means that the method is first order consistent in time and space. If $\theta = \frac{1}{2}$ it is second order in time, and if $w = \frac{1}{2}$ it is second order in space.

Differentiating the PDE (2.35) we find that $u_{tt} = (f')^2 u_{xx}$, so that the principal term in the truncation error can be written as

$$
f' \left[(\theta - \frac{1}{2})f'\Delta t + (w - \frac{1}{2})\Delta x\right] u_{xx} = \tag{2.38}
$$

$$f'\Delta x \left[(\theta - \frac{1}{2})\lambda + (w - \frac{1}{2}) \right] u_{xx} = \epsilon_{\mathrm{num}} u_{xx} \,. \tag{2.39}$$

So, we may see the difference scheme as a second order discretisation of a *modified equation* (or the *equivalent differential equation*)

$$u_t = f'\Delta x \left[(\theta - \frac{1}{2})\lambda + (w - \frac{1}{2}) \right] u_{xx} - f' u_x \,. \tag{2.40}$$

For this reason ϵ_{num} is also called the *numerical diffusion* of the difference scheme.

Remark:
Notice that the stability of the difference scheme as studied in Section 2.4.2 corresponds exactly with a *positive* numerical diffusion!

The numerical diffusion, however, can cause problems in a convection diffusion problem if the diffusion coefficient in the equation is (much) smaller than the numerical diffusion. The numerical diffusion may then overrule the 'physical' diffusion, and the discrete solutions are much more smeared than the true solution of the differential equation.

If we make computations with one of the above schemes, we notice:
(i) the numerical diffusion can have a significant effect;
(ii) the second order scheme (without numerical diffusion) shows an "overshoot", i.e in the numerical solutions maxima or minima appear that do not exist in the true solution;
(iii) instable scheme are useless.

Example 2.4.1
We solve the equation

$$u_t = 0.01\, u_{xx} - u_x$$

on the domain $t \geq 0$, $x \geq 0$, with initial condition $u = 0$ at $t = 0$, and boundary condition $u = 1$ at $x = 0$ (at the "inflow" boundary). We discretise the term u_{xx} by

$$u_{xx} \approx \frac{1}{2}(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}) + \frac{1}{2}(u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n}) \,.$$

The other terms are discretised by (2.29), and we solve the problem for different values of w and θ. In the Figures 2.10 and 2.11 we show the results for $t = 1$ (and in Figure 2.12 for $t = 0.5$). We take a fixed $\Delta x = 0.01$ and either $\Delta t = 0.05$ ($\lambda = 0.5$) or $\Delta t = 0.1$ ($\lambda = 1.0$). For comparison the true solution is also shown, as well as the solution of the modified equation

$$u_t = (0.01 + \epsilon_{\mathrm{num}}) u_{xx} - u_x \,.$$

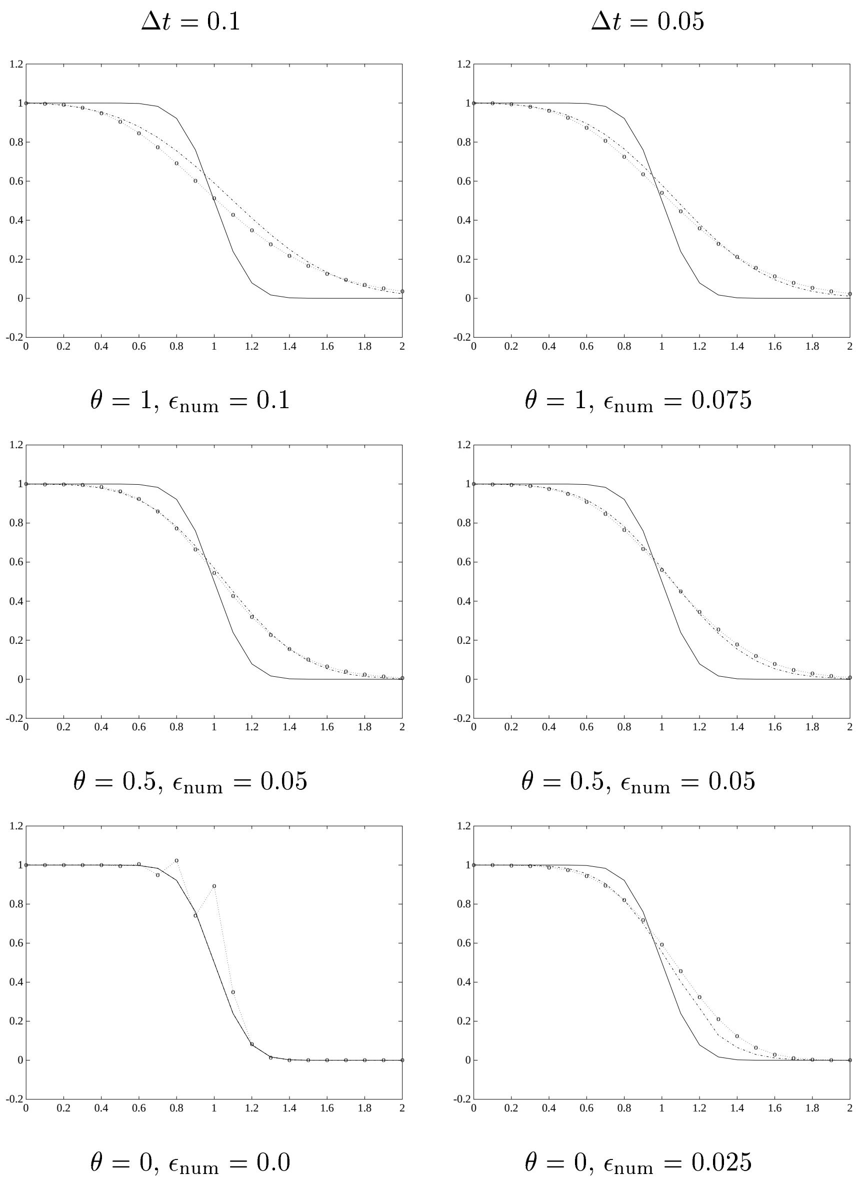


FIGURE 2.10. Numerical solution of the convection diffusion equation
In all figures the numerical solution is given for $t = 1$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with upwind space discretisation: $w = 1.0$, $\Delta x = 0.1$.

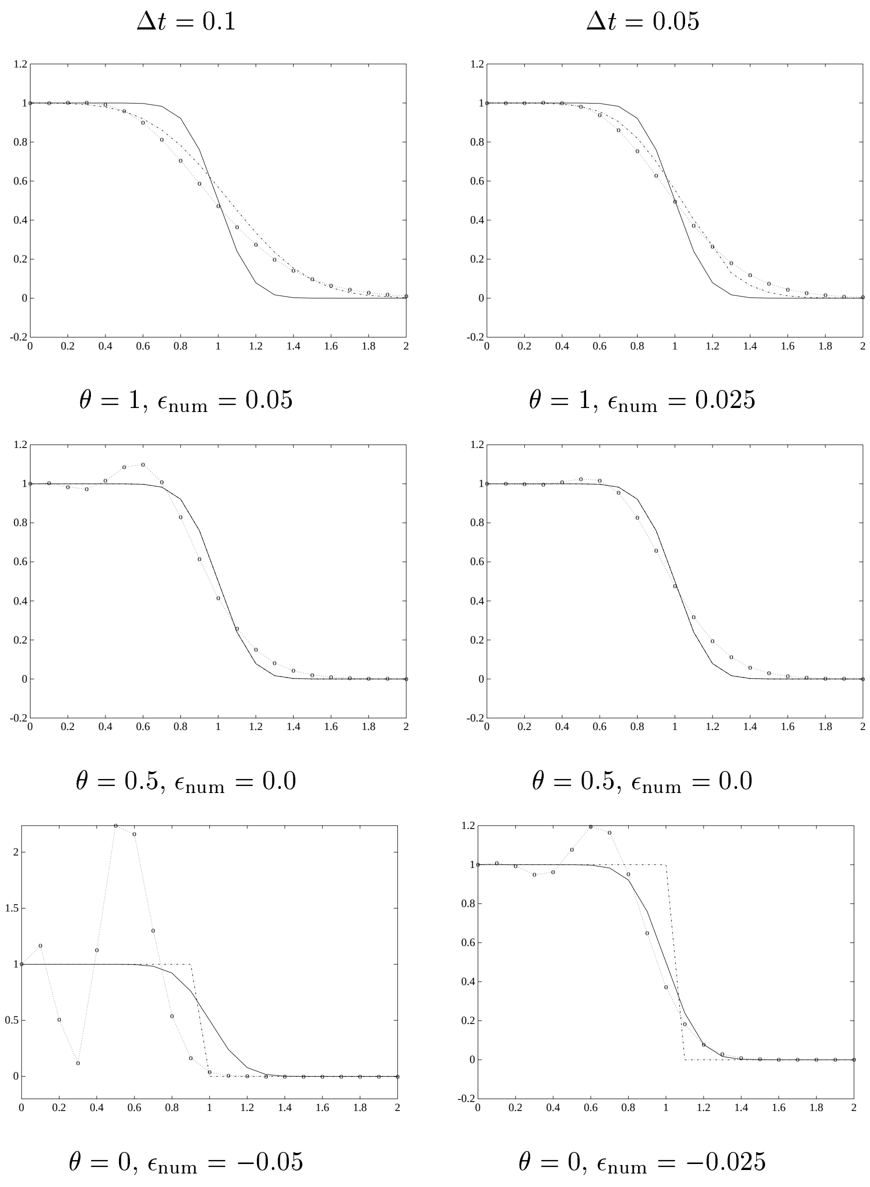
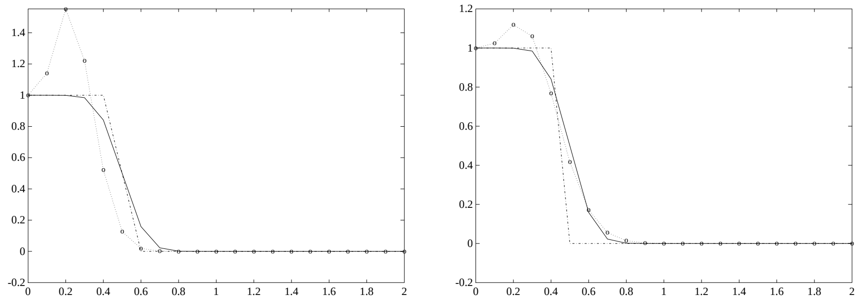


FIGURE 2.11. Numerical solution of the convection diffusion equation
In all figures the numerical solution is given for $t = 1$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with central space discretisation: $w = 0.5$, $\Delta x = 0.1$.



$\Delta t = 0.1$, $\epsilon_{\mathrm{num}} = -0.05$ $\Delta t = 0.05$, $\epsilon_{\mathrm{num}} = -0.025$

FIGURE 2.12. Numerical solution of the convection diffusion equation
In these figures the numerical solution is given for $t = 0.5$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with central space discretisation ($w = 0.5$), and forward Euler time-discretisation ($\theta = 0$), $\Delta x = 0.1$.

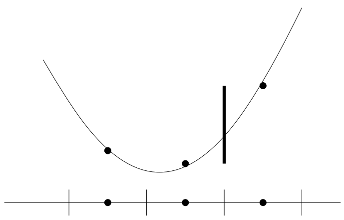


FIGURE 2.13. Non-conservative interpolation

2.5 Criteria for good discretisation methods

In this section we want to summarise a set of rules and give criteria that can help us to find good discretisation methods. The last three rules are taken from [41].

Rule 1. *If the number of nodal points (coefficients for the representation) increases we may expect (and we should require) that the discrete solutions converge to a true solution of the continuous problem.*

This property of a discretisation is usually mathematically formulated by an asymptotic statement: if the mesh-width h of the discretisation is sufficiently small, the discretisation error (in the solution) (in some norm) should vanish (at some rate) as h tends to zero.

Rule 2 *Often we wish more than only an asymptotic validity. We wish that a reasonably accurate solution is already obtained for a reasonably fine (= reasonably coarse) grid. We do not want to construct an extremely fine grid before we can be sure about the value of the approximation.*

This requirement is more difficult to formulate mathematically. Often it is expressed that we want to obtain a "physically realistic" solution. I.e. one requires that the discrete solution -already for a coarse grid- shows the same "global character" as the continuous solution.

For example, if it is known that the continuous solution is positive or monotone, one often requires that also the discrete solution is positive or monotone. E.g. (i) if concentrations of a chemical are computed, one doesn't like to find negative concentrations by approximation errors; (ii) for the stationary solution of a heat-conduction problem without sources or sinks, one requires that the temperature as computed does not show local extrema; (iii) if the continuous equation describes a conservation law, one requires that also the computed quantity is not created or does not disappear in the computation.

For the class of (hyperbolic) problems in conservation form it is relatively simple to satisfy the requirement that the discrete equations *satisfy the conservation law*, not only globally, for all Ω, but also locally, for each cell Ω_e of the discretisation of Ω.

A class of (elliptic) problems for which the requirement of "physical relevance" is usually not too hard to satisfy, are those problems that can be formulated as a minimisation problem. For such *symmetric elliptic problems* a convex functional exists that is minimised by the solution of the PDE.

Often the functional has a physical meaning (e.g. the energy left in the system) and, therefore, the minimising element in a subspace satisfies automatically the requirement of physical relevance. Some discretisation methods (finite element methods) find this minimising approximate solution.

Rule 3: Obeying the conservation law. A conservation law is most reliably discretised if the discrete system satisfies the law on each subelement of the discretisation separately. In order to satisfy the conservation law locally in a box method, the flux over box-interfaces should be computed consistently. I.e. the computed flux $\mathbf{J}(u) \cdot \mathbf{n}$ over the box interface should be the same when it represents the flow out of the one cell or when it represents the flow into the other. The explanation is simple: the amount of the conserved quantity that leaves the one cell should reappear (exactly) in the other cell.

Example 2.5.1
If we use a three-point, parabolic approximation to compute the flux over the right boundary of cell Ω_i (fig.2.13) then the requirement of conservation is not fulfilled. If we take the two-point linear approximation the conservation requirement is satisfied.

Example 2.5.2
See Section 2.3.1. In equation (2.12) we have (correctly) approximated $\mathbf{J}(u)$ by

$$\mathbf{J}(u) = k(x_{i+\frac{1}{2}})\frac{u_{i+1} - u_i}{x_{i+1} - x_i}$$

If in equation (2.12) we had approximated $k(x)$ by $k(x_i)$, then the discretisation would not have been conservative.

Rule 4: Trivial (constant) solutions should satisfy the discrete solution.

If a differential operator only contains derivatives of a dependent variable, then with an approximate solution u also $u +$ constant should be a solution of the discrete equations. I.e. the constant function should be in the kernel of the discrete operator. (The row-sum of the linearisation of the discrete operator should vanish.)

This rule can be seen as an instance for the more *general principle:* Seeking a good discretisation for a complex (intricate) operator, it makes sense to find a discretisation that -at least- satisfies basic requirements for most simplifications of the problem. (A good method for a problem is also a good method for any simplification of that problem.)

Rule 5: Guaranteed positivity of the solution. In order to guarantee the positivity of the solution (in the absence of sources and sinks) for the linearisation of the discrete operator, the diagonal and the off-diagonal elements of the Jacobian matrix should have different signs.

Explanation: If we write the linearised discrete equation for one space dimension as

$$a_i u_i = a_{i-1} u_{i-1} + a_{i+1} u_{i+1} + s_i \tag{2.41}$$

the coefficients a_i, a_{i-1} and a_{i+1} should have the same sign to guarantee that for all possible u_{i-1} and u_{i+1}. The same argument is simple to generalise for two or three dimensions.

For discretisations with a higher order of accuracy (than one) it is difficult (often impossible) to satisfy this requirement. For higher order discretisations more non-zero diagonals appear in the linearised operator.

The fifth rule allows an extension for the case that a source term is present. First, if the source term is neglected the discretisation should satisfy rule 5. Further, (ii) if there is a source term, it should not disturb the positivity (monotonicity)

Let $s(u)$ the source term in the equation $Lu = s(u)$ allow the linearisation $s(u) \approx s^C + s^L + u$, then the discretisation of $Lu = s(u)$ leads to (see sect (2.3.2))

$$(a_i - s_i^L)u_i = a_{i-1}u_{i-1} + a_{i+1}u_{i+1} + s_i^C \ ,$$

then s_i^L and a_i should have opposite signs.

3

Hyperbolic Conservation Laws

3.1 Scalar Conservation Laws

3.1.1 Introduction

In chapter 1 we have already seen typical examples of hyperbolic conservation laws. In its most simple scalar linear form, the hyperbolic conservation law is written as

$$u_t + a\, u_x = 0\,, \quad t\in\mathbb{R}^+,\ x\in\mathbb{R}, \tag{3.1}$$

where a is a constant number. This is an initial value problem, *IVP*, (*initial value problem*, *Cauchy problem*), with the initial condition

$$u(0,x) = u_0(x)\,. \tag{3.2}$$

We have seen that the solution is $u(t,x) = u_0(x - at)$. As time proceeds, the initial function shifts to the right ($a > 0$) or to the left ($a < 0$). The characteristics are computed from eq.(1.23)

$$\frac{1}{dt} = \frac{a}{dx} = \frac{0}{du}\,, \tag{3.3}$$

so that the characteristic direction ($\frac{1}{dt} = \frac{a}{dx}$) is

$$x = at + C,$$

and the solution along the characteristic ($\frac{1}{dt} = \frac{0}{du}$) is

$$u = \text{constant.}$$

More generally, with a smooth function $a(x)$, the characteristics are not straight lines anymore, but, as long as the characteristics don't intersect, the typical behaviour of the solution doesn't change. The solution has the property that at any point (t_1, x_1), $t_1 > 0$, the solution depends only on a single point $u_0(x_0)$ at $t = 0$. So, if we would change the initial data on a small subset $G \subset \mathbb{R}$, then the solution would change only in a corresponding strip. The set G is called the *domain of dependence* of this strip. However, for a system of equations, where more characteristic directions exist, the domain of dependence of a single point (t_1, x_1) is typically an interval at $t = 0$, as was also seen for the second order equation in section 1.1.2.

On the other hand, the domain in $\mathbb{R}^+ \times \mathbb{R}$ that depends on the solution in area G, is called the *range of influence* for G. Because hyperbolic equations have typically a finite propagation speed, this range of influence is bounded in space.

In the above considerations we assumed the function u to be sufficiently differentiable, i.e. $u \in C^1$. From the observation that (in the scalar case) the solution along a characteristic depends on a single value $u_0(x_0)$ only, we may also find solutions in the case where discontinuities are present in the initial solution $u_0(x)$. Then the discontinuity propagates along a characteristic. Such a generalisation is also readily extended for the vector case. Then, a discontinuity will propagate along all characteristics and we see that such a generalised solution *does* satisfy the integral form (1.35) of the differential equation.

The situation gets more interesting if the coefficient a depends on the solution u, or if we take the equation

$$u_t + f(u)_x = 0\,, \tag{3.4}$$

where $f(u)$ is a nonlinear function of u. This equation can also be written

$$u_t + f'(u)u_x = 0$$

and the *characteristic speed* is $f'(u)$. For convenience, we sometimes assume that f is a convex function (i.e. $f'' > 0$). This corresponds to what we will call *genuinely nonlinear*[1] for systems of equations in definition 3.2.6.

3.1.2 The Riemann problem

The conservation law (3.4) with piecewise constant initial function u_0, that has a single discontinuity is called the *Riemann problem.* As an example we study again the inviscid Burgers' equation (1.2.1). Without loss of generality, we take the discontinuity at $x = 0$. Hence the initial data are $u_0(x) = u_L$ for $x < 0$ and $u_0(x) = u_R$ for $x > 0$. In section 1.2.2 we have seen that such a discontinuity will move with a speed $(u_L + u_R)/2$. A solution with two constant states u_L and u_R and such a discontinuity is certainly a weak solution and satisfies the Rankine-Hugoniot relation. In the case that $u_L > u_R$ it is also clear that all the domain $\mathbb{R}^+ \times \mathbb{R}$ is in the range of influence of the initial data. However, that is not the case if $u_L < u_R$. Then, a wedge remains in which the origin of possible characteristics is not clear. This was shown in the examples 1.2.5 and 1.2.6. In fact we may construct an infinite number of (weak) solutions. If these solutions are differentiable, in the wedge they should satisfy $u(t, x) = h(x/t)$, and where they are discontinuous they should satisfy the Rankine-Hugoniot relation. By analogy to a Riemann problem in gas-dynamics the discontinuous solutions are called *shock waves*, and the continuous fan solutions are called *rarefaction waves*.

[1] This terminology can be understood because $f''(x) = 0$ (f' doesn't change), corresponds with a (local) linear behaviour.

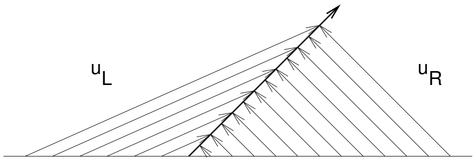


FIGURE 3.1. The generation of a shock for Burgers' equation

It is easily seen that a sufficiently differentiable solution satisfies

$$u(t,x) = \begin{cases} u_L, & x/t < f'(u_L), \\ h(x/t), & f'(u_L) < x/t < f'(u_R), \\ u_R, & x/t > f'(u_R), \end{cases} \tag{3.5}$$

where $h(w)$ is the solution of $f'(h(w)) = w$. Hence, for a general convex scalar Riemann problem (3.4) and $u_L < u_R$, the rarefaction wave solution is described by (3.5).

Notice that the *conserved quantity* in a conservation law is not uniquely expressed by the differential form. We see that the equations

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0 \tag{3.6}$$

and

$$(u^2)_t + \left(\frac{2}{3}u^3\right)_x = 0, \tag{3.7}$$

which are equivalent as differential forms, are not equivalent as conservation laws. Equation (3.6) expresses conservation of u, whereas equation (3.7) rather expresses conservation of u^2. The continuous solutions of (3.6) and (3.7) are identical, whereas possible shocks travel with different speeds, viz. the shock speed is $s = (f_R - f_L)/(u_R - u_L)$ that is $s = (u_L + u_R)/2$ for (3.6) and $s = \frac{2(u_R^3 - u_L^3)}{3(u_R^2 - u_L^2)}$ for (3.7).

3.1.3 The entropy function

In order to guarantee that the solution of the weak form of the differential equation corresponds with the physically relevant solution, an additional condition is required. Such possible conditions are known as entropy conditions. We have seen two of these already in (1.51) and (1.52) at the end of section 1.2.2. In this section we introduce the *entropy function* in order to formulate an entropy condition that is related to a (vanishing) "viscosity term" such as ϵu_{xx} in (1.32).

Assume that a function $\eta(u)$ exists, and a corresponding $\psi(u)$, such that

$$\eta(u)_t + \psi(u)_x = 0 \tag{3.8}$$

holds. Then, for a smooth u, we find

$$\eta'(u)u_t + \psi'(u)u_x = 0.$$

From the conservation law in the form $u_t + f'(u)u_x = 0$ we obtain

$$\psi'(u) = \eta'(u)f'(u). \tag{3.9}$$

For a scalar conservation laws many possible η and ψ may exist[2]. For a system of conservation laws, where η and ψ are still scalar functions, the relation between η and ψ reads[3]

$$\nabla_{\mathbf{u}}\psi(\mathbf{u}) = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}\ \nabla_{\mathbf{u}}\eta(\mathbf{u}).$$

This is a system of equations, where there are two variables η and ψ. If the system has too many equations, it may have no solution[4].

A convex[5]function $\eta(u)$ is called an *entropy function* for equation (3.4) if there exists a $\psi(u)$ so that $\psi'(u) = \eta'(u)f'(u)$ for all u. Then function $\psi(u)$ is called the *entropy flux.* (See also Definition 3.2.8). It directly follows that η is conserved for smooth solutions of (3.4), because (3.8) is satisfied for this solution u. For the solution of the viscous equation

$$u_t + f(u)_x = \epsilon u_{xx}\ ,$$

we can associate the (small but) positive viscosity term with an entropy inequality. This is seen as follows. Because the solution of the parabolic equation has no discontinuity, equation (3.9) leads to[6]

$$\eta(u)_t + \psi(u)_x = \epsilon\eta'(u)u_{xx}.$$

For all $\phi > 0$ with compact support we find

$$\int \phi(\eta(u)_t + \psi(u)_x)\, d\Omega = \epsilon \int \phi\eta'(u)u_{xx}d\Omega = -\epsilon \int (\phi\eta'(u))_x u_x d\Omega =$$

---

[2]E.g. for Burgers' equation $f(u) = \frac{1}{2}u^2$, and if we take $\eta(u) = u^2$ we easily construct $\psi(u) = \frac{2}{3}u^3 + c$.

[3]Here $\nabla_{\mathbf{u}}\psi(\mathbf{u})$ denotes $\left(\frac{\partial\psi}{\partial u_1}, \cdots, \frac{\partial\psi}{\partial u_n}\right)$.

[4]In the case of a system

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0\,, \tag{3.10}$$

where $A = d\mathbf{f}/d\mathbf{u}$ is symmetric, we know $\partial f_j/partial u_l = partial f_l/\partial u_j$. This symmetry relation allows the existence of a function $g(\mathbf{u})$, such that $\partial g/\partial u_l = f_l$. Now it is immediately clear that $\eta = \frac{1}{2}\sum_j u_j^2$ and $\psi = \sum_j u_j f_j - g$ are the entropy function and the entropy flux for (3.10).

[5]A convex $\eta(u)$ means that η is twice differentiable and $\eta_{uu} > 0$.

[6]$\epsilon\eta' u_{xx} = \eta'(u)(u_t + f(u)_x) = \eta_u u_t + \eta_u f_u u_x = \eta_u u_t + \psi_u u_x = \eta(u)_t + \psi(u)_x$.

$$= -\epsilon \int \phi_x \eta'(u) u_x \, d\Omega - \epsilon \int \phi \eta''(u) u_x u_x \, d\Omega$$

$$< -\epsilon \int \phi_x \eta'(u) u_x d\Omega \to 0,$$

for $\epsilon \to 0$. Or, in distributional sense, for functions u with compact support in $\mathbb{R}$ and $\phi \in C_0^\infty(\mathbb{R})$ we have $\eta(u)_t + \psi(u)_x \leq 0$.

The function $u(t,x)$ is called the entropy solution of $u_t + f(u)_x = 0$, if for an entropy function and its corresponding entropy flux, the following relation is satisfied (in the distributional sense)

$$\eta(u)_t + \psi(u)_x \leq 0\,. \tag{3.11}$$

Further, if u satisfies (3.11) and has compact support on $\mathbb{R}$, integrating 3.11) w.r.t. x gives $\int \eta(u)_t \, dx \leq 0$, which implies that $\int \eta(u) \, dx$ is a decreasing function of t. (Recommended reading: [35].)

3.2 Hyperbolic systems

3.2.1 Introduction

In this section we study a general first-order system of quasi-linear equations in two independent variables (time, and one space variable) of the form

$$\frac{\partial q}{\partial t} + \frac{\partial f(q)}{\partial x} = 0\,, \tag{3.12}$$

where $q = (q_1, ..., q_n)^T \in \mathbb{R}^n$ and $(t,x) \in \mathbb{R}^+ \times \mathbb{R}$. We assume that the vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^n$ is C^1. We define the $n \times n$ matrix $A(q) = \frac{df}{dq}(q)$.

Definition 3.2.1 The system (3.12) of first order PDEs is called hyperbolic if there exists a real diagonal matrix $D(q)$ and a non-singular real matrix $R(q)$ such that

$$A(q)R(q) = R(q)D(q)\,, \quad \forall q \in \mathbb{R}^n\;. \tag{3.13}$$

The column vectors of $R(q)$ are eigenvectors of $A(q)$, and the diagonal entries of $D(q)$ are the corresponding eigenvalues. We shall denote by $R_k(q)$ the k-th column vector of $R(q)$ and with $\lambda_k(q)$ the corresponding eigenvalue: $\lambda_k(q) = D_{kk}(q)$. Furthermore we shall assume that the eigenvalues $\lambda_k(q)$ have been labeled in increasing order $\lambda_1 \leq ... \leq \lambda_n$.

Example 3.2.2 *The linear case*
Suppose $f(q) = Aq$ where A is an $n \times n$ matrix, independent of q. Then (3.12) simplifies to

$$\frac{\partial q}{\partial t} + A\frac{\partial q}{\partial x} = 0\;. \tag{3.14}$$

Since the eigenvectors R_k are independent of q, solution $q = q(t,x)$ of (3.14) can now simply be decomposed with respect to the basis $\{R_1, ..., R_n\}$, i.e.

$$q = q(t,x) = \sum_{i=1}^{n} \alpha_i(t,x) R_i \,, \tag{3.15}$$

where $\alpha_i : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$. Substitution of this expression in (3.14) leads to

$$\frac{\partial q}{\partial t} + A\frac{\partial q}{\partial x} = \sum_{i=1}^{n} \left\{ \frac{\partial \alpha_i}{\partial t}(t,x) + \lambda_i \frac{\partial \alpha_i}{\partial x}(t,x) \right\} R_i = 0\,, \tag{3.16}$$

and, because the eigenvectors R_i are independent,

$$\frac{\partial \alpha_i}{\partial t}(t,x) + \lambda_i \frac{\partial \alpha_i}{\partial x}(t,x) = 0, \quad i = 1, ..., n. \tag{3.17}$$

The general solution of this equation is

$$\alpha_i(t,x) = \alpha_i^0(x - \lambda_i t)\,, \quad i = 1, ..., n, \tag{3.18}$$

with $\alpha_i^0 : \mathbb{R} \to \mathbb{R}$, and hence the solution of (3.14) is

$$q(t,x) = \sum_{i=1}^{n} \alpha_i^0(x - \lambda_i t) R_i \ . \tag{3.19}$$

So, the solution of the initial value problem on $\mathbb{R}^+ \times \mathbb{R}$

$$\begin{cases} \frac{\partial q}{\partial t} + A\frac{\partial q}{\partial x} = 0, \\ \\ q(0,x) = q_0(x)\,, \quad x \in \mathbb{R}\,, \end{cases} \tag{3.20}$$

is written (3.19), where the $\alpha_i^0(x)$ are such that

$$\sum_{i=1}^{n} \alpha_i^0(x) R_i = q_0(x) \ . \tag{3.21}$$

Hence, the solution of (3.20) is obtained by decomposing the initial function $q_0 : \mathbb{R} \to \mathbb{R}^n$ with respect to the basis $\{R_1, ..., R_n\}$.

3.2.2 The Riemann problem

The Riemann problem for (3.12) is important because it forms the basis for many schemes for its numerical solution. For instance, the *Godunov scheme* uses the exact solution of the Riemann problem for the numerical solution of the Euler equations. Other upwind schemes use approximate solutions of the Riemann problem.

Definition 3.2.3 The Riemann problem for a hyperbolic system (3.12) is the following initial value problem

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} f(q) = 0\,, \quad \text{for} \quad (t,x)\in\mathbb{R}^+ \times \mathbb{R}\,, \tag{3.22}$$

with

$$q(0,x) = \begin{cases} q_L, & \text{if } \ x < 0\,, \\ q_R, & \text{if } \ x > 0\,, \end{cases}$$

where q_L and q_R are constant states.

Theorem 3.2.4 Suppose there exists a unique solution $q = q(t,x)$ of the Riemann problem (3.22). Then the solution $q = q(t,x)$ can be written in similarity form $q(t,x) = h(x/t)$.
Proof: Define $q_\alpha(t,x) = q(\alpha t, \alpha x)$ with $\alpha\in\mathbb{R}^+$. Then it is easily verified that $q_\alpha(t,x)$ is also a solution of the Riemann problem. Hence, $q(t,x) = q(\alpha t, \alpha x)\ \ \forall\alpha\in\mathbb{R}^+$, so $q(t,x) = h(x/t)$. $\square$

Example 3.2.5
Consider the Riemann problem for a linear hyperbolic system. We follow the solution method outlined in example 3.2.2, and we decompose the initial states as $q_L = \sum_{i=1}^n \alpha_i R_i$, $q_R = \sum_{i=1}^n \beta_i R_i$. Hence,

$$q(0,x) = \sum_{i=1}^{n} \left(\beta_i H(x) + \alpha_i (1 - H(x))\right) R_i\,, \tag{3.23}$$

where $H : \mathbb{R} \to \mathbb{R}$ is the Heavyside function[7]. From (3.19) the solution of the Riemann-problem follows as

$$q(t,x) = \sum_{i=1}^{n} \{\beta_i H(x - \lambda_i t) + \alpha_i (1 - H(x - \lambda_i t))\} R_i\ . \tag{3.24}$$

This solution is illustrated in fig. 3.2 for $n = 3$ and $\lambda_1 < 0 < \lambda_2 < \lambda_3$.

As a corollary we notice that

$$\begin{aligned} q(t,0) &= \sum_{i=1}^{n} [\beta_i H(-\lambda_i t) + \alpha_i (1 - H(-\lambda_i t))] R_i \\ &= \sum_{i=1}^{n} (\text{ if } \ \lambda_i > 0 \ \text{ then } \ \alpha_i \ \text{ else } \ \beta_i \text{endif})\, R_i, \end{aligned}$$

[7] This Heavyside function is defined by $H(x) = 1$ if $x > 0$, $H(x) = 0$ if $x < 0$.

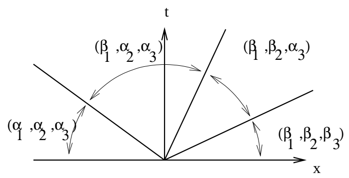


FIGURE 3.2. The solution of the Riemann problem for a linear hyperbolic system ($n = 3$).

and hence[8] (exercise)

$$f(q(t,0)) = A\,q(t,0) = A^+ q_L + A^- q_R \,.$$

In general, for a nonlinear hyperbolic system the solution of the Riemann-problem is not simple. But for certain initial states (q_L, q_R) the solution is relatively simple. In the remainder of this section we show how to treat such simpler solutions. For this purpose we introduce the following concept:

Definition 3.2.6 Consider the hyperbolic system (3.12) and the notation introduced in section 3.2.1. We call the eigenvalues $\lambda_k(q)$ and an eigenvector is *genuinely nonlinear* if

$$(\nabla\lambda_k(q),\ R_k(q)) \neq 0\,, \quad \forall q \in \mathbb{R}^n\,. \tag{3.25}$$

We call $\lambda_k(q)$ *linearly degenerate* if

$$(\nabla\lambda_k(q),\ R_k(q)) = 0\,, \quad \forall q \in \mathbb{R}^n\,. \tag{3.26}$$

Here (,) denotes the usual inner product in $\mathbb{R}^n$ and

$$\nabla\lambda_k(q) = (\frac{\partial\lambda_k}{\partial q_1}, \cdots, \frac{\partial\lambda_k}{\partial q_n})^T\,.$$

For the Euler equations (section 3.3) we can show that each eigenvalue $\lambda_k(q)$ is either genuinely nonlinear or linearly degenerate.

3.2.3 Simple wave solution of the Riemann problem

In the previous section we solved the Riemann problem for the linear system of equations (3.22). The solution consisted of a function that is constant in

[8]We use the notation $\lambda^+ = \max(\lambda, 0)$, and $\lambda^- = \min(\lambda, 0)$; with $\Lambda = \mathrm{diag}(\lambda_i)$ we have $\Lambda^+ = \mathrm{diag}(\lambda_i^+)$ and $\Lambda^- = \mathrm{diag}(\lambda_i^-)$; and with $AR = R\Lambda$ the eigenvector decomposition of Λ we have $A^+ = L\Lambda^+ R$ and $A^- = L\Lambda^- R$. We also use $|A| = A^+ - A^-$. This is the usual procedure to define functions for (normal) matrices, e.g. $\exp(A)$.

a number of sectors. These sectors are separated by discontinuities, each of which is associated with an eigenvalue of the linear system. Over these discontinuities, a number of quantities (associated with the other eigenvalues) remains constant. In order to analyse the Riemann problem in the nonlinear case, we first consider the part of a solution associated with a single eigenvalue. Here different possibilities exist. If the eigenvalue is linearly degenerate, the behaviour of the solution is much similar to the linear case: a discontinuity appears for each such eigenvalue, and the Rankine-Hugoniot relation is satisfied over that so called *contact discontinuity.* In the genuinely nonlinear case we consider again two possibilities: $\lambda(q_L) < \lambda(q_R)$ in which case we find a *rarefaction wave*, and $\lambda(q_L) > \lambda(q_R)$ in which case we encounter a *shock wave.* The solution around a shock wave is, again, described by the Rankine-Hugoniot relation, as we have seen in section 1.2.2. We first treat the rarefaction wave solution.

Suppose $\lambda_k(q)$ is a genuinely nonlinear eigenvalue. Then $R_k(q)$ can be normalised such that

$$(\nabla\lambda_k(q),\ R_k(q)) = 1 \quad \forall q\ . \tag{3.27}$$

For an arbitrary state q_L and $k\in\{1,\cdots,n\}$, we consider the following ordinary differential equation

$$\left\{\begin{array}{l} \frac{dq(\xi)}{d\xi} = R_k(q(\xi))\,, \quad \xi_L \leq \xi \leq \xi_R\,, \\ q(\xi_L) = q_L\,, \end{array}\right. \tag{3.28}$$

and we call the solution $q = \tilde{q}(\xi)$. In the space $\mathbb{R}^n$ of all states q, this solution describes a path that is everywhere tangent to the k-th eigenvector $R_k(q)$. On this path we define $q_R = \tilde{q}(\xi_R)$. (For a simple rarefaction wave this should be possible. The case of a more general q_R will be considered later.) Because

$$\frac{d}{d\xi}\lambda_k(\tilde{q}(\xi)) = \ \nabla\lambda_k(\tilde{q}(\xi))\cdot R_k(\tilde{q}(\xi)) = 1 \tag{3.29}$$

we have

$$\lambda_k(\tilde{q}(\xi)) = \xi\,, \tag{3.30}$$

and

$$\lambda_k(q_R) = \xi_R\,. \tag{3.31}$$

The path $\tilde{q}(\xi)$ described in the state space $\mathbb{R}^n$ is called the k-path through q_L.

By theorem 3.2.4, the solution $q(t,x)$ can be written in the form $q(t,x) = h(x/t)$, and hence, with $q(t,x) = \tilde{q}(x/t)$ we see

$$\frac{\partial q}{\partial t}(x,t) + \frac{\partial}{\partial x}f(q(x,t)) = \frac{\partial q}{\partial t}(x,t) + A(q(x,t))\frac{\partial q}{\partial x}(x,t,) \tag{3.32}$$

$$= -\frac{x}{t^2}R_k(q(x,t)) + \frac{1}{t}A(q(x,t))R_k(q(x,t)) \tag{3.33}$$

$$= -\frac{x}{t^2}R_k(q(x,t)) + \frac{1}{t}\lambda_k(q(x,t))R_k(q(x,t)) \; . \tag{3.34}$$

We see that $q(t,x)$ is a solution of (3.12) as soon as $x/t = \lambda_k(q(x,t))$. Hence, we find the solution by taking

$$q(t,x) = \begin{cases} q_L & \text{if} \quad x/t < \lambda_k(q_L), \\ \tilde{q}(x/t) & \text{if} \quad \lambda_k(q_L) < x/t < \lambda_k(q_R), \\ q_R & \text{if} \quad x/t > \lambda_k(q_R). \end{cases} \tag{3.35}$$

Now it is easily verified that $q(t,x)$ is the solution of the Riemann-problem. If $\lambda_k(q_L) = \xi_L < x/t < \lambda_k(q_R)$ then

$$\lambda_k(q(t,x)) \;=\; \lambda_k(\tilde{q}(x/t)) = x/t\,. \tag{3.36}$$

So $q(x,t)$ is the solution of the Riemann problem with initial states (q_L, q_R). This solution is called a kth simple wave (or rarefaction wave). An illustration of this solution is given in fig. 3.3.

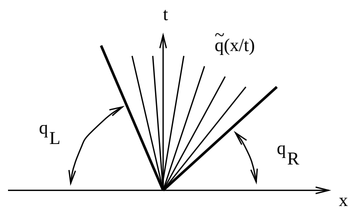


FIGURE 3.3. Simple wave solution of a Riemann problem

3.2.4 Contact discontinuity solution of the Riemann problem

Suppose $R_k(q)$ is a *linearly degenerate* eigenvector. Then

$$(\nabla\lambda_k(q),\; R_k(q)) = 0\,, \quad \forall q\in\mathbb{R}^n \; . \tag{3.37}$$

Let $\tilde{q}(\xi)$ be the solution of (3.28) and define $q_R = \tilde{q}(\xi_R)$. Because

$$\frac{d\,\lambda_k(\tilde{q}(\xi))}{d\xi} = \nabla\lambda_k(\tilde{q}(\xi)) \cdot R_k(\tilde{q}(\xi)) = 0 \tag{3.38}$$

we have $\lambda_k(\tilde{q}(\xi)) = \lambda_k(q_L) = \lambda_k(q_R)$, $\forall\xi\in[\xi_L, \xi_R]$. Presume

$$q(x,t) = \begin{cases} q_L\,, & \text{if} \;\; x/t < \lambda_k(q_L) = \lambda_k(q_R), \\ q_R\,, & \text{if} \;\; x/t > \lambda_k(q_L) = \lambda_k(q_R). \end{cases} \tag{3.39}$$

Now we see that $q(t,x)$ is discontinuous. We shall show that $q(t,x)$ is a solution of (the weak form of) the Riemann problem. An illustration of this solution is given in fig. 3.4.

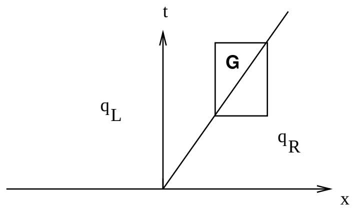


FIGURE 3.4. Contact discontinuity solution of a Riemann problem

Because $q(t,x)$ is discontinuous, the integral form of (3.12) has to be employed,

$$\int_{\partial G} \{q n_t + f(q) n_x\}\, d\sigma = 0\,, \tag{3.40}$$

with G an arbitrary area in the (t,x) space and $\mathbf{n} = (n_x, n_t)$ the outward unit normal an ∂G. It suffices to consider an infinitesimal rectangular volume with sides Δx and Δt having the discontinuity in its diagonal, cf. fig. 3.4. Equation (3.40) results in

$$f(q_R) - f(q_L) + \lambda_k(q_L)\ (q_L - q_R) = 0\,, \tag{3.41}$$

with $\frac{\Delta x}{\Delta t} = \lambda_k(q_L) = \lambda_k(q_R)$. These are the Rankine-Hugoniot relations. (Notice that q and f are now vectors!) We will show that (3.41) is satisfied.
Because for $\xi_L \leq \xi \leq \xi_R$

$$\frac{d}{d\xi}\left[f(\tilde{q}(\xi)) - \lambda_k(q_L)\tilde{q}(\xi)\right] \;=\; A(\tilde{q}(\xi))\frac{d\tilde{q}(\xi)}{d\xi} - \lambda_k(\tilde{q}(\xi))\frac{d\tilde{q}(\xi)}{d\xi} \tag{3.42}$$

$$= \; A(\tilde{q}(\xi))R_k(\tilde{q}(\xi)) - \lambda_k(\tilde{q}(\xi))R_k(\tilde{q}(\xi)) = 0\,, \tag{3.43}$$

we have

$$f(q_R) - \lambda_k(q_R)q_R = f(q_L) - \lambda_k(q_L)q_L\ . \tag{3.44}$$

Thus the jump relations (3.41) are indeed satisfied.

3.2.5 Shock waves

Beside contact discontinuities (λ_k linearly degenerate) and simple waves (λ_k genuinely nonlinear, $\lambda_L < \lambda_R$) there exists another elementary type of solutions of the Riemann problem, so called shock waves (λ_k genuinely nonlinear, $\lambda_L > \lambda_R$). Shock wave solutions are discontinuous and they satisfy the Rankine-Hugoniot relations and the entropy condition. The general solution of a Riemann problem is, under rather general circumstances, composed by simple waves, contact discontinuities and shock waves (see [7]). Here, we omit a detailed description of shock wave solutions because it would not contribute much to the understanding of the numerical solution methods that will be discussed.

3.2.6 Riemann invariants

We noticed in the previous section that, in the (t,x)-half plane, the solution of the Riemann problem consists of a number of sectors, where the solution is constant, and a number of discontinuities and rarefaction waves, one of these for each eigenvalue. In the linear case, we noticed that over the discontinuities some coefficients remain constant. In the nonlinear case the so called *Riemann invariants* take this role of such invariant quantities over discontinuities or rarefaction waves. As we shall see, Riemann invariants are not fixed numbers but functions of the state space $\mathbb{R}^n$.

Definition 3.2.7 Consider the hyperbolic system (3.12). Let $R_k(q)$ be the k-th eigenvector of $A(q) = \frac{df}{dq}(q)$. A k-Riemann invariant is a smooth function $\psi_k : \mathbb{R}^n \to \mathbb{R}$ such that

$$(\nabla\psi_k(q),\ R_k(q)) = 0\,, \quad \forall q{\in}\mathbb{R}^n\ . \tag{3.45}$$

By nature, a k-th Riemann invariant is a constant function along a k-path. Notice that if $R_k(q)$ is linearly degenerate, the corresponding eigenvalue $\lambda_k(q)$ is a Riemann invariant (see (3.26)). In general we may expect $n-1$ k-Riemann invariants whose gradients are linearly independent in $\mathbb{R}^n$. Riemann-invariants are useful for the construction of simple wave and contact discontinuity solutions for the Riemann problems.

If there are $n-1$ k-Riemann invariants $\psi_k^1, ..., \psi_k^{n-1}$, then it is easily seen that q_R, described by (3.28) is also described by

$$\psi_k^1(q_L) = \psi_k^1(q_R)\ ,\ \ldots\ ,\ \psi_k^{n-1}(q_L) = \psi_k^{n-1}(q_R)\,. \tag{3.46}$$

In the case of the Euler equations, the Riemann invariants are simple functions and, as we shall see in the section 3.4.1, formulas describing simple waves or contact discontinuities are easily obtained.

3.2.7 The entropy function

In this section the arguments given for a single conservation law in section 3.1.3 are generalised for systems of conservation laws.

Definition 3.2.8 A system of conservation laws (in d space dimensions)

$$\frac{\partial q}{\partial t} + \frac{\partial f^1(q)}{\partial x_1} + \cdots + \frac{\partial f^d(q)}{\partial x_d} = 0 \tag{3.47}$$

has an associated *entropy function* $U(u)$ and entropy flux*es* $F'u$, if there exists a scalar function $U : \mathbb{R}^n \to \mathbb{R}$, such that (i) U is convex (e.g. the matrix U_{uu} is symmetric positive definite) and (ii) U satisfies

$$U_u\, f_u^i = F_u^i\,, \quad \text{for } i = 1, ..., d, \tag{3.48}$$

for the entropy fluxes $F^i(u)$, $i = 1, ..., d$.

For smooth solutions u of (3.47) follows

$$\frac{\partial}{\partial t}U(u) + \text{div } F(u) = 0\,. \tag{3.49}$$

Limit solutions for small ϵ of

$$u_t + f_x(u) = \epsilon u_{xx}\,, \quad \epsilon > 0, \tag{3.50}$$

satisfy[9]

$$U(u)_t + F_x(u) \leq 0\,. \tag{3.51}$$

This inequality is called the *entropy inequality.*

Suppose u satisfies $U_t + F_x \leq 0$ and has compact support in x. Integration of $U_t + F_x \leq 0$ over $\mathbb{R}$ gives $\int U_t\,dx \leq 0$, which implies that $\int U_t\,dx$, is a decreasing function of t.

Definition 3.2.9 A hyperbolic system of conservation laws (3.47) is called *symmetric hyperbolic* if, by a transformation of dependent variables, it can be cast into the form

$$Pv_t + \sum_i B^i(v)\frac{\partial v}{\partial x_i},$$

where the matrices P and $B^i(v)$ are symmetric and P is positive definite.

Theorem 3.2.10 *Godunov's theorem.* If equation (3.12) can be symmetrised by introduction of a new variable v, (which implies that there exist scalar functions $q(v)$ and $r^i(v)$ with convex q, so that $\partial q/\partial v_k = u_k$ and $\partial r^i/\partial v_k = f^i_k$) then equation (3.12) has an associated entropy function

$$U(u) = q_v \cdot v - q(v)$$

and corresponding entropy fluxes

$$F^i(u) = r^i_v \cdot v - r^i(v).$$

Proof: exercise (see HLL/HCS , page 4). $\square$

Theorem 3.2.11 *Mock's theorem* If $U(u)$ is an entropy function for (3.12), then the variable set $\mathbf{v} = U_u$ symmetrises the equation.

Proof: exercise (see HLL/HCS , page 7). $\square$

---

[9] For a convex $U(u)$ we know $U_{uu}(u)u_xu_x > 0$ and, hence, $U_{xx} = (U_x)_x = (U_uu_x)_x = U_uu_{xx} + U_{uu}u_xu_x \geq U_uu_{xx}$. Further $U_t(u) + F_x(u) = U_uu_t + F_uu_x = U_uu_t + U_uf_uu_x = U_u(u_t + f_uu_x) = U_u(\epsilon u_{xx}) = \epsilon U_u(u_{xx}) = \epsilon(U_{xx} - U_{uu}u_xu_x) < \epsilon U(u)_{xx} \to 0$, as $\epsilon \to 0$ (in distributional sense).

The variables that symmetrise the system of equations are called the entropy variables. Further, we mention the following facts

1. The Euler equations (Example 1.1.4) have a *mathematical entropy function* $-\rho S$, with S the *physical entropy*
$$S(u) = \ln(\rho^{-\gamma} p)\,.$$
2. If u is a piecewise smooth solution with a discontinuity, then for the smooth parts (3.51) holds, whereas over the discontinuities we have
$$F(u_R) - F(u_L) - s(U(u_R) - U(u_L)) \leq 0\,, \tag{3.52}$$
where s is the shock speed.
3. If u_L and u_R are sufficiently close, then the entropy inequality (3.51) is equivalent with the condition: for some index $k \in \{1, 2, \cdots, n\}$ one has
$$\lambda_k(u_L) > s > \lambda_k(u_R)\,, \tag{3.53}$$
while
$$\lambda_{k-1}(u_L) < s < \lambda_{k+1}(u_R)\,. \tag{3.54}$$
This is the generalisation of (1.52) to systems of equations.

3.3 The Euler equations for inviscid flow

In the following two sections the *Euler equations*, mentioned earlier in Example 1.1.4, are studied in more detail. We consider the equations in two dimensions. This restriction is not essential and for ease of notation only.

We follow [49]. We work with a Cartesian coordinate system (x, y), and t represents the time. With $\rho = \rho(t, x, y)$, $u = u(t, x, y)$, $v = v(t, x, y)$ and $p = p(t, x, y)$ we denote density, velocity components in the x- and y-direction and pressure. These quantities are the so-called primitive variables. We consider an arbitrary open region $\Omega \subset \mathbb{R}^2$ and we take $\mathbf{n} = (n_1, n_2)^T$ the outward unit normal on the boundary $\Gamma = \partial\Omega$. We formulate the physical laws of conservation of mass, momentum and energy for the fluid in Ω.

3.3.1 Conservation of mass

The law of conservation of mass is given by
$$\frac{d}{dt}\int_\Omega \rho\, dV + \int_\Gamma \rho\, \mathbf{n}\cdot\mathbf{v}\, d\sigma\ = 0, \tag{3.55}$$
where $\mathbf{v} = (u, v)^T$ is the velocity, dV is a volume element, and $d\sigma$ a surface element. With $\mathbf{n}\cdot\mathbf{v}$ we denote the inner product: $\mathbf{n}\cdot\mathbf{v} = n_1 u + n_2 v$. With

Gauss' theorem for a vector field, assuming that $\rho\mathbf{v}$ is differentiable, we may write

$$\int_\Gamma \rho\, \mathbf{n}\cdot\mathbf{v}\, d\sigma = \int_\Omega \text{ div } (\rho\mathbf{v})\, dV\,.$$

Using the fact that $\Omega \subset \mathbb{R}^2$ is arbitrary we find the equation of continuity:

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0\,. \tag{3.56}$$

Equations (3.55) and (3.56) are the equation of continuity in integral and differential form respectively. Both forms are important.

3.3.2 Conservation of momentum

Assuming frictionless flow and absence of body forces, the law of conservation of momentum is given by

$$\frac{d}{dt}\int_\Omega \rho\mathbf{v} dV + \int_\Gamma \rho\mathbf{v}\,(\mathbf{n}\cdot\mathbf{v})\, d\sigma = -\int_\Gamma p\,\mathbf{n}\, d\sigma\,. \tag{3.57}$$

Equation (3.57) is a vector equation. The x-component is

$$\begin{aligned}\frac{d}{dt}\int_\Omega \rho u\, dV &= -\int_{\partial\Omega} \rho u\,(\mathbf{n}\cdot\mathbf{v})\, d\sigma - \left(\int_{\partial\Omega} p\mathbf{n} d\sigma\right)_x \\ &= -\int_\Omega \text{div } (\rho u\mathbf{v})\, dV - \left(\int_\Omega \nabla p dV\right)_x\,, \end{aligned} \tag{3.58}$$

where we have used the theorem of Gauss for both a vector- and scalar-field. The y-component of equation (3.57) is

$$\begin{aligned}\frac{d}{dt}\int_\Omega \rho v\, dV &= -\int_{\partial\Omega} \rho v\,(\mathbf{n}\cdot\mathbf{v})\, d\sigma - \left(\int_{\partial\Omega} p\mathbf{n}\, d\sigma\right)_y \\ &= -\int_\Omega \text{div } (\rho v\mathbf{v})\, dV - \left(\int_\Omega \nabla p\, dV\right)_y\,. \end{aligned} \tag{3.59}$$

Because Ω is arbitrary, we infer from equations (3.58) and (3.59)

$$\begin{aligned}&\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0\,, \\ &\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0\,, \end{aligned} \tag{3.60}$$

the momentum equations in differential form.

3.3.3 Conservation of energy

Let e denote the specific internal energy of the fluid. Internal and kinetic energy make up the the energy of the fluid and it is equal to $\rho e+\frac{1}{2}\rho(u^2+v^2)$ per unit of volume. We define the *total energy* as

$$E = \rho e + \frac{1}{2}\rho(u^2 + v^2) \ . \tag{3.61}$$

Under the assumptions of a nonviscous, non-heat-conducting fluid and absence of body forces, the law of conservation of energy is given by

$$\frac{d}{dt}\int_{\Omega} E\, dV + \int_{\partial\Omega} E\,\mathbf{n}\cdot\mathbf{v}\, d\sigma = -\int_{\partial\Omega} p\,\mathbf{n}\cdot\mathbf{v}\, d\sigma\,, \tag{3.62}$$

and, as before, by Gauss' theorem we derive the energy equation in differential form as

$$\frac{\partial}{\partial t}E + \frac{\partial}{\partial x}(E+p)u + \frac{\partial}{\partial y}(E+p)v = 0\,. \tag{3.63}$$

Combining the equations (3.56), (3.60) and (3.63) we find the Euler equations:

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho\\ \rho u\\ \rho v\\ E\end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix}\rho u\\ \rho u^2+p\\ \rho uv\\ (E+p)u\end{pmatrix} + \frac{\partial}{\partial y}\begin{pmatrix}\rho v\\ \rho uv\\ \rho v^2+p\\ (E+p)v\end{pmatrix} = 0\ . \tag{3.64}$$

These equations describe the flow of a nonviscous, non-heat-conducting fluid without body forces. Notice that there are 5 unknowns in 4 equations. An additional equation relates density, pressure and internal energy of the fluid, viz. the thermodynamical equation of state, which can be written as

$$p = p(\rho, e)\ . \tag{3.65}$$

For a *perfect gas* we have

$$e = c_v T, \quad p = \rho R T\,, \tag{3.66}$$

where T is the temperature, $R = c_p - c_v$ the gas constant, and c_v, c_p denote the specific heat at constant volume and constant pressure, respectively. Introducing the ratio of specific heats $\gamma = c_p/c_v$, for a perfect gas the thermodynamical equation of state reads

$$p = \frac{R\rho e}{c_v} = (\gamma-1)\rho e = (\gamma-1)\ (E - \frac{1}{2}\rho(u^2+v^2))\ . \tag{3.67}$$

For almost all aerodynamical problems (where the chemical composition of the air to be constant) we can assume the non-dimensional quantity γ to be constant ($\gamma = 1.4$).

3.3.4 Some general properties of the Euler equations

Here we introduce two important quantities: the total enthalpy H and the physical entropy s. We show that these quantities are constant along streamlines, under rather general circumstances. Further, we investigate what kind of discontinuities are possible in solutions of the steady Euler equations. It turns out that we can distinguish two types: shock waves and contact discontinuities.

The total enthalpy

The *enthalpy* h and the *total enthalpy* H are defined by

$$h = e + \frac{p}{\rho} , \quad H = h + \frac{1}{2}(u^2 + v^2) \ . \tag{3.68}$$

Using (3.61), we find

$$\rho H = E + p \ . \tag{3.69}$$

Hence, the energy equation (3.63) can also be written as

$$\frac{\partial}{\partial t}(\rho H) + \mathrm{div}\ (\rho H \mathbf{v}) = \frac{\partial p}{\partial t} \cdot \tag{3.70}$$

Combining this equation with the continuity equation (3.56) we find that (3.70) can be written as[10]

$$\frac{DH}{Dt} = \frac{\partial H}{\partial t} + u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y} = \frac{1}{\rho}\ \frac{\partial p}{\partial t} \ . \tag{3.72}$$

Thus, in the case of steady flow, where $\partial p/\partial t$ vanishes, the total enthalpy H is constant along streamlines.

The physical entropy

As we have combined the energy equation with the continuity equation, it is also possible to combine the momentum equations (3.60) with the continuity equation. Then we find

$$u_t + uu_x + vu_y = -\rho^{-1}\, p_x \ , \tag{3.73}$$

[10] In this equation we introduce the material derivative

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} \ . \tag{3.71}$$

The material derivative (or total derivative) expresses the rate of change of a property of a fluid particle.

$$v_t + uv_x + vv_y = -\rho^{-1}\, p_y\,. \tag{3.74}$$

By ordinary calculus we find immediately

$$uu_x + vu_y - \frac{1}{2}(u^2+v^2)_x = -\omega v, \tag{3.75}$$

$$uv_x + vv_y - \frac{1}{2}(u^2+v^2)_y = +\omega u, \tag{3.76}$$

with $\omega = v_x - u_y$ the vorticity of the flow. The *physical entropy* s is defined by[11]

$$T\,ds - pd(\rho^{-1}) = de\,, \tag{3.77}$$

so that

$$T\,ds + \rho^{-1}\,dp = dh\,, \tag{3.78}$$

and we get

$$\begin{array}{l} u_t + \frac{1}{2}(u^2+v^2)_x - \omega v = -\rho^{-1}\,p_x = T\,s_x - h_x\,,\\ v_t + \frac{1}{2}(u^2+v^2)_y + \omega u = -\rho^{-1}\,p_y = T\,s_y - h_y, \end{array} \tag{3.79}$$

$$\begin{array}{l} u_t - \omega v = T\,s_x - H_x\,,\\ v_t + \omega u = T\,s_y - H_y. \end{array} \tag{3.80}$$

So, for a steady state

$$T\nabla s = \nabla H - \omega \begin{pmatrix} v \\ -u \end{pmatrix}\,.$$

From this it follows that in the direction along $\mathbf{v}$ we have $T\mathbf{v}\cdot\nabla s = \mathbf{v}\cdot\nabla H = 0$ and it follows that also s is constant along streamlines,

$$\frac{Ds}{Dt} = 0, \tag{3.81}$$

as long as the flow is differentiable.

Hence, we have found the important result that the (physical) entropy of a fluid particle (in fact: the entropy in a small control volume following the fluid flow) remains constant. This result is only true for an inviscid non-conducting gas. If we take into account viscosity, a similar derivation shows that

$$\frac{Ds}{Dt} \geq 0\,. \tag{3.82}$$

Hence, in the case of viscous flow, the (physical) entropy of a fluid particle cannot decrease.

In the derivation of (3.81) we have assumed that the flow is smooth. Therefore, from (3.81) one cannot conclude that the (physical) entropy of

[11] For a perfect gas we can take $s = c_v \ln(p\rho^{\gamma})$. (exercise)

a fluid particle remains constant when the particle crosses a discontinuity. Indeed, there is an entropy jump at a shock wave. Because of the fact that in a viscous flow the physical entropy of a fluid particle cannot decrease and because all real fluids are, in fact, viscous, we know that the (physical) entropy of a flow, which passes a shock wave does not decrease. This is clearly related with the entropy condition. A true discontinuous solution of the inviscid Euler equations is the limit of a sequence of solutions of the Navier-Stokes equations (the equations that describe the conservation of mass, momentum and energy for viscous and heat-conducting flows) with vanishing viscosity.

3.4 Analysis of the Euler equations

In order later to derive numerical discretisation schemes for the Euler equations in two or three space dimensions, in this section we derive some of the properties of the Euler equations. Because generalisation to three space dimensions is straightforward as soon as the method for two dimensions is known, we restrict the treatment to d=2.

The Euler equations (3.64) can be written as

$$\frac{\partial}{\partial t}q + \frac{\partial}{\partial x}f(q) + \frac{\partial}{\partial y}g(q) = 0\,, \tag{3.83}$$

where

$$\begin{array}{ll} q & = (\rho, \rho u, \rho v, E)^T = (q_1, q_2, q_3, q_4)^T\,, \\ f(q) & = (\rho u, \rho u^2 + p, \rho u v, (E+p)u)^T \\ & = (q_2, q_2^2/q_1 + p, q_2 q_3/q_1, (q_4+p)q_2/q_1)^T\,, \\ g(q) & = (\rho v, \rho u v, \rho v^2 + p, (E+p)v)^T \\ & = (q_3, q_2 q_3/q_1, q_3^2/q_1 + p, (q_4+p)q_3/q_1)^T\,. \end{array} \tag{3.84}$$

and (see 3.67)

$$p = (\gamma - 1)\ (E - \frac{1}{2}\rho(u^2+v^2)) = (\gamma-1)\ (q_4 - (q_2^2+q_3^2)/(2q_1))\ . \tag{3.85}$$

For the construction of a numerical method the rotational invariance of the Euler equations is an important property.

Theorem 3.4.1 The Euler equations are *rotationally invariant* i.e.

$$\cos\phi\ f(q) + \sin\phi\ g(q) = T(\phi)^{-1} f(T(\phi)q) \tag{3.86}$$

for all $\phi \in \mathbb{R}$ and admissible states $q \in \mathbb{R}^4$, where $T(\phi)$ is the following rotation matrix:

$$T(\phi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\phi) & \sin(\phi) & 0 \\ 0 & -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.87}$$

Proof: The calculations to verify (3.86) are straightforward. □

In section 3.2.1 we have introduced the concept of hyperbolicity for a first-order system of quasi linear equations in the time and one space variable. Similarly, we define for more space variables:

Definition 3.4.2 System (3.83) of n equations is called *hyperbolic* (with respect to t) if there exist, for all $\phi \in \mathbb{R}$ and admissible states $q \in \mathbb{R}^n$, a real diagonal matrix $\mathcal{D}(q, \phi)$ and a non-singular matrix $\mathcal{R}(q, \phi)$ such that

$$\mathcal{A}(q,\phi)\mathcal{R}(q,\phi) = \mathcal{R}(q,\phi)\mathcal{D}(q,\phi)\,, \quad \forall q,\phi\,, \tag{3.88}$$

where

$$\mathcal{A}(q,\phi) = \cos\phi\,\frac{df(q)}{dq} + \sin\phi\,\frac{dg(q)}{dq} \equiv \cos\phi A(q) + \sin\phi B(q)\,. \tag{3.89}$$

Lemma 3.4.3 If there exists a real diagonal matrix $D(q)$ and a non-singular real matrix $R(q)$ such that

$$A(q)R(q) = R(q)D(q) \tag{3.90}$$

for all admissible states $q \in \mathbb{R}^4$, then the Euler equations (3.83) are hyperbolic.

Proof: By differentiating the rotational invariance relation (3.86) with respect to q, we have

$$\mathcal{A}(q,\phi) \;=\; \cos\phi A(q) + \sin\phi B(q) = T(\phi)^{-1} A(T(\phi)q)T(\phi)\,. \tag{3.91}$$

Hence,

$$\begin{aligned}\mathcal{A}(q,\phi) &= T(\phi)^{-1}\, R(T(\phi)q)\, D(T(\phi)q)\, R^{-1}(T(\phi)q)\, T(\phi)\\ &= (T(\phi)^{-1} R(T(\phi)q))\, D(T(\phi)q)\, (T(\phi)^{-1} R(T(\phi)q))^{-1}\,.\end{aligned} \tag{3.92}$$

By taking

$$\mathcal{R}(q,\phi) = T(\phi)^{-1} R(T(\phi)q),\;\; \mathcal{D}(q,\phi) = D(T(\phi)q) \tag{3.93}$$

we see that (3.88) is valid. □

Thus, the Euler equations are hyperbolic if the matrix $A(q) = \frac{df(q)}{dq}$ has 4 linearly independent eigenvectors. Using the relation

$$\nabla p = \left(\frac{\partial p}{\partial q_1}, \frac{\partial p}{\partial q_2}, \frac{\partial p}{\partial q_3}, \frac{\partial p}{\partial q_4}\right)^T = (\gamma - 1)\;(\frac{u^2+v^2}{2}, -u, -v, 1) \tag{3.94}$$

and the relations for total enthalpy H, eq. (3.69):

$$H = \frac{E+p}{\rho} = \frac{\gamma}{\gamma-1}\frac{p}{\rho} + \frac{u^2+v^2}{2}\,, \tag{3.95}$$

it is easily verified that

$$A(q) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -u^2 + \frac{\gamma-1}{2}(u^2+v^2) & (3-\gamma)u & -(\gamma-1)v & \gamma-1 \\ -uv & v & u & 0 \\ u(\frac{\gamma-1}{2}(u^2+v^2)-H) & H-(\gamma-1)u^2 & -(\gamma-1)uv & \gamma u \end{pmatrix}. \tag{3.96}$$

By the rotation invariance, a similar matrix is found for $B(q)$.

With the help of the matrix

$$M = \frac{\partial(\rho, \rho u, \rho v, E)}{\partial(\rho, u, v, p)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ u & \rho & 0 & 0 \\ v & 0 & \rho & 0 \\ \frac{1}{2}(u^2+v^2) & \rho u & \rho v & 1/(\gamma-1) \end{pmatrix} \tag{3.97}$$

for all $(n_1, n_2) = (\cos\phi, \sin\phi)$ we can write

$$M^{-1}(n_1 A + n_2 B)M = \begin{pmatrix} z & n_1\rho & n_2\rho & 0 \\ 0 & z & 0 & n_1/\rho \\ 0 & 0 & z & n_2/\rho \\ 0 & n_1\rho c^2 & n_2\rho c^2 & z \end{pmatrix}, \tag{3.98}$$

where $z = n_1 u + n_2 v$ is the flow velocity and

$$c = \sqrt{\gamma p/\rho} \tag{3.99}$$

is the *speed of sound*. From (3.98) it is directly seen that

$$\det(\mathcal{A}) = \det(n_1 A + n_2 B) = z^2(z^2 - c^2)\,. \tag{3.100}$$

The eigenvalues are $\lambda_1 = z - c$, $\lambda_2 = z$, $\lambda_3 = z$, $\lambda_4 = z + c$, and the corresponding eigenvectors of the matrix (3.98) are

$$\begin{pmatrix} \rho/c \\ -n_1 \\ -n_2 \\ \rho c \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -n_2 \\ n_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \rho/c \\ n_1 \\ n_2 \\ \rho c \end{pmatrix}. \tag{3.101}$$

The sign of $z^2 - c^2$ determines whether the flow is *subsonic* or *supersonic.* In all cases there are 4 independent eigenvectors.

Notice that $A(q)$ only depends on u, v and c. The eigenvalues and corresponding eigenvectors of $A(q)$ are

$$\lambda_1(q) = u - c,\ \lambda_2(q) = u,\ \lambda_3(q) = u,\ \lambda_4(q) = u + c \tag{3.102}$$

and

$$\begin{array}{rcl} R_1(q) & = & (1,\ u-c,\ v,\ H-cu)^T\ , \\ R_2(q) & = & \left(1,\ u,\ v,\ \frac{u^2+v^2}{2}\right)^T\ , \\ R_3(q) & = & (0,\ 0,\ 1,\ v)^T\ , \\ R_4(q) & = & (1,\ u+c,\ v,\ H+cu)^T\ . \end{array} \tag{3.103}$$

The eigenvectors are linearly independent and therefore we have found the following theorem.

Theorem 3.4.4 The Euler equations (3.83) are hyperbolic with respect to t.

In the following lemma we show that some of the eigenvalues are genuinely nonlinear, whereas others are linearly degenerate.

Lemma 3.4.5 The eigenvectors $R_k(q)$ given in (3.103) are genuinely nonlinear for $k = 1$ and 4, and linearly degenerate for $k = 2$ and 3.
Proof: Because

$$\nabla\lambda_2(q) = \nabla\lambda_3(q) = \nabla u = \frac{1}{\rho}(-u, 1, 0, 0) \tag{3.104}$$

it is immediately clear that $R_2(q)$ and $R_3(q)$ are linearly degenerate. Using definition (3.99) of the speed of sound c, we see that

$$\nabla c = \frac{\gamma}{2\rho c}\nabla p - \frac{c}{2\rho}\nabla\rho\,, \tag{3.105}$$

where $\nabla\rho = (1, 0, 0, 0)^T$ and ∇p is given by (3.94). It follows that

$$\begin{array}{ll} (\nabla\lambda_1(q), R_1(q)) & = (\nabla u, R_1(q)) - \frac{\gamma}{2\rho c}(\nabla p, R_1(q)) + \frac{c}{2\rho}(\nabla\rho, R_1(q)) \\ & = -\frac{c}{2\rho}(1+\gamma) < 0\ , \end{array} \tag{3.106}$$

and

$$\begin{array}{ll} (\nabla\lambda_4(q), R_4(q)) & = (\nabla u, R_4(q)) + \frac{\gamma}{2\rho c}(\nabla p, R_4(q)) - \frac{c}{2\rho}(\nabla\rho, R_4(q)) \\ & = +\frac{c}{2\rho}(1+\gamma) > 0\ . \end{array} \tag{3.107}$$

Hence, $R_1(q)$ and $R_4(q)$ are genuinely nonlinear. Thus, $R_1(q)$ and $R_4(q)$ correspond with simple waves while $R_2(q)$ and $R_3(q)$ correspond with contact discontinuities. $\square$

3.4.1 The Riemann problem for the 2D Euler equations

Here, we consider the Riemann problem for the Euler equations in one space dimension[12]:

$$\frac{\partial}{\partial t} q + \frac{\partial}{\partial x} f(q) = 0\,, \tag{3.108}$$

with

$$q(x,0) = \begin{cases} q_L\,, & x < 0\,, \\ q_R\,, & x > 0\,, \end{cases} \tag{3.109}$$

where q_L and q_R are given constant states, and q, $f(q)$ are defined in (3.84). We assume that there is a unique solution. We have already seen that $q(t,x) = \tilde{q}(x/t)$.

We shall seek pairs (q_L, q_R) for which the solution of the Riemann-problem (3.108) becomes simple, e.g. a simple rarefaction wave or contact discontinuity. In Lemma 3.4.5 we have seen which eigenvectors $R_k(q)$, $k = 1, \cdots, 4$, given in (3.103), are genuinely nonlinear or linearly degenerate.

Riemann invariants for the Euler equations

Riemann invariants are useful for the construction of a simple wave solution or a contact discontinuity solution of the Riemann problem (3.108), (3.109). In the following lemma, the Riemann invariants corresponding with the eigenvectors $R_k(q)$ are given.

Lemma 3.4.6 The Riemann invariants corresponding with the eigenvector $R_k(q)$, are given by ψ_k^m in the following list:

$$\begin{array}{lll} \psi_1^2 = u + \frac{2}{\gamma-1}c; & \psi_1^3 = v; & \psi_1^4 = p\rho^{-\gamma}; \\ \psi_4^2 = u - \frac{2}{\gamma-1}c; & \psi_4^3 = v; & \psi_4^1 = p\rho^{-\gamma}; \\ \psi_2^1 = u; & \psi_2^4 = p; & \\ \psi_3^1 = u; & \psi_3^4 = p. & \end{array} \tag{3.110}$$

We see that there exist 3 Riemann invariants for the eigenvalues λ_k, $k = 1, 4$ (as we did expect). The eigenvalue λ_k, $k = 2, 3$, is a double eigenvalue and has only two Riemann invariants.

Proof: The proof follows by a straightforward computation. $\square$

Thus, if the pairs (q_L, q_R) of the Riemann problem (3.108), (3.109) are such that

$$u_L + \frac{2}{\gamma - 1} c_L = u_R + \frac{2}{\gamma - 1} c_R \; ; \; v_R = v_L \; ; \; s_R = s_L \tag{3.111}$$

[12]The development of the solution of the Riemann problem will be essentially one-dimensional, but we consider still $q \in \mathbb{R}^4$.

where $s = p\rho^{-\gamma}$ denotes the (physical) entropy, and

$$u_L - c_L < u_R - c_R\,, \tag{3.112}$$

then a simple wave solution, corresponding with $R_1(q)$, exists, given by

$$\begin{array}{ll} q = q_L & \text{if } x/t < u_L - c_L\,, \\ \left.\begin{array}{l} u + \frac{2}{\gamma-1}c = u_L + \frac{2}{\gamma-1}c_L \\ v = v_L \\ s = s_L \\ u - c = x/t \end{array}\right\} & \text{if } u_L - c_L < x/t < u_R - c_R\,, \\ q = q_R & \text{if } x/t > u_R - c_R\,. \end{array} \tag{3.113}$$

Notice that $u - c = x/t$ follows from (3.36). This solution is also called the 1-rarefaction wave. If (q_L, q_R) are such that (3.111) holds, while $u_L - c_L > u_R - c_R$, then the solution given by (3.113) corresponds with a multi-valued solution. Then we speak of a compression wave. Although such a compression wave has no physical meaning, it will be shown in Chapter 4 that allowing compression waves, an approximate solution of the Riemann problem can be obtained which leads to an excellent upwind scheme for the Euler equations. This scheme was introduced by Osher in [40].

Similarly, if q_L and q_R are such that

$$u_L - \frac{2}{\gamma - 1}c_L = u_R - \frac{2}{\gamma - 1}c_R,\ v_L = v_R,\ s_R = s_L \tag{3.114}$$

and

$$u_L + c_L < u_R + c_R \tag{3.115}$$

then a simple wave solution, corresponding with $R_4(q)$ exists (the 4-rarefaction wave), given by

$$\begin{array}{ll} q = q_L & \text{if } x/t < u_L + c_L \\ \left.\begin{array}{l} u - \frac{2}{\gamma-1}c \;=\; u_L - \frac{2}{\gamma-1}c_L \\ v = v_L \\ s = s_L \\ u + c = x/t \end{array}\right\} & \text{if } u_L + c_L < x/t < u_R + c_R \\ q = q_R & \text{if } x/t > u_R + c_R \end{array} \tag{3.116}$$

If $u_L + c_L > u_R + c_R$, we have a multivalued 4-compression wave.

Finally, if q_L and q_R are such that $u_L = u_R$, $p_L = p_R$ then a contact discontinuity solution, corresponding with $R_2(q)$, $R_3(q)$ exists, given by

$$\left\{ \begin{array}{ll} q = q_L & \text{if } x/t < u_L = u_R\,, \\ q = q_R & \text{if } x/t > u_L = u_R\,. \end{array} \right. \tag{3.117}$$

To describe shock discontinuities in the flow, we have to rely on the Rankine-Hugoniot relations to determine the solution of the Riemann problem. In the following section we use these to study the case of steady solutions.

3.4.2 Discontinuous solutions of the steady Euler equations

Now we investigate under which conditions a flow field consists of two uniform flows, separated by a straight line l, can be a solution of the steady Euler equations. We denote by $q = (\rho, \rho u, \rho v, E)^T$ the state vector of the fluid. We choose the x-axis perpendicular to l and the y-axis along l. Hence, the question is reduced to: when is

$$q(x,y) = \begin{cases} q_L\,, & x < 0\,, \\ q_R\,, & x > 0\,, \end{cases} \tag{3.118}$$

a solution of the steady Euler equations. Because $q(x,y)$ is discontinuous we have to apply the Euler equations in integral form. We take a control volume with infinitesimal width but finite length along the discontinuity (see fig. 3.5). Application of the equations of continuity, momentum and

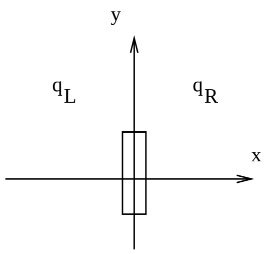


FIGURE 3.5. Discontinuous steady shock

energy gives:

$$\begin{aligned} \rho_L u_L &= \rho_R u_R\,, \\ \rho_L u_L^2 + p_L &= \rho_R u_R^2 + p_R\,, \\ \rho_L u_L v_L &= \rho_R u_R v_R\,, \\ \rho_L u_L H_L &= \rho_R u_R H_R\,. \end{aligned} \tag{3.119}$$

We consider two possibilities: either $u_L = 0$ or $u_L \neq 0$.

Suppose $u_L = 0$. Because density is always positive (3.119) is fulfilled if $u_R = 0$ and $p_L = p_R$. Hence, the flow field is composed of two uniformly constant flows with the same flow direction and pressure but with different densities and speeds. The discontinuity at the interface between the two flows is called a *contact discontinuity*[13] (or slip line). Notice that a fluid particle does not cross a contact discontinuity.

Suppose $u_L \neq 0$. Then $u_R \neq 0$, $v_L = v_R$ and $H_L = H_R$. Because $u_L = u_R$ implies $q_L = q_R$ i.e. a uniformly constant flow field, we may assume that

[13]Over this discontinuity the functions ψ_k^m, $m = 1, 4$, $k = 2, 3$, are invariant.

$u_L \neq u_R$. From the continuity equation in (3.119) it follows that u_L and u_R have the same sign. Therefore, without loss of generality, we suppose that $u_L > 0,\ u_R > 0$. This kind of discontinuity is called a shock wave. Notice that a fluid particle crosses a shock wave. From equation (3.119) we shall derive several jump relations.

From (3.119) we see that

$$u_L + \frac{p_L}{\rho_L u_L} = u_R + \frac{p_R}{\rho_R u_R} \tag{3.120}$$

or

$$u_L - u_R = \frac{c_R^2}{\gamma u_R} - \frac{c_L^2}{\gamma u_L} \ . \tag{3.121}$$

Because $H_L = H_R$, $v_L = v_R$, we also have (cf. eq. (3.95))

$$\frac{c_L^2}{\gamma - 1} + \frac{1}{2}u_L^2 = \frac{c_R^2}{\gamma - 1} + \frac{1}{2}u_R^2 \equiv \frac{1}{2}\frac{\gamma + 1}{\gamma - 1}c^{\star 2}$$

with $c^\star$ a constant. Combining these last two equations, one easily derives

$$u_L u_R = c^{\star 2} \ ,$$

which is known as the Prandtl relation.

Introducing $M = u/c,\ M^\star = u/c^\star$, the Prandtl relation becomes

$$M_R^\star M_L^\star = 1$$

and the relation

$$\frac{c^2}{\gamma - 1} + \frac{1}{2}u^2 = \frac{1}{2}\frac{\gamma + 1}{\gamma - 1}c^{\star 2}$$

results in a relation between M and $M^\star$

$$M^{\star 2} = \frac{(\gamma + 1)M^2}{(\gamma - 1)M^2 + 2} \ .$$

Now, the jump relations are easily derived. For instance

$$\frac{\rho_R}{\rho_L} = \frac{u_L}{u_R} = \frac{u_L^2}{u_L u_R} = M_L^{\star 2} = \frac{(\gamma + 1)M_L^2}{(\gamma - 1)M_L^2 + 2} \ . \tag{3.122}$$

The jump relation for the pressure becomes:

$$p_R - p_L = \rho_L u_L^2 - \rho_R u_R^2 = \rho_L u_L (u_L - u_R) = \rho_L u_L^2 \left(1 - \frac{u_R}{u_L}\right)$$

thus

$$\frac{p_R}{p_L} = 1 + \frac{2\gamma}{\gamma + 1}(M_L^2 - 1) \ . \tag{3.123}$$

The difference between the (physical) entropy in front of and behind the shock is

$$s_R - s_L = c_v \ \ln \ \left\{ \ \left(1 + \tfrac{2\gamma}{\gamma+1}(M_L^2 - 1)\right) \cdot \left(\tfrac{(\gamma-1)M_L^2+2}{(\gamma+1)M_L^2}\right)^\gamma \ \right\} \ . \qquad (3.124)$$

This is an important relation, because according to the entropy condition $s_R \geq s_L$. Using $s_R > s_L$ ($s_R = s_L \Rightarrow M_L = 1$, $u_R = u_L$), we can conclude that $M_L > 1$, hence, $M_L^\star > 1$, $M_R^\star < 1$ and $M_R < 1$. In the case of a normal shock ($v_L = v_R = 0$), M is the Mach number and the important conclusion can be drawn that in front of a normal shock the flow is supersonic ($M > 1$) and behind the shock the flow is subsonic ($M < 1$). Notice that this conclusion is a consequence of the entropy condition.

It is easily derived that, in the case of a weak shock i.e. $M_L = 1 + \epsilon$, $0 < \epsilon << 1$, we have

$$s_R - s_L = O(\epsilon^3) \ ; \ \frac{\Delta p}{p_L} = O(\epsilon) \ ; \ \frac{\Delta \rho}{\rho_L} = O(\epsilon) \ ; \ \frac{\Delta u}{u_L} = O(\epsilon) \qquad (3.125)$$

with $\Delta p = p_R - p_L$; $\Delta\rho = \rho_R - \rho_L$, $\Delta u = u_R - u_L$.

Thus, a small but finite change of pressure, causes corresponding first-order changes of density and velocity, but only a third-order change in entropy. Therefore, a weak shock produces a nearly isentropic change of state. This is an important result because the assumption that the fluid is isentropic (and isenthalpic) leads to a drastic simplification of the Euler equations.

3.5 Properties to preserve for a numerical scheme

Here we summarise some elementary properties of solutions of conservation laws that -if possible- we want to hold also for the numerical solutions that are obtained by the numerical schemes for

$$u_t + f(u)_x = 0 \,, \quad \Omega = \mathbb{R} \,.$$

These properties are:

1. The conservation property. For proper (a priori given) intervals $[a, b] \subset \mathbb{R}$ and $t_2 > t_1$, we require

$$\begin{array}{c} \int_a^b u(t_2, x)\, dx - \int_a^b u(t_1, x)\, dx \ = \\ \int_{t_1}^{t_2} f(u(t,a))\, dt - \int_{t_1}^{t_2} f(u(t,b))\, dt \,. \end{array} \qquad (3.126)$$

2. The maximum principle. For all $\Delta t \geq 0$,

$$\max_x u(t + \Delta t, x) \leq \max_x u(t, x) \,. \qquad (3.127)$$

3. Total variation diminishing:

$$TV(u(\cdot, t + \Delta t)) \leq TV(u(\cdot, t)), \tag{3.128}$$

   with $TV(w) = \int_{\mathbb{R}} |w_x(x)|\, dx$.

4. Stability in the L^1-norm:

$$\|u(\cdot, t + \Delta t)\|_1 \leq \|u(\cdot, t)\|_1, \tag{3.129}$$

   with $\|w\|_1 = \int_{\mathbb{R}} |w|\, dx$.

5. Decreasing mathematical entropy:

$$\int_{\mathbb{R}} U(t + \Delta t, x)\, dx \leq \int_{\mathbb{R}} U(t, x)\, dx\,. \tag{3.130}$$

   If the entropy function $U(\mathbf{v}) = \mathbf{v}\mathbf{v}^T$, for the entropy variables $\mathbf{v}$, then the last inequality implies the L^2-stability for the entropy variables.

4

Numerical Treatment of Hyperbolic Equations

4.1 Explicit difference schemes

4.1.1 Dissipation and dispersion error

In section 2.4.2 we have studied a two-parameter family of difference schemes for the equation

$$u_t + f(u)_x = 0\,, \tag{4.1}$$

where -for the analysis- we assumed $f(u)$ to be a linear function of u. The initial solution $u(0,x) = u_0(x)$, $x \in \mathbb{R}$, was given and $u(t,x)$, $t > 0$ had to be determined. We concluded that downwind discretisation leads to unstable schemes, whereas upwind is unconditionally stable if an appropriate implicit time discretisation is used, and the explicit upwind scheme is conditionally stable for small enough time steps. The explicit central discretisation is (unconditionally) unstable.

Now we will study more schemes and we will first consider the linear *system of equations*

$$\mathbf{u}_t + A\mathbf{u}_x = 0\,, \tag{4.2}$$

where $\mathbf{u} \in \mathbb{R}^n$. For convenience, we take $\Omega = \mathbb{R}$.

For the hyperbolic system we know that the eigenvalues of the matrix A are real. However, the eigenvalues can be positive as well as negative, and hence we have no unique meaning for the upwind or the downwind direction.

Before we generalise the idea of upwind schemes for systems of equations, we first study possibilities to modify the *explicit central scheme* in such a way that it can be used for our equations. The advantage of such a scheme is clear: no system of algebraic equations has to be solved in each time step and no knowledge about a preferred direction is necessary.

The original explicit central scheme is found in (2.29), with $w = 1/2$, $\theta = 0$:

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{2\Delta x}(f_{i-1}^n - f_{i+1}^n) = u_i^n + \frac{a\Delta t}{2\Delta x}(u_{i-1}^n - u_{i+1}^n)\,. \tag{4.3}$$

Knowing that each ℓ^2-function can be decomposed in elementary modes $e^{j\omega hi}$, with $j = \sqrt{-1}$, in section 2.4.2 we studied the behaviour of the scheme by setting $u_i^0 = e^{j\omega hi}$, and we found $u_i^n = \gamma^n e^{j\omega hi}$, for some complex amplification factor $\gamma = \gamma(\omega, \Delta x, \Delta t)$. For the differential equation, this procedure corresponds to taking $u_0(x) = e^{j\omega x}$, which allows the solution $u(t,x) = e^{-j\omega a t}e^{j\omega x}$. So, the ideal scheme has $\gamma = e^{-j\omega a \Delta t}$, this means that ideally $\text{abs}(\gamma) = 1$ and $\arg(\gamma) = -a\omega\Delta t$. This leads us to the definition of dissipation error and dispersion error of a numerical scheme.

Definition 4.1.1 The *dissipation error* of a linear, constant coefficient difference scheme for (4.1) is defined by $-\log(\,\text{abs}(\gamma(\omega,\Delta x,\Delta t)))$. The *dispersion error* of a linear, constant coefficient difference scheme for (4.1) is defined by $\arg(\gamma(\omega,\Delta x,\Delta t))/(-a\omega\Delta t) - 1$.

We further introduce some related definitions.

Definition 4.1.2 A one-step finite difference scheme for (4.2) with constant coefficients, and a specified $\Delta t/\Delta x$-ratio is called *stable* iff there is a constant K, (independent of ω, Δx or Δt) such that

$$\text{abs}(\gamma(\omega,\Delta x,\Delta t)) \;\le 1 + K\Delta t$$

for all ω, $0 < \Delta t \le \Delta t_0$.

This definition is easily understood if we consider the effect of integration over a fixed time interval $[0,T]$ with decreasing Δt. Let $T = n\Delta t$, then the total amplification factor at $t = T$, i.e. after n time steps, is bounded by

$$(1 + K\Delta t)^n = \left(1 + \frac{KT}{n}\right)^n \le e^{KT}\,.$$

I.e. the amplification over the interval $[0,T]$ is independent of Δt or Δx. If $\gamma(\omega,\Delta x,\Delta t)$ is independent of Δx or Δt (the $\Delta t/\Delta x$-ratio being fixed), then this condition can be replaced by

$$\text{abs}(\gamma(\omega)) \le 1.$$

Definition 4.1.3 A scheme is *dissipative of order* $2r$ iff there exists a positive K and C, independent of Δx or Δt, such that

$$\text{abs}(\gamma(\omega,\Delta x,\Delta t)) \le (1 - C\sin^{2r}(\omega\Delta x/2))\;\;(1 + K\Delta t)\;.$$

If $\gamma(\omega,\Delta x,\Delta t)$ is independent of Δt then this condition can be replaced by

$$|\gamma(\omega,\Delta x,\Delta t)| \;\le\; 1 - C\sin^{2r}(\omega\Delta x/2\,).$$

or, equivalently[1],

$$\gamma(\omega,\Delta x,\Delta t)^2 \;\le\; 1 - C\sin^{2r}(\omega\Delta x/2\,).$$

[1] C is a generic constant, that denotes different values at different places.

Example 4.1.4 *The central, explicit scheme*
For the scheme (4.3) we see that

$$\gamma(\omega, \Delta x, \Delta t) = 1 - j\,\frac{a\Delta t}{\Delta x}\sin(\omega\Delta x)\,,$$

i.e. the scheme is unstable.

A slight change can stabilise the scheme (4.3):

$$u_i^{n+1} = \frac{1}{2}(u_{i-1}^n + u_{i+1}^n) + \frac{a\Delta t}{2\Delta x}\,(u_{i-1}^n - u_{i+1}^n)\,. \tag{4.4}$$

This scheme is called the *Lax-Friedrichs scheme.* For this scheme we get

$$\gamma(\omega, \Delta x, \Delta t) = \cos(\omega\Delta x) - j\frac{a\Delta t}{\Delta x}\sin(\omega\Delta x),$$

or with $\theta = \omega\Delta x$, and $\lambda = a\Delta t/\Delta x$,

$$\gamma(\theta, \lambda) = \cos(\theta) + j\lambda\sin(\theta)\,,$$

i.e. the scheme is stable if

$$\lambda = \frac{a\Delta t}{\Delta x} \le 1.$$

This condition on the maximal allowed $\Delta t/\Delta x$-ratio implies that *the domain of dependence of the difference scheme should include the domain of dependence of the PDE.* This condition is usually called the *CFL condition* after a famous paper by Courant, Friedrichs and Lewy [4].

To compute the local discretisation error of the Lax-Friedrichs scheme, we proceed as in Section 2.4.2. We substitute a smooth solution $u(t, x)$ in the discrete operator (4.4), i.e. we set $u_m^k := u(t^k, x_m)$ where u is assumed a sufficiently smooth function. Expanding the result in a Taylor series about $u^\star = u(t^n, x_i)$ we obtain

$$\begin{aligned}
&\frac{1}{\Delta t}\left(u_i^{n+1} - \frac{1}{2}(u_{i-1}^n + u_{i+1}^n)\right) + \frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) \\
&= \frac{1}{\Delta t}\left((u^\star + \Delta t\, u_t^\star + \frac{(\Delta t)^2}{2}u_{tt}^\star + ...) - (u^\star + \frac{1}{2}(\Delta x)^2 u_{xx}^\star + ...)\right) \\
&\qquad + \frac{a}{2\Delta x}\left(2\Delta x\, u_x^\star + \frac{1}{3}\Delta x^3 u_{xxx}^\star + ...\right) = \\
&= u_t^\star + a u_x^\star + \frac{1}{2}(\Delta t\, u_{tt}^\star - \frac{\Delta x^2}{\Delta t}u_{xx}^\star) + \mathcal{O}(\Delta x^2 + \Delta t^2) \\
&= \frac{\Delta t}{2}\left(a^2 - (\Delta x/\Delta t)^2\right) u_{xx}^\star + \mathcal{O}(\Delta x^2 + \Delta t^2)\,.
\end{aligned}$$

If we take a constant mesh ratio $\Delta x/\Delta t$, which is natural for these hyperbolic equations, we see that the method is first order consistent (first order accurate).

4.1.2 The Lax-Wendroff and other schemes

It is possible to improve the accuracy and to get a one-step second order accurate explicit scheme based on three nodal points in the first time level. It is obtained by doing the space and the time discretisation at once. The initial state of the solution u_h^n, defined on the discrete grid is advanced over one time step. Using a second order approximation in time gives

$$u_h^{n+1} \approx u_h^n + \Delta t\,(u_h^n)_t + \frac{1}{2}(\Delta t)^2(u_h^n)_{tt}.$$

With the equation $u_t + au_x = 0$, and replacing the time derivatives by space differentials, we arrive at

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{\Delta x}\frac{1}{2}(u_{i+1}^n - u_{i-1}^n) + \left(\frac{a\Delta t}{\Delta x}\right)^2 \frac{1}{2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n)\,.$$

This is called the *Lax-Wendroff scheme.* The scheme can be analysed as the previous ones. It appears that the Lax-Wendroff scheme is the unique second-order space-centered discretisation on the three-point support for the linear one-dimensional convection equation (4.2). *There are many variants of the scheme that differ in the treatment of the nonlinearities or in the multi-dimensional aspects, but they all reduce to the same linear, one-dimensional form.*

Example 4.1.5 *Lax-Wendroff*
Computation of the amplification factor gives (with $\lambda = a\Delta t/\Delta x$)

$$\gamma = 1 - 2\lambda^2 \sin^2(\omega\Delta x/2) - j\,\lambda\,\sin(\omega\Delta x)$$

so that

$$\mathrm{abs}(\gamma) = \sqrt{1 - 4\lambda^2(1-\lambda)^2\sin^4(\omega\Delta x/2)}\,.$$

It follows that the stability condition reads $\lambda \le 1$ and that

$$\arg(\gamma) = \arctan\left(\frac{-\lambda\sin(\omega\Delta x)}{1 - 2\lambda^2\sin^2(\omega\Delta x/2)}\right)\,.$$

For small $\omega\Delta x$ this comes down to

$$\arg(\gamma) = \omega\,\Delta t\,a\;\{1 - \frac{1}{6}(\omega\Delta x)^2\,(1-3\lambda^2) + \mathcal{O}((\omega\Delta x)^4)\}$$

and the dispersion error is

$$-\frac{1}{6}(\omega\Delta x)^2(1-3\lambda^2) + \mathcal{O}((\omega\Delta x)^4)\,.$$

It is important to see that the Lax-Wendroff scheme satisfies the conservation property. It obeys the conservation law, Rule 3 in Section 2.5. This is immediately seen if we write the scheme as

$$u_i^{n+1} - u_i^n = -\frac{\Delta t}{\Delta x}(f^\star_{i+\frac{1}{2}} - f^\star_{i-\frac{1}{2}})\,, \tag{4.5}$$

with[2]

$$f^\star_{i+\frac{1}{2}} = \frac{f_i + f_{i+1}}{2} + \frac{1}{2}A_{i+\frac{1}{2}}[-\frac{\Delta t}{\Delta x}(f_{i+1} - f_i)]\,. \tag{4.6}$$

In view of possible generalisation to more dimensions, this last equation can also be written as

$$f^\star_{i+\frac{1}{2}} = f_{i+\frac{1}{2}} + \frac{1}{2}A_{i+\frac{1}{2}}\,\overline{\Delta u_{i+\frac{1}{2}}}\,, \tag{4.7}$$

where $f_{i+\frac{1}{2}} = (f_i + f_{i+1})/2$ and

$$\overline{\Delta u_{i+\frac{1}{2}}} = -\frac{\Delta t}{\Delta x}(f_{i+1} - f_i)\,.$$

All schemes in this section that we have seen for the scalar linear equation (4.1) can be used for the linear system (4.2) as well. Together with a few new schemes they are summarised in the following list.

i) The *central implicit scheme*

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{2\Delta x}A(\mathbf{u}_{i+1}^{n+1} - \mathbf{u}_{i-1}^{n+1})\,; \tag{4.8}$$

ii) The *Lax-Friedrichs scheme*

$$\mathbf{u}_i^{n+1} = \frac{1}{2}(\mathbf{u}_{i-1}^n + \mathbf{u}_{i+1}^n) - \frac{\Delta t}{2\Delta x}A(\mathbf{u}_{i+1}^n - \mathbf{u}_{i-1}^n)\,; \tag{4.9}$$

iii) The *Lax-Wendroff scheme*

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{2\Delta x}A(\mathbf{u}_{i+1}^n - \mathbf{u}_{i-1}^n) + \frac{\Delta t^2}{2\Delta x^2}A^2(\mathbf{u}_{i+1}^n - 2\mathbf{u}_i^n + \mathbf{u}_{i-1}^n)\,; \tag{4.10}$$

iv) The *Beam-Warming scheme* (similar to the Lax-Wendroff, but with a one-sided second order approximation of the first derivative)

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{2\Delta x}A(3\mathbf{u}_i^n - 4\mathbf{u}_{i-1}^n + \mathbf{u}_{i-2}^n) + \frac{\Delta t^2}{2\Delta x^2}A^2(\mathbf{u}_{i+1}^n - 2\mathbf{u}_i^n + \mathbf{u}_{i-1}^n)\,; \tag{4.11}$$

v) The *leapfrog scheme* (a central two-level method). In this scheme an approximate solution should be available on two time levels:

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^{n-1} - \frac{\Delta t}{2\Delta x}A(\mathbf{u}_{i+1}^n - \mathbf{u}_{i-1}^n)\,. \tag{4.12}$$

[2]Notice that for a linear problem, $f_i = Au_i$, it is irrelevant to know where the Jacobian $A = \partial f/\partial u$ is evaluated. The present notation, however, suggests direct generalisation to the nonlinear case.

4.1.3 Modified equations

To characterise the differences between the various schemes, we summarise the modified equations for a few schemes we have seen. For the Lax-Friedrichs scheme the modified equation is

$$\mathbf{u}_t + A\mathbf{u}_x = \frac{\Delta x^2}{2\Delta t}\left(I - \left(\frac{\Delta t}{\Delta x}A\right)^2\right)\mathbf{u}_{xx} + \mathcal{O}(\Delta x^2)\,. \tag{4.13}$$

In the scalar case this is $u_t + au_x = \nu u_{xx}$, with

$$\nu = \frac{\Delta x^2}{2\Delta t}(1-\lambda^2)\,.$$

This is a diffusive equation if the CFL-number $\lambda < 1$.

For the Lax-Wendroff scheme the modified equation is

$$\mathbf{u}_t + A\mathbf{u}_x = \frac{\Delta x^2}{6}A\left(\frac{\Delta t^2}{\Delta x^2}A^2 - I\right)\mathbf{u}_{xxx} + \mathcal{O}(\Delta x^3)\,. \tag{4.14}$$

In the scalar case this is $u_t + au_x = \mu u_{xxx}$, with

$$\mu = \frac{\Delta x^2}{6}(\lambda^2 - 1)\,.$$

This is the dispersive[3] equation and $\mu < 0$ if the CFL-number $\lambda < 1$. Notice that not only the phase velocity, but also the *group velocity*[4] plays a role.

---

[3]The linear *dispersion equation*

$$u_t + au_x = \mu u_{xxx}$$

can easily be analysed by Fourier analysis. Assume the solution to be of the form $u(t,x) = e^{j\omega x}e^{j\alpha(\omega)t}$, then we see $\alpha + a\omega = -\mu\omega^3$ or

$$u(t,x) = e^{j\omega x}e^{-j\omega t(a+\mu\omega^2)}\,,$$

i.e. this particular mode in the solution moves with a *phase velocity* $a + \mu\omega^2$. In other words, rapidly oscillating modes move faster in the positive direction if $\mu > 0$. The various modes in a general initial function $u_0(x)$ will be 'separated' as time proceeds.

[4]Let us consider the equation $u_t + au_x = 0$ with, for some given frequency ω_0, the initial data $u_0(x) = e^{j\omega_0 x}p(x)$, where $p(x)$ is a relatively smooth function (i.e. $\widehat{p}(\omega) \neq 0$ only for ω small). The solution of the problem for the convection equation (4.2) is

$$u(t,x) = e^{j\omega_0(x-at)}p(x-at)\,.$$

This function is called a *wave packet*, ω_0 is its *frequency* and $p(x)$ is its *envelope*. The dispersive equation $u_x + au_t = \mu u_{xxx}$ gives for this initial function the

For the Beam-Warming scheme the modified equation is

$$\mathbf{u}_t + A\mathbf{u}_x = \frac{\Delta x^2}{6} A \left(2I - \frac{3\Delta t}{\Delta x} A + \frac{\Delta t^2}{\Delta x^2} A^2 \right) \mathbf{u}_{xxx} \tag{4.15}$$

In the scalar case this is $u_t + au_x = \mu u_{xxx}$, with

$$\mu = \frac{\Delta x^2}{6}(2 - 3\lambda + \lambda^2)$$

This is a dispersive equation and $\mu > 0$ if the CFL-number $\lambda < 1$. In the numerical representation of a discontinuity, here the high frequency components will move ahead, in contrast with the Lax-Wendroff scheme, where they lag behind.

---

solution

$$\tilde{u}(t,x) = e^{j\omega_0(x-\alpha(\omega_0)t)} p(x - \gamma(\omega_0)t)\,,$$

where $\alpha(\omega_0)$ is the phase velocity, and $\gamma(\omega_0)$ is the *group velocity* of the wave packet:

$$\gamma(\omega) \approx \alpha(\omega) + \omega\alpha'(\omega)$$

This is easily seen as follows. Because $p(x) = u_0(x)e^{-j\omega_0 x}$ we have $\widehat{u}_0(\omega) = \widehat{p}(\omega - \omega_0)$ and the elementary mode $e^{j\omega x}$ in the initial function moves with its own velocity $\alpha(\omega)$ as $e^{j\omega(x-\alpha(\omega)t)}$, so that the time dependent solution can be written as

$$\begin{aligned}
u(t,x) &= \frac{1}{\sqrt{2\pi}} \int e^{j\omega(x-\alpha(\omega)t)} \widehat{u}_0(\omega) d\omega \\
&= \frac{1}{\sqrt{2\pi}} \int e^{j\omega(x-\alpha(\omega)t)} \widehat{p}(\omega - \omega_0) d\omega \\
&= \frac{1}{\sqrt{2\pi}} \int e^{j(\omega+\omega_0)(x-\alpha(\omega+\omega_0)t)} \widehat{p}(\omega) d\omega \\
&= e^{j\omega_0(x-\alpha(\omega_0)t)} \frac{1}{\sqrt{2\pi}} \int e^{-j\omega_0(x-\alpha(\omega_0)t)} e^{j(\omega+\omega_0)(x-\alpha(\omega+\omega_0)t)} \widehat{p}(\omega) d\omega \\
&\approx e^{j\omega_0(x-\alpha(\omega_0)t)} \frac{1}{\sqrt{2\pi}} \int e^{j\omega(x-\gamma(\omega_0)t)} \widehat{p}(\omega) d\omega \\
&= e^{j\omega_0(x-\alpha(\omega_0)t)} p(x - \gamma(\omega_0)t)\,.
\end{aligned}$$

Here

$$\alpha(\omega_0) + \omega_0\alpha'(\omega_0) = [\omega\alpha(\omega)]'(\omega_0) \approx \frac{(\omega_0 + \omega)\alpha(\omega_0 + \omega) - \omega_0\alpha(\omega_0)}{\omega} = \gamma(\omega_0)$$

because all relevant ω under the integral sign are small, as $p(x)$ is a smooth function.

4.2 Schemes in conservation form

4.2.1 Introduction

We have already seen that several discretisation schemes could be formulated in conservation form[5]:

$$u_i^{n+1} - u_i^n = -\frac{\Delta t}{\Delta x}(f^\star_{i+\frac{1}{2}} - f^\star_{i-\frac{1}{2}}) , \tag{4.16}$$

$$f^\star_{i+\frac{1}{2}} = f^\star(..., u_{i-1}, u_i, u_{i+1}, u_{i+2}, ...) .$$

For the Lax-Friedrichs scheme we have

$$f^\star(u_i, u_{i+1}) = \frac{\Delta x}{2\Delta t}(u_i - u_{i+1})\frac{1}{2}f_i^n + \frac{1}{2}f_{i+1}^n ,$$

where $f_i^n = f(u_i^n)$.

The form (4.16) is natural if we view u_i^n as an approximation to the average of $u(t,x)$ over a cell $\Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. Because we know that the solution $u(x,t)$ satisfies the integral form

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^{n+1})\,dx - \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^n)\,dx + \tag{4.17}$$

$$+\int_{t^n}^{t^{n+1}} f(u(x_{i-\frac{1}{2}},t))\,dt - \int_{t^n}^{t^{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt = 0$$

$$0 = \int (u_t + f_x)\,dV = \oint (n_t u + n_x f)\,d\sigma .$$

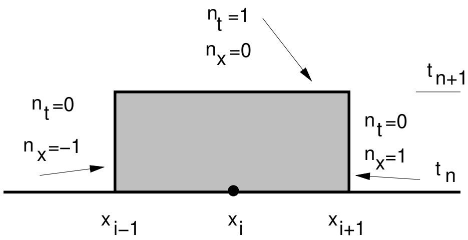


FIGURE 4.1. Conservation in the volume $[t^n, t^{n+1}] \times [x_{x-\frac{1}{2}}, x_{x+\frac{1}{2}}]$

[5]Because there is little difference between the treatment of the scalar equation and the system of equations, we do not continue the use of the (bold) vector notation in the case of systems.

If we associate u_i^n with $\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^n)\,dx/\Delta x$ we also obtain (4.16) by associating

$$f^\star_{i+\frac{1}{2}} = f^\star(u_i, u_{i+1}) \approx \int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}}, t))\,dt/\Delta t\,.$$

For consistency it is necessary that

$$f^\star(u,u) = f(u)\,,$$

and it is sufficient that for all v, w and u

$$|f^\star(v,w) - f(u)| \le K \max(|v-u|, |w-u|),$$

i.e. $f^\star$ is a Lipschitsch continuous function. The value of a consistent discretisation that satisfies the discrete conservation property lies in the following important theorem by Lax and Wendroff

Theorem 4.2.1 Let $u_m(x,t)$ denote a numerical approximation computed with Δx_m and Δt_m (so that $\Delta x_m \to 0$ and $\Delta t_m \to 0$ for $m \to \infty$) by means of a consistent and conservative method. Suppose that $u_m(x,t)$ converges to a function $u(x,t)$ as $m \to \infty$, then $u(x,t)$ is a weak solution of the conservation law.
Proof: See Lax-Wendroff [36] or [37, sect 12.4]. □

Notice that the above theorem doesn't say anything about the entropy condition or the existence of a limit function $u(t,x) = \lim_{m\to\infty} u_m(t,x)$!

4.2.2 Lax-Wendroff in more dimensions

In this section we make a few remarks about Lax-Wendroff type methods for the discretisation of hyperbolic systems of conservation laws in two space dimensions.

As we have seen before, there are three important ways to formulate such equations: the first is the differential form

$$u_t + f(u)_x + g(u)_y = 0 \quad \text{on} \quad \Omega \subset \mathbb{R}\,, \tag{4.18}$$

where u is the vector of conserved quantities and f and g describe the flux in the x- and the y-direction respectively; secondly we know the integral form, that is particularly useful if the solution is not differentiable:

$$\frac{\partial}{\partial t}\int_{\Omega_d} u\,dx\,dy + \oint_{\partial\Omega_d} (fn_x + gn_y)\,ds = 0 \quad \text{for all} \quad \Omega_d \subset \Omega, \tag{4.19}$$

and finally we have the quasi-linear form, that is obtained by differentiation of (4.18)

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial x} + \frac{\partial g}{\partial u}\cdot\frac{\partial u}{\partial y} = 0\,; \tag{4.20}$$

this equation is also written as

$$u_t + Au_x + Bu_y + = 0\,, \tag{4.21}$$

where $A = \partial f/\partial u$ and $B = \partial g/\partial u$.

For the discretisation of the hyperbolic system of conservation laws, (4.18) or its integral form (4.19), we have seen that two different approaches can be taken. First, the time and space discretisations can be made at once. This leads, for example, to discretisation schemes of Lax-Wendroff type. An initial state of the solution u_h^n, defined on a discrete grid, is advanced over one time step. Using a second order approximation in time gives,

$$u_h^{n+1} = u_h^n + \Delta t (u_h)_t + \frac{1}{2}(\Delta t)^2 (u_h)_{tt}.$$

With the equation (4.18) we arrive at

$$u_{ij}^{n+1} = u_{ij}^n - \Delta t (f_x + g_y)_{ij} + \frac{1}{2}(\Delta t)^2 \left\{ [A(f_x + g_y)]_x + [B(f_x + g_y)]_y \right\}_{ij} \,. \tag{4.22}$$

Using various difference approximations of the bracketed terms in the right-hand side, we may obtain different Lax-Wendroff type discretisations (cf [23, Chapter 17]).

One possible instance of (4.22) takes (4.5) as a starting point [23, Vol.II, p.256]:

$$u_{i,j}^{n+1} = u_{i,j}^n + \Delta u^{--}_{i+\frac12,j+\frac12} + \Delta u^{-+}_{i+\frac12,j-\frac12} + \Delta u^{+-}_{i-\frac12,j+\frac12} + \Delta u^{++}_{i-\frac12,j-\frac12}, \tag{4.23}$$

with $\Delta u^{--}_{i+\frac12,j+\frac12}$ etc. given by

$$\begin{aligned}
\Delta u^{\pm\pm}_{i+\frac12,j+\frac12} &= \frac{1}{4}\left(\overline{\Delta u}_{i+\frac12,j+\frac12} \pm \frac{\Delta t}{\Delta x}\overline{\Delta f}_{i+\frac12,j+\frac12} \pm \frac{\Delta t}{\Delta y}\overline{\Delta g}_{i+\frac12,j+\frac12} \right), \\
\overline{\Delta f}_{i+\frac12,j+\frac12} &= A_{i+\frac12,j+\frac12}\overline{\Delta u}_{i+\frac12,j+\frac12}\,, \\
\overline{\Delta g}_{i+\frac12,j+\frac12} &= B_{i+\frac12,j+\frac12}\overline{\Delta u}_{i+\frac12,j+\frac12}\,, \\
\overline{\Delta u}_{i+\frac12,j+\frac12} &= -\frac{\Delta t}{\Delta x}(f_{i+1,j+\frac12} - f_{i,j+\frac12}) - \frac{\Delta t}{\Delta y}(g_{i+\frac12,j+1} - g_{i+\frac12,j})\,,
\end{aligned}$$

where

$$f_{i,j+\frac12} = \frac12 f_{i,j} + \frac12 f_{i,j+1}\,, \text{ etc. .}$$

This formulation shows how the scheme can also be considered as the computation of a correction for a (dual) cell and the distribution of this correction over different neighbouring nodal points.

Another possible scheme is obtained by writing the Lax-Wendroff scheme as the *Richtmyer* two-step scheme. In one space dimension it reads

$$u_{i+\frac12}^{n+\frac12} = \frac12 (u_i^n + u_{i+1}^n) - \frac{\Delta t}{2\Delta x}(f_{i+1}^n - f_i^n)\,, \tag{4.24}$$

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}(f_{i+\frac{1}{2}}^{n+\frac{1}{2}} - f_{i-\frac{1}{2}}^{n+\frac{1}{2}})\,.$$

In fact this is a combination of a half step Lax-Friedrichs followed by a half step leapfrog scheme. In the linear case it combines to a single step Lax-Wendroff. We notice that in this form no explicit Jacobian $A = \partial f/\partial u$ is used. This Richtmyer form of the Lax-Wendroff scheme is easily extended to two dimensions:

$$\begin{aligned} u_{i+\frac{1}{2},j+\frac{1}{2}}^{n+\frac{1}{2}} = &\ \tfrac{1}{4}(u_{i,j}^n + u_{i+1,j}^n + u_{i,j+1}^n + u_{i+1,j+1}^n) \\ & -\tfrac{\Delta t}{2\Delta x}(f_{i+1,j+\frac{1}{2}}^n - f_{i,j+\frac{1}{2}}^n) - \tfrac{\Delta t}{2\Delta y}(g_{i+\frac{1}{2},j+1}^n - g_{i+\frac{1}{2},j}^n)\,, \\ u_{i,j}^{n+1} = u_{i,j}^n - &\frac{\Delta t}{\Delta x}(f_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - f_{i-\frac{1}{2},j}^{n+\frac{1}{2}}) - \frac{\Delta t}{\Delta y}(g_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} - g_{i,j-\frac{1}{2}}^{n+\frac{1}{2}}). \end{aligned} \tag{4.25}$$

where [23, Vol.II p.260]

$$f_{i+\frac{1}{2},j}^{n+\frac{1}{2}} = f\left(\frac{u_{i+\frac{1}{2},j+\frac{1}{2}}^{n+\frac{1}{2}} + u_{i+\frac{1}{2},j-\frac{1}{2}}^{n+\frac{1}{2}}}{2}\right)\ ,\ \text{etc.}\ .$$

Many other variants of the Lax-Wendroff scheme are possible, among which the McCormack scheme and Lerat's implicit schemes. However, we will not treat these schemes in these lectures. An comprehensive treatment is found in [23].

4.2.3 Steady state computations

A disagreeable property of most of these Lax-Wendroff discretisations is that, when by time stepping a stationary state is obtained such that $u_h^{n+1} = u_h^n$, the discrete stationary state still depends on Δt. This is caused by the fact that in general in (4.22) the discrete term with $(\Delta t)^2$ doesn't vanish.

Hence, a more satisfactory way to compute the steady state is by semi-discretisation. Here, first, a space discretisation is made for the partial differential equation (4.18) that we write in symbolic form

$$u_t = N(u)\,. \tag{4.26}$$

By the space semi-discretisation it is reduced to a large system of ODEs

$$\frac{\partial}{\partial t}u_h = N_h(u_h)\,. \tag{4.27}$$

Now, to find an approximation of the time-dependent solution of (4.27), any method for this integration of the system of ODEs can be used. The solution of the steady state can be computed by solving (4.27) until all transients have died out. Alternatively, we can avoid the ODEs (4.27) and solve the nonlinear system

$$N_h(u_h) = 0 \tag{4.28}$$

by other, more direct, means. In both cases (4.27) and (4.28) we find a steady approximation u_h, independent of the choice of a time step.

4.3 Upwind schemes

4.3.1 Introduction

We consider again the initial value problem for the scalar conservation law or a system of conservation laws, in *one* space dimension. We partition the domain $\Omega = [a,b]$ in N cells $\Omega_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, with $a = x_{\frac{1}{2}} < x_{1+\frac{1}{2}} < \ldots < x_{N+\frac{1}{2}} = b$. We consider equation (4.1) in its integral form (4.17)

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^{n+1})\,dx =$$

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^n)\,dx - \int_{t^n}^{t^{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt + \int_{t^n}^{t^{n+1}} f(u(x_{i-\frac{1}{2}},t))\,dt\,.$$

Identifying u_i^n with

$$u_i^n \sim \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t^n)\,dx \;/\; \Delta x \tag{4.29}$$

and $f_{i+\frac{1}{2}}^n$ with

$$f_{i+\frac{1}{2}}^n \sim f(u(x_{i+\frac{1}{2}},t))\,, \quad t\in[t^n,t^{n+1}] \tag{4.30}$$

we obtain the *explicit three-point schemes in conservation form*

$$u_i^{n+1} = u_i^n - p\,(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) \tag{4.31}$$

where $p = \Delta t/\Delta x$. To obtain a proper set of discrete equations we need some *numerical flux function* $f^\star(\cdot,\cdot)$ to relate the fluxes $f_{i+\frac{1}{2}}^n$ to the states u_i^n, $i = 1,\cdots,N$:

$$f_{i+\frac{1}{2}}^n = f^\star(u_i^n, u_{i+1}^n). \tag{4.32}$$

In this way a 3-point scheme is constructed. In fact, more complex schemes are constructed by taking

$$f_{i+\frac{1}{2}}^n = f^\star(u_{-k+1}^n,\cdots,u_i^n,u_{i+1}^n,\cdots,u_{i+k}^n).$$

We will return to such schemes later in Section 4.4. First we restrict ourselves to (4.32).

Obviously, in (4.31) the unknown u_i^n represents the approximation of the mean value of the unknown function $u(t,x)$ at time $t = t^n$ in the cell Ω_i. The function $f^\star(u,v)$, the *numerical flux function*, represents the flux

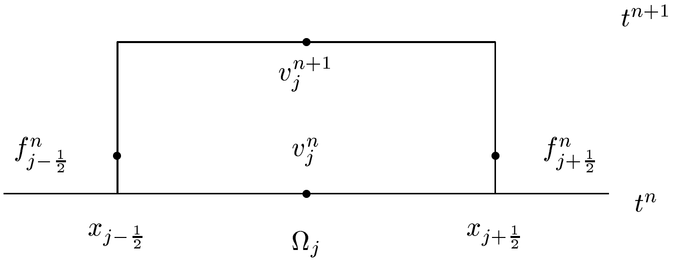


FIGURE 4.2. Fluxes and states

at the interface between two cells with the approximate solutions u and v respectively. If the solution in these cells were uniform this flux could be computed by solving the Riemann problem for these uniform states. This is what is done in the *Godunov scheme*. The solution $u(t^n, x)$ is assumed to be a piecewise constant function (constant on each Ω_i) and the flux $f_{i+\frac{1}{2}}^n$ is computed by solving the Riemann problem with $u_R = q_i$ and $u_L = q_{i+1}$. Then (4.31) determines the new approximation of $u(t,x)$ at $t = t^{n+1}$. In fact the numerical flux function can be seen as (some approximation of) a Riemann solver (it is called an *approximate Riemann solver*) and its behaviour determines the properties of the scheme. The numerical flux function is called to be *consistent* with the physical flux function $f(u)$ if it satisfies

$$f^\star(u,u) = f(u)\,.$$

4.3.2 The numerical entropy condition

Assuming that the system

$$u_t + f(u)_x = 0$$

has an entropy function $U(u)$, the difference scheme (4.31) is said to be *consistent with the entropy inequality* (3.51) if a following inequality exists

$$U_i^{n+1} \le U_i^n - p\,(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n)\,, \tag{4.33}$$

$$U_j^n = U(u_i^n)\,,$$
$$F_{i+\frac{1}{2}}^n = F(u_i^n, u_{i+1}^n)\,,$$

$F(u,v)$ is the *numerical entropy flux*, consistent with the entropy flux: $F(u,u) = F(u)$.

Theorem 4.3.1 Let the difference scheme (4.31) be consistent with the conservation law (4.1) and with the entropy inequality (3.51). Let u_i^n be the solution of (4.31), with the initial values u_i^0 given. For $t = t^n$ associate with the values u_i^n the piecewise constant approximate solution $v(t^n, x)$ so that $v(t^n, x) = u_i^n$ for $x \in \Omega_i$.

For a fixed $\Delta t/\Delta x = p$, let there exist the limit

$$\lim_{\Delta x \to 0} v(t, x) = u(t, x) \,.$$

Then the limit function $u(t, x)$ satisfies the weak form of the conservation law and the weak form of the entropy inequality.
Proof: See [17, Thm 1.1] □

Further, the total entropy $\sum_i U_i^{n+1} \leq \sum_i U_i^n$ is a decreasing function for increasing t^n.

The function $U(u)$ is determined modulo a linear function and hence it can be determined so that $U_i^n \geq 0$ for all i, n. From this, one may conclude the stability of the difference scheme: from (4.33) follows

$$\sum_i U_i^{n+1} \leq \sum_i U_i^n - pF^n_{\text{right}} + pF^n_{\text{left}} \,.$$

With the *Godunov scheme* the natural numerical approximation at $t = t^n$ is the piecewise constant function $v(t^n, x) = v_i^n$ on Ω_i. To calculate $v(t^{n+1}, \cdot)$, we first solve exactly the initial value problem for $t > t^n$

$$u_t + f(u)_x = 0$$

until $t = t^{n+1}$. Each discontinuity at a cell interface make a local Riemann problem. As long as $t^{n+1} - t^n < |\lambda_{\max}| \Delta x/2$, where $|\lambda_{\max}|$ is the largest signal speed, then no interaction takes place between neighbouring Riemann problems and $u(t, x)$ can be expressed in the initial conditions at $t = t^n$ from the most neighbouring cells only. With the Godunov scheme a new piecewise constant approximation is obtained at $t = t^{n+1}$ by taking new averages over all cells Ω_i at $t = t^{n+1}$.

The exact solution $u(t, x)$, $t \in [t^n, t^{n+1})$ satisfies the entropy condition

$$\begin{aligned}
&\int_{\Omega_i} U(u(t^{n+1}, x))\, dx \leq \\
&\int_{\Omega_i} U(u(t^n, x))\, dx - \Delta t F(u(t, x_{i+\frac{1}{2}})) + \Delta t F(u(t, x_{i-\frac{1}{2}}))
\end{aligned}$$

Since $U(u)$ is convex, we have $U(u)_{\text{mean}} \geq U(u_{\text{mean}})$, Jensens inequality, and hence

$$\frac{\int_{\Omega_i} U(u(t^{n+1}, x))\, dx}{\Delta x} \geq U\left(\frac{\int_{\Omega_i} u(t^{n+1}, x)\, dx}{\Delta x}\right) .$$

Combining the two inequalities we see that Godunov's scheme satisfies the entropy inequality (4.33).

4.3.3 The linear upwind scheme

From (4.31) we see that $f^n_{i+\frac{1}{2}}$ should represent an approximation of the flux at $x_{i+\frac{1}{2}}$ and that the numerical flux function should be a more or less reasonable recipe to determine this flux. It should be based on the state of the unknown function at the left and the right side of the interface between the cells Ω_i and Ω_{i+1}.

If we represent the numerical approximation of the state u by a piecewise constant function, as in the Godunov scheme, we can see the computation of the flux $f^n_{i+\frac{1}{2}} \approx f(u(t^n, x_{i+\frac{1}{2}}))$ as part of a Riemann problem: let for $t = t^n$ the approximate solution be piecewise constant on all cells, then for $t > t^n$, t sufficiently small, the neighbouring discontinuities will have no influence at $x = x_{i+\frac{1}{2}}$, and the flux at this interface will be determined by the solution of the Riemann problem (3.22) with $q_L = u^n_i$ and $q_R = u^n_{i+1}$.

If we consider the linear problem (4.2), then we found[6] in Example 3.2.5 that the flux at $x = x_{h+\frac{1}{2}}$ is given by

$$f_{i+\frac{1}{2}} = A^+ q_L + A^- q_R = A^+ u^n_i + A^- u^n_{i+1} \,.$$

For the piecewise constant approximation of $u(t,x)$ this flux is correct as long as $\Delta t < \Delta x/|\lambda_{\max}|$, with $\lambda_{\max}$ the largest eigenvalue of A.

This approach yields a natural numerical flux function for the linear problem (4.2), viz.

$$f^\star(u,v) = A^+ u + A^- v \,. \tag{4.34}$$

The resulting scheme is called the *Courant-Isaacson-Rees scheme.* (*CIR scheme*). It will appear to be first order accurate. We see immediately that the numerical flux is consistent:

$$f^\star(u,u) = A^+ u + A^- u = Au = f(u) \,.$$

The numerical scheme can now be written explicitly as

$$\begin{aligned} u^{n+1}_i &= u^n_i - p\,\{A^+(u^n_i - u^n_{i-1}) + A^-(u^n_{i+1} - u^n_i)\} \qquad (4.35)\\ &= \; pA^+ u^n_{i-1} + (I - p\,|A|\,)u^n_i - pA^- u^n_{i+1} \,. \end{aligned}$$

If the CFL-condition $p|\lambda_{\max}| \le 1$ is satisfied, all coefficients in the scheme are positive. It satisfies the Rule 5 as given at the end of Chapter 2. Such a scheme is called *monotone.* In the scalar case it is guaranteed that $\max|u^{n+1}_i| \le \max|u^n_i|$ and it is stable in the maximum norm.

Equation (4.35) can also be written as

$$\frac{1}{\Delta t}(u^{n+1}_i - u^n_i) + \frac{1}{\Delta x}\frac{1}{2}A(u^n_{i+1} - u^n_{i-1}) = \frac{1}{\Delta x}\frac{1}{2}|A|(u^n_{i+1} - 2u^n_i + u^n_{i-1}) \,. \tag{4.36}$$

[6]Notice that the eigenvalue decomposition $A = R\Lambda R^{-1}$ exists because we assume the system to be hyperbolic.

This is a second order approximation for the modified equation

$$u_t + Au_x = \frac{1}{2}\Delta x\,|A|\,(I - p|A|)\,u_{xx}\,. \tag{4.37}$$

(See also equation (2.40) with $\theta = 0$ and $w = 1$.) We see[7] that the scheme has a positive numerical diffusion for $p|\lambda_{\max}| < 1$. This is apparently a stability condition.

The amount of numerical dissipation is different for different components in the solution and vanishes for components related with an eigenvalue $\lambda = 0$. This effect may show up at a stationary discontinuity or rarefaction waves.

4.3.4 Nonlinear upwind schemes

We see that the Courant-Isaacson-Rees scheme is a generalisation for linear systems of the upwind scheme in Chapter 2 of hyperbolic equations. Now we want to generalise it for nonlinear problems. We will see that there are many different possibilities that all reduce to the same CIR scheme for the linear case.

Definition 4.3.2 A difference scheme in conservation form (4.31) is called an *upwind* (or *upstream*) scheme if:
(i) for small $\|u-v\|$ a linear approximation of $f(u,v)$ is given by A^+u+A^-v, i.e. (for some $u^\star$)

$$\begin{aligned} f(u,v) = \;& f(u^\star) + A^+(u^\star)(u-u^\star) + A^-(u^\star)(v-u^\star) \\ & + \mathcal{O}(\|(u-u^\star)\| + \|(v-u^\star)\|)\,; \end{aligned}$$

(ii) if all eigenvalues of $A(u) = \frac{df(u)}{du}$ are positive, then $f(u,v) = f(u)$, if they are all negative then $f(u,v) = f(v)$.

For a differentiable $f(u)$ we know

$$f\left(\frac{u+v}{2}\right) = \frac{f(u)+f(v)}{2} + \mathcal{O}(\|u-v\|),$$

and with $u^\star = (u+v)/2$ from (i) follows

$$f(u,v) = \frac{1}{2}(f(u)+f(v)) - \frac{1}{2}\left|A\left(\;\frac{u+v}{2}\;\right)\right|\,(v-u) + \mathcal{O}(\|u-v\|)\,. \tag{4.38}$$

We write the numerical flux as

$$f(u,v) = \frac{1}{2}f(u) + \frac{1}{2}f(v) - \frac{1}{2}d(u,v)\,. \tag{4.39}$$

---

[7] $\lambda_{\max}$ is the absolute largest eigenvalue of A; λ in equation (2.40) corresponds with $p\lambda_{\max}$.

Scheme	d(u,v)
van Leer	$\frac{1}{2}(\lvert A(v)\rvert + \lvert A(u)\rvert)\,(v-u)$
Huang-Vijayasundaran	$\left\lvert A\left(\frac{v+u}{2}\right)\right\rvert (v-u)$
Steger-Warming	$\lvert A(v)\rvert\, v - \lvert A(u)\rvert\, u$
Roe	$\lvert A(u,v)\rvert\,(v-u)$ where $A(u,v)$ is a real , normal matrix $A(u,v)\,(v-u) = f(v) - f(u)$ $A(u,u) = A(u)$
Osher	$\int_u^v \lvert A(w)\rvert dw$ various possible integration paths

TABLE 4.1. Various nonlinear upwind schemes

Because of consistency we always have

$$d(u,u) = 0\,, \tag{4.40}$$

and for an upwind scheme

$$d(u,v) = \;|A(u,v)|\;(v-u)\;+\;\mathcal{O}(\|u-v\|)\,, \tag{4.41}$$

where $|A(u,v)|$ is a matrix function of u and v that has non-negative eigenvalues and has the property $|A(u,u)| = |A(u)|$. The simplest form is $|A(u,v)| = |A(\frac{1}{2}(u+v))|$ or $|A(u,v)| = \frac{|A(u)|+|A(v)|}{2}$.

The term $d(u,v)$ can be seen as a contribution to the numerical dissipation in scheme (4.31):

$$\frac{u_i^{n+1} - v_i^n}{\Delta t} + \frac{1}{2\Delta x}(f(u_{i+1}^n) - f(u_{i-1}^n) = \tag{4.42}$$

$$= \frac{1}{2\Delta x}\left[d(u_{i+1}^n, u_i^n) + d(u_i^n, u_{i-1}^n)\right]\,.$$

Apparently, the various upwind schemes that are used for hyperbolic conservation laws, are characterised by their $d(u,v)$. We mention the more important ones in Table 4.1. It can be shown that the *Roe linearisation* $A(u,v)$ exists if the system of conservation laws has an entropy function [17, Thm.2.1].

An important property of the conservative upwind schemes is their ability to represent discontinuities (shocks). For time-dependent problems we saw in (4.37) that, for small enough Δt, the schemes show a positive numerical dissipation, which means that physically relevant solutions are found. If we want to compute steady shocks, the we have to be careful: for steady shocks holds

$$f(u_R) = f(u_L)$$

and

$$F(u_R) \leq F(u_L)\,.$$

In order *not* to smear the shock, we want little diffusion (cf. (4.42) and hence $d(u_R, u_L) = 0$ over a physical shock. I.e. we want for all relevant u and v

$$f(u) = f(v) \text{ and } F(v) \leq F(u) \;\Rightarrow\; d(u,v) = 0\,,$$

$$f(u) = f(v) \text{ and } F(v) > F(u) \;\Rightarrow\; d(u,v) \neq 0\,.$$

This would exclude unphysical shocks.

In Roe's scheme, when u and v correspond to a stationary discontinuity, then it follows $f(u) = f(v)$, and $v - u$ is a null-vector of $A(u,v)$. Consequently $d(u,v) = 0$, whether or not the entropy condition $F(u) > F(v)$ is satisfied. Therefore, Roe's upwind differencing scheme may admit non-physical solutions. There exist proposals for additional viscosity like terms to $d(u,v)$ that reject non-physical solutions and still have a good resolution of the physical solutions.

4.3.5 Flux splitting schemes

Flux splitting schemes are schemes for which the numerical flux f can be written as

$$f(u,v) = f^+(u) + f^-(v)\,. \tag{4.43}$$

For consistency is required

$$f(u) = f^+(u) + f^-(u)\,.$$

For an upwind scheme $f^+(u)$ should take care of the information from left and $f^-(u)$ for the information from right.

With the definition $f^a(u) = f^+(u) - f^-(v)$ we can identify (4.43) with (4.39) by

$$d(u,v) = f^a(v) - f^a(u)\,,$$

and in the linear case we have $f^a(w) = |A|w$.

The *Steger-Warming scheme* is clearly a flux splitting scheme with $f^a(w) = |A(w)|w$. For the Euler equations, that are homogeneous of degree one, we have

$$f(w) = A(w)w$$

and we may write

$$d(u,v) = \frac{|A(u)| + |A(v)|}{2}(v-u) + \frac{|A(u)| - |A(v)|}{2}(u+v)\,.$$

For values u and v with small $\|u-v\|$ but $\operatorname{sign}(\lambda_k(u)) \neq \operatorname{sign}(\lambda_k(v))$ for some k, the term $d(u,v)$ is not a smooth function. This may cause unwanted effects in the computed solutions near such transitions (e.g. at sonic points). These unwanted effects are sometimes repaired by replacing $|A(u)|$ by a $\chi(A(u))$, where $\chi(s)$ is some smooth function that approximates the function $|s|$.

A most popular flux splitting scheme is *Van Leer's flux splitting scheme.* The splitting is essentially based on mathematical requirements for the split flux ([60, 57]):

(i) the split fluxes are continuously differentiable. (Steger-Warming fluxes are not differentiable, which shows up in inaccurate representations of sonic and stagnation points.)

(ii) for subsonic flow the Jacobians of the split fluxes have a zero eigenvalue (which accounts for a sharp representation of steady shocks).

Relaxing the second constraint a little, Van Leer obtains for the Euler equations a one-parameter family of continously differentiable split fluxes that are different only in the energy-flux splitting [60]. This family includes the scheme proposed by Hänel [13, 12], which satisfies

$$f^{\pm}_{\text{energy}} = f^{\pm}_{\text{mass}}\, H\,,$$

where H is the specific total enthalpy.

Although much work has been done in the field of flux splitting schemes, now these schemes are generally found unsatisfactory, because contact discontinuities and slip surfaces are diffused by these schemes, even when they are stationary and aligned with the grid [60]. This is in contrast with flux difference splittings such as Roe's and Osher's.

Following Van Leer [60] this can be easily seen (e.g. from the transverse-momentum flux across the slip-layer as computed with flux vector splitting

$$f_{\text{momentum}} = \rho v u = (f^{+}_{\text{mass}} + f^{-}_{\text{mass}})\frac{u_B + u_T}{2} - (f^{+}_{\text{mass}} - f^{-}_{\text{mass}})\frac{u_B - u_T}{2}\,.$$

When the net mass flux $f_{\text{mass}} = f^{+}_{\text{mass}} + f^{-}_{\text{mass}}$ vanishes, the second term in the momentum flux still causes numerical diffusion.

4.3.6 Osher's scheme

In Osher's scheme we have

$$d(u,v) = \int_u^v |A(w)|\, dw = \int_0^1 |A(w(s))|\, \dot{w}(s)\, ds\,. \tag{4.44}$$

Here s represents the parametrisation of the integration path:

$$w(s) = \int_0^s \dot{w}(t)\, dt\,;\quad w(0) = u,\, w(1) = v\,.$$

The matrix $A(w)$ allows a decomposition

$$A(w) = R(w)\,\Lambda(w)\,R^{-1}(w)$$

with eigenvalues $\lambda_i(w)$ and eigenvectors $R_i(w)$.

The result of the integration is dependent on the choice of the integration path. In Osher's paper [40], the integration path is selected such that $\dot{w}(s) = R_i(w(s))$, i.e. the integration takes place in the direction of the eigenvectors. Now the result only depends on the number of subpaths (in general $= n$, the number of equations) and the order in which they are selected. For some ordering of the eigenvalues we get

$$\begin{aligned} d(u,v) &= \textstyle\sum_{k=1}^{n} \int_{s=(k-1)/n}^{s=k/n} |A(w(s))|\, R_k(w(s))\, ds \\ &= \textstyle\sum_{k=1}^{n} \int_{s=(k-1)/n}^{s=k/n} R_k(w(s))\ |\lambda_k(w(s))|\, ds \\ &= \textstyle\sum_{k=1}^{n}\ \operatorname{sign}(\lambda_k(w(s))) \int_{s=(k-1)/n}^{s=k/n} R_k(w(s))\, \lambda_k(w(s))\, ds \\ &= \textstyle\sum_{k=1}^{n}\ \operatorname{sign}(\lambda_k(w(s))) \int_{s=(k-1)/n}^{s=k/n} A(w(s))\, \dot{w}(s)\, ds \\ &= \textstyle\sum_{k=1}^{n}\ \operatorname{sign}(\lambda_k) \left[f(w_{k/n}) - f(w_{(k-1)/n} \right] . \end{aligned} \tag{4.45}$$

Here, we assumed that $\lambda_k(w(s))$ does not changes sign on the k-th subpath. We see that the sum concatenates if the eigenvalues λ_k do not change sign, contributions to the sum are found only as $f(u(s))|_{s=0}$, $f(u(s))|_{s=1}$, and $2f(u(s))$ for those s where $\lambda_k(u(s))$ changes sign.

For the evaluation of Osher's scheme the values $u_{k/n}$ are essential, and for their computation Riemann invariants (see Section 3.4.1) play an important role. It is easily seen that on the k-th subpath with Riemann invariants ψ_k^m the $(n-1)$ states $u_{k/n}$, $k = 1, ..., n-1$, are determined by he equations

$$\psi_k^m(u_{k/n}) = \psi_k^m(u_{(k-1)/n})\; k = 1, ..., n-1,\; m = 1, ..., n-1,\; m \neq k\,.$$

Boundary conditions

If we consider the equation $u_t + f(u)_x = 0$ on a finite interval $\Omega = (a,b) \subset \mathbb{R}$, then we need not only initial conditions $u(x,0) = u_0(x)$, but also boundary conditions for $x = a$ and $x = b$. Which boundary condition is needed is not immediately clear. For the scalar linear problem

$$u_t + \alpha u_x = 0$$

we need a boundary condition at $x = a$ if $\alpha > 0$, and a condition at $x = b$ if $\alpha < 0$.

For a linear hyperbolic system, we may decompose the matrix A and we can distinguish solution components that are in the span(R_k) for all eigenvectors R_k with positive λ_k, and those in the span(R_k) for all eigenvectors with negative λ_k. The first set of components should be determined by the left boundary conditions, the second by the right boundary conditions.

This means that for each $\lambda_k > 0$ a relation $B_k(u(a,t)) = 0$ is needed, and similarly for each $\lambda_k < 0$ a relation $B_k(u(b,t)) = 0$. These relations should not be degenerate in the sense that they lack the information needed; i.e. we should not have $\frac{dB_k(u)}{du} R_k = 0$. Thus, we see that for each entering characteristic we need a boundary condition.

If the system is not linear, we can formulate similar conditions by linearising the problem in a small (x,t)-neighbourhood of the solution.

4.3.7 Osher' scheme for the Euler equations

The computation of the numerical flux function is essentially as described in the sections 3.4.1 and 4.3.6. A few details, however, have to be taken into account.

For the genuinely nonlinear eigenvalues, (e.g. $k=1$ or $k=4$ in the 2D Euler equations) the eigenvalue $\lambda_k(u)$ can change sign (at most once) in the k-th subpath, $k = 1,4$. If this is the case we have $\lambda_k(u_{k/n})\lambda_k(u_{(k-1)/n}) < 0$. In order to compute

$$d(u,v) = \sum_{k=1}^{n} \int_{s=(k-1)/n}^{s=k/n} R_k(w(s))\, |\lambda_k(w(s))|\, ds\,, \tag{4.46}$$

we also have to determine $s^\star$ such that $\lambda(w(s^\star)) = 0$. I.e. $w(s^\star)$ is the state on the k-th sub-path where the eigenvalue $\lambda_k(w(s))$ changes sign. Along the k-th sub-path the Riemann invariants $\psi_k^m(w(s))$ are constant, for all $k \neq m$. This yields another nonlinear system hat can be solved. It appears that it contributes to the numerical flux with

$$\begin{aligned} &\textstyle\int_{s=(k-1)/n}^{s=k/n} R_k(w(s))\, |\lambda_k(w(s))|\, ds \\ =\;& \operatorname{sign}(\lambda(w_{k-1}) \textstyle\int_{s=(k-1)/n}^{s=s^\star} R_k(w(s))\, \lambda_k(w(s))\, ds \\ &\;+\; \operatorname{sign}(\lambda(w_k) \textstyle\int_{s=s^\star}^{s=k/n} R_k(w(s))\, \lambda_k(w(s))\, ds \\ =\;& \operatorname{sign}(\lambda(w_k) \left[f(w_{(k-1)/n}) - 2f(w_{s^\star}) + f(w_{k/n}) \right]\,. \end{aligned} \tag{4.47}$$

We notice that we we can distinguish different possible variants of the Osher scheme: the original O-scheme [40], where the subpaths, going from the left state u_L to the right state u_R run along the eigenvalues λ_k, with k=4,3,2,1; and another variant, the P-variant, where we run along the eigenvalues λ_k, with k=1,2,3,4.

4.3.8 Two-dimensional discretisation for the Euler equations

When the numerical flux function is known, it is simple to construct a first order accurate discretisation for the steady Euler equations in two dimensions. Making use of the rotational invariance of the Euler equations,

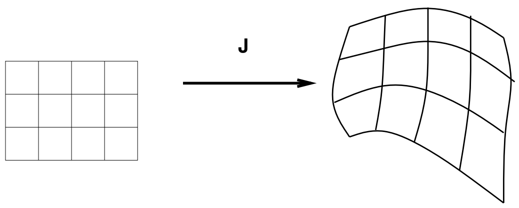


FIGURE 4.3. The smooth mapping J

essentially all the techniques for discretisation in one space dimension can also be used in the two (or three) dimensional case.

We write the Euler equations (3.83) in symbolic form

$$\frac{\partial}{\partial t} q + N(q) = 0\,. \tag{4.48}$$

We consider a bounded domain $\Omega \subset \mathbb{R}^2$ that can be partitioned in quadrilateral cells Ω_{ij}. I.e. we assume that there exists a smooth mapping $J : \widehat{\Omega} \to \Omega$ that maps a rectangular domain $\widehat{\Omega} \subset \mathbb{R}^2$ onto the domain Ω. The vertices of a regular partitioning of $\widehat{\Omega}$ in small rectangles $\widehat{\Omega_{ij}}$ are mapped by J to the vertices of quadrangles Ω_{ij}. The discrete equations

$$\frac{\partial}{\partial t} q_h + N_h(q_h) = 0\,. \tag{4.49}$$

are derived from the weak form of the conservation laws as follows. We denote the flux vector by $\mathbf{f}(q)$, with $\mathbf{f}(q) = (f(q), g(q))^T$, and for each Ω_{ij} we write

$$\begin{aligned}
\int_{\Omega_{ij}} \frac{\partial}{\partial t} q\, dx\, dy + \int_{\Omega_{ij}} \operatorname{div}\, \mathbf{f}(q)\, dx\, dy &= 0\,, \\
\frac{\partial}{\partial t} \int_{\Omega_{ij}} q\, dx\, dy + \oint_{\partial\Omega_{ij}} \mathbf{f}(q) \cdot \mathbf{n}\, ds &= 0\,, \\
V_{ij}\, \frac{\partial}{\partial t} q_{ij} + \sum_{k=N,E,S,W} \int_{\partial\Omega_{ij,k}} \mathbf{f}(q) \cdot \mathbf{n}\, ds &= 0\,.
\end{aligned}$$

where q_{ij} represents the mean value of $q(x,y)$ on Ω_{ij} (cf. equation (4.29), and V_{ij} is the volume (area) of cell Ω_{ij}; $\partial\Omega_{ij,k}$ is the k-th edge of cell Ω_{ij}. The discrete equations are now

$$V_{ij}\, \frac{\partial}{\partial t} q_{ij} + \sum_{k=N,E,S,W} l_{ij,k} f_{ij,k}(q_{ij}, q_{ij,k}) = 0\,, \tag{4.50}$$

$l_{ij,k}$ is the length of $\partial\Omega_{ij,k}$; $q_{ij,k}$ represents the state in the k-th neighbouring cell $\Omega_{ij,k}$; $f_{ij,k}(q_{ij}, q_{ij,k})$ is the numerical flux over the edge $\partial\Omega_{ij,k}$ between cell Ω_{ij} to cell $\Omega_{ij,k}$.

For the numerical flux $f_{ij,k}(q_{ij}, q_{ij,k})$ we can use any numerical flux function $f(u,v)$. By the rotational invariance of the Euler equations we can construct a local coordinate axis perpendicular to $\partial\Omega_{ij,k}$ that takes the role of the one-dimensional coordinate axis. (The fact that in more than one dimension a piecewise constant initial situation will cause interactions between waves from neighbouring edges at cell corners is simply neglected.)

With the operator T as defined in (3.87), we simply can write

$$f_{ij,k}(q_{ij}, q_{ij,k}) = T_{ij,k}^{-1} f(T_{ij,k} q_{ij}, T_{ij,k} q_{ij,k}) \,, \tag{4.51}$$

where $T_{ij,k}$ is the rotation matrix T associated with $\partial\Omega_{ij,k}$. Inserting (4.51) into (4.50) and dividing (4.50) by V_{ij}, we obtain the semi-discretisation (4.49). The discrete equations for the steady Euler equations are written

$$N_h(q_h) = 0 \,.$$

Several methods are available to solve this (large) system of nonlinear equations, see e.g. Section 4.5. The system is a nonlinear system with $4N$ unknown, where N is the number of cells in the grid. To see the structure of the system, we consider its linearisation $dN_h(q_h)/dq_h$. We notice that the matrix is block-diagonal with 5 diagonals of which each element is a 4×4-matrix. Each (block) row or column in the matrix corresponds with a particular cell and each (block) off-diagonal element corresponds with (the flux towards) a neighbour cell. For Osher's numerical flux function (e.g.) it can be shown that all off-diagonal elements have non-positive eigenvalues. From the conservation property it is seen that (for all interior cells) column sums vanish.

4.4 Second-order discretisations

This section is essentially taken from [49].

4.4.1 Introduction

The space discretisation of the steady Euler equations described in sections 4.3.8 is first-order accurate, as will be seen in section 4.4.2. It is highly desirable to improve the order of accuracy. In the smooth part of the flow field first-order accuracy is too low for practical purposes, while oblique (with respect to the mesh) shocks and contact discontinuities are very much smeared out because of the viscosity hidden in the first-order scheme. Therefore, we improve the order of accuracy. This will be done in such a way that we also steepen oblique discontinuities, taking care that we do not introduce over- or under-shoots.

The order of accuracy can be improved easily. The space discretisation is determined by the flux computation at the control volume boundaries.

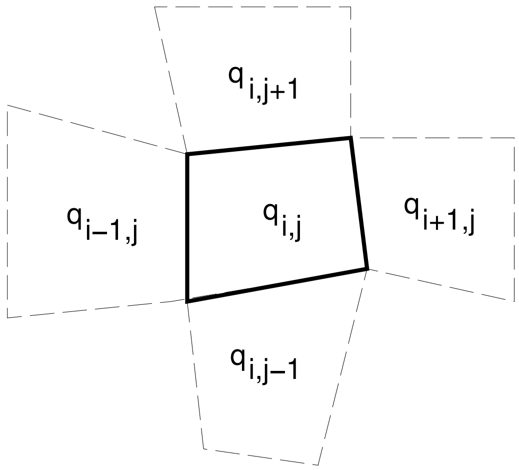


FIGURE 4.4. Finite volume $\Omega_{i,j}$

For the first-order scheme, the flux $f_{i+\frac{1}{2},j}$ at the interior control volume boundary $\partial\Omega_{i+\frac{1}{2},j}$ is (see 4.51):

$$f_{i+\frac{1}{2},j} = T^{-1}_{i+\frac{1}{2},j} f(T_{i+\frac{1}{2},j} q_{i,j},\ T_{i+\frac{1}{2},j} q_{i+1,j}), \tag{4.52}$$

where $f(u,v)$ is the numerical flux function. The order of accuracy is improved by taking

$$f_{i+\frac{1}{2},j} = T^{-1}_{i+\frac{1}{2},j} f(T_{i+\frac{1}{2},j} q^L_{i+\frac{1}{2},j},\ T_{i+\frac{1}{2},j} q^R_{i+\frac{1}{2},j}) \tag{4.53}$$

where $q^L_{i+\frac{1}{2},j}$ and $q^R_{i+\frac{1}{2},j}$ denote more accurate approximations of the fluid state at the left and the right side of the volume boundary $\partial\Omega_{i+\frac{1}{2},j}$ (see Figure 4.6). These values are obtained by a more accurate interpolation of the (mean) cell state values.

Second-order accuracy can be obtained by, for example, the κ-schemes introduced by Van Leer [59]:

$$q^L_{i+\frac{1}{2},j} = q_{i,j} + \frac{1+\kappa}{4}(q_{i+1,j} - q_{i,j}) + \frac{1-\kappa}{4}(q_{i,j} - q_{i-1,j})\ , \tag{4.54}$$

$$q^R_{i+\frac{1}{2},j} = q_{i+1,j} + \frac{1+\kappa}{4}(q_{i,j} - q_{i+1,j}) + \frac{1-\kappa}{4}(q_{i+1,j} - q_{i+2,j})\ ,$$

with $\kappa \in [-1,1]$. For $\kappa = -1,\ \kappa = 0,\ \kappa = \frac{1}{3}$ and $\kappa = 1$ we find respectively: the *fully one-sided upwind scheme*, the *Fromm scheme*, the *upwind biased scheme* (third-order accurate for 1D problems as we shall see in the next subsection) and the *central scheme*. A disadvantage of these κ-schemes is that near discontinuities spurious non-monotonicity (wiggles or over- and undershoots) appears [33].

The space discretisation corresponding with (4.53) is sometimes called the *projection-evolution approach* [59]. The projection stage is the computation of the states $q^L_{i+\frac{1}{2},j}$ and $q^R_{i+\frac{1}{2},j}$, whereas the evolution stage is the computation of the flux by the approximate Riemann solver $f(u,v)$, and the updating of the values q_{ij}.

In subsection 4.4.2 the accuracy of the space discretisation corresponding with (4.53) is considered. In subsection 4.4.3 a monotonicity concept is introduced and it is shown that it is possible to construct a monotone second-order accurate scheme. A solution of a monotone second-order scheme has the desired properties: second-order accuracy in the smooth part of the flow field and steepening of oblique discontinuities without introducing spurious non-monotonicity.

The analysis concerns a general nonlinear scalar hyperbolic conservation law. Without the complexity of hyperbolic systems, the analysis is more complete and more transparent. It appears that it is essential to compute the states $q^L_{i+\frac{1}{2},j}$, $q^R_{i+\frac{1}{2},j}$ by nonlinear interpolation. The nonlinear part of the interpolation is called a *limiter*. The results of the scalar analysis is generalised in a straightforward manner to the Euler equations.

4.4.2 Accuracy on a smooth mesh

Consider the Euler equations

$$\frac{\partial}{\partial t}q + \frac{\partial}{\partial x}f(q) + \frac{\partial}{\partial y}g(q) = 0\,, \tag{4.55}$$

on an open domain $\Omega \subset \mathbb{R}^2$, q, $f(q)$ and $g(q)$ are given in (3.84). The domain Ω is subdivided into equal rectangular volumes[8], which we denote by $\Omega_{i,j}$, $(i,j)\in\{1,\cdots,M,\ 1,\cdots,N\}$ so that

1. $\Omega = \bigcup_{i,j}\Omega_{i,j}$;
2. $\Omega_{i,j}$, $\Omega_{i\pm1,j}$, $\Omega_{i,j\pm1}$ are neighbouring volumes;
3. $(x_{i+\frac{1}{2},j+\frac{1}{2}},\ y_{i+\frac{1}{2},j+\frac{1}{2}})$ is the common vertex of the volumes $\Omega_{i,j}$, $\Omega_{i+1,j}$, $\Omega_{i,j+1}$ and $\Omega_{i+1,j+1}$.

We write (4.55) as

$$q_t + N(q) = 0\,, \tag{4.56}$$

[8]In [49] the same derivation is made for regular quadrilateral volumes and general domains that can be mapped onto a rectangle by means of a smooth mapping J.

$$
\begin{tikzcd}
X \arrow[r, "N"] \arrow[d, "R_h"'] & Y \arrow[d, "\overline{R_h}"] \\
X_h \arrow[r, "N_h^p"] & Y_h
\end{tikzcd}
$$

FIGURE 4.5. Spaces and operators

and the steady Euler equations as

$$N(q) = 0 \ . \tag{4.57}$$

Here $N : X \to Y$ is a nonlinear operator, $X \subset [L^2(\Omega)]^4$ is the space that contains all possible fluid states and $Y = [L^2(\Omega)]^4$ is the space of rates of change (for these states).

Define the finite dimensional vector spaces X_h and Y_h by

$$X_h = Y_h = \{ q_{i,j} \in \mathbb{R}^4 \mid i = 1 \cdots M, \ j = 1 \cdots N \} \ . \tag{4.58}$$

A relation between the spaces X and X_h, Y and Y_h is obtained by introducing $R_h : X \to X_h$ and $\overline{R_h} : Y \to Y_h$:

$$(R_h q)_{i,j} = (\overline{R_h} q)_{i,j} = \frac{1}{h^2} \int \int_{\Omega_{i,j}} q(\xi, \eta) \, d\xi d\eta \tag{4.59}$$

for any $q \in X = Y = [L^2(\Omega)]^4$. Thus $(R_h q)_{i,j}$ is the mean value of q in $\Omega_{i,j}$. We define the accuracy of a discretisation of (4.57) as follows.

Definition 4.4.1 A *p-th order accurate* discretisation of (4.57) is an associated problem:

$$N_h^p(q_h) = 0 \ , \tag{4.60}$$

where $N_h^p : X_h \to Y_h$ has the property that for all sufficiently smooth $q \in X$

$$(N_h^p R_h q)_{i,j} - (\overline{R_h} N q)_{i,j} = \mathcal{O}(h^p) \ . \tag{4.61}$$

The relation between the various spaces and mappings in the discretisation is summarised in Figure 4.5. Notice that

$$
\begin{aligned}
(\overline{R_h} N q)_{i,j} = \tfrac{1}{h^2} \{ \quad & \textstyle\int_{\partial\Omega_{i+\frac{1}{2},j}} f \ (q((i+\frac{1}{2})h, \eta)) \, d\eta \\
& - \textstyle\int_{\partial\Omega_{i-\frac{1}{2},j}} f \ (q((i-\frac{1}{2})h, \eta)) \, d\eta \\
& + \textstyle\int_{\partial\Omega_{i,j+\frac{1}{2}}} g \ (q(\xi, (j+\frac{1}{2})h)) \, d\xi \\
& - \textstyle\int_{\partial\Omega_{i,j-\frac{1}{2}}} g \ (q(\xi, (j-\frac{1}{2})h)) \, d\xi \ \} \ ,
\end{aligned} \tag{4.62}
$$

where $\partial\Omega_{i,j+\frac{1}{2}}$ etc., denote the boundaries of the control volume $\Omega_{i,j}$, defined by $\partial\Omega_{i,j+\frac{1}{2}} = \overline{\Omega_{i,j}} \cap \overline{\Omega_{i+1,j}}$ etc..

To perform our analysis, we introduce the operator $\tilde{N}_h : X \to Y_h$ defined by

$$(\tilde{N}_h q)_{i,j} = \frac{1}{h}\{f(\bar{q}_{i+\frac{1}{2},j}) - f(\bar{q}_{i-\frac{1}{2},j}) + g(\bar{q}_{i,j+\frac{1}{2}}) - g(\bar{q}_{i,j-\frac{1}{2}})\}\,, \qquad (4.63)$$

and

$$\bar{q}_{i+\frac{1}{2},j} = \frac{1}{h}\int_{\partial\Omega_{i+\frac{1}{2},j}} q((i+\frac{1}{2})h,\eta)d\eta\ . \qquad (4.64)$$

Thus $\bar{q}_{i+\frac{1}{2},j}$ is the mean value of $q(\xi,\eta)$ at the control volume boundary $\partial\Omega_{i+\frac{1}{2},j}$. In a similar way $\bar{q}_{i-\frac{1}{2},j}, \bar{q}_{i,j+\frac{1}{2}}$ and $\bar{q}_{i,j-\frac{1}{2}}$ are defined. It can be shown by elementary interpolation theory that

$$(\tilde{N}_h q)_{i,j} - (\overline{R_h} N q)_{i,j} = \mathcal{O}(h^2) \qquad (4.65)$$

for all sufficiently smooth $q \in X$.

In the more-dimensional case, the right-hand-side of (4.65) is not zero, due to the fact that at a cell boundary the mean flux differs from the flux computed in the mean state. However, for one-dimensional problems we have $\tilde{N}_h = \overline{R_h} N$.

We only consider discretisations in which states are interpolated[9] (see e.g. (4.54)). From (4.65) we conclude that the order of accuracy of such discretisations is at most two. (For 1D problems the order of accuracy can be higher.)

From (4.65) we see that if $N_h^p : X_h \to Y_h,\ p = 1,2$ is such that

$$(N_h^p R_h q)_{i,j} - (\tilde{N}_h q)_{i,j} = \mathcal{O}(h^p) \qquad (4.66)$$

then also

$$(N_h^p R_h q)_{i,j} - (\overline{R_h} N q)_{i,j} = \mathcal{O}(h^p)\ . \qquad (4.67)$$

This means that we may approximate $\tilde{N}_h$ instead of $\overline{R_h} N$. Therefore, let us look more carefully at $\tilde{N}_h$. Using the rotational invariance, we see from (4.63) that

$$\begin{aligned}(\tilde{N}_h q)_{i,j} = \tfrac{1}{h} \quad & \{T^{-1}_{i+\frac{1}{2},j} f(T_{i+\frac{1}{2},j}\bar{q}_{i+\frac{1}{2},j}) - T^{-1}_{i-\frac{1}{2},j} f(T_{i-\frac{1}{2},j}\bar{q}_{i-\frac{1}{2},j}) \\ & + T^{-1}_{i,j+\frac{1}{2}} f(T_{i,j+\frac{1}{2}}\bar{q}_{i,j+\frac{1}{2}}) - T^{-1}_{i,j-\frac{1}{2}} f(T_{i,j-\frac{1}{2}}\bar{q}_{i,j-\frac{1}{2}})\}\,,\end{aligned} \qquad (4.68)$$

where $T_{i+\frac{1}{2},j} = T(\phi_{i+\frac{1}{2},j})$, with $\phi_{i+\frac{1}{2},j}$ the angle of $\partial\Omega_{i+\frac{1}{2},j}$ with the x-axis, etc..

[9] It is also possible to interpolate fluxes (flux components) more accurately.

Notice that $l_{i+\frac{1}{2},j}$ is the length of the boundary $\partial\Omega_{i+\frac{1}{2},j}$ and $(\cos\phi_{i+\frac{1}{2},j}, \sin\phi_{i+\frac{1}{2},j})$ is the outward unit normal on $\partial\Omega_{i+\frac{1}{2},j}$, directed from $\Omega_{i,j}$ to $\Omega_{i+1,j}$ (under the assumption that the Jacobian of the mapping $J > 0$).

We define $N_h^p : X_h^p \to Y_h^p$ as follows

$$
\begin{array}{rll}
(N_h^p q_h)_{i,j} = \frac{1}{h}\{ & T^{-1}_{i+\frac{1}{2},j} & f(T_{i+\frac{1}{2},j} q^L_{i+\frac{1}{2},j},\ T_{i+\frac{1}{2},j} q^R_{i+\frac{1}{2},j}) \\
- & T^{-1}_{i-\frac{1}{2},j} & f(T_{i-\frac{1}{2},j} q^L_{i-\frac{1}{2},j},\ T_{i-\frac{1}{2},j} q^R_{i-\frac{1}{2},j}) \\
+ & T^{-1}_{i,j+\frac{1}{2}} & f(T_{i,j+\frac{1}{2}} q^L_{i,j+\frac{1}{2}},\ T_{i,j+\frac{1}{2}} q^R_{i,j+\frac{1}{2}}) \\
- & T^{-1}_{i,j-\frac{1}{2}} & f(T_{i,j-\frac{1}{2}} q^L_{i,j-\frac{1}{2}},\ T_{i,j-\frac{1}{2}} q^R_{i,j-\frac{1}{2}})\}\ ,
\end{array}
\tag{4.69}
$$

where $q^L_{i+\frac{1}{2},j}$, $q^R_{i+\frac{1}{2},j}$ etc. are obtained by interpolation of the states $\{q_{i,j}\} = q_h$ and $f(u,v)$ is the numerical flux function. Now we can establish the following theorem.

Theorem 4.4.2 Let $q \in X$ be sufficiently smooth. Define $q_h \in X_h$ by $q_h = R_h q$ and

$$
q_{i,j} = (R_h q)_{i,j} = \frac{1}{h^2} \int\int_{\Omega_{i,j}} q(\xi,\eta)\, d\xi d\eta\ . \tag{4.70}
$$

Define the mean values of q at the control volume boundaries $\bar{q}_{i+\frac{1}{2},j}$ by (4.64), and a similar formula for $\bar{q}_{i,j+\frac{1}{2}}$.

Define the states $q^L_{i+\frac{1}{2},j}$, $q^R_{i+\frac{1}{2},j}$ etc. by interpolation of the states $\{q_{i,j}\}$ such that

$$
\begin{array}{rcl}
q^R_{i+\frac{1}{2},j} - q^R_{i-\frac{1}{2},j} & = & \bar{q}_{i+\frac{1}{2},j} - \bar{q}_{i-\frac{1}{2},j} + \mathcal{O}(h^{p+1})\ , \\
q^L_{i+\frac{1}{2},j} - q^L_{i-\frac{1}{2},j} & = & \bar{q}_{i+\frac{1}{2},j} - \bar{q}_{i-\frac{1}{2},j} + \mathcal{O}(h^{p+1})\ , \\
q^R_{i+\frac{1}{2},j} - \bar{q}_{i+\frac{1}{2},j} & = & \mathcal{O}(h^p)\ , \\
q^L_{i+\frac{1}{2},j} - \bar{q}_{i+\frac{1}{2},j} & = & \mathcal{O}(h^p)\ ,
\end{array}
\tag{4.71}
$$

with $p = 1$ or $p = 2$, and let similar formulae hold for the j-direction. Then N_h^p defined by (4.69) is a pth-order accurate discretisation.
Proof: For the proof we refer to [49, Thm.2.5.2a] □

First, we consider two linear interpolation strategies:
(I1)

$$
q^L_{i+\frac{1}{2},j} = q_{i,j}\ ;\ q^R_{i+\frac{1}{2}} = q_{i+1,j}\ , \tag{4.72}
$$

and: (I2)

$$
q^L_{i+\frac{1}{2},j} = q_{i,j} + \frac{1+\kappa}{4}(q_{i+1,j} - q_{i,j}) + \frac{1-\kappa}{4}(q_{i,j} - q_{i-1,j})\ , \tag{4.73}
$$

$$
q^R_{i+\frac{1}{2},j} = q_{i+1,j} + \frac{1+\kappa}{4}(q_{i,j} - q_{i+1,j}) + \frac{1-\kappa}{4}(q_{i+1,j} - q_{i+2,j})\ . \tag{4.74}
$$

By using truncated Taylor series, we can show that interpolation (I1) is first-order accurate ($p = 1$) and interpolation (I2) is second-order accurate ($p = 2$). Further, the interpolation (I2) is third order accurate ($p = 3$) if $\kappa = 1/3$. For details we refer to [49, p.77,78]. Hence, we have found the following result.

Theorem 4.4.3 . Scheme (4.69) is first-order accurate for interpolation (I1) and second-order accurate for interpolation (I2). Furthermore, interpolation (I2) with $\kappa = \frac{1}{3}$ yields a third-order accurate scheme in 1-D.

The schemes using interpolation (I2) are called *Van Leer's κ-schemes*.

4.4.3 Monotonicity and second-order accuracy

Solutions of Van Leer's κ-schemes suffer from spurious oscillations (wiggles) in the neighbourhood of discontinuities. The purpose of this section is to study second-order schemes which do not exhibit such oscillations. These schemes is called monotone.

Several monotonicity concepts occur in the literature. They are all based on the scalar conservation law

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \ , \\ u(x,0) = u_0(x) \ , \end{cases} \tag{4.75}$$

which is assumed to be discretised in conservation form as (4.31):

$$u_i^{n+1} = u_i^n - p\,(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) \tag{4.76}$$

where $p = \frac{\Delta t}{\Delta x}$ and

$$f_{i+\frac{1}{2}}^n = f^\star(u_{i-l+1}^n, \cdots, u_{i+m}^n) \tag{4.77}$$

is the numerical flux function, satisfying the consistency condition

$$f^\star(u,\ u, \cdots, u) = f(u) \ . \tag{4.78}$$

The main reason for considering the scalar conservation law (4.75) is the following monotonicity property:

For any weak solution of (4.75) (see [35], [14]), we have

(i) No new local maximum or minimum can appear for $t > 0$.

(ii) The value of a local maximum is nonincreasing, that of a local minimum is nondecreasing.

and therefore

(iii) The total variation, defined by

$$TV[u(t)] := \sup \sum_i |u(x_{i+1}, t) - u(x_i, t)| \ , \tag{4.79}$$

is a nonincreasing function of time t.

The common, well known definition of a *monotone scheme* is due to Harten, Hyman and Lax [15]. They call the finite difference scheme (4.76) monotone if u_i^{n+1} is a monotone nondecreasing function of its $l + m + 1$ arguments $u_{i-l}^n, \cdots, u_{i+m}^n$. They were able to prove the following theorem.

Theorem 4.4.4 (cf.[15]). Assume that the finite difference scheme (4.76) is monotone in the sense of Harten, Hyman and Lax. Assume that, as Δt and Δx tend to zero, $p = \Delta t/\Delta x$ is constant, and u_i^n converges boundedly almost everywhere to some function $u(x, t)$. Then, according to the theorem of Lax and Wendroff [36] $u(x, t)$ is a weak solution of (4.75), moreover an entropy condition is satisfied for all discontinuities of u, i.e. $u(x, t)$ is the unique physically significant solution.

Unfortunately, a scheme which is monotone in the sense of Harten, Hyman and Lax is only first-order accurate [15]. To allow higher order of accuracy, Harten [14] introduced a weaker concept of monotonicity: scheme (4.76) is called *TVD* (or *total variation diminishing*) if

$$TV(u_h^{n+1}) \leq TV(u_h^n) \ , \tag{4.80}$$

where

$$TV(u_h^n) = TV(\{u_i^n\}) = \sum_{i=-\infty}^{\infty} |u_i^n - u_{i-1}^n| \ . \tag{4.81}$$

A grid function u_h is called *monotone* if for all i

$$\min(u_{i-1},\ u_{i+1}) \leq u_i \leq \max(u_{i-1},\ u_{i+1}) \ . \tag{4.82}$$

Following Harten [14], scheme (4.76) is called *monotonicity preserving* if monotonicity of u^{n+1} follows from monotonicity of u^n.

The relation between the above three properties is given by the following theorem.

Theorem 4.4.5 (cf. Harten [14]).

1. A scheme which is monotone in the sense of Harten, Hyman and Lax is TVD.

2. A TVD scheme is monotonicity preserving.

Remark:
It is well known (see [9] [14]) that a linear scheme

$$u_i^{n+1} = \sum_{k=-m}^{m} c_k u_{i+k}^n \tag{4.83}$$

is monotonicity preserving if and only if [10]

$$c_k \geq 0 \quad -m \leq k \leq m\,. \tag{4.84}$$

Hence, any *linear* monotonicity-preserving scheme is monotone in the sense of Harten, Hyman and Lax (and consequently first-order accurate). By Theorem 4.4.5, any *linear* TVD scheme will be only first-order accurate. Hence, only nonlinear schemes can be second-order and TVD. For the construction of such schemes, Harten's lemma [14] plays a fundamental role.

Lemma 4.4.6 (cf. Harten [14]). Consider a discretisation of (4.75) given by

$$u_i^{n+1} = u_i^n + p\,A_{i+\frac{1}{2}}^n(u_{i+1}^n - u_i^n) + p\,B_{i-\frac{1}{2}}^n(u_{i-1}^n - u_i^n)\ , \tag{4.85}$$

where $p = \Delta t/\Delta x$ and

$$A_{i+\frac{1}{2}}^n = A(\cdots,\ u_{i-1}^n,\ u_i^n,\ u_{i+1}^n, \cdots)\ . \tag{4.86}$$

---

[10] Proof: (i) If $c_k \geq 0$ for every k and $u_i^n - u_{i-1}^n \geq 0$ for all i, then

$$u_i^{n+1} - u_{i-1}^{n+1} = \sum_{k=-m}^{m} c_k u_{i+k}^n - \sum_{k=-m}^{m} c_k u_{i-1+k}^n$$

$$= \sum_{k=-m}^{m} c_k (u_{i+k}^n - u_{i+k-1}^n) \geq 0\ .$$

The case of nonincreasing u_h^n is handled similarly.
(ii) Conversely, supposing that $c_{k_0} < 0$, then for the particular function

$$u_i^n = \begin{cases} 1, & i \geq k_0, \\ 0, & i < 0, \end{cases}$$

we obtain

$$u_0^{n+1} - u_{-1}^{n+1} = \sum_{k=-m}^{m} c_k (u_k^n - u_{k-1}^n) = c_{k_0}(u_{k_0}^n - u_{k_0-1}^n) = c_{k_0} < 0$$

and monotonicity is not preserved.

$$B^n_{i-\frac{1}{2}} = B(\cdots,\ u^n_{i-1},\ u^n_i,\ u^n_{i+1},\cdots)\ . \tag{4.87}$$

If the coefficients $A^n_{i+\frac{1}{2}},\ B^n_{i+\frac{1}{2}}$ satisfy

$$A^n_{i+\frac{1}{2}} \geq 0,\quad B^n_{i+\frac{1}{2}} \geq 0,\quad 1 - p\,A^n_{i+\frac{1}{2}} - p\,B^n_{i+\frac{1}{2}} \geq 0\ , \tag{4.88}$$

then scheme (4.85) is TVD.
Proof: For the proof we refer to [49]. (Hint: consider $|u^{n+1}_i - u^{n+1}_{i-1}|$ and show $TV(u^{n+1}_h \leq TV(u^n_h.)$ □

The usefulness of this lemma for the construction of higher order TVD schemes is shown in [53]. However, Goodman and LeVeque [10] have obtained the following negative result: every conservative TVD scheme for scalar hyperbolic conservation laws in two space dimensions, is at most first-order accurate [10]. This result forces us to use a monotonicity concept weaker than TVD.

Consider the nonlinear scalar conservation law in two space dimensions:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) + \frac{\partial}{\partial y} g(u) = 0\ . \tag{4.89}$$

Assume that (4.89) is discretised on an equidistant mesh with mesh size h. Consider a discretisation of (4.89) given by

$$\begin{aligned} u^{n+1}_{i,j} = u^n_{i,j} &+ p\,A^n_{i+\frac{1}{2},j}(u^n_{i+1,j} - u^n_{i,j}) + p\,B^n_{i,j+\frac{1}{2}}(u^n_{i,j+1} - u^n_{i,j}) \\ &+ p\,C^n_{i-\frac{1}{2},j}(u^n_{i-1,j} - u^n_{i,j}) + p\,D^n_{i,j-\frac{1}{2}}(u_{i,j-1} - u^n_{i,j})\ , \end{aligned} \tag{4.90}$$

where $p = \Delta t/\Delta x$ and

$$\begin{aligned} A^n_{i+\frac{1}{2},j} &= A\ (\cdots, u^n_{i-1,j}, u^n_{i,j}, u_{i+1,j}, \cdots)\ , \\ B^n_{i,j+\frac{1}{2}} &= B\ (\cdots, u^n_{i,j-1}, u^n_{i,j}, u_{i,j+1}, \cdots)\ , \\ C^n_{i-\frac{1}{2},j} &= C\ (\cdots, u^n_{i-1,j}, u^n_{i,j}, u_{i+1,j}, \cdots)\ , \\ D^n_{i,j-\frac{1}{2}} &= D\ (\cdots, u^n_{i,j-1}, u^n_{i,j}, u_{i,j+1}, \cdots)\ . \end{aligned} \tag{4.91}$$

We introduce the following monotonicity concept:

Definition 4.4.7 . Scheme (4.90) is called *monotone* if

$$A^n_{i+\frac{1}{2},j} \geq 0\ ;\ B^n_{i,j+\frac{1}{2}} \geq 0\ ;\ C^n_{i-\frac{1}{2},j} \geq 0\ ;\ D^n_{i,j-\frac{1}{2}} \geq 0\ , \tag{4.92}$$

and if $A^n_{i+\frac{1}{2},j},\ B^n_{i,j+\frac{1}{2}},\ C^n_{i-\frac{1}{2},j},\ D^n_{i,j-\frac{1}{2}}$ are uniformly bounded, i.e. there exists a $K > 0$ such that for all (i,j)

$$A^n_{i+\frac{1}{2},j} \leq K\ ;\ B^n_{i,j+\frac{1}{2}} \leq K\ ;\ C^n_{i-\frac{1}{2},j} \leq K\ ;\ D^n_{i,j-\frac{1}{2}} \leq K \tag{4.93}$$

This monotonicity concept is not to be confused with what we called monotonicity in the sense of Harten, Hyman and Lax. Monotonicity is weaker than TVD in more than one space dimension. In two dimensions we define the total variation of u_h as

$$TV(u_h) = \sum_{i,j} \{|u_{i,j} - u_{i-1,j}| + |u_{i,j} - u_{i,j-1}|\} \ , \tag{4.94}$$

and u_h is called monotone if for all (i,j)

$$\begin{aligned} &\min(u_{i-1,j}, u_{i+1,j}, u_{i,j-1}, u_{i,j+1}) \le u_{i,j} \ , \\ &u_{i,j} \le \max(u_{i-1,j}, u_{i+1,j}, u_{i,j-1}, u_{i,j+1}) \ . \end{aligned} \tag{4.95}$$

Theorem 4.4.8 .

1. In one space dimension a monotone scheme is TVD.
2. In two space dimensions a monotone scheme is not necessarily TVD.

Proof: (1.) For scheme (4.85) we have

$$0 \le A^n_{i+\frac{1}{2},j} \le K \ ; \ 0 \le B^n_{i+\frac{1}{2},j} \le K \ . \tag{4.96}$$

For p sufficiently small ($p \le \frac{1}{2K}$) it is easily seen that (4.88) is fulfilled. Hence, (4.85) is TVD.
(2.) Consider the following grid function

$$u^n_{i,j} = \begin{cases} 1 & \text{for } (i,j) = (1,0) \ , \\ 0 & \text{for } (i,j) \ne (1,0) \ , \end{cases} \tag{4.97}$$

and let in (4.90)

$$A^n_{i+\frac{1}{2},j} = B^n_{i,j+\frac{1}{2}} = C^n_{i-\frac{1}{2},j} = D^n_{i,j-\frac{1}{2}} = 0 \quad \forall (i,j) \ne (0,0) \ , \tag{4.98}$$

$$A_{\frac{1}{2},0} = 1 \ ; \ A_{0,\frac{1}{2}} = 0 \ ; \ C_{-\frac{1}{2},0} = 0 \ ; \ D_{0,-\frac{1}{2}} = 0 \ . \tag{4.99}$$

Then

$$u^{n+1}_{i,j} = u^n_{i,j} \quad \forall (i,j) \ne (0,0) \ , \tag{4.100}$$

$$u^{n+1}_{0,0} = p \ . \tag{4.101}$$

Hence, $TV(u^n_h) = 4$, $TV(u^{n+1}_h) = 4 + 2p$. Because $p > 0$ we find $TV(u^{n+1}_h) > TV(u^n_h)$. Thus we have found a monotone scheme which is not TVD. □

Nevertheless, the monotonicity concept of definition (4.4.7) has some use, as shown by the following theorem.

Theorem 4.4.9 If scheme (4.90) is monotone then a steady state solution of (4.90) is monotone.

Proof: From (4.90) we see that for a steady state solution $\{u_{i,j}\}$ we have

$$u_{i,j} = \frac{A_{i+\frac{1}{2},j}u_{i+1,j} + B_{i,j+\frac{1}{2}}u_{i,j+1} + C_{i-\frac{1}{2},j}u_{i-1,j} + D_{i,j-\frac{1}{2}}u_{i,j-1}}{A_{i+\frac{1}{2},j} + B_{i,j+\frac{1}{2}} + C_{i-\frac{1}{2},j} + D_{i,j-\frac{1}{2}}} \tag{4.102}$$

which, due to the positivity of the coefficients, proves the theorem immediately. □

Thus, a monotone scheme guarantees that in a steady solution spurious oscillations do not occur. There is no contradiction between monotonicity and second-order accuracy (neither in one nor in more dimensions). This will be shown by constructing a second-order monotone scheme.

Consider (4.89) and suppose that the flux functions $f(u)$ and $g(u)$ can be split in a forward and a backward flux:

$$\begin{aligned} f(u) &= f^+(u) + f^-(u)\ , \\ g(u) &= g^+(u) + g^-(u)\ , \end{aligned} \tag{4.103}$$

where

$$\begin{cases} \frac{d}{du}f^+(u) \geq 0\ , & \frac{d}{du}f^-(u) \leq 0, \quad \forall u \in \mathbb{R}\ , \\ \frac{d}{du}g^+(u) \geq 0\ , & \frac{d}{du}g^-(u) \leq 0, \quad \forall u \in \mathbb{R}\ . \end{cases} \tag{4.104}$$

A finite volume discretisation on a regular square grid is given by

$$\begin{aligned} u^{n+1}_{i,j} = \ & u^n_{i,j} + \\ & + \tfrac{\Delta t}{h}[\{f^+(u^{L\ n}_{i-\frac{1}{2},j}) + f^-(u^{R\ n}_{i-\frac{1}{2},j})\} - \{f^+(u^{L\ n}_{i+\frac{1}{2},j}) + f^-(u^{R\ n}_{i+\frac{1}{2},j})\}] \\ & + \tfrac{\Delta t}{h}[\{g^+(u^{L\ n}_{i,j-\frac{1}{2}}) + g^-(u^{R\ n}_{i,j-\frac{1}{2}})\} - \{g^+(u^{L\ n}_{i,j+\frac{1}{2}}) + g^-(u^{R\ n}_{i,j+\frac{1}{2}})\}]\ , \end{aligned} \tag{4.105}$$

where

$$\begin{aligned} u^{L\ n}_{i+\frac{1}{2},j} &= u^n_{i,j} + \tfrac{1}{2}\psi(R^n_{i,j})(u^n_{i,j} - u^n_{i-1,j})\ , \\ u^{R\ n}_{i-\frac{1}{2},j} &= u^n_{i,j} + \tfrac{1}{2}\psi(\tfrac{1}{R^n_{i,j}})(u^n_{i,j} - u^n_{i+1,j})\ , \\ u^{L\ n}_{i,j+\frac{1}{2}} &= u^n_{i,j} + \tfrac{1}{2}\psi(S^n_{i,j})(u^n_{i,j} - u^n_{i,j-1})\ , \\ u^{R\ n}_{i,j-\frac{1}{2}} &= u^n_{i,j} + \tfrac{1}{2}\psi(\tfrac{1}{S^n_{i,j}})(u^n_{i,j} - u^n_{i,j+1})\ . \end{aligned} \tag{4.106}$$

and

$$R^n_{i,j} = \frac{u^n_{i+1,j} - u^n_{i,j}}{u^n_{i,j} - u^n_{i-1,j}}\ ;\ S^n_{i,j} = \frac{u^n_{i,j+1} - u^n_{i,j}}{u^n_{i,j} - u^n_{i,j-1}}\ ; \tag{4.107}$$

and $\psi : \mathbb{R} \to \mathbb{R}$ is a continuous function called the *limiter*. The value $u^n_{i,j}$ is a numerical approximation of the mean value of u in cell (i,j) at time $t = n\Delta t$, the values $u^{L\ n}_{i+\frac{1}{2},j}$, $u^{R\ n}_{i+\frac{1}{2},j}$ are approximations of

$$\frac{1}{h}\int_{(j+\frac{1}{2})h}^{(j-\frac{1}{2})h} u((i+\frac{1}{2})h, \eta, n\Delta t)\, d\eta$$

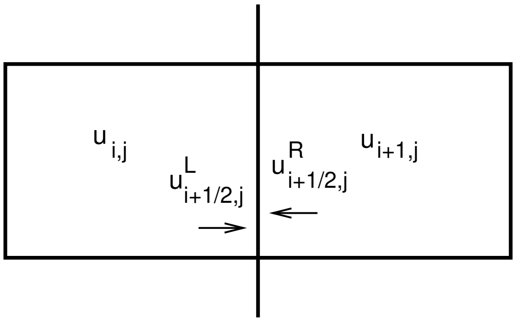


FIGURE 4.6. Location of the various variables in the space discretisation

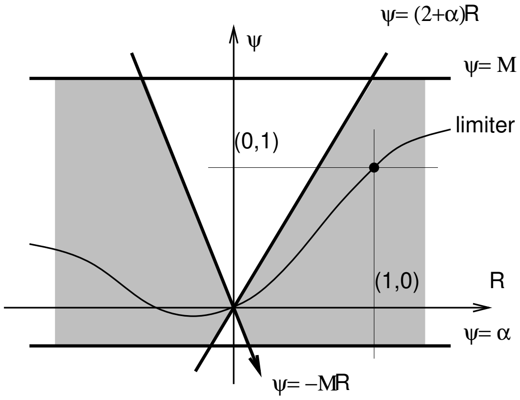


FIGURE 4.7. Monotonicity region for $\psi(R)$.

.

at the left and right side of the cell wall $(i+\frac{1}{2},j)$ respectively. See Figure 4.6. The limiter $\psi = \psi(R)$ is introduced in the discretisation in order to construct a monotone, spatially second-order scheme. The limiter is a function of the neighbouring gradients. Notice that $\psi \equiv 0$ corresponds to the first-order upwind scheme, while $\psi \equiv 1$ yields the fully one-sided second-order upwind scheme ($\kappa = -1$, see (4.54)). We already know that, for a monotone scheme, $\psi(R)$ has to be a nonlinear function.

We wish to show under what conditions scheme (4.105) is monotone. It can be easily seen that scheme (4.105) can be written as (4.90) and conditions for monotonicity are easily derived. The result is stated in the following theorem.

Theorem 4.4.10 [49, Thm.2.5.3f] If the limiter $\psi = \psi(R)$ has the properties that there exist constants $M\in(0,\infty), \alpha\in[-2,0]$ such that $\alpha \leq \psi(R) \leq M$; $-M \leq \frac{\psi(R)}{R} \leq 2+\alpha, \forall R\in\mathbb{R}$, then (4.105) is a monotone scheme.

This theorem implies that the graph of ψ is restricted to a particular area in the (R-ψ)-plane (cf. Figure 4.7). The requirements for the theorem imply that $\psi(0) = 0$.

Now, we investigate under which conditions scheme (4.105) is second-order accurate with respect to the space discretisation. Define

$$\begin{array}{rcl} \tilde{u}^L_{i+\frac{1}{2},j} & = & u_{i,j} + \frac{1}{2}(u_{i,j} - u_{i-1,j}) \ , \\ u^L_{i+\frac{1}{2},j} & = & u_{i,j} + \frac{1}{2}\psi(R_{i,j}) \ (u_{i,j} - u_{i-1,j}) \ , \\ \tilde{u}^R_{i-\frac{1}{2},j} & = & u_{i,j} + \frac{1}{2}(u_{i,j} - u_{i+1,j}) \ , \\ u^L_{i-\frac{1}{2},j} & = & u_{i,j} + \frac{1}{2}\psi(\frac{1}{R_{i,j}}) \ (u_{i,j} - u_{i-1,j}) \ , \end{array} \tag{4.108}$$

and similarly for $\tilde{u}^{L,R}_{i,j\pm\frac{1}{2}}$ and $u^{L,R}_{i,j\pm\frac{1}{2}}$. (Notice that $\tilde{u}$ corresponds with $\psi \equiv 1$, the fully one-sided second-order upwind scheme, $\kappa = -1$.)

Theorem 4.4.11 . If the limiter $\psi = \psi(R)$ is constructed such that

$$\begin{array}{rcl} u^R_{i+\frac{1}{2},j} - u^R_{i-\frac{1}{2},j} & = & \tilde{u}^R_{i+\frac{1}{2},j} - \tilde{u}^R_{i+\frac{1}{2},j} + \mathcal{O}(h^{p+1}) \ , \\ u^L_{i+\frac{1}{2},j} - u^L_{i-\frac{1}{2},j} & = & \tilde{u}^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i+\frac{1}{2},j} + \mathcal{O}(h^{p+1}) \ , \\ u^R_{i+\frac{1}{2},j} - \tilde{u}^R_{i+\frac{1}{2},j} & = & \mathcal{O}(h^p) \ , \\ u^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i+\frac{1}{2},j} & = & \mathcal{O}(h^p) \ , \end{array} \tag{4.109}$$

with $p = 1$ or $p = 2$, and where $u^R_{i+\frac{1}{2},j}$, $u^L_{i+\frac{1}{2},j}$, $\tilde{u}^R_{i+\frac{1}{2},j}$, $\tilde{u}^L_{i+\frac{1}{2},j}$ etc., are given by (4.108), then scheme (4.105) is p-order accurate with respect to the space discretisation.

Proof: This theorem is a direct consequence of Theorems 4.4.2 and 4.4.3. □

From (4.108) we see that

$$u^L_{i+\frac{1}{2},j} = \tilde{u}^L_{i+\frac{1}{2},j} + \frac{1}{2}(\psi(R_{i,j}) - 1) \ (u_{i,j} - u_{i-1,j}) \ . \tag{4.110}$$

Because

$$\psi(R) - 1 = \psi(1) - 1 + \psi'(1)(R-1) + \frac{1}{2}\psi''(\rho)(R-1)^2 \ , \tag{4.111}$$

with $\psi' = d\psi/dR$, $\psi'' = d^2\psi/dR^2$ and ρ between 1 and R and

$$(R_{i,j} - 1) \ (u_{i,j} - u_{i-1,j}) = u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \tag{4.112}$$

we find that the assumption

$$\psi(1) = 1 \tag{4.113}$$

leads to

$$\begin{aligned} u^L_{i+\frac{1}{2},j} = \quad & \tilde{u}^L_{i+\frac{1}{2},j} + \tfrac{1}{2}\psi'(1)(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \\ & + \tfrac{1}{4}\psi''(\rho)(R_{i,j} - 1)(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \ , \end{aligned} \tag{4.114}$$

with $\rho \in [1, R_{i,j}]$. Assume that $|\psi'|$ is uniformly bounded, then

$$u^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i+\frac{1}{2},j} = \mathcal{O}(h^2) \ , \tag{4.115}$$

$$u^L_{i+\frac{1}{2},j} - u^L_{i-\frac{1}{2},j} = \tilde{u}^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i-\frac{1}{2},j} + \mathcal{O}(h^2) \ . \tag{4.116}$$

Similar relations can be derived for $u^R_{i+\frac{1}{2},j} - \tilde{u}^R_{i+\frac{1}{2},j}$, $u^R_{i+\frac{1}{2},j} - \tilde{u}^R_{i-\frac{1}{2},j}$. Hence, scheme (4.105) is first-order accurate with respect to the space discretisation. Furthermore, by assuming that $\frac{\partial u}{\partial x} \neq 0$, we see that

$$R_{i,j} - 1 = \mathcal{O}(h) \ . \tag{4.117}$$

Then, assuming $|\psi''|$ is uniformly bounded, (4.114) leads to

$$u^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i+\frac{1}{2},j} = \mathcal{O}(h^2) \ , \tag{4.118}$$

$$u^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i-\frac{1}{2},j} = \tilde{u}^L_{i+\frac{1}{2},j} - \tilde{u}^L_{i-\frac{1}{2},j} + \mathcal{O}(h^3) \ , \tag{4.119}$$

and we see that the scheme is even second-order accurate in space. Thus we have got the following theorem.

Theorem 4.4.12 .

1. If $\psi(1) = 1$ and $|\psi'|$ is uniformly bounded, then scheme (4.105) is first-order accurate in space.

2. Assume furthermore that $|\psi''|$ is uniformly bounded. Away from points where $\frac{\partial u}{\partial x} = 0$ or $\frac{\partial u}{\partial y} = 0$ the space discretisation is second-order accurate.

Notice that scheme (4.105) is linear when $f(u)$ and $g(u)$ are linear and $\psi(R) = a + bR$; $a, b \in \mathbb{R}$. However, note that $\psi(R) = a + bR$ does not satisfy the requirements of Theorems 4.4.10 and 4.4.12, as we should expect.

Examples of limiters combining the property of second-order accuracy and monotonicity are:

Example 4.4.13 *The Van Leer limiter*
The Van Leer limiter [53] [55] [56]. is defined by:

$$\psi_{VL}(R) = \frac{R + |R|}{|R| + 1} \ . \tag{4.120}$$

By taking $M = 2$ and $\alpha = 0$ it is easily seen that this limiter satisfies the monotonicity restriction of Theorem 4.4.10. Because $\psi_{VL}(1) = 1$, second-order accuracy is obtained.

Example 4.4.14 *The Van Albada limiter*
The Van Albada limiter [1]. is defined by:

$$\psi_{VA}(R) = \frac{R^2 + R}{R^2 + 1} \ . \tag{4.121}$$

By taking $M = 2$ and $\alpha = -\frac{1}{2}$ it is easily seen that this limiter combines monotonicity with second-order accuracy. Notice that $\psi_{VA} \in C^{\infty}(\mathbb{R})$. Another advantage of this limiter is that $\lim_{R\to\pm\infty} \psi_{VA}(R) = 1$. This implies that at discontinuities the scheme renders the fully one-sided upwind scheme ($\kappa = -1$), which is a natural scheme at discontinuities.

For a review of other limiters see [53]. However, consulting [53], notice that a limiter $\phi(r)$ in [53] is related to $\psi(R)$ by $R = 1/r$; $\psi(R) = R\phi(r)$. A limiter $\phi(r)$ in [53] is only identical with $\psi(R)$ if $\psi(R)/R = \psi(1/R)$.

Remark:
It has been observed [56] that second-order accuracy can be achieved by assuming a linear distribution in each cell, rather than the uniform distribution that is associated with first-order schemes. In a cell, a linear distribution in the x direction is achieved if

$$u^L_{i+\frac{1}{2},j} - u_{i,j} = u_{i,j} - u^R_{i-\frac{1}{2},j} \ , \tag{4.122}$$

and similarly in the y-direction. Using (4.108), this means

$$\psi(R_{i,j})(u_{i,j} - u_{i-1,j}) = \psi(\frac{1}{R_{i,j}})(u_{i+1,j} - u_{i,j}) \ , \tag{4.123}$$

or, equivalently

$$\psi(\frac{1}{R_{i,j}}) = \psi(R_{i,j})R_{i,j} \ . \tag{4.124}$$

So, if a limiter satisfies

$$\psi(\frac{1}{R}) = \frac{\psi(R)}{R} \quad \forall R \in \mathbb{R} \tag{4.125}$$

we can speak of linear distributions in each cell. It can be verified that both ψ_{VL} and ψ_{VA} possess this property.

If a limiter satisfies (4.125) we have the following theorem:

Theorem 4.4.15 If a limiter $\psi = \psi(R)$ has the property $\psi(\frac{1}{R}) = \frac{\psi(R)}{R}$, and if $\psi_{\max} - \psi_{\min} \leq 2$, where

$$\psi_{\max} = \max_{R\in\mathbb{R}}(\psi(R)); \ \psi_{\min} = \min_{R\in\mathbb{R}}(\psi(R)), \tag{4.126}$$

then scheme (4.105) is monotone.
Proof: See [49, Thm.2.5.3i] □

Remark:
The κ-schemes (4.54) can be written as (4.108) by taking

$$\psi(R) = \frac{1-\kappa}{2} + \frac{1+\kappa}{2} R \ . \tag{4.127}$$

The κ-schemes correspond with a linear limiter and, therefore, are not monotone. A class of limiters which combines monotonicity and second-order accuracy is derived from the κ-schemes by

$$\psi_\kappa(R) = \left(\frac{1-\kappa}{2} + \frac{1+\kappa}{2} R \right) \ \phi(R) \ , \tag{4.128}$$

where

$$\phi(R) = \frac{2R}{R^2+1} \ . \tag{4.129}$$

Notice that $\phi(1) = 1$, $\phi'(1) = 0$ and therefore the ψ_κ limiters resemble the κ-schemes in the smooth part of the flow field, where $R \approx 1$. With respect to accuracy, the $\psi_{1/3}$ limiter is the best. Notice that $\psi_0(R)$ is the Van Albada limiter.

The use of limiters in the second-order space discretisation of the steady Euler equations is straightforward. Let $q^{L\ (k)}_{i+\frac{1}{2},j}$ and $q^{R\ (k)}_{i+\frac{1}{2},j}$ be the $k-$th component ($k = 1, 2, 3, 4$) of $q^L_{i+\frac{1}{2},j}$ and $q^L_{i+\frac{1}{2},j}$ (see (4.53)). Then we take

$$q^{L\ (k)}_{i+\frac{1}{2},j} = q^{(k)}_{i,j} + \frac{1}{2} \psi_\kappa (R^{(k)}_{i,j}) \ (q^{(k)}_{i,j} - q^{(k)}_{i-1,j}) \ , \tag{4.130}$$

$$q^{R\ (k)}_{i+\frac{1}{2},j} = q^{(k)}_{i+1,j} + \frac{1}{2} \psi_\kappa \left(\frac{1}{R^{(k)}_{i+1,j}}\right) \ (q^{(k)}_{i+1,j} - q^{(k)}_{i+2,j}) \ , \tag{4.131}$$

where

$$R^{(k)}_{i,j} = \frac{q^{(k)}_{i+1,j} - q^{(k)}_{i,j}}{q^{(k)}_{i,j} - q^{(k)}_{i-1,j}} \ . \tag{4.132}$$

Thus the limiter $\psi_\kappa(R)$ is applied to each component of the states $\{q_{i,j}\}$ separately. Notice that it makes a difference how the vector q is decomposed into components. E.g. for the Euler equations we can take $q = (\rho, u, v, p)$ or $q = (u, v, c, \ln(\rho^{-\gamma} p))$.

Remark:
In case $\Omega_{i,j}$ is a boundary volume, so that, for example $\partial\Omega_{i+\frac{1}{2},j}$ is part of the domain boundary, no limiter can be used to compute $q^L_{i+\frac{1}{2},j}$ and $q^R_{i-\frac{1}{2},j}$. In this case one escapes by using e.g. a simple linear interpolation, i.e.

$$q^L_{i+\frac{1}{2},j} = q_{i,j} + \frac{1}{2}(q_{i,j} - q_{i-1,j}) \ , \tag{4.133}$$

$$q^R_{i-\frac{1}{2},j} = q_{i,j} - \frac{1}{2}(q_{i,j} - q_{i-1,j}) \,. \tag{4.134}$$

The boundary conditions, together with the state $q^L_{i+\frac{1}{2},j}$, are used to compute the boundary state $q^R_{i+\frac{1}{2},j}$, by considering the Riemann boundary problem. The flux $f_{i+\frac{1}{2},j}$ at $\partial\Omega_{i+\frac{1}{2},j}$ is computed by (4.53).

4.4.4 ENO schemes

To describe[11] the *ENO* (or *essentially non-oscillating*) schemes [18], consider the scalar one-dimensional problem, and a conservative approximation of the spatial operator given by

$$N(u)_j = -\frac{1}{\Delta x}(f^\star_{j+\frac{1}{2}} - f^\star_{j-\frac{1}{2}}) \,. \tag{4.135}$$

Here, the numerical flux $f^\star$ is assumed consistent:

$$f^\star_{j+\frac{1}{2}} = f^\star(u_{j-l}, \cdots, u_{j+k})\,;\, f^\star(u, \cdots, u) = f(u)\,.$$

The conservative scheme 4.135, which characterises the $f^\star$ divided difference as an approximation to $f(u)_x$, suggests that $f^\star$ can be identified with an appropriate function h satisfying

$$f(u(x)) = \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} h(\xi)\, d\xi \,.$$

If H is any primitive of h, then h can be computed from H'. The primitive H itself can be constructed by a (Newton) divided difference method, beginning with differences of order one, since the constant term is arbitrary. The necessary divided differences of H, of a given order, are expressed as constant multiples of those of f of order one lower. After the polynomial Q of degree $r+1$ has been constructed, set

$$f^\star_{j+\frac{1}{2}} = \frac{d}{dx}Q(x)|_{x=x_{j+\frac{1}{2}}} \,, \tag{4.136}$$

to obtain an r-th order method. The construction is based on an adaptive stencil in the following sense:

- One begins with an appropriate starting point to the left or the right of the current *cell* by means of upwinding as determined by the sign of the derivative of the selected flux.

[11] This section is derived from [27].

- As the order of the divided differences is increased, the divided differences themselves determine the stencil: the *smaller* divided difference is chosen from two possible choices at each stage, ensuring a smoothest fit.
- Lax-Friedrichs building blocks or Roe building blocks can both be used. For the latter, in cells with sonic points, a local Lax-Friedrichs building block is used to avoid expansion shocks.

Steady states are reached by explicit time stepping of arbitrary order; nonstandard high order Runge-Kutta methods exist which preserve nonlinear stability of the first order Euler forward version under suitable CFL time step restrictions (cf. [45], [46].

4.5 Multiple grid methods

When a multiple grid technique is used to solve the system of nonlinear (differential) equations (4.27) or (4.28), we assume the existence of a nested set of grids. Usually this nesting is such that a set of 2×2 cells in a fine mesh forms a single cell in the next coarser one (no staggered grids). The coarser grids are used to effect the acceleration of a basic iterative (time marching or relaxation) procedure on the finest grid.

Slightly generalising the equations (4.27) or (4.28) to

$$\frac{\partial}{\partial t} q_h = N_h(q_h) - r_h \tag{4.137}$$

or

$$N_h(q_h) = r_h, \tag{4.138}$$

where r_h denotes a possible correction or source term, we can write the basic iterative (or time-marching) procedure (the *relaxation procedure* or *smoothing procedure*) as

$$q_h^{(n+1)} \leftarrow \mathcal{S}^h\,(q_h^{(n)}, r_h). \tag{4.139}$$

The usual coarse grid acceleration algorithm is as follows: starting with an approximation $q_h^{(k)}$ on the finest mesh, and some approximation $q_{2h}^{(0)}$ on the next coarser (e.g. $q_{2h}^{(0)} = R_{2h,h} q_h^{(k)}$), first an approximate solution is found for the coarse grid problem

$$N_{2h}(q_{2h}) = N_{2h}(q_{2h}^{(0)}) - \overline{R}_{2h,h}(N_h(q_h^{(n)} - r_h)), \tag{4.140}$$

and then the value $q_h^{(k)}$ is updated by

$$q_h^{(k+1)} = q_h^{(k)} + P_{h,2h}(q_{2h} - q_{2h}^{(0)})\,. \tag{4.141}$$

The combination of (4.140) and (4.141) is the *coarse grid correction* (CGC) step. The solution q_{2h} of (4.140) can be approximated e.g. by an (accelerated) iteration process on the $2h$-grid again.

We shall see later in this section that, besides this usual coarse grid acceleration procedure, the coarser grids sometimes play a different role in the acceleration process [39],[32].

A multigrid *FAS cycle* for the solution of (4.138) now consists of the following steps:

(0) start with an approximate solution q_h.

(1) improve q_h by application of p nonlinear *(pre-) relaxation* iterations to $N_h(q_h) = r_h$.

(2) if the present grid is not the coarsest, improve q_h by application of one coarse-grid-correction step, where the approximation of (4.140) is effected by σ FAS-cycles to this coarser grid problem; if the present grid is the coarsest, simply skip to (3).

(3) improve q_h by application of q nonlinear *(post-) relaxation* iterations to $N_h(q_h) = r_h$.

4.5.1 Ni's multiple grid method

A paper by Ni [39] was among the first to apply a multigrid acceleration to the (isenthalpic) Euler equations. He uses the following time-stepping procedure as a basic iteration. Starting with an initial state $q_h^{(n)}$, where the values $q_{ij}^{(n)}$ are given at the grid points, he first computes the following quantities, by means of a control volume centered integration method with fluxes interpolated from corner values:

$$\begin{aligned} \Delta q_{i+\frac{1}{2},j+\frac{1}{2}} = \; & -\tfrac{1}{2}\tfrac{\Delta t}{\Delta x}[(F_{i+1,j} - F_{ij}) + (F_{i+1,j+1} - F_{i,j+1})] \\ & -\tfrac{1}{2}\tfrac{\Delta t}{\Delta y}[(G_{i,j+1} - G_{ij}) + (G_{i+1,j} - G_{i+1,j+1})]\,, \end{aligned} \tag{4.142}$$

where $F_{ij} = F(q_{ij}^{(n)})$ etc..

These increments then are distributed over the mesh points, using direction-weighted means (cell-increments are distributed over mesh-point values):

$$\Delta q_{ij} = \frac{1}{4}\sum_{l=\pm1}\sum_{k=\pm1}\left[I - k\frac{\Delta t}{\Delta x}A_{i+\frac{k}{2},j+\frac{l}{2}} - l\frac{\Delta t}{\Delta y}B_{i+\frac{k}{2},j+\frac{l}{2}}\right]\Delta q_{i+\frac{k}{2},j+\frac{l}{2}}\,, \tag{4.143a}$$

$$q_{ij}^{(n+1)} = q_{ij}^{(n)} + \Delta q_{ij}\,. \tag{4.143b}$$

By the use of the Jacobian matrices A and B, this distribution formula has a kind of upwind effect, but for transonic or supersonic cases an artificial damping is still necessary.

Symbolically, this time stepping process (4.142)-(4.143b) is described as:

$$\text{compute} \quad \Delta q_h^{\text{cell}}\,, \tag{4.144}$$

with cell values

$$\Delta q_{i+1/2,j+1/2} \approx -\Delta t \int_{\partial\Omega_{i+1/2,j+1/2}} (f\,n_x + g\,n_y)\,ds\,/(\Delta x\,\Delta y)\,;$$

$$q_h^{(n+1)} := q_h^{(n)} + D_h\,\Delta q_h^{\text{cell}}\,. \tag{4.145}$$

The operator D_h is the distribution operator that transfers the cell centered corrections to the grid points by means of (4.143b).

The coarse grid acceleration as introduced in [39] by Ni deviates from the usual coarse grid scheme (4.140)-(4.141). In [39] the coarse grid correction is obtained by first computing corrections at coarser cells, $\Delta q_{2h}^{\text{cell}}$. This can be done by restriction of Δq_h to the $2h$-grid. Then the corrections $\Delta q_{2h}^{\text{cell}}$ are distributed to the coarser mesh points similar to (4.143b), and the coarse grid correction is interpolated to the fine grid. Thus, here the coarse grid correction reads

$$\Delta q_{2h}^{\text{cell}} := R_{2h,h}\,\Delta q_h^{\text{cell}}\,, \tag{4.146a}$$

$$q_h^{(n+1)} := q_h^{(n)} + P_{h,2h} D_{2h}\,\Delta q_{2h}^{\text{cell}}\,. \tag{4.146b}$$

where $P_{h,2h}$ is a (bi-)linear interpolation operator.

Since the coarse grid corrections are based on fine grid residuals, it is obvious that the possible convergence to a steady state yields a solution of the system (4.28).

In the same way the correction procedure can be repeated on progressively coarser grids. Therefore, in (4.146a), $2h$ should be replaced by $2^m h$. We notice that, in contrast with the usual MG-method, here the corrections on the different levels can be computed independent of each other. This yields the possibility to compute all coarse grid corrections, $m = 1, ..., L$, in parallel and to form the correction

$$q_h^{(n+1)} = q_h^{(n)} + \sum_{m=1}^{L} P_{2,2^m h}\, D_{2^m h}\,\Delta q_{2^m h}^{\text{cell}}\,. \tag{4.147}$$

at once [54]. When optimal use of modern multi-processor computers is to be made, it is also possible to perform both computations (4.144) and (4.146a) in parallel.

We see that there are still possibilities to form different variants in the Ni-type multigrid Euler solver. First, any other Lax-Wendroff-type time-marching procedure can be used for (4.144). In [3], [32] Johnson applies the popular MacCormack scheme. Further, in (4.146a) various restrictions, $R_{2h,h}$, can be used. It transfers the values of the fine grid corrections to a single value for each control volume in the coarser grid. Injection of the correction in the main point of the corresponding cell is often used, but also weighted averages are an obvious choice.

Heuristically, the elucidation for the accelerating effect of the corrections (4.140) is, that these coarse grid corrections may move disturbances of the steady state over the distance of many mesh cells in one time step, whereas the accuracy of the final solution is only determined by the finest grid. Apparently, it is also necessary that the Lax-Wendroff schemes used in combination with this coarse grid correction are (by the choice of a suitable Δt or otherwise) sufficiently dissipative to reduce the high frequency disturbances that are present in the initial approximation and those introduced during the process by the interpolation in (4.141). Up to now, no complete mathematical theory has been developed to explain and to quantify the amount of acceleration, which is clearly found in the many computations that use the described method.

4.5.2 Methods based on semidiscretisation and time stepping

When only the solution of the steady state is to be computed, the time-accurate integration of the system of ODEs is wasteful. The convergence of (4.26) to steady state is slow. However, the desire to have a procedure that solves transient as well as steady state problems, coding convenience, or the restrictions imposed by the optimal use of vector computers may be a reason to prefer time-stepping methods. When no time accuracy is desired, many devices are known to accelerate the integration process (cf. [44]). For the solution of the Euler equations, these devices include: (i) local time-stepping, which means that the step size in the integration process may differ over different parts of the domain $\Omega^\star$, such that we have the same CFL number per cell[12]; (ii) enthalpy damping, where a-priori knowledge about the behaviour of the enthalpy over $\Omega^\star$ is used (e.g. H constant over $\Omega^\star$); (iii) residual smoothing, (iv) implicit residual averaging, and (v) implicit corrected viscosity acceleration [7]. In residual smoothing and implicit residual averaging the fact is used that instability effects appear first for high frequencies, so that larger time steps are possible when the residual is smooth.

For all explicit integration methods, stability requirements set a limit to the size of the possible time steps (CFL limits). Implicit integration procedures can be unconditionally stable, but they require the solution of a (nonlinear) system in each individual time step.

An important code, based on a time-stepping method has been developed

[12]Or even characteristic time stepping, which means that the time step may differ per local equation, so that we have the same CFL number for each separate equation.

by Jameson, Schmidt and Turkel [25]. They use an explicit time-stepping method of Runge-Kutta type. This *multistage time-stepping procedure* is a specially adapted Runge-Kutta method, where the hyperbolic (=convective) and the parabolic (=dissipative) parts of $N_h(q_h)$ are treated separately. The Runge-Kutta coefficients in the k-stage Runge-Kutta schemes (k= 3,4), are selected not only for their large stability bounds, but also with the aim to improve the damping of the high frequency modes. In the k stages of the Runge-Kutta process, the updating of the dissipative part is frozen at the first stage. This saves a substantial part of the computational effort.

The multigrid scheme used by *Jameson* [24] is a FAS sawtooth cycle with $q = 1$. The restriction $R_{2h,h}$ ($\overline{R}_{2h,h}$) is defined by volume-weighted averaging of the states (respectively summation of changes of states). The prolongation $P_{h,2h}$ is defined by bilinear interpolation. The basic smoothing procedure is the "multistage time-stepping scheme". On the coarser grids the stability bounds for the time step, which are $\mathcal{O}(h)$, allow larger time steps. On each grid the time step is varied locally to yield a fixed CFL number per cell (local time stepping), and the same CFL number is used on all grids, so that progressively larger time steps are used after each transfer to a coarser grid. As for Ni's method, the reasoning is that disturbances from the steady state will be more rapidly expelled from the domain $\Omega^\star$ by the larger time steps. The interpolation of corrections back to the fine grid introduces high frequency errors, which cannot be rapidly expelled. These errors should be locally damped. Hence, to obtain a fast rate of convergence, the time-stepping process should rapidly damp the high frequency errors.

In [30] Jespersen mentions an interesting theorem on the use of the multigrid process in combination with a time-stepping procedure. This theorem asserts the following. Let $N_h(q_h) = 0$ be a space discretisation of $N(q) = 0$, which is consistent, i.e.

$$N_h(R_h(q)) - \overline{R}_h N(q) = \mathcal{O}(h)\,, \tag{4.148}$$

and let the time-stepping procedure be consistent in time

$$q_h^{(n+1)} = q_h^{(n)} + \Delta t_(n)[N_h(q_h^{(n)}) - r_h] + \mathcal{O}((\Delta t_{(n)})^2)\,. \tag{4.149}$$

If we consider the sawtooth algorithm, with $q = 1$, $p = 0$, $\sigma = 1$, and if P_h and R_h satisfy an approximation property (i.e. for a smooth function q the prolongation and restriction in the state space are such that $P_h R_h q - q = \mathcal{O}(h)$), then the multigrid algorithm on L grids is a consistent, first-order in time, discretisation of (4.26) with time step $\Delta t_{\text{tot}} = \sum_{j=1,\ldots,L} \Delta t_j$.

This theorem formalises in a sense the heuristic reasoning that on coarser grids the deviations from steady state can be expelled faster by the use of larger time steps. This may suggest that more, say $k > 1$, steps on the coarser grids would improve the convergence even more. However, the theorem regards consistency; stability is not considered. In the same paper [30] Jespersen shows by an example that convergence is lost when a large number of relaxations is made on the coarse grid. In fact a strong stability condition of the form $\Delta t/\Delta x \leq \mathcal{O}(k^{-1})$ seems to appear.

4.5.3 Fully implicit methods

Most methods so far developed are based on the concept of integrating the equations (4.26) in time until a steady state is reached. If we are only interested in a possible solution of the steady state equation and assume that this solution is unique, we may disregard the time-dependence completely. Further, assuming that a suitable space discretisation takes into account the proper characteristic directions, we can restrict ourselves simply to the solution of the nonlinear system (4.28) or

$$N_h(q_h) = r_h\,. \tag{4.150}$$

Also, if the time-dependent system (4.137) is solved by means of an implicit time-stepping method -in order to circumvent the stability bounds on Δt-, we have to solve systems (4.150) at each step time step. As soon as we mix time-dependent solution with these implicit solution methods and give up time accuracy for (4.150), there is little or no difference between these time stepping procedures and (nonlinear) relaxation methods.

Starting with the nonlinear system (4.150), two direct multigrid approaches are open. We can either apply the nonlinear multiple grid algorithm (FAS) directly to the system (4.150) or we may apply linearisation (Newton's method) and use the linear version of multiple grid for the solution of the resulting linear systems. Jespersen [29] gives an extensive recital of the (dis)advantages of both approaches. Both have been used with success for the Euler equations.

Linearisation has been used by Jespersen [28] and Mulder [38]; the nonlinear FAS procedure is used by Steger [50], Jespersen [28], Hemker-Spekreijse [48] [21], and Dick [6].

In all these papers upwind discretisations have been used. In [28], [50] the Steger-Warming scheme is used; [38] uses the differentiable Van Leer flux-splitting method; [21], [21] use Osher's flux difference splitting, and [6] uses Lombard's flux difference splitting. (In [5] Dick also considers Roe's flux difference splitting for the one-dimensional Euler equations.)

When Newton's method is applied for linearisation, it may be difficult to start in the domain of contraction of the iteration. Therefore, Mulder [38] introduces the so-called Switched Evolution Relaxation (SER) scheme, which is a chimera of a forward Euler time-stepping and a Newton method:

$$\left[\frac{1}{\Delta t}\,I - \frac{\partial}{\partial q}N_h(q_h^{(n+1)})\right]\,(q_h^{(n+1)} - q_h^{(n)}) \;=\; N_h(q_h^{(n)})\,. \tag{4.151}$$

For $\Delta t \to 0$, this gives the simple time stepping procedure; for $\Delta t \to \infty$, (4.151) is equivalent to Newton's method. In the actual computation Δt varies, depending on the size of the residual, such that (4.151) is initially a time stepping procedure and becomes Newton's method in the final stages of the solution process.

In a FAS procedure, a natural way to obtain an initial estimate is - of course- the use of Full Multi-Grid (FMG) [2]. The initial estimate is obtained by interpolation from the approximate solution on the coarser grid(s). For many problems this process gives very good results, even if one starts with rough approximations on a really coarse grid.

4.6 References

[1] G. D. Van Albada, B. Van Leer, and W. W. Roberts. A comparative study of computational methods in cosmic gasdynamics. *Astron. Astrophys.*, 108:76–84, 1982.

[2] A. Brandt. Guide to multigrid development. In W. Hackbusch and U. Trottenberg, editors, *Multigrid Methods*, volume 960 of *Lect. Notes in Mathematics*, pages 220–312. Springer Verlag, 1982.

[3] R. V. Chima and G. M. Johnson. Efficient solution of the Euler and Navier Stokes equations with a vectorized multi-grid algorithm. Technical Report 83-1893, AIAA, 1983.

[4] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928.

[5] E. Dick. A multigrid method for the Cauchy-Riemann and steady state Euler equations based on flux difference splitting. In W. Hackbusch, editor, *Efficient Solution of Elliptic Systems*, pages 20–37. Vieweg Verlag, 1984.

[6] E. Dick. A multigrid method for steady Euler equations, based on flux-difference splitting with respect to primitive variables. In W. Hackbusch, editor, *Robust Multi-Grid Methods*, pages 69–85. Vieweg Verlag, 1988.

[7] J. A. Essers. Explicit and implicit corrected viscosity schemes for the computation of steady transonic flows. In H. Viviand, editor, *Proceedings of the 4th GAMM-Conference on Numerical Methods in Fluid Mechanics.* Vieweg Verlag, 1982.

[8] G.E. Forsythe and W.R. Wasow. *Finite-Difference Methods for Partial Differential Equations.* Wiley, New York, 1960.

[9] S.K. Godunov. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Math. Sbornik.*, 47:271–306, 1959.

[10] J.B. Goodman and R.J. LeVeque. On the accuracy of stable schemes for 2d scalar conservation laws. *Math. Comp.*, 45:156, 1985.

[11] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* B. G. Teubner, Stuttgart, 1986.

[12] D. Hänel and R. Schwane. An implicit flux-vector splitting scheme for the computation of viscous hypersonic flow. AIAA Paper 89-0274, 1989.

[13] D. Hänel, R. Schwane, and G. Seider. On the accuracy of upwind schemes for the solution of the navier-stokes equations. AIAA Paper 87-1105-CP, 1987.

[14] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49:357–393, 1983.

[15] A. Harten, J.M. Hyman, and P.D. Lax. On finite difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.*, 29:297–322, 1976.

[16] A. Harten, P. D. Lax, and B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, 25:35–61, 1983.

[17] A. Harten, P. D. Lax, and B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, 25:35–61, 1983.

[18] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes. *SIAM J. Numer. Anal*, 24:279–309, 1987.

[19] P.W. Hemker and B. Koren. Defect correction and nonlinear multigrid for the steady Euler equations. Technical Report NM-N8801, CWI, 1988.

[20] P.W. Hemker and B. Koren. Efficient multi-dimensional upwinding for the steady Euler equations. Technical Report NM-R9107, CWI, 1991.

[21] P.W. Hemker and S.P. Spekreijse. Multiple grid and Osher's scheme for the efficient solution of the steady Euler equations. *Appl. Num. Math.*, 2:475– 493, 1986.

[22] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 1: Fundamentals of Numerical Discretisation*. J. Wiley, 1988.

[23] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 2: Computational Methods fir Inviscid and Viscous Flow*. J. Wiley, 1990.

[24] A. Jameson. Solution of the Euler equations for two dimensional transonic flow by a multigrid method. *Appl. Math. and Comput.*, 13:327–355, 1983.

[25] A. Jameson, W. Schmidt, and E. Turkel. Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes. Technical Report 81-1259, AIAA, 1981.

[26] R. Jeltsch. Properties of discrete hyperbolic conservation laws. Technical Report 88-54, Inst. f. Geometrie u. Praktische Mathematik, RWTH, Aken, 1988.

[27] J.W. Jerome and Chi-Wanh Shu. Energy models for one-carrier transport in semiconductor devices. Technical Report ICASE report No. 91-78, NASA, Langley Research Center, Hampton, Virginia 23665-5225, 1991.

[28] D. C. Jespersen. Design and implementation of a multigrid code for the Euler equations. *Appl. Math. and Comput.*, 13:357–374, 1983.

[29] D. C. Jespersen. *Recent developments in multigrid methods for the steady Euler equations.* von Karman Inst., Rhode-St.Genese, Belgium, 1984. Lecture Notes of a course given at VKI, March 12-16, 1984.

[30] D. C. Jespersen. A time-accurate multiple-grid algorithm. Technical Report 85-1493-CP, AIAA, 1985.

[31] C. Johnson, A. Szepessy, and P. Hansbo. Defect correction and nonlinear multigrid for the steady Euler equations. Technical Report 21, Chalmers Univ. Techn., 1987.

[32] G. M. Johnson. Accelerated solution of the steady state Euler equations. In W.G. Habashi, editor, *Recent Advances in Numerical methods in Fluids. Vol.4.*, Recent advances in numerical methods in fluids. Pineridge Press, 1983.

[33] B. Koren. Evaluation of second order schemes and defect correction for the multigrid computation of airfoil flows with the steady Euler equations. Technical report, CWI, 1986. Accepted for publication in J.Comp.Phys.

[34] B. Koren. *Multigrid and Defect Correction for the Steady Navier-Stokes equations, Applications to aerodynamics.* Number 74 in CWI Tract series. CWI, Amsterdam, 1991.

[35] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, volume 11 of *Regional Conference Series in Applied Mathematics.* SIAM Publication, Philadelphia, 1973.

[36] P.D. Lax and B. Wendroff. Systems of conservation laws. *Comm. Pure Appl. Math.*, 13:217–237, 1960.

[37] R.J. LeVeque. *Numerical Methods for Conservation Laws.* Lectures in Mathematics. Birkhäuser, Basel, 1990.

[38] W. A. Mulder. Multigrid relaxation for the Euler equations. *J. Comput. Phys.*, 60:235 – 252, 1985.

[39] Ron-Ho Ni. A multiple grid scheme for solving the Euler equations. *AIAA Journal*, 20:1565–1571, 1982.

[40] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38:339–374, 1982.

[41] S. V. Patankar. *Numerical Heat Transfer and Fluid Flow.* Hemisphere McGraw-Hill, 1980.

[42] D. W. Peaceman. *Fundamentals of Numerical Reservoir Simulation.* Elsevier Science Publ., 1977.

[43] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial Value Problems.* Interscience Publ., 1967.

[44] W. Schmidt and A. Jameson. Euler solvers as an analysis tool for aircraft aerodynamics. In W.G. Habashi, editor, *Advances in Computational Transonics*, Swansea. Pineridge Press.

[45] C.W. Shu and S.J. Osher. Efficient implementation of essentially non-oscillatory shock capturing schemes. *J. Comput. Phys.*, 77:439–471, 1988.

[46] C.W. Shu and S.J. Osher. Efficient implementation of essentially non-oscillatory shock capturing schemes, II. *J.Comp.Phys.*, 83:32–78, 1989.

[47] J. Smoller. *Shock Waves and Reaction Diffusion Equations*, volume 258 of *Grundlehren der mathematische Wissenschaften.* Springer Verlag, 1983.

[48] S.P. Spekreijse. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comput.*, 49:135–155, 1987.

[49] S.P. Spekreijse. *Multigrid Solution of the Steady Euler Equations*, volume 46 of *CWI Tracts.* CWI, 1988.

[50] J. L. Steger. A preliminary study of relaxation methods for the inviscid conservative gasdynamics equations using flux splitting. Technical Report NASA-CR-3415, NASA, 1981.

[51] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations.* Wadswordth and Brooks, Pacific Grove, Calif., 1989.

[52] R. Struijs, H. Deconinck, and G. Coussement. Multidimensional upwind schemes for the compressible flow equations. Technical report, Von Karman Institute, 1989.

[53] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J.Numer.Anal.*, 21:995–1011, 1984.

[54] J. M. Swisshelm and G. M. Johnson. Numerical simulation of three-dimensional flowfields using the Cyber 205. In R.W. Numrich, editor, *Supercomputer Applications*. Plenum Publ. Comp., 1985.

[55] B. Van Leer. Towards the ultimate conservative difference scheme II. monotonicity and conservation combined in a second order scheme. *J. Comput. Phys.*, 14:361–370, 1974.

[56] B. Van Leer. Towards the ultimate conservative difference scheme IV. a new approach to numerical convection. *J. Comput. Phys.*, 23:276–299, 1977.

[57] B. Van Leer. Flux vector splitting for the euler equations. In *Procs. 8th Intern. Conf. Num. Meth. Engng*, 1982.

[58] B. Van Leer. On numerical dispersion by upwind differencing. Technical Report 85-51, Delft Univ. Techn., 1985.

[59] B. Van Leer. Upwind difference methods for aerodynamic problems governed by the Euler equations. In S. Osher B.E. Engquist and R.C.J. Somerville, editors, *Large-Scale Computations in Fluid Mechanics*, volume 22 of *Lectures in Applied Mathematics*, pages 327–336 (vol2), Providence, Rhode Island, 1985 1984. American Mathematical Society.

[60] B. Van Leer. Flux vector splitting for the 1990's. Technical report, Univ. Michigan, 1990.

Index