Some aspects of high resolution numerical methods for hyperbolic systems of conservation laws,with applications to gas dynamics

Paul Arminjon

INRIA

CENTRE
SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 520

SOME ASPECTS OF HIGH RESOLUTION NUMERICAL METHODS FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS, WITH APPLICATIONS TO GAS DYNAMICS

Paul ARMINJON

Avril 1986

*SOME ASPECTS OF HIGH RESOLUTION NUMERICAL METHODS FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS, WITH APPLICATIONS TO GAS DYNAMICS**

Paul ARMINJON

Département de mathématiques et de statistique
Université de Montréal

Dedicated to Professor Lothar Collatz on the occasion of his 75th birthday.

ABSTRACT

This paper discusses several important aspects of finite difference methods for hyperbolic systems of conservation laws, from first order upwind to second-order TVD schemes based on flux limiters and flux splitting, or on Godunov's method. Particular attention is given to the recent schemes of Sweby, Davis, and Goodman-LeVeque, which are second-order TVD versions of the Lax-Wendroff or Godunov schemes. An essential role is played by Harten's criterion for Total Variation Diminishing schemes.

---

* This research was supported by the Alexander von Humboldt Foundation (Bonn, Germany); the Institut National de Recherche en Informatique et Automatique (Sophia Antipolis, France), and the Natural Sciences and Engineering Research Council of Canada.

RÉSUMÉ

Ce rapport étudie certains aspects importants des méthodes numériques pour la résolution des systèmes hyperboliques de lois de conservation. Procédant à partir des schémas à dérivation amont du premier ordre, il conduit aux récentes méthodes TVD d'ordre deux de Sweby, Davis, et Goodman-LeVeque, en passant par les schèmas classiques de Lax-Wendroff, MacCormack, Godounov et les éléments de la théorie TVD de Harten.

SOME ASPECTS OF HIGH RESOLUTION NUMERICAL METHODS
FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS,
WITH APPLICATIONS TO GAS DYNAMICS

by

Paul ARMINJON

Université de Montréal

1. *INTRODUCTION: HYPERBOLIC SYSTEMS OF CONSERVATION LAWS AND NUMERICAL METHODS*

We consider the numerical resolution of hyperbolic systems of conservation laws, in one space dimension

$$U_t + F(U)_x = 0 \tag{1.1}$$

or in several space dimensions

$$U_t + \sum_{i=1}^{3} F_i(U)_{x_i} = 0 \tag{1.1'}$$

where $U = (u_1,\ldots,u_m)^T$ and $F = (f_1(U),\ldots,f_m(U))^T$ are m-component column vectors depending on time t and one or several spatial coordinates x or x_i $(i = 1,2,3)$. Introducing the jacobian matrix $A = \partial F/\partial U$ in the case of one space variable x, we can also write (1) in the form of a quasi-linear system

$$U_t + A(U)U_x = 0 \ . \tag{1.2}$$

Since solutions of (1) or (2) are known to display, in certain circumstances, discontinuities which, in the case of the eulerian equations of gas dynamics, are called shocks or contact discontinuities (see for instance [10], [30], [62]), both the applied mathematician and the engineer are interested in the conception of numerical methods which are able to give a precise approximation of the solution in the neighbourhood of discontinuities and away from them, while respecting some stability ([47], [20]) and efficiency requirements.

First order explicit difference schemes are restricted by their lack of precision and the strong limitation imposed on the time mesh Δt by stability requirements. Centered second order explicit methods, which very often give satisfactory results, generally only do so once one has adjusted some problem-dependent parameters associated with appropriate techniques designed to eliminate the undesirable oscillations observed near the discontinuities (The pseudo-viscosity method of von Neumann and Richtmyer [40], the artificial viscosity of Lax and Wendroff [29], [47], the anti-diffusion method of Boris and Book [6], and the artificial compression method (ACM) of Harten [25]. See also [53] for a review of these and a few other methods).

On the other hand, one can distinguish between centered difference schemes, such as for instance the Lax-Wendroff scheme, and upstream-centered schemes, as the Courant-Isaacson-Rees scheme [11]("C.I.R."), or the flux-splitting scheme of Steger and Warming [55], and many extensions and applications to finite difference and finite element (van Leer [33], [34], [35], Harten-Lax-van Leer [27], Angrand et al. [1], Angrand-Dervieux [2].

Upstream-centered schemes, where the new value u_j^{n+1} is computed with the help of grid-points x_j belonging to those grid-intervals only, from which, physically, the relevant information might effectively come, generally have smaller phase errors and sharper shock resolution than centered schemes with the same

order of accuracy. In most cases, they contain a rather important amount of "numerical viscosity", which has two primary consequences: it inhibits the formation of numerical oscillations in the neighbourhood of shocks and other discontinuities, and it has a smearing effect which more or less smoothes out these discontinuities, which may be completely masked in some unfavourable cases (see [55] or fig. 2.2).

Ideally, one would like to find a scheme which is

(i) monotonicity preserving, thanks to the benefic action of an appropriate amount of numerical (or artificial) viscosity

(ii) second (or higher) - order accurate

Property (i) guarantees that no numerical oscillations will be generated by the scheme. In the case of artificial viscosity (i.e. deliberately added, without respecting the original differential equations), one has to adjust and fine-tune some parameters which are problem-dependent (as with the Lax-Wendroff scheme, for instance); in the case of numerical viscosity (which generally originates from upwinding and flux-splitting) there is a strong risk of accuracy loss due to the smearing effect mentioned above, so that one of the main goals of recent research has been to find a scheme which, while being second-order occurate away from the shocks, has a built-in mechanism (independent of the particular problem under consideration) to nearly of totally eliminate oscillations near the discontinuities.

The numerous studies and experiments which have been carried recently seem to lead to the conclusion that for practical purposes, the design of a suitable scheme ultimately rests on a clever compromise, performed with the help of an efficient shifting strategy, between monotonicity preservation, with the

risk of accuracy loss and shock smearing, and higher order accuracy.

Among the methods which seem to best take advantage of the upstream-centering principle is Godunov's method ([19], [20], [21], [27]) and its higher-order extensions, for which the upwinding lies rather deeply buried in the process of resolution of Riemann problems, and which certainly is one of the safest first-order methods, although one might object that the need to solve a Riemann problem at each grid-point (and time step) (and at each cell-interface in several space dimensions) makes it a little awkward and expensive.

In his quest of the "ultimate conservative scheme", van Leer [35] has extended Godunov's scheme to an interesting second-order Godunov-type method, capable of very sharp shock resolution, and Hancock [23] has presented a valuable simplification of van Leer's scheme, which has recently been extended to two space dimensions by Fezoui [15], with some modifications.

In van Leer's scheme, monotonicity is strictly preserved with the help of slope-limiters, while in [15] the basic idea is to try and obtain a nearly monotonicity preserving scheme, to avoid the excessive dissipation observed, for irregular grids in two space dimensions, with the Hancock-van Leer algorithm; the numerical examples displayed there still show some slight oscillations, but very good shock resolution.

Recently a large number of papers have concentrated on a new family of schemes, called Total Variation Diminishing (TVD) schemes by Harten [64], which include van Leer's method and the more restricted family of monotone schemes (written in conservation form), and form a sub-class of the family of monotonicity preserving schemes (see van Leer [32], [34], Harten [64], Osher-Chakravarthy [44], [45]; [43], [42]; Goodman-LeVeque [22], Davis [12], Sweby [56], Yee, Warming and Harten [61], and the references there).

The obvious advantage of monotone, TVD or monotonicity preserving schemes is the elimination of spurious oscillations in the neighbourhood of discontinuities. While Godunov's method, although monotonicity preserving, was only first-order accurate, the above mentioned work is leading to high resolution schemes (at least 2^{nd}-order accurate away from shocks or "sonic" points, in one space dimension), which bring great improvements in all numerical tests which had been used earlier.

In this report, we shall try to give a clear account of some of this important contribution to the subject, and show how an essential part of this theory consists of either

(i) improving on Godunov's scheme by starting from a more accurate initialization process and forcing monotonicity preservation with slope limiters ([34], [22], [43]) or function limiters ([15])

or

(ii) starting from the Lax-Wendroff scheme, analyse it as an upstream-centered first order scheme plus a second-order anti-dissipative term the effect of which should be constrained to regions of smooth flow via flux limiters, while it will be nearly suppressed near the shocks in order to smother the tendency of the scheme to generate oscillations ([56], [12]).

A number of numerical tests are presently in progress for one and two-dimensional compressible flow problems. We shall report on them in a forthcoming publication.

The content of this report is as follows. In the next section, we shall motivate the introduction of upstream-centered schemes and present Steger and Warming's extension of the Courant-Isaacson-Rees scheme to the Euler equations of ideal compressible flow. Section 3 gives a short description of the now classical second-order schemes of Lax and Wendroff, MacCormack, Beam and Warming, for which one has to use a specific strategy to damp out the oscillations, and Section 4 presents some basic facts about TVD schemes, following Harten [64]. Sections 5, 6 and 7 give a rather detailed introduction to the schemes of Sweby, Davis and Goodman-LeVeque, with a quick description of the highly competitive schemes of van Leer and Chakravarthy-Osher.

2. *UPSTREAM-CENTERED SCHEMES. THE EULER EQUATIONS OF IDEAL COMPRESSIBLE FLOW. FLUX SPLITTING.*

A. *Upwinding and upstream-centered schemes*

We consider, to simplify the presentation of the basic concepts, the one-dimensional scalar linear wave (or convection) equation

$$u_t + au_x = 0 \ , \ a = \text{constant}, \quad \begin{matrix} -\infty < x < +\infty \\ 0 \le t \end{matrix} \tag{2.1}$$

with initial condition

$$u(x,0) = u_0(x) \ , \ u_0 : \text{a given function of } x \tag{2.2}$$

and either finite boundary conditions

$$u(x_L,t) = u_L \ , \ u(x_R,t) = u_R \quad -\infty < x_L < x_R < +\infty \tag{2.3}$$

or periodic boundary condition

$$u(x+L,t) = u(x,t) \quad \text{for some} \quad L > 0 \quad \text{and all} \quad x\ ,\ t \geq 0 \tag{2.3'}$$

The first difference scheme one would normally think of is the "centered" (in space) scheme based on explicit, forward time differencing:

$$u_j^{n+1} = u_j^n - \frac{a\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n) \tag{2.4}$$

where $x_j = j\Delta x$, $j = 0,\pm 1,\ldots$, $t^n = n\Delta t$, $n = 0,1,2,\ldots$, and u_j^n is a numerical approximation to the value of the solution $u(x_j,t^n)$ at $x = x_j$, $t = t^n$.

Unfortunately, this scheme is unconditionally unstable, as can be seen by introducing in (2.4) von Neumann's assumption of individual harmonics of the form

$$\bar{u}_j^n = \rho^n e^{ik(j\Delta x)} \tag{2.5}$$

where ρ is the complex "amplification factor" and k is the wave number of the corresponding harmonic of $u_j^n = \sum_{k=-\infty}^{\infty} \rho^n(k) e^{ikj\Delta x}$; here the form of the Fourier coefficient $\rho^n = \rho^n(k)$ comes from basic properties of difference equations. This leads here to

$$\rho = \rho(k) = 1 - \nu(e^{ik\Delta x} - e^{-ik\Delta x}) = 1 - 2i\nu \sin(k\Delta x)$$

with $\nu = a\Delta t/\Delta x$, so that the amplitude $|\rho|$ of this harmonic is always larger than 1, allowing for unbounded growth of the numerical solution u_j^n , while the exact solution is nothing but the translated image of the initial function.

For $a > 0$ the "upwind" scheme

$$u_j^{n+1} = u_j^n - \frac{a\Delta t}{\Delta x}(u_j^n - u_{j-1}^n) \tag{2.6}$$

is stable under the celebrated "C.F.L." condition of Courant, Friedrichs and Lewy (1928)

$$\nu = \frac{a\Delta t}{\Delta x} \leq 1 \quad \text{(C.F.L.)} \tag{2.7}$$

as we can easily check, using the same von Neumann approach. For $a < 0$, the upstream-centered scheme

$$u_j^{n+1} = u_j^n - a\lambda(u_{j+1}^n - u_j^n) \ , \ \lambda = \Delta t/\Delta x \tag{2.6'}$$

is stable under the same (C.F.L.)-condition (but with $|a|$ instead of a).

Since in a system of quasi-linear conservation equations, some of the characteristic speeds will be positive and some negative, Courant, Isaacson and Rees [11] conceived a scheme which can switch the direction of differencing in order to always use upstream values (i.e. information coming from upstream rather than downstream): For the scalar equation (2.1) they set

$$u_j^{n+1} = \begin{cases} u_j^n - a\lambda(u_j^n - u_{j-1}^n) & \text{if } a > 0 \\ u_j^n - a\lambda(u_{j+1}^n - u_j^n) & \text{if } a < 0 \end{cases} \quad \text{(C.I.R.)} \tag{2.8}$$

One can readily verify that this scheme computes the new value of u at x_j (at time $(n+1)\Delta t$) as the linear interpolate, at $x_j - a\Delta t$, of old values (at time $n\Delta t$) at the adjacent grid points x_{j-1} and x_j $(a>0)$ or x_j and x_{j+1} $(a<0)$. For the linear advection equation (2.1), the interpolated value at $x_j - a\Delta t$ would be exactly transmitted to the point (x_j, t^{n+1}) by the analytic solution $u(x,t) = u(x-at,0) = u_0(x-at)$ where $u_0(x)$ is the initial value function.

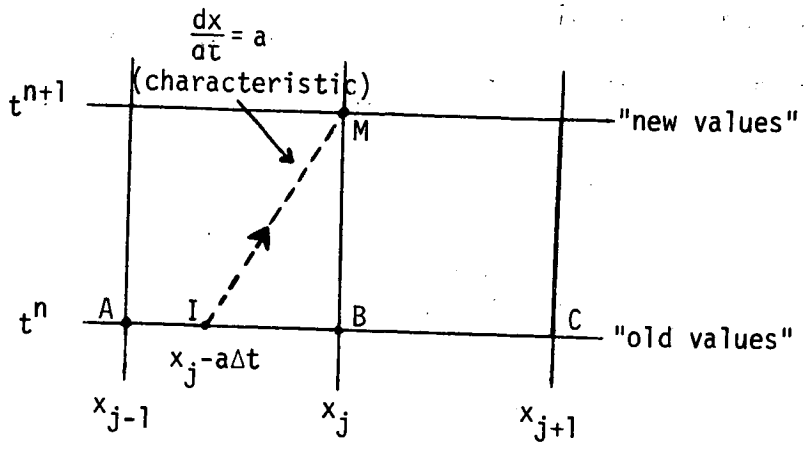

Fig 2.1 The Courant-Isaacson-Rees scheme (for the case $a > 0$)

In particular, it is clear from fig. 1 that the value u_j^{n+1} given by (2.6) is <u>exact</u> if the mesh ratio $\lambda = \Delta t/\Delta x$ and the characteristic slope $1/a$ are such that the interpolation point I, from which the characteristic line through M departs, coincides with the left-end gridpoint A, so that the "interpolated" value at I is exactly equal to $u^n(A) = u_{j-1}^n$, which will then be carried on exactly to grid-point M both by our scheme (2.6) and by the wave propagation mechanism of equation (2.1) through the (therewith) characteristic line $AM = IM$, i.e. (2.6) is exact for

$$\frac{\Delta t}{\Delta x} = \frac{1}{a} \quad \text{or} \quad \frac{a\Delta t}{\Delta x} = 1 \ . \tag{2.9}$$

Obviously, the full extent of the C.I.R. scheme only comes into play when the wave propagation speed a depends on (x,t,u) and may change its sign. Before applying it first to a linear system

$$U_t + AU_x = 0 \ , \tag{2.10}$$

with constant coefficients, and then presenting Steger and Warming's extension to the system of the one-dimensional equations of gas dynamics, let us rewrite the

C.I.R. scheme in order to reveal its close association with the aforementioned principle of pseudo-viscosity: introducing the positive and negative parts

$$a^+ = \max(a,0) = \frac{1}{2}(a+|a|)$$
$$a^- = \min(a,0) = \frac{1}{2}(a-|a|) \tag{2.11}$$

where a and $|a|$ also satisfy

$$a = a^+ + a^-$$
$$|a| = a^+ - a^- \tag{2.11'}$$

the C.I.R. scheme (2.8) now becomes

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}[a^+(u_j^n-u_{j-1}^n)+a^-(u_{j+1}^n-u_j^n)] \tag{2.12}$$

or by (2.11)

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}[(a+|a|)(u_j^n-u_{j-1}^n)+(a-|a|)(u_{j+1}^n-u_j^n)]$$

i.e.

$$(u_j^{n+1})_{C.I.R.} = u_j^n - \frac{a\Delta t}{2\Delta x}(u_{j+1}^n-u_{j-1}^n) + \frac{|a|\Delta t}{2\Delta x}(u_{j+1}^n-2u_j^n+u_{j-1}^n) \tag{2.13}$$

Therefore the C.I.R. scheme appears as a first order scheme consistent with (2.1) but also containing a second order term

$$\frac{|a|\Delta t\cdot\Delta x}{2}\cdot\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{(\Delta x)^2} \tag{2.14}$$

which can be interpreted as a <u>numerical viscosity</u>, with a sign which does not change with the direction of propagation of the wave solution, so that its effect always tends to "slow down" the propagation of the physical wave satisfying (2.1), whether it be to the right or to the left side of the x-axis.

The "switch" between A and C , for the second interpolation node used in (2.8) and fig. 1 to define $u(I) = u(M)$, is performed by the absolute value sign in (2.13): if $a > 0$ ($a<0$) the interpolation nodes are A , B (B,C) .

We shall now make two comments on the C.I.R. scheme.

(i) The (C.F.L.)-stability condition (2.7) has the following physical meaning (see fig. 1): the "numerical characteristic" IM should not start from a point I located entirely outside of the domain of dependence of the solution at the new point M ; since $u(M) = u(x_j,(n+1)\Delta t) = u(x_j-a\Delta t,n\Delta t) \equiv u(I)$ (exactly), point I should belong to the (closed) interval AB (BC if $a<0$) of points beeing used in the computation of the numerical approximation of u at point M ; for this to be true we must have

$$\text{slope AM} \leq \text{slope IM} \equiv \frac{1}{a} \quad \text{by definition of point} \quad I \equiv (x_j-a\Delta t,t^n)$$

therefore

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{a} \quad \text{i.e. the C.F.L. condition} \tag{2.7}$$

One says that the numerical domain of dependence must contain points from the physical domain of dependence, or more precisely it must contain the physical domain of dependence.

(ii) The C.I.R. scheme can also be written as a three-point scheme in the form

$$u_j^{n+1} = (1-|a|\lambda)u_j^n + (\frac{a\lambda}{2}+\frac{|a|\lambda}{2})u_{j-1}^n - (\frac{a\lambda}{2}-\frac{|a|\lambda}{2})u_{j+1}^n \tag{2.15}$$

i.e.

$$u_j^{n+1})_{C.I.R.} = a^+\lambda u_{j-1}^n + (1-|a|\lambda)u_j^n - a^-\lambda u_{j+1}^n \tag{2.16}$$

So if the C.F.L. stability condition $|a|\lambda \leq 1$ is satisfied, all three coefficients in the right member are non-negative. The C.I.R. scheme is then said to be "monotonicity preserving" (see for instance [21] p. 41, [64]); it transforms a monotonic grid-function $\{u_j^n\}$ into a monotonic function $\{u_j^{n+1}\}$. From (2.16) we can easily prove the stability of the C.I.R. scheme ([21], p. 42); we have (all coefficients beeing ≥ 0 under the C.F.L. condition (2.7))

$$|u_j^{n+1}| \leq (a^+\lambda+(1-|a|\lambda)-a^-\lambda)\max|u_j^n| \tag{2.16}$$

and therefore

$$\max_j|u_j^{n+1}| \leq \max_j|u_j^n| \tag{2.17}$$

since $a^+ - a^- = |a|$, which means stability in the maximum norm. This property holds for all linear monotonicity preserving schemes

$$\{u_j^n\} \to \{Lu_j^n\} = \sum_{\ell=-k}^{+k} c_\ell u_{j+\ell}^n \quad \text{which satisfy} \quad c_\ell \geq 0\ ,\ \sum_{-k}^{+k} c_\ell = 1\ . \tag{2.17'}$$

Although monotone schemes and linear monotonicity preserving schemes play an essential role in this theory, they are only first order accurate ([19], [24], [64]) and one of the main goals of current research is to design higher order accurate schemes with some kind of monotonicity (see Section 4, and [64], [32], [15], [50]). We shall examine some nonlinear, second order accurate monotonicity preserving schemes in Sections 5, 6, 7.

If we now turn to the linear system (2.10) with constant coefficient matrix $A = (a_{ij})$, we can extend the C.I.R. scheme (2.8) by considering its equivalent form (2.13). The hyperbolicity of system (2.10) means that there exists a similarity transformation

$$A \longrightarrow T^{-1}AT = \Lambda = (a_i\delta_{ij}) \tag{2.18}$$

which diagonalizes A ; the a_i are the eigenvalues of A , which are real. Introducing the characteristic variables w^i $(i = 1,\ldots,m)$ by

$$w = (w^i) = T^{-1}u \tag{2.19}$$

reduces (2.10) to the diagonal (canonical) form

$$w_t + \Lambda w_x = 0 \quad \text{(characteristic equivalent of (2.10))} \tag{2.20}$$

and we can easily apply the C.I.R. scheme (2.13) to each of the obtained decoupled equations: introducing first the matrix $|\Lambda| = \text{diag}[|a_1|,\ldots,|a_m|]$, i.e. $|\Lambda|_{ij} \equiv |a_i|\delta_{ij}$, we write

$$w_j^{n+1} = w_j^n - \frac{\Delta t}{2\Delta x}\Lambda(w_{j+1}^n - w_{j-1}^n) + \frac{\Delta t}{2\Delta x}|\Lambda|(w_{j+1}^n - 2w_j^n + w_{j-1}^n) . \tag{2.21}$$

Returning to the original variables u_i , (2.21) can now be written

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}A(u_{j+1}^n - u_{j-1}^n) + \frac{\Delta t}{2\Delta x}|A|(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \tag{2.21'}$$

with $|A| \equiv T|\Lambda|T^{-1}$. We shall have stability if each individual component of this scheme satisfies the C.F.L. condition, and therefore if

$$(\max_i |a_i|)(\frac{\Delta t}{\Delta x}) \leq 1 \tag{2.7'}$$

Let us notice that the use of scheme (2.21') requires the generally tedious process of diagonalizing the matrix A . This will be a source of imprecision for big systems (m large) but in the case of the inviscid gas dynamics equations, the eigenvalues a_i and the similarity transformation matrix T are easily obtained and allow for an easy implementation, as we shall now see.

B. The Euler equations of ideal compressible flow

In many situations the mathematical description of compressible flow given by the Navier-Stokes equations, can be greatly simplified by neglecting the kinematic viscosity $\nu = \mu/\rho$ of the gas and considering the one-dimensional Euler equations of ideal (inviscid) compressible flow; in conservative dependent variables, they take the form of a hyperbolic system of conservation laws $U_t + F(U)_x = 0$ where

$$U = \begin{pmatrix} \rho \\ \rho u \\ e \end{pmatrix} \equiv \begin{pmatrix} \rho \\ m \\ e \end{pmatrix} \quad \text{and} \quad F(U) = \begin{pmatrix} \rho u \\ p+\rho u^2 \\ (e+p)u \end{pmatrix} = \begin{pmatrix} m \\ p + \frac{m^2}{\rho} \\ (e+p)\frac{m}{\rho} \end{pmatrix} \tag{2.22}$$

Here ρ, u and p are the "primitive" (or physical) variables: density, velocity and pressure, e is the total energy per unit volume, given by

$$e = \rho\varepsilon + \rho\frac{u^2}{2} \quad (= \rho\varepsilon + \frac{m^2}{2\rho} \text{ in conservative variables}) \tag{2.23}$$

where ε is the specific internal energy (per unit mass), related to p and ρ by an equation of state

$$p = p(\rho,\varepsilon) \,. \tag{2.24}$$

Until here F does not appear to be only a function of the conservative variables ρ, ρu, e, but let us now make use of the fact that for a perfect gas

$$\varepsilon = \frac{1}{\gamma-1}\frac{P}{\rho} \quad \text{or} \quad P = (\gamma-1)\rho\varepsilon \tag{2.25}$$

where $\gamma = C_p/C_v$, the ratio of specific heats. This enables us to eliminate p and ε from (2.22) with the help of (2.23) and (2.25):

$$p = (\gamma-1)(e - \frac{m^2}{2\rho}) \ . \tag{2.26}$$

Introducing this in (2.22), we can express the flux vector $F(U)$ as

$$F(U) = \begin{pmatrix} m \\ (\gamma-1)e + (3-\gamma)m^2/2\rho \\ \gamma em/\rho - (\gamma-1)m^3/(2\rho^2) \end{pmatrix} . \tag{2.27}$$

The jacobian matrix A takes the form

$$A \equiv \frac{\partial F}{\partial U} = \begin{pmatrix} 0 & 1 & 0 \\ (\gamma-3)u^2/2 & (3-\gamma)u & \gamma-1 \\ (\gamma-1)u^3 - \gamma eu/\rho & \gamma e/\rho - 3(\gamma-1)u^2/2 & \gamma u \end{pmatrix} \tag{2.28}$$

and its eigenvalues are

$$a_1 = u \ , \ a_2 = u + c \ , \ a_3 = u - c \tag{2.29}$$

where $c = (\frac{\partial p}{\partial \rho})^{1/2} = (\frac{\gamma p}{\rho})^{1/2}$ is the local speed of sound, a function of the dependent variables ρ , m , e . These eigenvalues can be all positive or all negative (for supersonic flow, i.e. if $|u| > c$), or of different signs if $|u| < c$ (subsonic flow). In the latter case, our preliminary study shows that no globally upwinded difference scheme like the direct extension of (2.6):

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x} A(U_j^n - U_{j-1}^n) \tag{2.6'}$$

could be stable. What we need is an extension of the switching device built in the C.I.R. scheme (2.21) - (2.21') to the case of non-linear flux vectors $F(U)$. We shall see that for the gas dynamics equations, a property of homogeneity of $F(U)$ allows a straightforward generalization of (2.21'). But at the present

time we can only observe that a mere extension of (2.13) to the non-linear system (1) requires separate treatment of the positive eigenvalues of $A = \partial F/\partial U$ and the negative ones, and therefore of the corresponding parts F^+ and F^- of the nonlinear flux vector.

C. *Steger and Warming's flux splitting for the Euler equations*

To extend the C.I.R. scheme (21') to the quasi-linear hyperbolic system of the Euler equations of gas dynamics (ideal compressible flow), Steger and Warming [55] have designed an ingenious method of "flux splitting" and made use of a special property of the Euler equations

$$U_t + F(U)_x = 0 \tag{1.1}$$

with U, $F(U)$ given by (2.22).

For these equations, the flux vector $F(U)$ is a homogeneous function of degree one of the dependent variable vector U, so that $F(\alpha U) = \alpha F(U)$ for any real number α, provided that the equation of state of the gas is of the form

$$p = \rho \cdot \varphi(\varepsilon) \tag{2.30}$$

(p: pressure, ρ: density, ε: specific internal energy i.e. per unit mass).

In this case, Euler's theorem (see Courant, Differential Calculus, Volume II) shows that

$$F = AU \equiv \left(\frac{\partial F}{\partial U}\right)U \ . \tag{2.31}$$

This property remains valid for the Eulerian equations in 2 and 3 space dimensions:

$$U_t = \sum_{i=1}^{3} F_i(U)_{x_i} \quad \text{with} \quad F_i = A_i U \equiv \left(\frac{\partial F_i}{\partial U}\right)U \qquad 1 = 1,2,3 \ . \tag{2.31'}$$

Using the assumed homogeneity and the hyperbolicity of our original system (1.1), one can split the flux vector F in the following manner. Using (2.18) and (2.31) one can write

$$F = AU = (T\Lambda T^{-1})U \quad \text{with} \quad \Lambda = \text{diag}[a_1,\ldots,a_m] \tag{2.32}$$

and since every eigenvalue a_i of A can be represented as

$$a_i = a_i^+ + a_i^- \tag{2.33}$$

with

$$a_i^+ = \tfrac{1}{2}(a+|a|) \ , \ a_i^- = \tfrac{1}{2}(a-|a|) \tag{2.33'}$$

one can split the diagonal matrix Λ into its positive and negative parts:

$$\Lambda = \Lambda^+ + \Lambda^- \tag{2.34}$$

where

$$\Lambda^+ = \text{diag}[a_i^+] \quad \text{and} \quad \Lambda^- = \text{diag}[a_i^-] \equiv \begin{pmatrix} a_1^- & & 0 \\ & \ddots & \\ 0 & & a_m^- \end{pmatrix} \tag{2.35}$$

This leads to

$$F = T(\Lambda^+ + \Lambda^-)T^{-1}U \tag{2.36}$$

and therefore to the flux splitting

$$F = A^+U + A^-U \equiv F^+ + F^- \tag{2.36'}$$

where A^+ , A^- are defined by

$$A^+ \equiv T\Lambda^+T^{-1} \ , \ A^- \equiv T\Lambda^-T^{-1} \tag{2.36"}$$

and

$$F^+ = A^+U \ , \ F^- = A^-U \tag{2.36'}$$

with

$$A = A^+ + A^- \ . \tag{2.37}$$

Now the eigenvalues of A^+ are ≥ 0 , those of A^- are ≤ 0 , meaning that we can now handle each of these matrices and corresponding fluxes by an appropriate upwinding scheme just as the C.I.R. scheme did for equation (2.1). The matrix T has been computed by R.F. Warming and R.M. Beam [60]; see also [58]. Using the splitting $a_i = a_i^+ + a_i^-$ with (2.33') we obtain

$$\begin{aligned} a_1^+ &= \frac{u+|u|}{2} \ , \ a_1^- = \frac{u-|u|}{2} \\ a_2^+ &= \frac{u+c+|u+c|}{2} \ , \ a_2^- = \frac{u+c-|u+c|}{2} \\ a_3^+ &= \frac{u-c+|u-c|}{2} \ , \ a_3^- = \frac{u-c-|u-c|}{2} \ . \end{aligned} \tag{2.38}$$

Steger and Warming give the corresponding flux vectors F^+ and F^- , for the particular case of subsonic flow where $0 \leq u \leq c$, as

$$F^+ = \frac{\rho}{2\gamma}\begin{pmatrix} 2\gamma u + c - u \\ 2(\gamma-1)u^2 + (u+c)^2 \\ (\gamma-1)u^3 + \frac{(u+c)^3}{2} + \frac{(3-\gamma)(u+c)c^2}{2(\gamma-1)} \end{pmatrix} \tag{2.39}$$

$$F^- = \frac{\rho}{2\gamma}\begin{pmatrix} u-c \\ (u-c)^2 \\ \frac{(u-c)^3}{2} + \frac{(3-\gamma)(u-c)c^2}{2(\gamma-1)} \end{pmatrix} \qquad (2.39')$$

and naturally, if the flow is <u>supersonic</u> (towards x^+) i.e. if $u > c$

$$F^+ = F \ , \ F^- = 0 \ . \qquad (2.39")$$

On the basis of this splitting of the flux vector, a variety of useful numerical schemes can be obtained for the Euler equations of gas dynamics, both explicit and implicit.

D. <u>*Explicit direct extension of the C.I.R. scheme for the gas dynamics equation*</u>

This is Steger and Warming's first scheme:

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}[\Delta_-(F_j^+)^n + \Delta_+(F_j^-)^n] \qquad (2.40)$$

where

$$\Delta_- F_j \equiv F_j - F_{j-1} \ , \ \Delta_+ F_j \equiv F_{j+1} - F_j \ , \ F_j^n \equiv F(U_j^n) \ . \qquad (2.40')$$

This scheme, the direct explicit extension of the C.I.R. scheme to the gas dynamics equations, can be written in "conservation form" (Harten, Lax, van Leer [27]),

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}(f_{j+1/2}^n - f_{j-1/2}^n) \qquad (2.41a)$$

if we define

$$f^n_{j+1/2} \equiv f(U^n_j, U^n_{j+1}) \equiv (F^n_j)^+ + (F^n_{j+1})^- \equiv [F(U^n_j)]^+ + [F(U^n_{j+1})]^- \ . \qquad (2.41b)$$

Here $f^n_{j+1/2}$ is the "numerical flux" vector, depending on F , U^n_j , U^n_{j+1} ; it should satisfy the consistency condition

$$f(U,U) = F(U) \qquad (2.41c)$$

which is readily verified:

$$f(U,U) = [F(U)]^+ + [F(U)]^- \equiv F(U) \ .$$

As the C.I.R. scheme, Steger and Warming's scheme (2.40) can be written as a centered scheme, if we define

$$\tilde{F} = |A|U \quad \text{with} \quad |A| = T|\Lambda|T^{-1} \qquad (2.42)$$

and

$$|\Lambda| = \text{diag}[|a_1|,\ldots,|a_n|]$$

as we had done earlier for linear system (2.10) with constant matrix A , and eigenvalues a_i . Setting

$$F^+ = \tfrac{1}{2}(F+\tilde{F}) \ , \ F^- = \tfrac{1}{2}(F-\tilde{F}) \qquad (2.43)$$

we obtain the centered form of Steger and Warming's scheme:

$$U^{n+1}_j = U^n_j - \tfrac{\lambda}{2}(F^n_{j+1}-F^n_{j-1}) + \tfrac{\lambda}{2}(\tilde{F}^n_{j+1}-2\tilde{F}^n_j+\tilde{F}^n_{j-1}) \ . \qquad (2.44)$$

For the C.I.R. scheme, the second order difference term, which plays the role of a numerical viscosity ("dissipative term"), was

$$+ \tfrac{\lambda}{2}\,|A|(u^n_{j+1}-2u^n_j+u^n_{j-1}) \ . \qquad (2.45)$$

For Steger and Warming's scheme, this dissipative term is

$$\frac{\lambda}{2}(\tilde{F}_{j+1}^{n}-2\tilde{F}_{j}^{n}+\tilde{F}_{j-1}^{n}) \qquad (2.46)$$

which cannot be reduced to the form (2.45) since the matrix $|A|$, in the flux vector $\tilde{F}_j^n = |A|U_j^n$, is no longer constant.

Steger and Warming's scheme, which is first-order accurate in space and time, is <u>stable</u> for linear stability theory (i.e. if the coefficients of $A(U)$ are locally "frozen") (and this guarantees that the numerical solution remains bounded whenever the exact solution is bounded) if and only if the C.F.L. - like condition

$$|a_i^{\pm}|\lambda \leq 1 \quad \lambda \equiv \Delta t/\Delta x \qquad (2.47)$$

is satisfied by all eigenvalues $a_i = a_i^+ + a_i^-$ of $A = \partial F/\partial U$.

To give an idea of the precision level of this scheme, we shall reproduce here Steger and Warming's results for the so-called shock-tube problem (see Sod [53]; [55] for a description and the exact initial values of the parameters p, ρ; here $p_L/p_R = 10$, $\lambda = \Delta t/\Delta x = 0.4$ and the C.F.L. number was $\nu \cong 0.95$, i.e. below stability limit). The continuous curve on fig. 2.2 shows the exact solution (see [21] for instance for a description of the solution procedure), and the circles are numerical values obtained with Steger and Warming's first order explicit scheme (2.40). From left to right, one encounters a constant state ($U = U_L$) a rarefaction wave (not too badly reproduced), a constant state (completely masked or "missed"), a contact discontinuity (fully "smeared"), a constant state (rather acceptable), a shock wave (smeared) and a constant state ($U = U_R$, well aproximated)

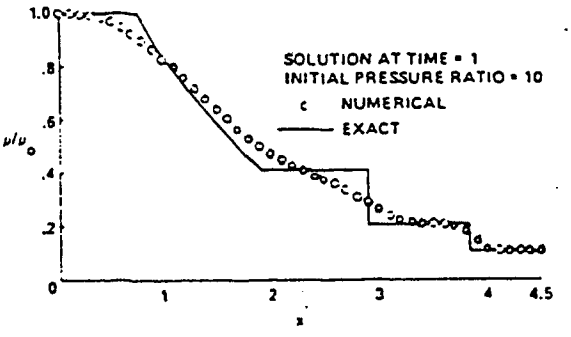

Fig. 2.2 Shock-tube problem with Steger-Warming's first order explicit upwind (flux-splitting) scheme (Courtesy of Steger and Warming, J. Comp. Phys. (1981))

These rather disappointing results suggest that despite the desirable monotonicity, obviously achieved by Steger and Warming's scheme which inherited the C.I.R.'s property of monotonicity preservation, we are still in need of a more accurate scheme, should we ever attempt resolution of the real 3 dimensional problems of compressible flow past physically relevant obstacles (wing profiles, engine nacelles, etc.). In the next section we shall review a few prototypes of second order accurate schemes, and explain why they also call for some improvement.

3. *THE SECOND ORDER ACCURATE SCHEMES OF LAX AND WENDROFF, MAC CORMACK, WARMING AND BEAM*

For the linear scalar wave equation $u_t + au_x = 0 \quad (a>0)$, we can apply Taylor expansions to the exact solution at gridpoints (x_j,t^n):

$$u(x_j,t^{n+1}) = u(x_j,t^n) + \Delta t(\frac{\partial u}{\partial t})^n_j + \frac{(\Delta t)^2}{2}(\frac{\partial^2 u}{\partial t^2})^n_j + O(\Delta t^3) \ . \qquad (3.1)$$

Using the differential equation, we have $u_t = -au_x$,

$u_{tt} = (-au_x)_t = (-au_t)_x = a^2u_{xx}$; and introducing this in (3.1) we get, after

differencing in space and replacing $u(x_j,t^n)$ by the numerical approximation u_j^n

$$u_j^{n+1} = u_j^n - \frac{a\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n) + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \qquad (3.2)$$

which is the Lax-Wendroff scheme ([29]) for the linear convection /wave equation.

Extending this to the nonlinear hyperbolic system (1.1) leads to the classical Lax-Wendroff scheme:

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{2\Delta x}[F(U_{j+1}^n) - F(U_{j-1}^n)] + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 [A_{j+1/2}^n \{F(U_{j+1}^n) - F(U_j^n)\} - A_{j-1/2}^n \{F(U_j^n) - F(U_{j-1}^n)\}] \qquad (3.3)$$

in which $A_{j+1/2}^n \equiv A\left(\frac{U_j^n + U_{j+1}^n}{2}\right)$ (A evaluated at $(U_j^n + U_{j+1}^n)/2$) .

The second order term comes from repeated use of the differential equation:

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial U}{\partial t}\right) = \frac{\partial}{\partial t}\left(-\frac{\partial F}{\partial x}\right) = \frac{-\partial}{\partial x}\left(\frac{\partial F}{\partial t}\right) = -\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial U}\frac{\partial U}{\partial t}\right) = +\frac{\partial}{\partial x}\left(A\frac{\partial F}{\partial x}\right) .$$

For a constant coefficient system $U_t + AU_x = 0$, the Lax-Wendroff scheme takes the simpler form

$$U_j^{n+1} = U_j^n - \frac{A\Delta t}{2\Delta x}(U_{j+1}^n - U_{j-1}^n) + \frac{1}{2}\left(\frac{A\Delta t}{\Delta x}\right)^2 (U_{j+1}^n - 2U_j^n + U_{j-1}^n) . \qquad (3.4)$$

To avoid time-consuming matrix multiplications, Richtmyer has designed a two-step version of the Lax-Wendroff scheme:

$$U_{j+1/2}^{n+1/2} = \frac{1}{2}(U_j^n + U_{j+1}^n) - \frac{\Delta t}{2\Delta x}(F_{j+1}^n - F_j^n) \qquad (3.5a)$$

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}(F_{j+1/2}^{n+1/2} - F_{j-1/2}^{n+1/2}) \qquad (3.5b)$$

2 step Lax-Wendroff, or Richtmyer scheme

All versions of the L.W. scheme are second order accurate ([47]), as the derivation

of (3.2), (3.4) shows. In the linear case where $F(U) = AU$, A constant matrix, the Richtmyer scheme reduces to the L.W. scheme (3.4); in this linear case, stability is obtained if

$$(\max_i |a_i|)\frac{\Delta t}{\Delta x} \leq 1 \qquad \text{(C.F.L.)} \qquad (3.6)$$

To give an idea of the problems associated with the Lax-Wendroff scheme, we reproduced calculations indicated in [21] for the initial value problem

$$u_t + u_x = 0$$

$$u(x,0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases} \qquad (3.7)$$

performed with (3.2) and $a\lambda = a\frac{\Delta t}{\Delta x} = 0.1$ (well below the C.F.L. stability limit; here $a = 1$). The results reveal the most important drawback of the Lax-Wendroff scheme, which tends to generate oscillations in the neighbourhood of discontinuities.

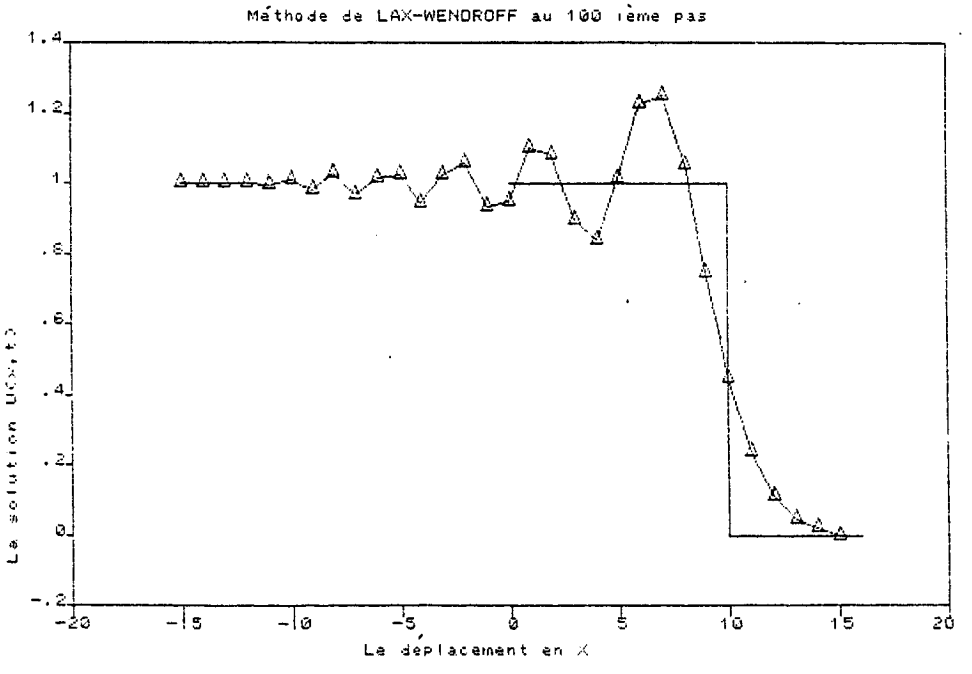

Fig. 3.1 The Lax-Wendroff scheme $u_t + u_x = 0$

at time $t = 100\Delta t$

One well established remedy against this difficulty is the introduction, on the right-hand side of scheme (3.3), of von Neumann and Richtmyer's artificial viscosity ([40], [29]). For $U_t + F(U)_x = 0$, Lax and Wendroff's artificial viscosity consists in a second order difference-type term

$$+ \frac{\Delta t}{2\Delta x}[Q_{j+1/2}\cdot\Delta U_{j+1/2} - Q_{j-1/2}\cdot\Delta U_{j-1/2}] \tag{3.8}$$

which is added to the right-hand side of (3.3), where $Q_{j+1/2} = Q_{j+1/2}(U_j, U_{j+1})$ is a matrix taylored to give appropriate dissipative action when there is strong variation between the adjacent states U_j , U_{j+1} , and to have negligible effect when the gradient is small: in this manner the oscillations are broken near the shocks, while second-order accuracy is maintained everywhere else. For the scalar conservation equation $u_t + f(u)_x = 0$, the Lax-Wendroff artificial viscosity takes the form

$$\frac{\Delta t}{2\Delta x}[\kappa|a_{j+1} - a_j|\Delta u_{j+1/2} - \kappa|a_j - a_{j-1}|\Delta u_{j-1/2}] \tag{3.9}$$

where $a_j = a(u_j) = (df/du)_j$, and κ is a dimensionless constant of the order of 1 . Although this leads to an identically vanishing artificial viscosity term in the constant coefficient case $(a(u) = a = \text{const.})$, and therefore cannot bring any improvement for problem (3.7), it has been found to lead to much better shock profiles in many nonlinear applications; nevertheless, the parameter κ requires a special adjustment in each particular case (see [47] for a detailed discussion), which is not desirable for engineering design computations.

In 1969, MacCormack presented another 2-step, second order accurate variant of the Lax-Wendroff scheme, which takes the following form for $U_t + F(U)_x = 0$:

(a) Predictor step $\overline{U_j^{n+1}} = U_j^n - \lambda\Delta F_{j+1/2} \equiv U_j^n - \lambda(F_{j+1} - F_j)$ (3.10)

(downwind if all a_i's are ≥ 0) .

(b) Corrector step

$$U_j^{n+1} = \frac{1}{2}(U_j^n + \overline{U_j^{n+1}}) - \frac{\lambda}{2}\Delta \overline{F_{j-1/2}^{n+1}} = \frac{1}{2}(U_j^n + \overline{U_j^{n+1}}) - \frac{\lambda}{2}(\overline{F_j^{n+1}} - \overline{F_{j-1}^{n+1}}) \tag{3.10}$$

(upwind if all a_i's are ≥ 0) .

This method generally yields better results than the L.W. or Richtmyer schemes, while enjoying the same advantage as the latter as regards operation counts. For a comparison of these and van Leer's method, see [54]. Since the tendency to generate oscillations near the shocks is still present, we shall give a detailed description of an upwind-type MacCormack scheme defined and tested in [55].

Instead of (3.10), Steger and Warming start from the following version of the MacCormack scheme:

$$\text{Predictor} \quad \overline{U_j^{n+1}} = U_j^n - \lambda \nabla_x F_j^n \quad \text{(upwind if all } a_i\text{'s} \geq 0) \tag{3.11a}$$

$$\text{Corrector} \quad U_j^{n+1} = \frac{1}{2}(U_j^n + \overline{U_j^{n+1}}) - \frac{\lambda}{2}\Delta_x \overline{F_j^{n+1}} \quad \text{(downwind if all } a_i\text{'s} \geq 0) \tag{3.11b}$$

(here $\nabla_x U_j = U_j - U_{j-1}$, $\Delta_x U_j = U_{j+1} - U_j$, $\delta_x U_j = U_{j+1/2} - U_{j-1/2}$) . Notice that (3.11) can be written

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}[\nabla_x F_j^n + \Delta_x \overline{F_j^{n+1}}] \tag{3.11c}$$

thus revealing the nearly trapezoidal nature of MacCormack's scheme (with respect to time integration). The corrector (3.11b) can be put in the form

$$U_j^{n+1} = \frac{1}{2}(U_j^n + \overline{U_j^{n+1}}) - \frac{\lambda}{2}[(\overline{F_{j+1}^{n+1}} - \overline{F_j^{n+1}}) - (\overline{F_j^{n+1}} - \overline{F_{j-1}^{n+1}}) + (\overline{F_j^{n+1}} - \overline{F_{j-1}^{n+1}})]$$

$$= \frac{1}{2}(U_j^n + \overline{U_j^{n+1}}) - \frac{\lambda}{2}[F(U_{j+1}^n - \lambda \nabla_x F_{j+1}^n) - 2F(U_j^n - \lambda \nabla_x F_j^n) + F(U_{j-1}^n - \lambda \nabla_x F_{j-1}^n)] - \frac{\lambda}{2}\nabla_x \overline{F_j^{n+1}}$$

which becomes, by Taylor's expansion, to within third-order differences:

$$U_j^{n+1} = \frac{1}{2}(U_j^n + U_j^{\overline{n+1}}) - \frac{\lambda}{2}\nabla_x F_j^{\overline{n+1}} - \frac{\lambda}{2}\delta_x^2 F_j^n \qquad (3.12)$$

or, after shifting from x_j to x_{j-1} in the second-order difference:

$$U_j^{n+1} = \frac{1}{2}(U_j^n + U_j^{\overline{n+1}}) - \frac{\lambda}{2}\nabla_x F_j^{\overline{n+1}} - \frac{\Delta t}{2}\Delta x \cdot \frac{\nabla_x^2 F_j^n}{(\Delta x)^2} + O(\Delta x)^2 \ .$$

Neglecting these higher-order truncation errors leads to a completely backward form of the MacCormack scheme:

$$U_j^{\overline{n+1}} = U_j^n - \lambda \nabla_x F_j^n \qquad (3.13a)$$

$$U_j^{n+1} = \frac{1}{2}(U_j^n + U_j^{\overline{n+1}}) - \frac{\lambda}{2}\nabla_x F_j^{\overline{n+1}} - \frac{\Delta t \cdot \Delta x}{2}\frac{\nabla_x^2 F_j^n}{(\Delta x)^2} \qquad (3.13b)$$

modified, backward MacCormack scheme

(3.13) is fully upwind if all eigenvalues a_i are ≥ 0 .

Examining (3.12) or (3.13b) we see that the last term in both cases can be interpreted as a negative viscosity or anti-dissipative term. We shall see in Section 5 that the Lax-Wendroff scheme can also be interpreted as the superposition of a first-order, upwind-type scheme plus a second-order anti-dissipative term. It is this anti-diffusive term which is responsible for the oscillations observed near the shocks. Observing that in the general case (e.g. for subsonic flow governed by the Euler equations) the eigenvalues are of both signs, Steger and Warming, applying their flux splitting principle, deduced a fully upwind split-flux version of MacCormack's scheme, which we shall write as

$$U_j^{\overline{n+1}} = U_j^n - \lambda[\nabla_x (F_j^n)^+ + \Delta_x (F_j^n)^-] \qquad (3.14a)$$

$$U_j^{n+1} = \frac{1}{2}(U_j^n + U_j^{\overline{n+1}}) - \frac{\lambda}{2}[\nabla_x (F_j^{\overline{n+1}})^+ + \Delta_x (F_j^{\overline{n+1}})^-] - \frac{\lambda}{2}[\nabla_x^2 (F_j^n)^+ + \Delta_x^2 (F_j^n)^-] \qquad (3.14b)$$

This split-flux, second-order scheme gives a substantial improvement on Steger and Warming's first-order split-flux scheme (2.40): the anti-dissipative tendency of (3.13b) is damped by the flux-splitting, and the shock-tube problem leads to nearly monotonic density profiles, with some undershoot/overshoot, and a good improvement of the shock and rarefaction wave resolution. The contact discontinuity is still spread on too many grid-intervals. For comparison, we reproduce in figures 3.2 and 3.3 Steger and Warming's results for the classical MacCormack scheme (3.11), and the split-flux scheme (3.14). The oscillations, for the former, are beyond the acceptability threshold. A reasonable improvement can be obtained by introducing an appropriate artificial viscosity (see e.g. [46]).

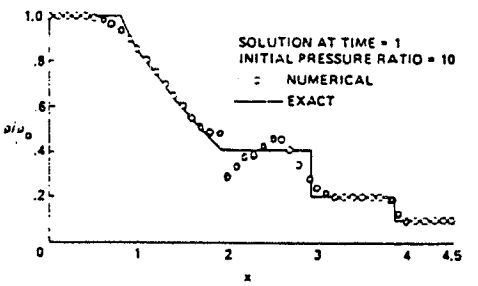

Fig. 3.2 Shock tube problem solved with MacCormack's explicit scheme (3.11) (Courtesy of Steger and Warming, J. Comp. Phys. (1981)).

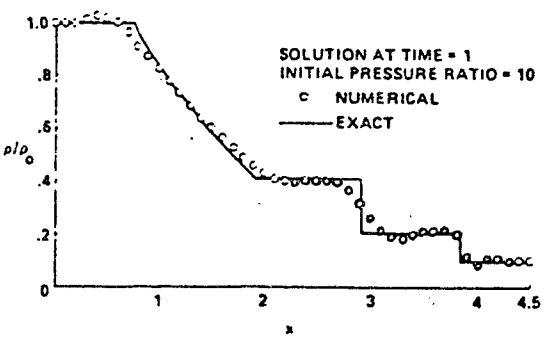

Fig. 3.3 Shock tube problem solved with the explicit second-order, fully upwind, flux-split version of MacCormack's method (Courtesy of Steger and Warming, J. Comp. Phys. (1981)).

The last second-order scheme we shall mention in this section is Beam and Warming's implicit scheme, motivated by stability considerations ([5], [60]); one simple form of this scheme is obtained by using the trapezoidal rule for the time integration

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{2}[\delta_x F_j^n + \delta_x F_j^{n+1}]$$

or, using the homogeneity property of the flux function (true for the Euler equations):

$$U_j^{n+1} + \frac{\Delta t}{2}\delta_x(A_j^{n+1}\cdot U_j^{n+1}) = U_j^n - \frac{\Delta t}{2}\delta_x(A_j^n \cdot U_j^n) \ .$$

Here δ_x is a discrete finite-difference approximation of the spatial differentiation operator $\frac{\partial}{\partial x}$. Subtracting $\frac{1}{2}\Delta t\delta_x F_j^n = \Delta t\delta_x(A^n U_j^n)/2$ on both sides gives

$$U_j^{n+1} + \frac{\Delta t}{2}\delta_x[A_j^{n+1}U_j^{n+1} - A_j^n U_j^n] = U_j^n - \Delta t\delta_x F_j^n$$

which is linearized, in Beam and Warming's method, into

$$U_j^{n+1} + \frac{\Delta t}{2}\delta_x[A_j^n(U_j^{n+1} - U_j^n)] = U_j^n - \Delta t\delta_x F_j^n$$

or equivalently

$$[I + \frac{\Delta t}{2}\delta_x A_j^n \cdot](U_j^{n+1} - U_j^n) = -\Delta t\delta_x F_j^n \quad \text{(Beam-Warming)} \tag{3.15}$$

Flux-splitting can also be introduced here, but for the shock tube problem, this did not give significant improvement on the upwind, flux-split version of MacCormack's scheme (3.14); in fact the shock and contact discontinuity are more smeared, and the only gain is a better rarefaction wave (see [55], fig. 2 and 5). The main advantage, though, lies in the unconditional stability of the Beam-Warming scheme, which is an advantage for steady-state computations by time-relaxation. Moreover, the Beam-Warming scheme leads to approximate factorization methods which are useful in 2 and 3 space dimensions.

Up to this point, we still have to improve on the accuracy of the shocks and rarefactions, since none of the above methods seems to give completely satisfactory results. In the next sections, we shall present several essential improvements, all founded on Godunov's and van Leer's concepts on monotonicity or Harten's idea of "Total Variation Diminishing" schemes.

4. *THE TVD-IDEA: MONOTONE, TVNI, AND MONOTONICITY PRESERVING SCHEMES*

To simplify the discussion we shall consider here a scalar conservation law

$$u_t + f(u)_x = 0 \quad \text{or} \quad u_t + a(u)u_x = 0 \quad (a(u) \equiv \frac{df(u)}{du}) \tag{4.1a}$$

with initial conditions

$$u(x,0) = u_0(x) \qquad -\infty < x < +\infty \, . \tag{4.1b}$$

The main reason for considering this single conservation law rather than system (1.1) is the following monotonicity property, valid for scalar conservation laws: For any weak solution (see Lax [30]) of the scalar conservation law (4.1), we have

(M1) No new local maximum or minimum can appear for $t > 0$

(M2) The value of a local maximum is non increasing, that of a local minimum is nondecreasing

and therefore

(M3) The total variation $TV[u(t)] \equiv \sup \sum_j |u(x_{j+1},t)-u(x_j,t)|$ is a nonincreasing function of time t .

The reason for this is the property of constancy of u, in the (x,t)-plane, along the integral curves of the ordinary differential equations

$$\frac{dx}{dt} = a(u)(x,t) \tag{4.2}$$

called the characteristic curves of equation (4.1) (indeed, following one such curve C we can write $(\frac{du}{dt})_C = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = u_t + a(u)u_x = 0$ by (4.1)). As the density profile in convergent divergent nozzles (see e.g. [21], [4]), velocity profiles in the shock tube problem ([53], [55]) or transonic flow past airfoils, all suggest, this property of monotonicity is no longer valid for hyperbolic systems of conservation laws. It is nevertheless a fundamental milestone in the quest for higher order accurate methods for these problems; a numerical scheme which generates large oscillations for the scalar law (4.1) is unlikely to behave better for the Euler equations.

To obtain numerical approximations of the solution of (4.1), let us now introduce explicit $(\ell+m+1)$-point finite difference schemes in conservation from (2.41a)

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2} - h_{j-1/2}) \equiv H(u_{j-\ell}^n, u_{j-\ell+1}^n, \ldots, u_j^n, \ldots, u_{j+m}^n) \tag{4.3a}$$

where

$$h_{j+1/2} \equiv h(u_{j-\ell+1}, u_{j-\ell+2}, \ldots, u_j, \ldots, u_{j+m}) \tag{4.3b}$$

is the so-called numerical flux function, satisfying a consistency condition

$$h(v,v,\ldots,v) = f(v) \quad \text{(value of the original flux function)} . \tag{4.3c}$$

We shall occasionaly use Godunov's shorthand notation, whenever no confusion is likely to happen:

$$u_j^n = u_j \ , \ u_j^{n+1} = u^j \tag{4.4}$$

thus concentrating on the single time step $t^n \to t^{n+1}$.

Denoting by $L : \{u_j^n\} \to \{u_j^{n+1}\} = L\{u_j^n\}$ the finite difference operator performing this time step, we say, following Harten ([64]) that the scheme (4.3) is total variation nonincreasing (TVNI) if for all function v of bounded total variation we have

$$TV(L \cdot v) \leq TV(v) \tag{4.5}$$

where the total variation means the total discrete (grid-dependent) variation

$$TV(v) \equiv TV\{v_j\} = \sum_{j=-\infty}^{+\infty} |v_j - v_{j-1}| \equiv \sum_{j=-\infty}^{\infty} |\Delta v_{j-1/2}| \tag{4.6}$$

A finite difference scheme (4.3) is said to be monotone if the function H is a monotone nondecreasing function of each of its $(\ell+m+1)$ arguments; it is called monotonicity preserving if the transition from $\{u_j^n\}$ to $\{u_j^{n+1}\} = L\{u_j^n\}$ preserves the sense of variation of the mesh function $\{u_j^n\}$, i.e. if

$$\operatorname{sgn}(Lu_j^n - Lu_{j-1}^n) \equiv \operatorname{sgn}(u_j^{n+1} - u_{j-1}^{n+1}) = \operatorname{sgn}(u_j^n - u_{j-1}^n) \quad \text{for all } j . \tag{4.7}$$

The relation between the above three properties is given by the following Theorem 4.1 (cf. Harten [64])

(i) a monotone scheme (4.3) is TVNI

(ii) a TVNI scheme is monotonicity preserving

We note here that the introduction of schemes in conservation form, due to Godunov ([19], p. 286) and elaborated in [29], leads to numerical solutions which, if convergent, automatically define a weak solution. If the numerical scheme also satisfies an appropriate "entropy condition" (see [30]), then the limit of the numerical solutions verifies Oleinik's entropy conditions, which guarantees

unicity and convergence to the physical solution. We shall not elaborate on this problem here, and refer to [30], [13], [42], [43] and the literature quoted there.

On the other hand, monotonicity and bounds on the total variation are useful to prove convergence results directly, without assuming satisfaction of an entropy condition (For initial data $u_0(x)$ with bounded total variation, convergence to the unique entropy-satisfying weak solution is obtained by compactness arguments, starting from the observation that $TV\{u_j^n\}$ is a non-increasing function of time t^n, for a monotone scheme, by property (i) of Theorem 4.1.)

As mentioned in Harten [64], and earlier in [24], a monotone scheme (4.3) for $u_t + f(u)_x = 0$ approximates with second-order accuracy the solution of the viscous modified equation

$$u_t + f(u)_x = \Delta t[\beta(u,\lambda)u_x]_x \tag{4.8a}$$

where

$$\beta(u,\lambda) \equiv \frac{1}{2\lambda^2}\Big[\sum_{k=-\ell}^{+m} k^2\Big(\frac{\partial H}{\partial v_k}\Big)(v_{-\ell},v_{-\ell+1},\ldots,v_0,\ldots,v_m) - \lambda^2 a^2(u)\Big] \tag{4.8b}$$

satisfies, for a monotone scheme,

$$\beta(u,\lambda) \geq 0\ ,\ \beta(u,\lambda) \not\equiv 0\ . \tag{4.8c}$$

Since the "numerical viscosity" term $\Delta t[\beta(u,\lambda)u_x]_x$ does not vanish identically, a monotone scheme can be no better than first order accurate. Moreover the same is true for linear monotonicity preserving schemes

$$u_j^{n+1} = \sum_{k=-m}^{+m} c_k u_{j+k}^n \tag{4.9}$$

Godunov ([19], p. 275) has proved that a scheme (4.9) transforms a monotone discrete

function $\{u_j^n\}$ into a monotone function $\{u_j^{n+1}\}$ if and only if all coefficients c_k are nonnegative.

(*PROOF*. (a) If $c_k \geq 0$ for every k and $\{u_j^n\}$ is increasing, then all differences $u_j^n - u_{j-1}^n$ are ≥ 0, therefore

$$u_j^{n+1} - u_{j-1}^{n+1} = \sum_{k=-m}^{m} c_k u_{j+k}^n - \sum_{-m}^{m} c_k u_{j-1+k}^n = \sum c_k (u_{j+k}^n - u_{j+k-1}^n) \geq 0 .$$

(b) Conversely, supposing that $c_{k_o} < 0$ leads, for the particular function

$$u_j^n = \begin{cases} 1 & \text{if } j \geq k_o \\ 0 & \text{if } j < k_o \end{cases}$$

to

$$u_o^{n+1} - u_{-1}^{n+1} = \sum_{-m}^{m} c_k (u_k^n - u_{k-1}^n) = c_{k_o} (u_{k_o}^n - u_{k_o-1}^n) = c_{k_o} < 0$$

and monotonicity preservation is violated.)

It follows from the definition and Godunov's result that a linear monotonicity preserving scheme (4.9) is necessarily monotone and therefore at most first-order accurate by (4.8). By Theorem (4.1), any linear TVNI scheme will also be first-order accurate. The only way around this obstacle is the design of nonlinear schemes which are second-order accurate and TVNI , or monotonicity preserving. For this, Harten's lemma ([64]) plays a fundamental role.

We start from a difference scheme in conservation from

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2} - h_{j-1/2}) = u_j^n - \lambda \left[(h_{j+1/2} - f(u_j^n)) + (f(u_j^n) - h_{j-1/2})\right] \quad (4.10)$$

and we first suppose, for simplicity, that $h_{j+1/2} = h(u_j^n, u_{j+1}^n)$ (three-point scheme) where h is a continuously differentiable function of both arguments.

We can write (4.10)

$$u_j^{n+1} = u_j^n - \lambda[\{(h(u_j,u_{j+1})-h(u_j,u_j)\}+\{h(u_j,u_j)-h(u_{j-1},u_j)\}]$$

$$= u_j^n - \lambda[(u_{j+1}-u_j)C_+(u_j,u_{j+1})+(u_j-u_{j-1})C_-(u_{j-1},u_j)]$$

by the mean value theorem. Defining

$$C_{j-1/2} \equiv \lambda C_-(u_{j-1},u_j) \ , \ D_{j+1/2} \equiv -\lambda C_+(u_j,u_{j+1})$$

we obtain a very useful form of our scheme (4.10):

$$\underline{u_j^{n+1} = u_j^n - C_{j-1/2}\Delta u_{j-1/2}^n + D_{j+1/2}\Delta u_{j+1/2}^n} \quad (\text{Harten's form of numerical scheme (4.3)}) \quad (4.11)$$

The basic tool of TVNI or TVD-scheme theory is then

LEMMA 4.1. (Harten, [64]) If the coefficients $C_{j-1/2}$, $D_{j+1/2}$ in a difference scheme written in Harten's form (4.11) satisfy

$$\begin{cases} C_{j-1/2} \geq 0 \ , \ D_{j+1/2} \geq 0 & (4.12a) \\ 0 \leq C_{j+1/2} + D_{j+1/2} \leq 1 & (4.12b) \end{cases}$$

then the scheme is TVNI (total variation nonincreasing).

This lemma is valid for any scheme of the form (4.11), for instance 5 point schemes. Harten has designed a method allowing any first-order, 3 point TVNI/TVD scheme to be converted into a second-order, 5 point TVNI/TVD scheme. The basic principle is the observation that, due to the invariance of the solution of (4.1) along characteristics, u(x,t) is independent of the particular flux function f(u) appearing in the differential equation $u_t + f(u)_x = 0$ (away from shocks), and only depends on the initial data function $u_o(x)$, provided the flux function f(u) satisfies the following initial characteristic velocity condition

(ICVC): The characteristic velocity $a_f(u) = a(u) = \frac{df}{du}$ takes the same initial values $a(u_o)$, at the initial time $t_o = 0$, for all members of the flux family $F = \{f\}$ under consideration. Notice that this property is improperly stated in [61] p. 113 line -8, where the ICVC is not assumed, since modifying the flux function f into $f + g$ with $[f'(u)+g'(u)]_{u_o} \neq f'(u)\big|_{u_o}$ i.e. with $g'(u_o) \neq 0$ does affect the solution: the characteristic curves are modified from $\frac{dx}{dt} = a(u_o) = f'(u_o)$ to $\frac{dx}{dt} = a(u_o) + g'(u_o) \neq a(u_o)$. Nevertheless the whole argumentation in [61] remains correct, thanks to the fact that $g(u) \to 0$ as $\Delta x \to 0$, and $g'(u_o) \to 0$.

Harten then propose a "Q-scheme", in conservation form:

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}-h_{j-1/2}) \tag{4.13a}$$

where (limiting our description to the explicit subcase of [61])

$$h_{j+1/2} = \frac{1}{2}[f_j+f_{j+1}-Q(a_{j+1/2})\Delta u_{j+1/2}] \tag{4.13b}$$

and Q is an appropriate function of λ and $a_{j+1/2}$ defined by

$$a_{j+1/2} = \begin{cases} \Delta f_{j+1/2}/\Delta u_{j+1/2} & \text{if } \Delta u_{j+1/2} \neq 0 \\ \left(\frac{df}{du}\right)_j \equiv a(u_j) & \text{if } \Delta u_{j+1/2} = 0 \end{cases} \tag{4.13c}$$

called the coefficient of numerical viscosity, for the obvious reason that the scheme becomes

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(f_{j+1}^n-f_{j-1}^n) + \frac{\lambda}{2}[Q(a_{j+1/2})\Delta u_{j+1/2}-Q(a_{j-1/2})\Delta u_{j-1/2}] \tag{4.14}$$

or

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(f_{j+1}^n-f_{j-1}^n) + \frac{\lambda}{2}Q(u_{j+1}^n-2u_j^n+u_{j-1}^n)$$

if Q were constant. The last term is a viscosity term if, as is the case under stability CFL-like conditions, $Q \geq 0$.

Choosing $Q(a_{j+1/2}) = |a_{j+1/2}|$ (resp. $\lambda(a_{j+1/2})^2$) leads to the extension, to the case of non-constant characteristic speed $a_{j+1/2}$, of the C.I.R. scheme (2.13), called G.C.I.R. (resp. to the Lax-Wendroff scheme).

Harten's method consists in observing, by truncation error analysis and comparison with the Lax-Wendroff scheme, that (4.13) gives a second-order accurate approximation to the solutions of the modified equation

$$u_t + f(u)_x = \Delta x(\sigma(a)u_x)_x \tag{4.15a}$$

with

$$\sigma(a) \equiv \frac{1}{2}Q(a) - \frac{\lambda}{2}a^2 \tag{4.15b}$$

In other words, (4.13) is a second-order approximation to solutions of

$$u_t + (f-g)_x = 0 \tag{4.16a}$$

with

$$g(u) \equiv \Delta x \cdot \sigma(a)u_x \tag{4.16b}$$

It can then be proved that applying the first-order scheme (4.13) to the modified equation

$$u_t + (f+g)_x = 0 \tag{4.17}$$

leads to numerical solutions which are, in the limit when $\Delta x \to 0$, asymptotically equal to second-order approximations of the equation

$$u_t + [(f+g)-g]_x = 0 \quad \text{i.e.} \quad u_t + f(u)_x = 0 \ . \tag{4.1}$$

HEURISTIC PROOF. Indeed, the characteristic field $\frac{dx}{dt} = f'(u) + g'(u)$ of the modified-flux equation tends to the characteristic field $\frac{dx}{dt} = f'(u)$ of the original equation $u_t + f(u)_x = 0$, as $\Delta x \to 0$, since $g(u) = O(\Delta x)$ and therefore $g'(u) \equiv \gamma(u) = O(\Delta x)$, and the initial ICVC condition is satisfied in the limit as $\Delta x \to 0$. Considering the differential equation (4.17) instead of $u_t + f(u)_x = 0$, we see by comparison with (4.15) that scheme (4.13) will give second-order approximations to solutions of the modified equation

$$u_t + (f+g)_x = \Delta x(\sigma(a+\gamma)u_x)_x \tag{4.18a}$$

where

$$a + \gamma \equiv f'(u) + g'(u) . \tag{4.18b}$$

Now (4.18a) can be written

$$u_t + f_x + g_x = \Delta x\{[\sigma(a) + \frac{\partial\sigma}{\partial a} \cdot \gamma + O(\Delta x)^2]u_x\}_x$$

$$= \Delta x[\sigma(a)u_x]_x + \Delta x[\frac{\partial\sigma}{\partial a} \cdot O(\Delta x)u_x]_x + O(\Delta x)^3 u_{xx}$$

since $\gamma = O(\Delta x)$, whence

$$u_t + f_x + g_x = g_x + O(\Delta x)^2 \quad \text{or} \quad u_t + f(u)_x = O(\Delta x)^2 . \tag{4.19}$$

Therefore the scheme (4.13) applied to the modified flux function $f + g$ will give second-order accurate approximations of the solution of the original equation (4.1).

Since the additional flux function g must be a differentiable function of u, the effective numerical flux, for the modified-flux equation (4.17), will be obtained by an interpolation-type, 4 point-formula ([25], [26]), thus transforming the original scheme (4.13) into a 5-point, second-order accurate scheme which can be proved to be TVNI/TVD, under Harten's specific assumptions for the

choice of the function Q, with the help of Harten's lemma 4.1. We refer the reader to [64], [61] for a more detailed description of this ingenious conversion from 1st-order to 2nd-order TVNI, explicit or implicit, conservation schemes.

These schemes are extended to hyperbolic systems of conservation laws (1.1) by first considering the case of constant coefficient hyperbolic systems

$$U_t + AU_x = 0 \tag{4.20}$$

and applying the above theory to each of the scalar decoupled conservation equations obtained for the characteristic variables as follows. By the assumed hyperbolicity, the eigenvalues $a^i (i = 1,\ldots,m)$ of A are real, there exists an orthonormal basis of right-eigenvectors R^i, and the matrix

$$R = (R^1,\ldots,R^i,\ldots,R^m) \tag{4.21}$$

is invertible, with

$$R^{-1}AR = \text{diag}[a_1,\ldots,a_m] \equiv \Lambda . \tag{4.22}$$

Defining the characteristic variables as

$$W = R^{-1}U \tag{4.23}$$

leads to a diagonal system of decoupled conservation laws

$$W_t + \Lambda W_x = 0 \quad \text{or} \quad \frac{\partial w^i}{\partial t} + a_i \frac{\partial w^i}{\partial x} = 0 \quad i = 1,\ldots,m \tag{4.24}$$

to which the previous TVD procedure can be applied.

For nonlinear systems (1.1), or $U_t + A(U)U_x = 0$, with nonconstant jacobian matrix A, we have mentioned earlier that the total variation no longer needs to be nonincreasing as t increases, and the direct extension of

the scalar TVD theory is not possible. Nevertheless, one can apply the principle of local freezing of the coefficients (see [61]) and obtain TVD-like schemes which behave very much like TVD-schemes do for the scalar equation as regards monotonicity. The numerical tests displayed in [64], [61] suggest that Harten's method is a very powerful tool, leading to a family of TVD schemes able to give both very sharp shock resolution and the desired monotonicity; see e.g. the very convincing piecewise monotone density profile obtained by both the explicit and the implicit TVD-like schemes in [61], p. 119, fig. 4.2, for the quasi-one-dimensional nozzle problem, as opposed to the oscillatory profile obtained with Beam and Warming's second-order implicit scheme, or the smeared shock computed with Steger and Warming's first order split-flux algorithm.

Before introducing Sweby's method of generating higher order schemes using flux limiters, let us mention as a first application of Harten's lemma that the C.I.R. scheme is TVNI under the C.F.L. condition $|a|\lambda = |a|\Delta t/\Delta x \leq 1$, since it can be written

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(|a|+a)\Delta u_{j-1/2}^n + \frac{\lambda}{2}(|a|-a)\Delta u_{j+1/2}^n \qquad (2.12')$$

and Harten's condition (4.12a) is obviously satisfied with

$$C_{j-1/2} = \frac{\lambda}{2}(|a|+a) \geq 0 \ , \ D_{j+1/2} = \frac{\lambda}{2}(|a|-a) \geq 0 \qquad (4.25)$$

while

$$C_{j+1/2} + D_{j+1/2} \equiv |a|\lambda \ .$$

On the other hand, the Lax-Wendroff scheme does not satisfy Harten's sufficient condition (4.12), and is not monotonicity preserving, as Godunov ([19], p. 274) has shown as follows: considering

$$u_t - u_x = 0 \tag{4.26a}$$

with discrete initial data

$$\begin{aligned} u_j &= 0 \quad \text{for} \quad j \le 0 \\ u_j &= 1 \quad \text{for} \quad j \ge 1 \end{aligned} \tag{4.26b}$$

The Lax-Wendroff scheme gives for the first time-step:

$$u_j^1 = u_j^0 + \frac{\lambda}{2}(u_{j+1}^0 - u_{j-1}^0) + \frac{\lambda^2}{2}(u_{j+1}^0 - 2u_j^0 + u_{j-1}^0) \tag{4.27}$$

therefore

$$u_j^1 = 0 \quad \text{for all} \quad j \le -1 \ , \ u_0^1 = \frac{\lambda+\lambda^2}{2} \ , \ u_1^1 = 1 + \frac{\lambda}{2} - \frac{\lambda^2}{2}$$

and

$$u_j^1 = 1 \quad \text{for} \quad j \ge 2 \ .$$

Now for $\lambda = \Delta t/\Delta x < 1$ i.e. for a strictly stable scheme, we have $u_0^1 < 1$, $u_1^1 > 1$ and $u_2^1 = 1$, and the monotonicity of the initial function is already broken at the first time step!

We note at this point that in Godunov's work (1959), written in 1956, the scheme (4.27) is naturally not quoted under the names of Lax and Wendroff, as the now famous paper [29] appeared in 1960. Godunov only considered the particular form (4.27), for the special case of the constant coefficient linear scalar wave equation (4.26a), to motivate his considerations on monotonicity preservation; [29] was the first analytical and general presentation, for the general nonlinear hyperbolic system (1.2), of what is now the LW scheme. It is interesting to note that monotonicity was at the heart of Godunov's motivation in his attempt to construct a scheme which would closely reproduce the solutions of the Euler equations.

5. *SECOND-ORDER SCHEMES WITH FLUX LIMITERS. SWEBY'S SCHEME*

A. *Sweby's construction of 2nd-order TVD schemes derived from the Lax-Wendroff scheme*

We consider the simple case of the linear scalar wave equation

$$u_t + au_x = 0 \ , \ a > 0 \ \text{(constant)} \ . \tag{5.1}$$

We can rewrite the Lax-Wendroff scheme (4.27), to enhance its anti-dissipative character, as

$$u_j^{n+1} = u_j^n - \lambda a(u_j^n - u_{j-1}^n) - \frac{(1-\nu)\nu}{2}[(u_{j+1}^n - u_j^n) - (u_j^n - u_{j-1}^n)] \tag{5.2}$$

or in Sweby's notation

$$u^j = u_j - \lambda a \Delta u_{j-1/2} - \lambda \Delta_-[\frac{(1-\nu)\nu}{2\lambda}\Delta u_{j+1/2}] \tag{5.3}$$

which can be written in conservation form

$$u^j = u_j - \lambda[(au_j + \frac{(1-\nu)\nu}{2\lambda}\Delta u_{j+1/2}) - (au_{j-1} + \frac{(1-\nu)\nu}{2\lambda}\Delta u_{j-1/2})] \ . \tag{5.4}$$

This last equation reveals that the numerical flux of the LW scheme

$$h_{j+1/2} = au_j + \frac{(1-\nu)}{2}a\Delta u_{j+1/2} = au_j + \frac{1-\nu}{2}\Delta(au)_{j+1/2} \tag{5.5}$$

can be viewed as the sum of the numerical flux au_j^n of the first-order upwind scheme (2.6), known to be highly dissipative, and an additional flux

$$\frac{(1-\nu)}{2}a\Delta u_{j+1/2} = \frac{1-\nu}{2}\Delta(au)_{j+1/2} \tag{5.6}$$

the effect of which is to introduce in the right-hand side a second-difference term with a coefficient $-a(1-\nu)/2$ or $-\nu(1-\nu)/2\lambda$ which is <u>negative</u> under the CFL stability condition $\nu \leq 1$ (since $a > 0$) , thus playing the role of a <u>negative viscosity</u>, or <u>anti-diffusion</u>, in the Lax-Wendroff scheme.

It is then tempting to try and construct some sort of intermediate between these two schemes, by introducing in the anti-diffusion term a so-called "flux-limiter" φ_j depending on the ratio of the two adjacent increments $u_j - u_{j-1}$, $u_{j+1} - u_j$:

$$\varphi_j = \varphi_j(r_j) \quad \text{with} \quad r_j = \frac{u_j - u_{j-1}}{u_{j+1} - u_j} \tag{5.7}$$

to obtain a flux-limited version of the Lax-Wendroff scheme

$$u^j = u_j - \lambda\Delta(au)_{j-1/2} - \lambda\Delta_-\left[\varphi_j(r_j)\frac{a(1-\nu)}{2}\Delta u_{j+1/2}\right] \tag{5.8}$$

In order to obtain a TVD scheme, we have to choose the limiter $\varphi(r_j)$ in such a manner that the scheme becomes dissipative in the immediate vicinity of shocks, but the joint action of the "sensor" r_j and the limiter φ_j should not be too spread out if we want to keep sharp shock profiles. Since the first thing we want to eliminate is the tendency of the LW scheme to generate oscillations, we impose

$$\varphi(r) = 0 \quad \text{for} \quad r \le 0 \tag{5.9}$$

i.e. if $(u_j - u_{j-1})(u_{j+1} - u_j) \le 0$ we add no anti-diffusion at all, so that our scheme reduces to the highly dissipative first-order upwind scheme, in an attempt to break the threatening oscillations detected by the sensor.

On the other hand, away from shocks we would like our scheme to display the second-order accuracy of the LW scheme, and for this reason we shall try to add as much of the anti-diffusive flux as we can while maintaining the TVD character. In order to be able to apply Harten's lemma 4.1, we rewrite (5.8) in Harten form: Sweby proposes, to this effect, to choose

$$C_{j-1/2} = a\lambda + \frac{\lambda\Delta_-\left[\varphi(r_j)\frac{(1-\nu)}{2}\Delta(au)_{j+1/2}\right]}{\Delta u_{j-1/2}} \ , \ D_{j+1/2} \equiv 0 \ . \tag{5.10}$$

Since the characteristic speed a has been assumed constant, we get

$$C_{j-1/2} = \nu[1 + \frac{1-\nu}{2}(\frac{\varphi(r_j)}{r_j} - \varphi(r_{j-1}))] \tag{5.11}$$

so that imposing a bound Φ on the magnitude of the inside bracket

$$|\frac{\varphi(r_j)}{r_j} - \varphi(r_{j-1})| \leq \Phi \quad \text{(to be specified later)} \tag{5.12}$$

leads to the following bounds for $C_{j-1/2}$:

$$\nu[1 - \frac{1-\nu}{2}\Phi] \leq C_{j-1/2} \leq \nu[1 + \frac{1-\nu}{2}\Phi] \tag{5.13}$$

By Harten's lemma, the scheme will be TVD if $0 \leq C_{j-1/2} \leq 1$, and therefore if

$$0 \leq 1 - \frac{1}{2}(1-\nu)\Phi \quad \text{and} \quad \nu(1 + \frac{1}{2}(1-\nu)\Phi) \leq 1 . \tag{5.14}$$

From the first inequality we must have

$$(1-\nu)\Phi \leq 2$$

which imposes, for $0 \leq \nu \leq 1$:

$$\Phi \leq 2 \tag{5.15}$$

This condition also guarantees that the second inequality (5.14) is satisfied for all $\nu \in [0,1]$. From (5.9), (5.12) and (5.15) we conclude that the limiter φ must satisfy

$$\begin{cases} 0 \leq \max[\frac{\varphi(r)}{r}, \varphi(r)] \leq 2 \\ \varphi(r) = 0 \quad r \leq 0 \end{cases} \tag{5.16}$$

with $\varphi \geq 0$ for all r to maintain the anti-dissipative character of the second-difference term.

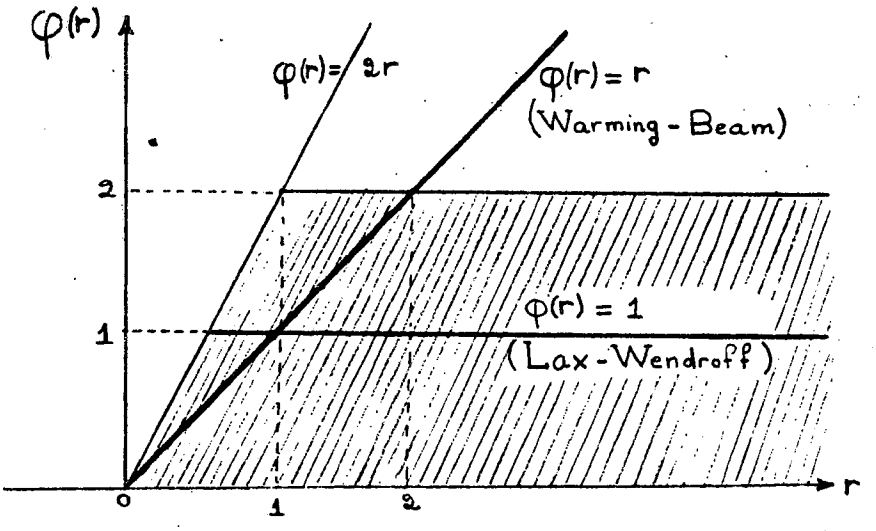

Fig. 5.1 TVD region for the limited LW schemes

Figure 5.1 shows the region, in the $(r,\varphi(r))$-plane , where the graph of the limiter function $\varphi(r)$ must be located to guarantee a TVD scheme. As expected, the graph of the constant limiter $\varphi \equiv 1$ of the Lax-Wendroff scheme is not wholly contained in the TVD region. For the Warming and Beam upwind scheme ([59])

$$u_j^{n+1} = u_j^n - a\lambda(u_j^n - u_{j-1}^n) - \frac{1}{2} a\lambda(1-\nu)(u_j^n - 2u_{j-1}^n + u_{j-2}^n) \qquad (5.17)$$

which can be written

$$u_j^{n+1} = u_j^n - a\lambda\Delta u_{j-1/2} - \lambda\Delta_-\left[r_j \frac{(1-\nu)}{2} a\Delta u_{j+1/2}^n\right] \qquad (5.17')$$

which is also second-order accurate but uses nodes $j-2$, $j-1$, j instead of $j-1$, j , $j+1$, the flux limiter, in the notation (5.8), is $\varphi(r) = r$; this also reaches out of the TVD region. Following Fromm's approach ([16], [17]) which leads to a "zero-average phase error scheme" by alternating one step with the explicit split-flux upwind version of MacCormack's scheme (3.14) and one step with the symmetric explicit centered MacCormack scheme (3.11), Sweby constructs

a convex combination of the Lax-Wendroff and Warming-Beam schemes, in order to generate, by a clever "tuning" of the corresponding flux limiter, a scheme which would tentatively enjoy both the monotonicity properties of upwind schemes and the second-order accuracy of these two schemes. Multiplying (5.3) by $(1-\theta)$, (5.17') by θ and adding leads to the scheme

$$u^j = u_j - \nu\Delta u_{j-1/2} - \Delta_-\{\tfrac{1-\nu}{2}\nu[1-\theta+\theta r_j]\Delta u_{j+1/2}\} \tag{5.18}$$

which is a limited LW scheme of the form (5.8) with flux limiter

$$\varphi(r_j) = (1-\theta) + r_j\theta = 1 + \theta(r_j-1) \tag{5.19}$$

i.e.

$$\varphi(r_j) = (1-\theta)\varphi_{LW}(r_j) + \theta\varphi_{WB}(r_j) \tag{5.19'}$$

with

$$\varphi_{LW}(r) \equiv 1 \quad \text{and} \quad \varphi_{WB}(r) = r\ . \tag{5.19"}$$

Sweby suggests to have $0 \le \theta(r) \le 1$, as the numerical experiments performed with non-convex combinations seemed to show too much compression, i.e. the opposite of too much dissipation: sine wave initial data tended to become square waves (notice, though, that due to the characteristic property of $u_t + f(u)_x = 0$, the decreasing part of the initial function will always tend to generate a shock; see Lax [30] for details).

Now applying Harten's Lemma, we can verify that the scheme (which is by construction guaranteed to be second-order accurate) will be TVD if the graph of $\varphi(r) = 1 + \theta(r-1)$ is contained in the hatched region of Fig. 5.2.

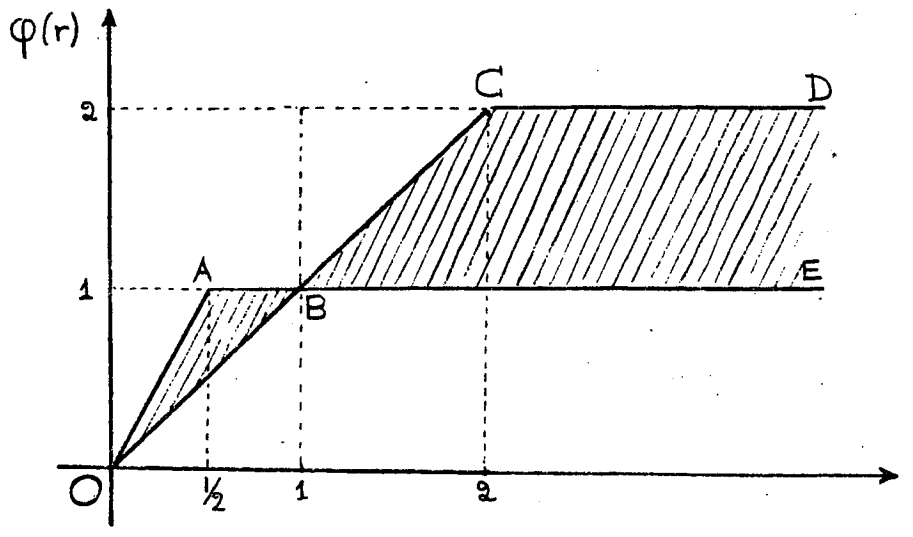

Fig. 5.2 Second-order TVD region for Sweby's 4-point scheme (5.18).

PROOF. (a) For $\theta = 1$ we shall obtain a lower bound for $\varphi(r)$, for $0 \le r \le 1$, namely $\varphi_{min} = 1 + r - 1 = r$, thus justifying the boundary segment OB.

(b) Still with $0 \le r \le 1$, the largest possible values of $\varphi(r)$ are obtained for $\theta = 0$, thus giving $\varphi(r) = 1$, but we still have to fulfil condition (5.16), which gives $\varphi(r) \le 2r$; this accounts for OA and AB.

(c) For $r > 1$, $\varphi(r) = 1 + \theta(r)(r-1)$ must remain ≥ 0, to maintain the anti-diffusive character of the flux in (5.8). The maximum values are those for which $\theta(r) \equiv 1$, i.e. $\varphi \equiv r$, leading to BC; CD comes from the second restriction in (5.16): $\varphi(r) \le 2$.

(d) Finally for $r > 1$ we cannot have $\varphi < 1$ since $\theta \ge 0$ gives $1 + \theta(r-1) \ge 1$, whence BE.

We shall now consider the limiters of Sweby, van Leer, Chakravarthy and Osher.

1. Sweby's limiter. Sweby introduces a flux limiter $\varphi_\Phi(r)$ depending on a parameter Φ :

$$\varphi_\Phi(r) = \max[0,\min(\Phi r,1),\min(r,\Phi)] \quad \text{with} \quad 1 \le \Phi \le 2 \ . \tag{5.20}$$

The graph of φ_Φ sweeps the whole second-order TVD region of Fig. 5.2 as Φ increases from 1 to 2, see Fig. 5.3.

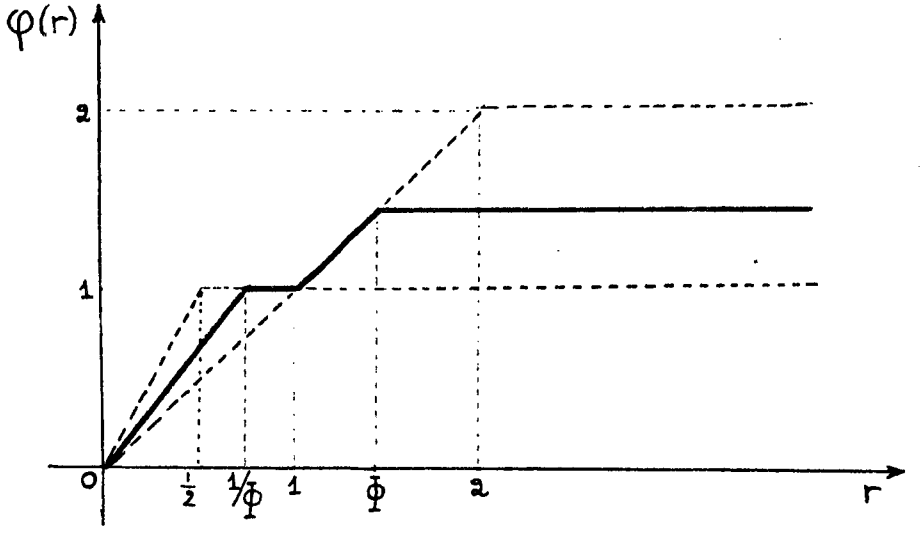

Fig. 5.3 Graph of Sweby's limiter.

As the limiter of the LW and Warming-Beam schemes, Sweby's limiter passes through the point (1,1) i.e. $\varphi_\Phi(1) = 1$; this reflects the fact that if the adjacent gradients $u_j - u_{j-1}$ and $u_{j+1} - u_j$ are of the same order, the solution u_j^n is not approaching a shock in the neighbourhood of x_j , and therefore we can leave the LW scheme do its second-order accurate work without having to worry about a potential oscillation. Moreover, for the scalar conservation law $u_t + f(u)_x = 0$, Oleinik's entropy condition (see [41]; [67], p. 251) (for convex flux i.e. $f'' > 0$)

$$\frac{u(x+\alpha,t)-u(x,t)}{\alpha} \le \frac{E}{t} \quad \text{for all} \quad \alpha > 0 \ , \ t > 0 \tag{5.21}$$

shows, fixing $t > 0$ and going from $x = -\infty$ through the shock to $x = +\infty$, that the jump at a shock can only be downward, i.e. $u_L > u_R$; indeed if $u_L < u_R$, taking arbitrarily small values of the increment α would lead to arbitrarily large values of the left-hand side, contradicting (5.2).

Considering, then, the case of a downward shock with a numerical approximation showing the shock located in or close to the interval $[x_{j-1},x_j]$, and moving from $x = -\infty$ through the shock toward $x = \infty$, we have, typically, the following situation (see fig. 5.4):

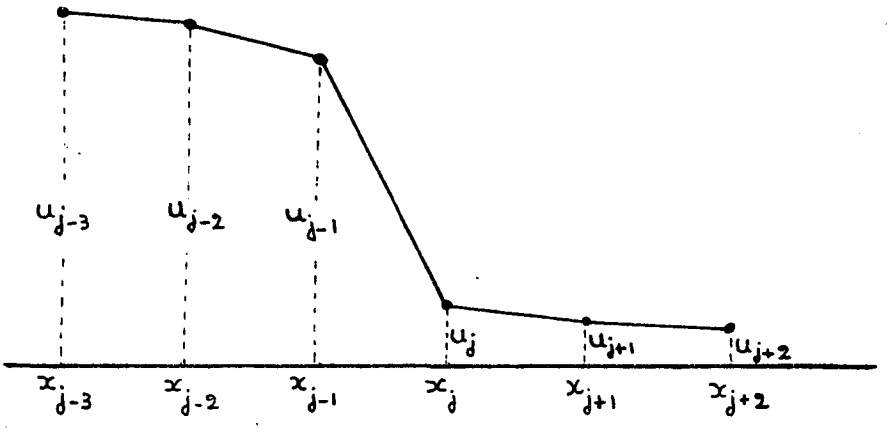

Fig. 5.4 Downward shock for $u_t + f(u)_x = 0$.

(i) $0 < r_{j-2} = (u_{j-2}-u_{j-3})/(u_{j-1}-u_{j-2}) < 1$ with r_{j-2} close to 1. The limiter-point $(r,\varphi(r))$ will be in triangle OAB (Fig. 5.2), close to (1,1) (If $r_{j-2} < 0$ we take $\varphi(r_{j-2}) = 0$).

(ii) $0 < r_{j-1} << 1$, and we take $\varphi(r_{j-1})$ close to 0 to obtain a local dissipative effect.

(iii) $r_j >> 1$ and normally, with the original limited version (5.8) of the LW scheme, we should take $\varphi(r_j)$ very small to get a dissipative effect, since we are still in the vicinity of the shock. But using the blended scheme (5.18), which contain a built-in component of the upwind, dissipative scheme of Warming and Beam, we actually use a limiter $\varphi_S = 1 + (r-1)\theta > 1$ since $\theta > 0$ and $r >> 1$; but the point $(r,\varphi(r))$ is still in the TVD region of Fig. 5.2.

(iv) if $r_{j+1} > 1$ but $r_{j+1} = O(1)$ we have an effective limiter $\varphi_S > 1$ as in case (iii), while if $r_{j+1} < 0$ we take $\varphi_{j+1} = 0$.

Owing to these remarks, we see that Sweby's limiter should give fairly good results, since it fulfils all conditions obtained in (i) - (iv) .

Returning to the original limited LW scheme (5.8), which is a symmetric scheme with nodes x_{j-1} , x_j , x_{j+1} , we see that the above considerations would suggest taking a flux limiter with a behaviour of the following form

r	$-\infty$	-	0	+	$1-\varepsilon$	1	$1+\varepsilon$	10	r	$+\infty$
$\varphi(r)$	0	0	0	+	$O(1-\varepsilon)$	1	$O(1+\varepsilon)$	$O(\frac{1}{10})$	$O(\frac{1}{r})$	0

(5.22)

Here $\varphi(r)$ has to decrease to zero as r increases, since large values of r announce the immediate vicinity of a shock and call for a dissipative mechanism.

On this basis, we suggest using for the limiter in (5.8) one of the exponential type functions familiar in the kinetic theory of gases

$$\varphi_{1,p}(r) = e \cdot r^p \cdot \exp(-r^2) = r^p \exp(1-r^2) \qquad (5.23)$$

with a preference for $p = 1,2$ or alternately

$$\varphi_{2,p}(r) = r^p \exp[-(r-1)^2] \tag{5.24}$$

Choosing $p = 2$ for $\varphi_{1,p}$ gives $\varphi_{1,2}(0) = 0$, $\varphi_{1,2}(1) = 1$, $\varphi'_{1,2}(1) = 0$, and continuous differentiability of the limiter from $-\infty$ to $+\infty$ (including the origin), which might be of some value. For this limiter, one would expect a reasonably compressive effect away from shocks, and diffusivity near the shocks. Comparative numerical tests for Burgers' equation and for problems in gas dynamics will be described elsewhere.

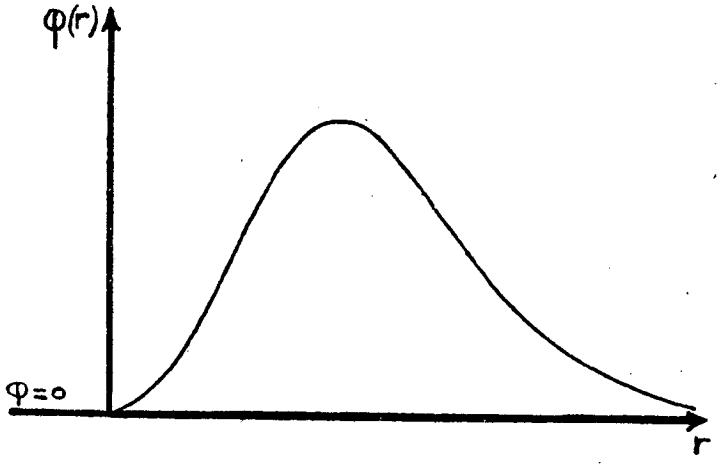

Fig. 5.5 Graph of the proposed limiter $\varphi_{1,2}(r)$.

2. van Leer's limiter. Inspired by an idea of Fromm, who constructed the arithmetic average of the Lax-Wendroff and Warming-Beam schemes (5.2) and (5.17), van Leer [32] used a weighted average of non conservative and flux-limited forms of these two schemes to obtain a monotonicity preserving, flux-limited version of Fromm's scheme, which can be written in conservation form ([32], p. 364 for an explanation) or equivalently in the following way, where the 4-point dependency is evident:

$$u^j = u_j - \nu(u_j-u_{j-1}) - \frac{\nu}{4}(1-\nu)[(u_{j+1}-u_j)-(u_{j-1}-u_{j-2})] \tag{5.25a}$$

$$+ \frac{\nu}{4}(1-\nu)[S(\theta_j)(u_{j+1}-2u_j+u_{j-1})-S(\theta_{j-1})(u_j-2u_{j-1}+u_{j-2})]$$

where θ_j is van Leer's sensor (or "smoothness monitor")

$$\theta_j = \frac{\Delta u_{j+1/2}}{\Delta u_{j-1/2}} \equiv \frac{u_{j+1}-u_j}{u_j-u_{j-1}} \equiv \frac{1}{r_j} \tag{5.25b}$$

with $\theta_j = 1$ if both $\Delta u_{j-1/2}$ and $\Delta u_{j+1/2}$ vanish and

$$S(\theta) \equiv \frac{|\theta|-1}{|\theta|+1} \tag{5.25c}$$

In order to reduce this to the form (5.8) of a flux-limited LW scheme, Sweby rewrites (5.25) as

$$u^j = u_j - \nu\Delta u_{j-1/2} - \Delta_-\{\varphi_{j2}^{1}(1-\nu)\nu\Delta u_{j+1/2}\} \tag{5.26a}$$

with

$$\varphi_j = \frac{1-S(\theta_j)}{2} + \frac{1+S(\theta_j)}{2\theta_j} \tag{5.26b}$$

Using (5.26b) one can express van Leer's limiter as a function of the sensor r :

$$\varphi_j \equiv \varphi_{VL}(r_j) = \frac{r+|r|}{1+|r|} \tag{5.27}$$

This limiter, which is in the class of blended limiters (5.19), has a graph which is a smooth curve in the second-order TVD region of Fig. 5.2, with

$$\varphi_{VL}(0) = 0 \quad \text{and} \quad \varphi_{VL}(1) = 1 \ .$$

It leads to a very robust scheme with sharp shock resolution, and can be taken as the prototype of efficient limiters. By our (oscillation damping) convention to take $\varphi(r) = 0$ for $r \le 0$, it can be written

$$\varphi_{VL}(r) = \begin{cases} \frac{2r}{1+r} & \text{for } r \geq 0 \\ 0 & \text{for } r \leq 0 \end{cases} \tag{5.28}$$

and is a continuously differentiable function of r except at $r = 0$, satisfying the symmetry property

$$\varphi_{VL}(\tfrac{1}{r}) = \tfrac{1}{r}\varphi_{VL}(r) \tag{5.28'}$$

which guarantees that the scheme handles positive or negative gradients in the same way (no preferred direction of wave propagation).

Looking at its graph in the middle of the 2nd-order TVD region and at the good monotonic shock profiles it gives,(see e.g. [56], [32] confirms the old saying "in medio stat virtus".

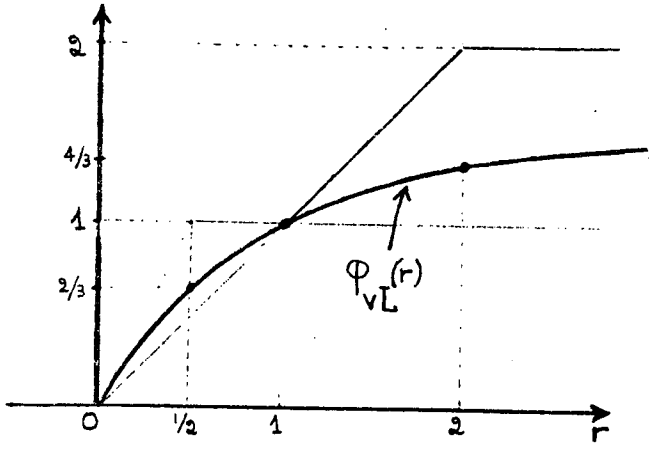

Fig. 5.6 van Leer's flux limiter (5.28) for the 4-point scheme (5.26.

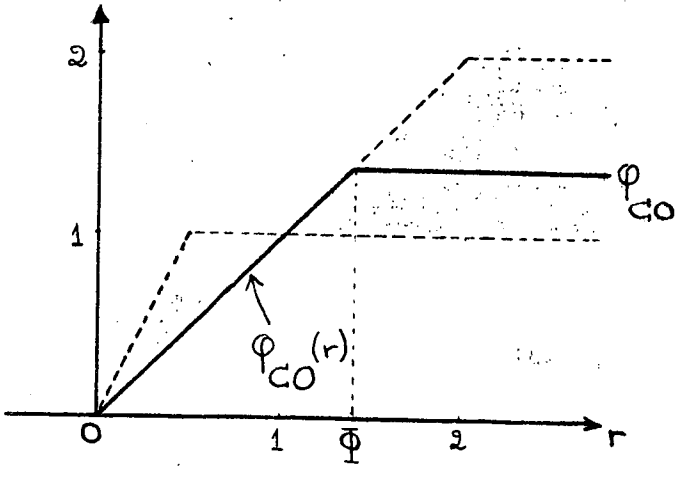

Fig. 5.7 Chakravarthy and Osher's limiter

3. The flux-limiter of Chakravarthy and Osher. Considering again the linear wave equation

$$u_t + au_x = 0 \quad (a > 0\ ,\ \text{constant})$$

and following the same Taylor-expansion approach as Lax and Wendroff (see equation (3.1)), Chakravarthy and Osher [8] use the upwind difference approximations

$$\begin{aligned} u_x &\cong (3u_j - 4u_{j-1} + u_{j-2})/(2\Delta x) \\ u_{xx} &\cong (u_j - 2u_{j-1} + u_{j-2})/(\Delta x)^2 \end{aligned} \tag{5.29}$$

to obtain a (space and time) second-order accurate scheme which they rewrite as

$$u^j = u_j - \nu(u_j - u_{j-1}) - \frac{\nu^2}{2}(u_j - 2u_{j-1} + u_{j-2}) - \frac{\nu}{2}[(u_j - u_{j-1}) - (u_{j-1} - u_{j-2})]\ . \tag{5.30}$$

The first three terms in the right-hand side would make the scheme 1st-order accurate in space and 2nd-order accurate in time, and the last term can be viewed as a correction to reinstate 2nd-order accuracy. As it plays the role of a negative viscosity term (anti-diffusive), one can verify that (5.30) is not TVD. To obtain a second-order accurate TVD scheme, the second-order difference of the correction term $[\Delta u_{j-1/2} - \Delta u_{j-3/2}]$ is modified into a term

$$\overline{\Delta u_{j-1/2}} - \overline{\Delta u_{j-3/2}} \ . \tag{5.31}$$

where the modified increments are obtained by a slope-limiting procedure

(i) If $\Delta u_{j-1/2} \cdot \Delta u_{j+1/2} \le 0$ we take $\overline{\Delta u_{j-1/2}} = 0$

(ii) If $\Delta u_{j-1/2} \cdot \Delta u_{j+1/2} > 0$ and if $|\Delta u_{j-1/2}| \le \Phi|\Delta u_{j+1/2}|$ we take $\overline{\Delta u_{j-1/2}} = \Delta u_{j-1/2}$

(iii) If $\Delta u_{j-1/2} \cdot \Delta u_{j+1/2} > 0$ and if $|\Delta u_{j-1/2}| > \Phi|\Delta u_{j+1/2}|$ we take $\overline{\Delta u_{j-1/2}} = \Phi \Delta u_{j+1/2}$.

This procedure can be shown to be equivalent to either

$$\overline{\Delta u_{j-1/2}} = \Delta u_{j+1/2} \cdot \max(0,\min[\frac{\Delta u_{j-1/2}\cdot\Delta u_{j+1/2}}{\Delta u_{j+1/2}\cdot\Delta u_{j+1/2}},\Phi]) \qquad (5.32a)$$

or

$$\overline{\Delta u_{j-1/2}} = \Delta u_{j-1/2} \cdot \max(0,\min[\frac{\Delta u_{j+1/2}\cdot\Delta u_{j-1/2}}{\Delta u_{j-1/2}\cdot\Delta u_{j-1/2}}\cdot\Phi,1]) \ . \qquad (5.32b)$$

The parameter Φ is the slope-limiter, chosen between 1.0 and 2.0 in [8] while in Goodman-LeVeque [22], its value is more severely fixed at 1.0 (see Section 7). In terms of flux-limiters as discussed earlier, this procedure corresponds to take a flux-limiter

$$\varphi_{CO}(r) = \max(0,\min[r,\Phi]) \quad 1 \le \Phi \le 2 \qquad (5.33)$$

in a 4-point blended scheme of the form (5.8) - (5.18). This limiter, represented in Fig. 5.7, is not symmetric in the sense of (5.28'), but makes the scheme of Chakravarthy and Osher TVD (and 2nd-order accurate).

5.B. Extension to the nonlinear equation $u_t + f(u)_x = 0$.

In order to generalize the schemes obtained in Section 5.A for $u_t + au_x = 0$ to the nonlinear conservation equation 4.1, Sweby resorts to the notion of E-schemes introduced by Osher [42], for semi-discrete approximations of the form

$$u_t = -\frac{1}{\Delta x}(h^E_{j+1/2} - h^E_{j-1/2}) \ . \qquad (5.34)$$

A scheme (5.34) is an E-scheme if the inequality

$$\mathrm{sgn}(u_{j+1}-u_j)[h^E_{j+1/2} - f(u)] \le 0 \qquad (5.35)$$

holds for all u between u_j and u_{j+1} (these two values being included). In [42], these E-schemes are shown to lead to numerical solutions which converge to the physically correct solution, supposed to satisfy an entropy condition (see [30]) excluding rarefaction shocks, whence Osher's notation.

In the present framework, let us consider three-point difference schemes in conservation form

$$u^j = u_j - \frac{\Delta t}{\Delta x}(h^E_{j+1/2} - h^E_{j-1/2}) \tag{5.36a}$$

where the numerical flux

$$h^E_{j+1/2} = h^E(u_j, u_{j+1}) \tag{5.36b}$$

satisfies Osher's E-condition (5.35). An example of such E-schemes is given by monotone schemes. The advantage of E-schemes is their convergence to the solution which satisfies the entropy condition (see [68]); but their main drawback is the same as that of monotone schemes: they can only be first-order accurate, and shocks are smeared. Osher and Chakravarthy [45], followed by Sweby [56], have used flux-limited E-schemes to obtain 2nd-order accurate TVD schemes. An important example of an E-scheme is the Engquist-Osher scheme ([14]), defined by

$$h^{EO}_{j+1/2} = f^+_j + f^-_{j+1} + f(\bar{u}) \tag{5.37}$$

where $\bar{u}$ is the sonic point of f i.e. $f'(\bar{u}) = 0$, and

$$f^+_j \equiv \int_{\bar{u}}^{u_j} \chi(s) f'(s)\,ds \quad \left(\begin{array}{l}\text{integral of the positive part of}\\ \text{the characteristic velocity}\end{array}\right) \tag{5.38a}$$

$$f^-_j \equiv \int_{\bar{u}}^{u_j} (1-\chi(s)) f'(s)\,ds \quad \left(\begin{array}{l}\text{integral of the negative part of}\\ \text{the characteristic velocity}\end{array}\right) \tag{5.38b}$$

with

$$\chi(s) = \begin{cases} 1 & \text{if } f'(s) > 0 \\ 0 & \text{if } f'(s) \le 0 \,. \end{cases}$$

Since the characteristic velocity $a(u) = f'(u)$ is no longer constrained, as in Section 5.A, to be of one sign only, the extension of the previous theory requires the introduction of specific signed flux differences and signed local Courant numbers: let

$$(\Delta f_{j+1/2})^+ \equiv -[h_{j+1/2} - f(u_{j+1})] \,, \quad \nu^+_{j+1/2} = \frac{\lambda(\Delta f_{j+1/2})^+}{\Delta u_{j+1/2}} \tag{5.39a}$$

$$(\Delta f_{j+1/2})^- \equiv [h_{j+1/2} - f(u_j)] \,, \quad \nu^-_{j+1/2} = \frac{\lambda(\Delta f_{j+1/2})^-}{\Delta u_{j+1/2}} \,. \tag{5.39b}$$

We have

$$(\Delta f_{j+1/2})^+ + (\Delta f_{j+1/2})^- = f(u_{j+1}) - f(u_j) \equiv \Delta f_{j+1/2} \,, \tag{5.39c}$$

and for an E-scheme we get, by (5.35)

$$\nu^+_{j+1/2} \ge 0 \,, \quad \nu^-_{j+1/2} \le 0 \,. \tag{5.40}$$

Moreover, defining

$$\nu_{j+1/2} = \nu^+_{j+1/2} + \nu^-_{j+1/2} \tag{5.41a}$$

we obtain

$$\nu_{j+1/2} = \lambda[(\Delta f_{j+1/2})^+ + (\Delta f_{j+1/2})^-]/\Delta u_{j+1/2} = \frac{\lambda \Delta f_{j+1/2}}{\Delta u_{j+1/2}} \tag{5.41b}$$

which has the typical form of a discrete CFL-number. The signed local Courant numbers will allows us to introduce an appropriate flux-splitting in the extension of Sweby's scheme (5.8) - (5.18) to $u_t + f(u)_x = 0$, and by-pass the computation

of the characteristic velocity $f'(u)$, which is discretized by (5.41b) and split by (5.41a).

We can now use these signed local CFL-numbers to reduce our E-scheme (5.36) - (5.35) to Harten's form, in order to obtain sufficient TVD conditions; we have

$$u^j = u_j - \lambda[(h_{j+1/2}-f(u_j))-(h_{j-1/2}-f(u_j))] = u_j - \lambda(\Delta f_{j+1/2})^- - \lambda(\Delta f_{j-1/2})^+$$

or

$$u^j = u_j - \nu^+_{j-1/2}\,\Delta u_{j-1/2} - \nu^-_{j+1/2}\Delta u_{j+1/2}\,. \tag{5.42}$$

This is Harten's form if we define

$$C_{j-1/2} = \nu^+_{j-1/2}\,,\quad D_{j+1/2} = -\nu^-_{j+1/2}\,. \tag{5.43}$$

Therefore condition (4.12a) is satisfied for an E-scheme, and condition (4.12b) becomes

$$0 \le \nu^+_{j+1/2} - \nu^-_{j+1/2} \le 1 \tag{5.44}$$

for a TVD fully discretized E-scheme (5.36). Notice that (5.44) has the form and dimensions of a CFL condition. Indeed if we had only defined $\nu_{j+1/2} = \lambda(\Delta f_{j+1/2})/\Delta u_{j+1/2}$ and split it into its positive and negative parts

$$(\nu_{j+1/2})^+ \equiv \max(0,\nu_{j+1/2})\,,\quad (\nu_{j+1/2})^- = \min(0,\nu_{j+1/2})\,,$$

then $(\nu_{j+1/2})^+ - (\nu_{j+1/2})^-$ would denote $|\nu_{j+1/2}|$ and the condition would read

$$|\nu_{j+1/2}| = \lambda|a(u)_{j+1/2}| \le 1 \quad (\text{with } a(u)_{j+1/2} \equiv \Delta f_{j+1/2}/\Delta u_{j+1/2})\,;$$

but of course it has a slightly different meaning here, although away from the sonic point u, we do have either $\nu^-_{j+1/2} = 0$ or $\nu^+_{j+1/2} = 0$, and (5.44) then exactly reduces to the CFL condition $\lambda|a(u)_{j+1/2}| \le 1$.

For the Engquist-Osher scheme, we have

$$(\Delta f_{j+1/2})^-_{EO} = f^+_j + f^-_{j+1} + f(\bar{u}) - f_j = f^+_j + f^-_{j+1} + f(\bar{u}) - [f(\bar{u}) + \int_{\bar{u}}^{u_j} f'(s)ds]$$

$$= f^+_j + f^-_{j+1} - \int_{\bar{u}}^{u_j} \chi(s)f'(s) - \int_{\bar{u}}^{u_j} [1-\chi(s)]f'(s)ds = f^-_{j+1} - f^-_j$$

$$= \int_{\bar{u}}^{u_{j+1}} (1-\chi(s))f'(s)ds - \int_{\bar{u}}^{u_j} [1-\chi(s)]f'(s)ds = \int_{u_j}^{u_{j+1}} [1-\chi(s)]f'(s)ds$$

as well as

$$(\Delta f_{j+1/2})^+_{EO} = -(f^+_j + f^-_{j+1} + f(\bar{u}) - [f(\bar{u}) + \int_{\bar{u}}^{u_{j+1}} f'(s)ds])$$

$$= f^+_{j+1} - f^+_j = \int_{u_j}^{u_{j+1}} \chi(s)f'(s)ds$$

so that

$$\nu^+_{j+1/2} - \nu^-_{j+1/2} = \frac{\lambda}{\Delta u_{j+1/2}} \int_{u_j}^{u_{j+1}} [2\chi(s)-1]f'(s)ds = \frac{\lambda}{\Delta u_{j+1/2}} \int_{u_j}^{u_{j+1}} |f'(s)|ds$$

and we get the inequality

$$\nu^+_{j+1/2} - \nu^-_{j+1/2} \leq \lambda \sup_s |f'(s)| .$$

The Engquist-Osher is therefore TVD under the following (sufficient) CFL-like condition

$$\lambda \sup_s |f'(s)| \leq 1 \quad \text{(TVD EO-scheme)} . \tag{5.45}$$

Now to extend Sweby's scheme, we replace au by $f(u)$ in (5.1) and rewrite (5.8) as

$$u^j = u_j - \lambda \Delta f_{j-1/2} - \lambda \Delta_- [\varphi(r_j)\tfrac{1-\nu}{2}\Delta f_{j+1/2}] \tag{5.46}$$

Since the first part of the right-hand side corresponds to the upwing (monotonicity preserving) scheme, if $\nu_{j-1/2} = \lambda a(u)_{j-1/2} \geq 0$, we shall consider replacing it by an E-scheme (also monotonicity preserving under condition (5.44)) and assuming temporarily that we are away from the sonic point $\bar{u}$, and therefore that, say $\nu_{j+1/2} = \nu^+_{j+1/2}$, we should consider the following to be a reasonable extension of (5.46)

$$u^j = u_j - \lambda(h^E_{j+1/2}-h^E_{j-1/2}) - \lambda\Delta_-[\varphi(r^+_j)\alpha^+_{j+1/2}(\Delta f_{j+1/2})^+] \qquad (5.47)$$

where Sweby proposes to choose (see the motivation below)

$$r^+_j = \alpha^+_{j-1/2}(\Delta f_{j-1/2})^+/\alpha^+_{j+1/2}(\Delta f^+_{j+1/2}),\ \alpha^+_{j+1/2} = \tfrac{1}{2}(1-\nu^+_{j+1/2}) \ . \qquad (5.48)$$

Comparing (5.47) and (5.46) shows that $f_{j+1} - f_j$ has been replaced[1] by $-(h^E_{j+1/2}-f_{j+1})$, and choosing $h^E_{j+1/2} = h^{EO}_{j+1/2}$ to gain some insight, this becomes $f^+_{j+1} - f^+_j$ by previous calculations, which reduces to $f_{j+1} - f_j$ away from the sonic point for positive characteristic velocity; this fully justifies the transition from (5.46) to (5.47), in this particular case.

It remains to consider the general case where $f'(u)$ is of both signs in the regions of interest, with the sonic point $\bar{u}$ between u_j and u_{j+1} for instance. Taking into account the difference in sign between $(\Delta f_{j+1/2})^- = h^E_{j+1/2} - f(u_j)$ and $(\Delta f_{j+1/2})^+ = -(h^E_{j+1/2}-f(u_{j+1}))$ leads to Sweby's following extension of (5.8):

$$u^j = u_j-\lambda\Delta_-h^E_{j+1/2}-\lambda\Delta_-[\varphi(r^+_j)\alpha^+_{j+1/2}(\Delta f_{j+1/2})^+-\varphi(r^-_{j+1})\alpha^-_{j+1/2}(\Delta f_{j+1/2})^-] \qquad (5.49)$$

where

$$\alpha^-_{j+1/2} = \tfrac{1}{2}(1+\nu^-_{j+1/2}) \ , \ r^-_j = \frac{\alpha^-_{j+1/2}(\Delta f_{j+1/2})^-}{\alpha^-_{j-1/2}(\Delta f_{j-1/2})^-} \quad (\text{for } u_t + f(u)_x = 0) \qquad (5.48')$$

---

[1] In the last term of (5.46).

To justify (5.48), (5.49) and the minus sign of the second term in the bracket, consider the reciprocal of the situation discussed above, i.e. the case $\nu_{j+1/2} = \nu^-_{j+1/2}$, for the Engquist-Osher scheme, after first writing the Lax-Wendroff scheme for the case $u_t + au_x = 0$, $a < 0$ in the following form:

$$u^j = u_j - \nu\Delta u_{j+1/2} + \frac{\nu(1+\nu)}{2}(u_{j+1}-2u_j+u_{j-1}) \quad \underset{a<0}{\text{(Lax-Wendroff)}} . \tag{5.50}$$

Introducing an upwind biased flux limiter $\varphi = \varphi(r^-_{j+1})$ with

$$r^-_j \equiv \frac{\Delta u_{j+1/2}}{\Delta u_{j-1/2}} = \frac{1}{r^+_j} \equiv \frac{1}{r_j} \quad \text{(for } u_t + au_x = 0 \text{ , } a < 0) \tag{5.51}$$

leads to Sweby's form of the flux-limited, upwind biased LW scheme for $u_t + au_x = 0$ $(a < 0)$ (the reciprocal of (5.8) - (5.11) where a was > 0) :

$$u^j = u_j - a\lambda\Delta u_{j+1/2} + \lambda\Delta_-\{\varphi(r^-_{j+1})\frac{1+\nu}{2}\Delta(au)_{j+1/2}\} \tag{5.52}$$

or using Harten's form

$$u^j = u_j + \nu\{-1+\frac{1+\nu}{2}[\varphi(r^-_{j+1}) - \frac{\varphi(r^-_j)}{r^-_j}]\}\Delta u_{j+1/2} = 0 . \tag{5.53}$$

Defining

$$\alpha^-_{j+1/2} = \frac{1}{2}(1+\nu^-_{j+1/2}) \tag{5.54}$$

and replacing ν by $\nu^-_{j+1/2} \equiv \lambda(\Delta f_{j+1/2})^-/\Delta u_{j+1/2}$ and $\Delta(au)_{j+1/2}$ in (5.52) by $(\Delta f_{j+1/2})^- \equiv +[h_{j+1/2}-f(u_j)]$ then leads to the second term (negative flux) in the bracket of (5.49) (and this is again justified by considering the case of the Engquist-Osher scheme, where $(\Delta f_{j+1/2})^- = f^-_{j+1} - f^-_j$, which reduces to $f_{j+1} - f_j$ away from the sonic point for negative characteristic velocity). To motivate the change of notation from (5.7), (5.46), (5.51) and (5.52) to (5.47), (5.48), (5.49), and at the same time to obtain Harten's form of equation (5.49), we first observe that (by 5.39),

$$\lambda\Delta_- h^E_{j+1/2} = \lambda[h^E_{j+1/2} - h^E_{j-1/2}] = \lambda[h^E_{j+1/2} - f(u_j)] - \lambda[h^E_{j-1/2} - f(u_j)]$$

$$= \lambda(\Delta f_{j+1/2})^- + \lambda(\Delta f_{j-1/2})^+ = \nu^-_{j+1/2}\Delta u_{j+1/2} + \nu^+_{j-1/2}\Delta u_{j-1/2} \tag{5.54}$$

and apply flux-splitting to make use of both (5.8) and (5.52); we assume homogeneity of the flux, i.e. $f(u) = a(u) \cdot u = f'(u) \cdot u$, and we try to factorize $\Delta u_{j-1/2}$ in the part of (5.49) (the form of which is already justified by our previous 2 cases and flux-splitting) corresponding to increasing flux (positive $a(u)$) , and $\Delta u_{j+1/2}$ for the decreasing flux (negative $a(u)$) , thus respecting Steger and Warming's flux-splitting technique:

$$u^j = u_j - \nu^+_{j-1/2}\Delta u_{j-1/2} - \nu^-_{j+1/2}\Delta u_{j+1/2}$$

$$- [\varphi(r^+_j)\alpha^+_{j+1/2}\frac{\lambda(\Delta f_{j+1/2})^+}{\Delta u_{j-1/2}} - \varphi(r^+_{j-1})\alpha^+_{j-1/2}\frac{\lambda(\Delta f_{j-1/2})^+}{\Delta u_{j-1/2}}]\Delta u_{j-1/2}$$

$$+ [\varphi(r^-_{j+1})\alpha^-_{j+1/2}\frac{\lambda(\Delta f_{j+1/2})^-}{\Delta u_{j+1/2}} - \varphi(r^-_j)\alpha^-_{j-1/2}\frac{\lambda(\Delta f_{j-1/2})^-}{\Delta u_{j+1/2}}]\Delta u_{j+1/2}$$

i.e.

$$u^j = u_j - \nu^+_{j-1/2}\{1+\alpha^+_{j-1/2}[\varphi(r^+_j)\frac{\alpha^+_{j+1/2}}{\alpha^+_{j-1/2}}\cdot\frac{\nu^+_{j+1/2}}{\nu^+_{j-1/2}}\cdot\frac{\Delta u_{j+1/2}}{\Delta u_{j-1/2}} - \varphi(r^+_{j-1})]\}\Delta u_{j-1/2}$$

$$- \nu^-_{j+1/2}\{1-\alpha^-_{j+1/2}[\varphi(r^-_{j+1}) - \varphi(r^-_j)\frac{\alpha^-_{j-1/2}}{\alpha^-_{j+1/2}}\cdot\frac{\nu^-_{j-1/2}}{\nu^-_{j+1/2}}\cdot\frac{\Delta u_{j-1/2}}{\Delta u_{j+1/2}}]\}\Delta u_{j+1/2}$$

or using (5.48), which is therewith justified:

$$u^j = u_j - \nu^+_{j-1/2}\{1+\alpha^+_{j-1/2}[\frac{\varphi(r^+_j)}{r^+_j} - \varphi(r^+_{j-1})]\}\Delta u_{j-1/2}$$

$$- \nu^-_{j+1/2}\{1-\alpha^-_{j+1/2}[\varphi(r^-_{j+1}) - \frac{\varphi(r^-_j)}{r^-_j}]\}\Delta u_{j+1/2} \tag{5.55}$$

Harten's form of Sweby's scheme, general case $u_t + f(u)_x = 0$

Applying Harten's lemma 4.1 and the bound (5.12), Sweby shows that (5.49) is TVD if

$$\lambda \sup_s |f'(s)| \leq \frac{2}{3} \tag{5.56}$$

We observe here that (5.55) is the exact superposition of (5.10) - (5.53) obtained for increasing and decreasing flux, respectively.

As we intend to present in a forthcoming paper [63] detailed calculations with extensions of the different above versions of second-order flux-limited TVD schemes to gas-dynamical problems, we shall only summarize here some of the numerical experiments performed by Sweby, to help the reader to assess the great improvement in the accuracy due to the introduction of flux-limiters.
In [56], Sweby compares, for the linear wave equation $u_t + au_x = 0$ and square wave or $\sin^2$-wave initial data, the following schemes:

(i) first order upwind, without limiter (i.e. the C.I.R. or Steger-Warming's scheme)

(ii) Sweby's scheme (5.18) -(5.8) with limiter (5.20) and $\Phi = 1$ (i.e. Roe's minmod limiter φ_1 , the lower boundary of the 2^{nd}-order TVD region of Fig. 5.2)

(iii) same, with Chakravarthy and Osher's limiter (5.33) and $\Phi = 2$ (i.e. $\varphi_{CO} = \max[0,\min(r,2)]$)

(iv) same, with limiter (5.20) and $\Phi = 2$ (i.e. Roe's highly compressive limiter φ_2 "superbee", the upper boundary of the 2^{nd}-order TVD region of Fig. 5.2)

(v) same, with van Leer's limiter (5.28) i.e. $\varphi_{VL}(r) = 2r/1+r$.

Fig. 5.8, borrowed form [56], shows the extent of the accuracy gain achieved with all limited schemes, compared with the first-order upwind scheme;

the continuous line represents the exact solution. We feel that for these two initial data, the five schemes considered here are numbered, in first approximation, by increasing order of quality, although it is hard to judge, for instance, whether the over-compressive $\sin^2$ maximum for φ_2 is not compensated, with respect to φ_{VL}, by the better φ_2-square wave. The essential observation lies in the obvious improved accuracy and the monotonic character of the numerical solutions: as expected from the underlying theory, the oscillations have been completely removed.

Among the four limiters, φ_1-minmod is the least anti-diffusive (least compressive) as can be seen from Fig. 5.2 and therefore gives the highest level of smearing; φ_2-superbee is the most anti-diffusive (compressive) and leads to very sharp "shocks" and some squaring effect for the $\sin^2$-wave problem. Chakravarthy and Osher's φ_{CO}, with its unsymmetric location in the 2^{nd}-order TVD region, and lack of symmetry in the sense of (5.28'), indeed produces slightly asymmetric profiles, which seems to justify its ranking in third position, despite the sharp "shocks" and very good overall precision.

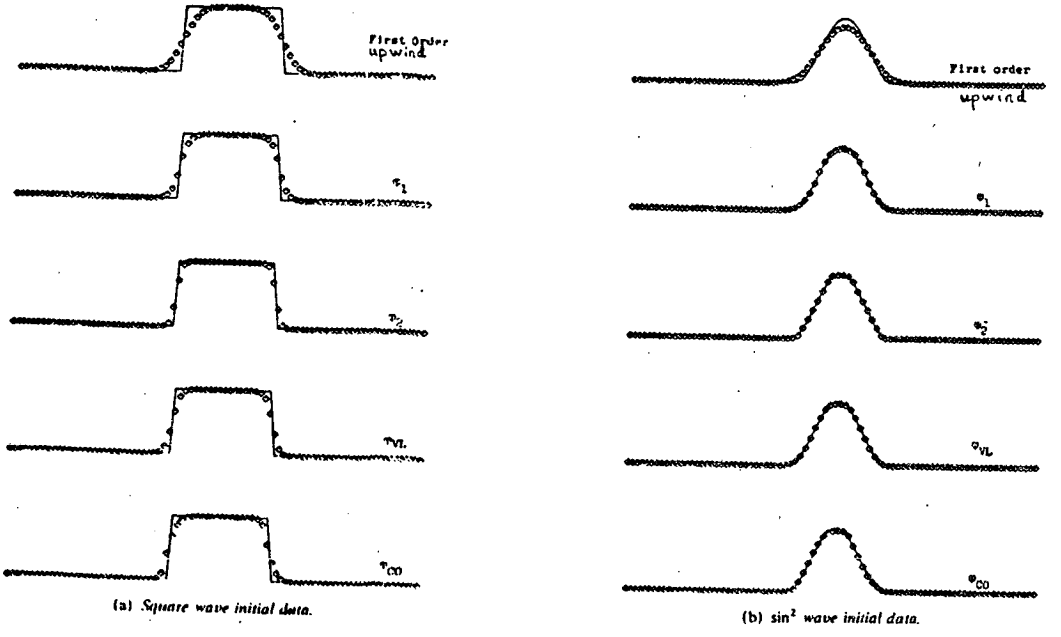

Fig. 5.8 Numerical solutions for $u_t + au_x = 0$, courtesy of Sweby [56] SIAM J. for Num. Anal. vol. 21 (1984).

6. *DAVIS' MODIFICATION OF SWEBY'S SCHEME*

Observing that the Lax-Wendroff scheme (5.3) contains too much anti-diffusion, and is not TVD , Davis [12] has obtained a TVD scheme by adding to the Lax-Wendroff scheme an artificial viscosity term which is, at first, chosen to be upstream-centered, but has a form independent of the problem under consideration (unlike usual artificial viscosity terms, which involve some problem dependent parameters). His scheme contains Sweby's scheme as a particular case. He then eliminates the upstream-centering after comparing the cases of increasing and decreasing flux, to obtain a second-order accurate, non-upstream centered, TVD scheme, with an ar tificial viscosity term which does not require the introduction of problem-dependent parameters.

6.A. Considering first the case $a > 0$ for $u_t + au_x = 0$ and adding an upstream-biased (see the indices of r) artificial viscosity term

$$K^+_{j+1/2}(r^+_j)\Delta u_{j+1/2} - K^+_{j-1/2}(r^+_{j-1}) \ , \quad \text{with} \quad r^+_j \equiv \Delta u_{j-1/2}/\Delta u_{j+1/2} \tag{6.1}$$

to the Lax-Wendroff scheme (3.2) written as (cf. (5.3))

$$u^j = u_j - \nu\Delta u_{j-1/2} - \Delta_-[\tfrac{\nu}{2}(1-\nu)\Delta u_{j+1/2}] \tag{6.2}$$

gives, after factorizing $\Delta u_{j-1/2}$ to enhance the case of positive velocity in Harten's form and prepare for the flux-splitting in the general case of positive or negative velocity (we use (5.11) with $\varphi \equiv 1$ for the LW term):

$$u^j = u_j - \nu[1 + \tfrac{1}{2}(1-\nu)(\frac{1}{r^+_j} - 1)]\Delta u_{j-1/2} + [\frac{K^+_{j+1/2}}{r^+_j} - K^+_{j-1/2}]\Delta u_{j-1/2} \tag{6.3}$$

which is in Harten's form with

$$C_{j-1/2} = \nu[1 + \tfrac{1}{2}(1-\nu)(\frac{1}{r^+_j} - 1)] - [\frac{K^+_{j+1/2}}{r^+_j} - K^+_{j-1/2}] \ , \ D_{j+1/2} = 0 \ . \tag{6.4}$$

For Sweby's scheme we had obtained

$$C_{j-1/2} = \nu[1 + \frac{1}{2}(1-\nu)\{\frac{\varphi(r_j^+)}{r_j^+} - \varphi(r_{j-1}^+)\}] \ , \ D_{j+1/2} = 0 \quad \text{(Sweby, } a > 0) \ . \tag{5.11}$$

Imposing equality between these two forms leads to the condition

$$\frac{\nu}{2}(1-\nu)[\frac{1-\varphi(r_j^+)}{r_j^+} - \{1-\varphi(r_{j-1}^+)\}] = \frac{K_{j+1/2}^+}{r_j^+} - K_{j-1/2}^+$$

which is obviously satisfied if we take for Davis' coefficient

$$K_{j+1/2}^+ = \frac{\nu}{2}(1-\nu)[1-\varphi(r_j^+)] \ . \tag{6.5}$$

For this choice of the artificial viscosity coefficient, Davis' scheme

$$u^j = u_j - \nu\Delta u_{j-1/2} - \Delta_-[\frac{\nu}{2}(1-\nu)\Delta u_{j+1/2}] + \Delta_-[K_{j+1/2}^+\Delta u_{j+1/2}] \tag{6.6}$$

coincides with Sweby's scheme.

6.B. Considering now the case $a < 0$ for $u_t + au_x = 0$, and adding an upstream-biased artificial viscosity

$$K_{j+1/2}^-(r_{j+1}^-)\Delta u_{j+1/2} - K_{j-1/2}^-(r_j^-)\Delta u_{j-1/2} \tag{6.7}$$

to the Lax-Wendroff scheme obtained by taking $\varphi \equiv 1$ in Sweby's scheme (5.53)

$$u^j = u_j - \nu(1 - \frac{1+\nu}{2}[\varphi(r_{j+1}^-) - \frac{\varphi(r_j^-)}{r_j^-}])\Delta u_{j+1/2} \quad \text{(Sweby, } a < 0) \tag{6.8}$$

(where we have now factored $\Delta u_{j+1/2}$ for upwinding considerations) leads to Davis' scheme for negative characteristic velocity:

$$u^j = u_j - \nu(1 - \frac{1+\nu}{2}[1 - \frac{1}{r_j^-}])\Delta u_{j+1/2} + [K^-_{j+1/2} - \frac{K^-_{j-1/2}}{r_j^-}]\Delta u_{j+1/2} \tag{6.9}$$

here

$$r_j^- = (\Delta u_{j+1/2})/(\Delta u_{j-1/2}) \ . \tag{6.10}$$

Davis' scheme can again be made identical to Sweby's scheme by choosing

$$K^-_{j+1/2} = \frac{\nu}{2}(1+\nu)[\varphi(r^-_{j+1})-1] \ . \tag{6.11}$$

6.C.

By applying Steger and Warming's flux-splitting idea, Davis has shown that one can combine the above two cases into one Lax-Wendroff-type scheme, with an upstream-centered artificial viscosity. Defining new coefficients of artificial viscosity

$$K^+_{j+1/2} = \begin{cases} \frac{\nu}{2}(1-\nu)[1-\varphi(r_j^+)] & \text{if } a > 0 \\ 0 & \text{if } a \le 0 \end{cases} \tag{6.12a}$$

$$K^-_{j+1/2} = \begin{cases} \frac{\nu}{2}(1+\nu)[\varphi(r^-_{j+1})-1] & \text{if } a < 0 \\ 0 & \text{if } a \ge 0 \end{cases} \tag{6.12b}$$

and adding, according to the flux-splitting principle, both artificial viscosity terms (6.1) - (6.7) with their built-in upstream-biasing (see the indices of r in (6.1) - (6.7), and (6.12)) to the Lax-Wendroff scheme in its original, fully centered form (3.2), we obtain Davis' form of a Lax-Wendroff scheme with an upstream-centered artificial viscosity

$$u^j = u_j - \nu(u_{j+1}-u_{j-1}) + \frac{\nu^2}{2}(u_{j+1}-2u_j+u_{j-1})$$

$$+ [K^+_{j+1/2}(r^+_j)+K^-_{j+1/2}(r^-_{j+1})]\Delta u_{j+1/2} - [K^+_{j-1/2}(r^+_{j-1})+K^-_{j-1/2}(r^-_j)]\Delta u_{j-1/2} \tag{6.13}$$

Now in the case of constant or piecewise constant characteristic velocity a, one of the following cases takes place (fig. 6.1 shows the corresponding typical characteristic segments):

(i) $K^+_{j+1/2} \neq 0$, $K^+_{j-1/2} \neq 0$ and $K^-_{j+1/2} = K^-_{j-1/2} = 0$

(ii) $K^+_{j+1/2} = K^+_{j-1/2} = 0$ and $K^-_{j+1/2} \neq 0$, $K^-_{j-1/2} \neq 0$

(iii) $K^+_{j+1/2} \neq 0$, $K^+_{j-1/2} = 0$ and $K^-_{j+1/2} = 0$, $K^-_{j-1/2} \neq 0$

(iv) $K^+_{j+1/2} = 0$, $K^+_{j-1/2} \neq 0$ and $K^-_{j+1/2} \neq 0$, $K^-_{j-1/2} = 0$

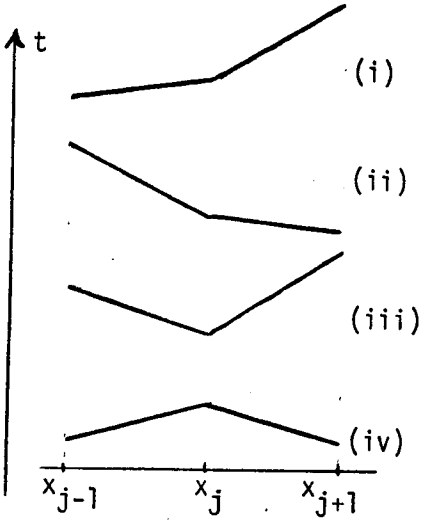

Fig. 6.1 Typical characteristic segments in the (x,t)-plane ($\frac{dx}{dt} = a$)

In case (i) we could replace ν by $|\nu|$ in (6.12) without any change in the artificial viscosity (6.13). In case (iii) the artificial viscosity is

$$Q = \frac{1}{2}\nu_{j+1/2}(1-\nu_{j+1/2})(1-\varphi^+_j)\Delta u_{j+1/2} - \frac{1}{2}\nu_{j-1/2}(1+\nu_{j-1/2})(\varphi^-_j-1)\Delta u_{j-1/2} \tag{6.14}$$

$$\text{with } \nu_{j+1/2} > 0 \text{ , } \nu_{j-1/2} < 0$$

and we obtain the same result if we write this as

$$Q = \frac{1}{2}|\nu_{j+1/2}|(1-|\nu_{j+1/2}|)(1-\varphi^+_j)\Delta u_{j+1/2} - \frac{1}{2}|\nu_{j-1/2}|(1-|\nu_{j-1/2}|)(1-\varphi^-_j)\Delta u_{j-1/2} \tag{6.15}$$

This motivates Davis to redefine the artificial viscosity coefficients as

$$K^{+}_{j+1/2} = \frac{1}{2}|\nu|(1-|\nu|)[1-\varphi(r^{+}_{j})] \tag{6.16a}$$

$$K^{-}_{j+1/2} = \frac{1}{2}|\nu|(1-|\nu|)[1-\varphi(r^{-}_{j+1})] \tag{6.16b}$$

in an attempt to eliminate the task (easy here, for the linear wave equation, but tedious for systems of nonlinear hyperbolic equations) of determining which direction is upwind at each gridpoint.

Obviously, using this new artificial viscosity in scheme (6.13) introduces an additional viscosity into what was, for the linear wave equation, Sweby's scheme with flux-splitting and upwinding. Considering for instance case (i), where ν is positive on both intervals $[x_{j-1},x_j]$, $[x_j,x_{j+1}]$, we have a modified artificial viscosity

$$Q_{new} = [\frac{\nu(1-\nu)}{2}(1-\varphi^{+}_{j}) + \frac{\nu(1-\nu)}{2}(1-\varphi^{-}_{j+1})]\Delta u_{j+1/2} - [\frac{\nu(1-\nu)}{2}(1-\varphi^{+}_{j-1}) + \frac{\nu(1-\nu)}{2}(1-\varphi^{-}_{j})]\Delta u_{j-1/2} \tag{6.17a}$$

as compared with the original artificial viscosity given by (6.12):

$$Q_{old} = \frac{\nu(1-\nu)}{2}[1-\varphi^{+}_{j}]\Delta u_{j+1/2} - \frac{\nu(1-\nu)}{2}[1-\varphi^{+}_{j-1}]\Delta u_{j-1/2} \tag{6.17b}$$

so that

$$Q_{new} - Q_{old} = \frac{\nu}{2}(1-\nu)[(1-\varphi^{-}_{j+1})\Delta u_{j+1/2} - (1-\varphi^{-}_{j})\Delta u_{j-1/2}] \ . \tag{6.18}$$

Now if $0 < \varphi^{-}_{j+1} < 1$ and $0 < \varphi^{-}_{j} < 1$ (near or approaching a shock) we shall therefore have

$$Q_{new} - Q_{old} = O[\frac{\nu(1-\nu)}{2}(u_{j+1}-2u_j+u_{j-1})] > 0 \quad \text{here} \ . \tag{6.18'}$$

If we are using, for example, the limiter φ_D proposed by Davis:

$$\varphi_D(r) = \begin{cases} \min(2r,1) & \text{if } r > 0 \\ 0 & \text{if } r \le 0 \end{cases} \tag{6.19}$$

we see that near the origin of Fig. 5.1, which generally corresponds to the vicinity of a shock (i.e. for small values of r_j^-), we have $\varphi_j^- = 2r_j^- < 1$, $1 - \varphi(r_j^-) > 0$ and thus $Q_{new} > Q_{old}$ while for $r_j^- \ge \frac{1}{2}$, $\varphi_j^- = 1$ and the artificial viscosities are the same. Near $r = 1$, $\varphi = 1$ by (6.19) (see Fig. 5.1), which corresponds to a region of smooth flow and the difference (6.18) should be negligible or zero.

These considerations suggest that the modified Davis scheme (6.13) - (6.16) should be very roughly equal to Sweby's scheme (6.13) - (6.12) away from shocks, but have a tendency to slightly more smear the shocks. A close look at the "shocks" of Fig. 2b, 2c of Davis' paper seems to confirm this (the shock appears to be spread on approximately 7 points (Sweby) and 9 points (Davis)); even at that, the difference is only marginal, and Davis' scheme seems to be an extremely interesting compromise between high accuracy with the accompanying complexity, and nearly as high an accuracy with a much simpler implementation.

Our analysis of the additional artificial viscosity (6.18) in case (i) of Davis' scheme also finds confirmation in Fig. 3.b (Sweby), 3.c (Davis) of [12]; the "entropy glitch" or "kink" in the expansion region has been nearly eliminated by the additional viscosity of Davis' scheme. On the other hand, Davis' scheme beeing nearly identical with Sweby's scheme for $r > \frac{1}{2}$ by (6.18) (at least in case (i)), it enjoys its second-order accuracy (inherited from the Lax-Wendroff and Warming-Beam schemes) away from the shocks. Finally, let us note that using Harten's lemma, Davis also proved that scheme (6.13) - (6.16) with limiter (6.19) is TVD for $|\nu| < 1$.

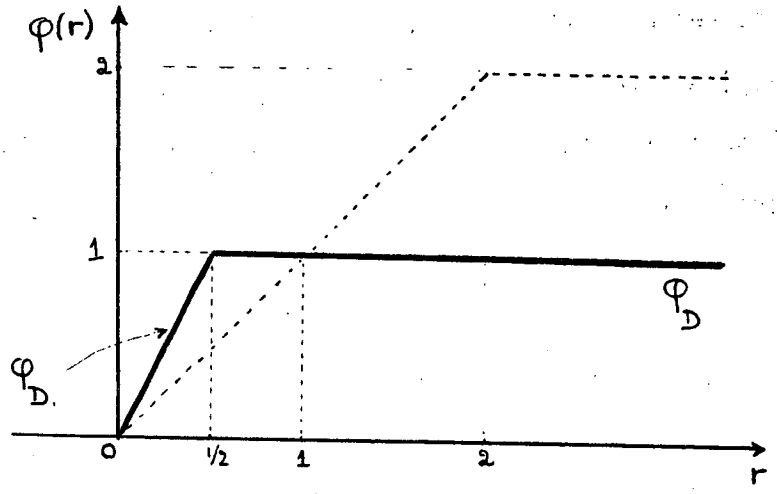

Fig. 6.2 Davis' limiter $\varphi_D(r) = \min(2r,1)$.

7. *GODUNOV AND LEVEQUE'S GEOMETRIC APPROACH TO A GODUNOV-TYPE 2ND-ORDER TVD METHOD*

7.A. *Godunov's method* ([19]).

We consider, for simplicity, the scalar conservation equation

$$u_t + f(u)_x = 0 \ , \ u(x,0) = u_0(x) \tag{7.2}$$

where the flux function $f(u)$ will be assumed to be convex, i.e. $f''(u) > 0$. Godunov's first idea consists in replacing the initial function $u_0(x)$ by a piecewise constant noted u_j in the interval $I = [x_{j-1/2}, x_{j+1/2}]$:

$$u_j = \frac{1}{x_{j+1/2} - x_{j-1/2}} \int_{x_{j-1/2}}^{x_{j+1/2}} u_0(x)dx \ . \tag{7.2}$$

To compute the numerical approximation u^j to $u(x_j,\Delta t)$, Godunov solves (exactly, for the scalar equation, or by an iteration, for system (1.1)) each of

the Riemann problems defined, at the interval interfaces $x_{j-1/2}$, $x_{j+1/2}$, by the replacement (7.2), and thus obtains a solution noted $w(x,t = \Delta t)$ at time Δt . To complete the cycle from $t = 0$ to $t = \Delta t$, one finally averages $w(x,\Delta t)$ on each interval I_j , thus defining the numerical solution u^j at time Δt :

$$u^j = \frac{1}{x_{j+1/2}-x_{j-1/2}} \int_{x_{j-1/2}}^{x_{j+1/2}} w(x,\Delta t)dx \ . \tag{7.3}$$

Rather than computing directly this integral, Godunov uses the integral form of the conservation law

$$\oint_{\partial R_j^n} [udx-f(u)dt] = 0 \tag{7.4}$$

where ∂R_j^n is the boundary of the typical grid-rectangle R_j^n in the (x,t)-plane (Fig. 7.1)

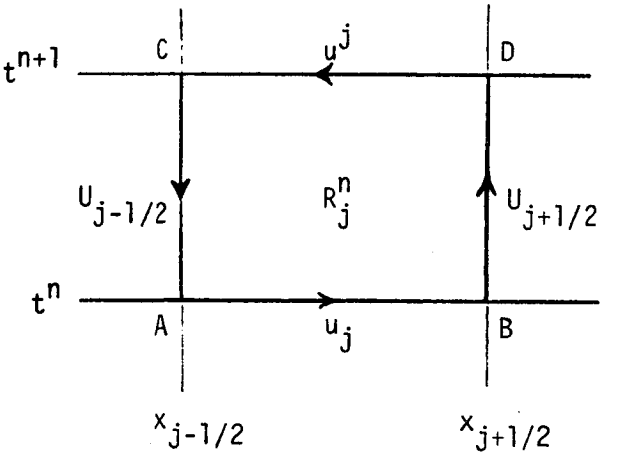

Fig. 7.1 Godunov's integration contour

This leads to Godunov's scheme in conservation form

$$u^j = u_j - \frac{\Delta t}{\Delta x}(F_{j+1/2}-F_{j-1/2}) \tag{7.5}$$

where the numerical flux $F_{j+1/2}$ is defined by

$$F_{j+1/2} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(w(x_{j+1/2},t))dt \quad \begin{pmatrix} n=0 \text{ for the first time} \\ \text{step described above} \end{pmatrix} \tag{7.6}$$

and $w(x_{j+1/2},t)$ is the solution of the Riemann problem at $(x_{j+1/2},t^n)$, which is known to depend only on the adjacent states u_j, u_{j+1} and on $(x-x_{j+1/2})/(t-t^n)$, and is therefore constant along the side BD of R_j^n (For the scalar equation (7.1), this is an immediate consequence of the properties of characteristic lines, see (4.2), or [67]; for the Euler equations of one-dimensional gas dynamics, see [19], [21], Section 13; [30], p. 28). This constant value will be denoted $U_{j+1/2}$, following [19]. Accordingly, we shall write

$$u^j = u_j - \lambda(F_{j+1/2}-F_{j-1/2}) \quad \text{with} \quad F_{j+1/2} \equiv \frac{1}{\Delta t}\int f(U_{j+1/2})dt = f(U_{j+1/2}) \tag{7.7}$$

valid for the transition from (x_j,t^n) to (x_j,t^{n+1}). Again using characteristic theory, we find that away from the sonic point $\bar{u}$,

$$U_{j+1/2} = \begin{cases} u_j & \text{if } f'(u_j) > 0 \text{ and } f'(u_{j+1}) > 0 \\ u_{j+1} & \text{if } f'(u_j) < 0 \text{ and } f'(u_{j+1}) < 0 \end{cases} \tag{7.8}$$

while in the vicinity of the sonic point, say if $u_j < \bar{u} < u_{j+1}$ with $f'(\bar{u}) = 0$,

$$u = u(\xi) \quad \text{with} \quad \xi = (x-x_{j+1/2})/(t-t^n) \tag{7.9a}$$

which is the solution (unique if $f'' > 0$) of the fixed-point equation (see [67], p. 302)

$$f'(u(\xi)) = \xi \tag{7.9b}$$

and then we take

$$U_{j+1/2} = u(\xi = 0) \tag{7.9c}$$

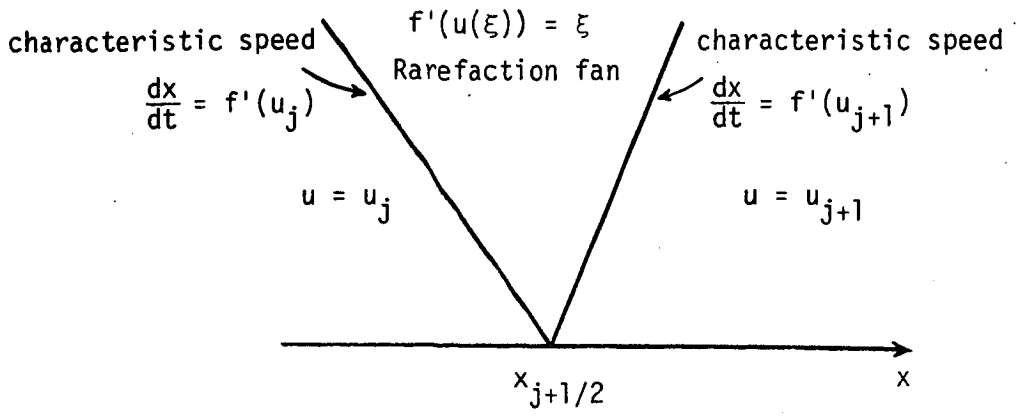

Fig. 7.2 Riemann problem for $u_t + f(u)_x = 0$ in the sonic case with $u_j < \bar{u} < u_{j+1}$ $(f'' > 0)$

On the other hand if $u_{j+1} < \bar{u} < u_j$, we have the "sonic shock" case, and the value $U_{j+1/2}$ depends on the sign of the Rankine-Hugoniot shock speed

$$s = \frac{dx}{dt} = \frac{f(u_j)-f(u_{j+1})}{u_j-u_{j+1}} \tag{7.10}$$

we take

$$U_{j+1/2} = \begin{cases} u_j & \text{if } f(u_j) > f(u_{j+1}) \quad \text{(forward shock)} \\ u_{j+1} & \text{if } f(u_j) < f(u_{j+1}) \quad \text{(backward shock)} \end{cases} \tag{7.11}$$

Godunov's solution procedure is valid whenever the time step Δt is small enough to prevent the adjacent Riemann problems to interfere, which will certainly be the case if

$$\nu \equiv \frac{\Delta t}{\Delta x} \max|f'(u)| \leq \frac{1}{2} . \tag{7.12}$$

Moreover, using Harten's lemma, we can show that Godunov's method is TVD, away from sonic points, provided that $|\nu| \leq 1$; we write

$$u^j = u_j - \lambda[f(U_{j+1/2})-f(U_{j-1/2})] = u_j - \lambda[f(U_{j+1/2})-f(u_j)+f(u_j)-f(U_{j-1/2})] \quad (7.13)$$

with $U_{j+1/2}$ defined by (7.8). Supposing, for example, that $\bar{u} < u_{j-1} < u_j < u_{j+1}$ we get (since $f'' > 0$) $f'(u_j) > 0$ and by (7.8), $U_{j+1/2} = u_j$, $U_{j-1/2} = u_{j-1}$,

$$u^j = u_j - \lambda([f(u_j)-f(u_{j-1})]/\Delta u_{j-1/2})\Delta u_{j-1/2} \quad (7.14)$$

and Harten's coefficients are

$$C_{j-1/2} = \lambda \frac{f(u_j)-f(u_{j-1})}{u_j-u_{j-1}} , \quad D_{j+1/2} = 0$$

so that if $\lambda \max|f'(u)| \le 1$, the method is TVD (the other cases can be treated similarly) under the assumption that we are away from the sonic point $\bar{u}$.

Notice that using (7.14) we can also show directly, with Godunov's criterion (4.9), that Godunov's method is monotonicity preserving, defining

$$a_{j-1/2} = (\Delta f_{j-1/2})/(\Delta u_{j-1/2}) \quad (>0 \text{ by our assumptions}) \quad (7.15a)$$

we write (7.14) as

$$u^j = u_j(1-\lambda a_{j-1/2}) + \lambda a_{j-1/2} u_{j-1} \quad (7.15b)$$

and both coefficients are ≥ 0 provided $\lambda a_{j-1/2} < 1$ (since by assumption $a_{j-1/2} \ge 0$) . In fact (7.15) shows that for (7.1), Godunov's method reduces to the C.I.R. scheme applied to the mesh average $u_{j+1/2}$ of u , rather than to the nodal values of usual difference schemes.

7.B. *Goodman and LeVeque's geometric approach to high resolution TVD schemes*

In Godunov's method, which is only first-order accurate, the main source of error is the loss of information due to the averaging process taking place at

the beginning of each time-step. Following van Leer's idea ([34], [35]), who replaced the piecewise constant approximation of $u_0(x)$ by piecewise first or second-order (Legendre or other) polynomials, thus obtaining second and third-order accurate schemes, Goodman and LeVeque [22] replace the initial function $w(x,t^n)$[1] by a piecewise linear, slope-limited, function $v(x,t^n)$ which is not necessarily continuous at mesh interfaces: on the j-th cell $I_j = (x_{j-1/2},x_{j+1/2})$, it is defined by

$$v(x,t^n) = u_j + s_j(x-x_j) \qquad x_{j-1/2} < x < x_{j+1/2} \tag{7.16}$$

where u_j is Godunov's cell average on I_j.

As Godunov's method, where $s_j \equiv 0$, is TVD for $|\nu| \le 1$, the problem here is to choose the slopes s_j in such a manner that the total variation $TV\{v(x,t^n)\}$ is not larger than that of Godunov's initial function. In [22], this is accomplished by choosing

$$s_j = \begin{cases} \operatorname{sgn}(u_{j+1}-u_j)\min\left[\dfrac{|u_{j+1}-u_j|}{h}, \dfrac{|u_j-u_{j-1}|}{h}\right] & \text{if } \Delta u_{j-1/2} \cdot \Delta u_{j+1/2} > 0 \quad (7.17a) \\ 0 & \text{if } \Delta u_{j-1/2} \cdot \Delta u_{j+1/2} \le 0 \quad (7.17b) \end{cases}$$

We note that (7.17b) does the same for slopes as Sweby's flux limiter ($\varphi(r_j) = 0$ if $r_j \le 0$) did for the numerical flux in (5.8), (5.16); (7.17a) is much stricter than van Leer's limiter for 2^{nd}-order accurate schemes ([34], p. 289) where s_j is allowed to be up to

$$s_j^{VL} = \begin{cases} \dfrac{1}{h}\operatorname{sgn}(\bar{\Delta}_j u) \cdot \min[2|u_j-u_{j-1}|, |\bar{\Delta}_j u|, 2|u_{j+1}-u_j|] \\ \qquad \text{if } \operatorname{sgn}(u_j-u_{j-1}) = \operatorname{sgn}(\bar{\Delta}_j u) = \operatorname{sgn}(u_{j+1}-u_j) \quad (7.18) \\ 0 \quad \text{otherwise} \end{cases}$$

where u_j is the average value of $u(x,t^n)$ (computed) on I_j at time t^n, and $\bar{\Delta}u_j/\Delta x$ is the average gradient of u on I_j at t^n, which may be

---

1 Obtained at the previous time step by the method of Section 7.A.

determined by least-squares fitting of the solution $u(x,t^n)$ just obtained at the previous time step (before the slope limiting process (7.18)). See [34] for details and figures showing the effect of van Leer's slope limiter. We also note that (7.17) corresponds to the case $\Phi = 1$ of Chakravarthy and Osher's limiter (5.32).

Goodman and LeVeque's solution procedure would then consist of

(1) Computing with the help of (7.17) the piecewise linear initial distribution $v(x,t^n)$ from the mesh averages u_j defined by (7.3) but at $t = t^n$

(2) solving (7.1), exactly or approximately, using the piecewise linear "initial" distributions $v(x,t^n)$, to obtain a numerical solution $w(x,t^{n+1})$

(3) averaging $w(x,t^{n+1})$ to define the mesh averages u_j.

If step 2 were performed <u>exactly</u>, this algorithm would certainly be TVD by (7.17) and the properties of characteristics for $u_t + f(u)_x = 0$. Leaning on van Leer's work [34] one could also prove that this method would be second-order accurate away from sonic points or extremas of u. But the main obstacle would be to solve (7.1) exactly with piecewise linear initial data, which is a lot more delicate than solving Riemann problems for (7.1).

To get around this difficulty, Goodman and LeVeque proposed an elegant variant, which consists in steps (1) to (3) above, but after modifying first the flux function $f(u)$, which is replaced by a piecewise linear function $g(u)$, thus enabling us to easily determine the numerical flux function $F_{j\pm1/2}$ defined as in Godunov's method by (7.6) with g instead of f. The method proceeds as follows. First we compute the slopes s_j by (7.17) and determine the piecewise

linear "initial" distribution $v(x,t^n)$ as in (7.16). We define values at the boundaries of the mesh I_j by

$$U_j^{\pm} = u_j \pm s_j(\Delta x/2) \tag{7.19}$$

We now observe that due to the slope limiting (7.17), the four points, U_j^- , U_j^+ , U_{j+1}^- and U_{j+1}^+ are in monotonic order: either

$$\text{(i)}\ \ U_j^- \le U_j^+ \le U_{j+1}^- \le U_{j+1}^+ \ \text{ or } \ \text{(ii)}\ \ U_j^- \ge U_j^+ \ge U_{j+1}^- \ge U_{j+1}^+ \ . \tag{7.20}$$

The approximate flux function $g(u)$ is now defined as the piecewise linear function which interpolates $f(u)$ at these points; the characteristic speed $a(u) = f'(u)$ will be replaced by the discrete slopes of the interpolate $g(u)$:

$$g_j' = \begin{cases} \dfrac{f(U_j^+)-f(U_j^-)}{U_j^+-U_j^-} & \text{if } \ s_j = \dfrac{U_j^+-U_j^-}{\Delta x} \neq 0 \\ f'(u_j) & \text{if } \ s_j = 0 \end{cases} \qquad \text{at } \ x_j \ , \ x_{j+1} \tag{7.21}$$

and problem (7.1) for $u(x,t)$ is now replaced locally on each mesh interval I_j by the approximate problem

$$v_t + g(v)_x = 0 \quad \text{for} \quad t^n \le t \le t^{n+1} \tag{7.22a}$$

with initial condition

$$v(x,t = t^n) = v(x,t^n) \quad \text{(our piecewise linear function (7.16))} \tag{7.22b}$$

To solve (7.22), we write (7.22a) as

$$v_t + g'(v)v_x = 0 \quad t^n \le t \le t^{n+1} \ . \tag{7.23a}$$

LEMMA 7.1. For $t^n \le t \le t^{n+1}$, $g'(v(x_{j+1/2},t))$ is constant if the CFL condition

$$\lambda \max|f'(u)| \le 1 \tag{7.24}$$

is satisfied.

PROOF. Assume for instance that we are in case (i) of (7.20), and away from the sonic point $\bar{u}$: $f(\bar{u}) = 0$. $g'(v(x_{j+1/2},t))$ is the slope of the piecewise linear function $g(u)$ at $u = v(x_{j+1/2},t)$, so if $v(x_{j+1/2},t)$ belongs to the interval (U_j^-,U_j^+) , g' will have the constant value $[f(U_j^+)-f(U_j^-)\ /(U_j^+-U_j^-)]$. But on each interval where g' is constant, say $U_j^- < u < U_j^+$, $v(x,t)$ is the solution of a linear wave equation $v_t + av_x = 0$ with a = constant (provisorily assumed) and initial values $v(x,t = t^n) = v(x,t^n)$ defined by (7.16). We immediately get $v(x,t) = v[x-a(t-t^n),t^n]$. Since $f' \neq 0$ (assumed) and we are in case (i), we have $f' > 0$, $a \equiv g_j'$ defined by (7.21) is > 0 , and

$$v(x_{j+1/2},t) = v[x_{j+1/2}-a(t-t^n),t^n] = U_j^+ - g_j'(t-t^n)s_j \tag{7.25}$$

provided that the CFL-like condition $g'(t-t^n) \le \Delta x$ holds to ensure that $x_{j+1/2} - g_j'(t-t^n) \ge x_{j-1/2}$ so that

$$U_j^- < v(x_{j+1/2}-g_j'(t-t^n),t^n) < U_j^+ \tag{7.26}$$

thus guaranteeing that $g' = [f(U_j^+)-f(U_j^-)]/(U_j^+-U_j^-)$ for $t^n \le t \le t^{n+1}$. We note the importance of the CFL condition in this derivation. □

Using (7.16), (7.19) and Lemma 7.1, and analogous deductions for the case (ii) of (7.20) with $f' < 0$ (again away from the sonic point u), we get

$$v(x_{j+1/2},t) = \begin{cases} U_j^+ - s_j g_j'(t-t^n) & \text{if } f' > 0 \qquad (7.27a) \\ U_{j+1}^- - s_{j+1} g_{j+1}'(t-t^n) & \text{if } f' < 0 \qquad (7.27b) \end{cases}$$

and for the piecewise linear interpolation $g(u)$ of f :

$$g(u) = \begin{cases} f(U_j^+) + (u-U_j^+)g_j' & \text{for } u \text{ between } U_j^- \text{ and } U_j^+ \quad (7.28a) \\ f(U_{j+1}^-) + (u-U_{j+1}^-)g_{j+1}' & \text{for } u \text{ between } U_{j+1}^- \text{ and } U_{j+1}^+ . \quad (7.28b) \end{cases}$$

We are now able to compute the numerical flux $G(U;j+1/2)$ corresponding to problem (7.22), in the context of Godunov's method (7.3) to (7.6):

$$G(U;j+1/2) \equiv \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} g[v(x_{j+1/2},t)]dt \quad (7.29)$$

Applying (7.27a) in case (i) with $f' > 0$, and the definition of $g(u)$, we have

$$g[v(x_{j+1/2},t)] = f(U_j^+) - [U_j^+ - v(x_{j+1/2},t)]g_j' = f(U_j^+) - s_j(t-t^n)(g_j')^2 \quad (7.30)$$

so that

$$G(U;j+1/2) = f(U_j^+) - s_j(g_j')^2 \frac{\Delta t}{2} \equiv G_{j+1/2}^- \quad \text{if} \quad f' > 0 \quad (7.31a)$$

and similarly

$$G(U;j+1/2) = f(U_{j+1}^-) - s_{j+1}(g_{j+1}')^2 \frac{\Delta t}{2} \equiv G_{j+1/2}^+ \quad \text{if} \quad f' < 0 . \quad (7.31b)$$

These formulas lead to the numerical approximation of our solution at time t^{n+1}, for points such that the initial values u_{j-1}, u_j, u_{j+1} are away from the sonic point $\bar{u}$: they are the piecewise constants u^j defined by the conservation scheme

$$u^j = u_j - \lambda[G(U;j+1/2) - G(U;j-1/2)] . \quad (7.32)$$

Goodman and LeVeque's method can now be summarized in the following algorithm:

ALGORITHM 7.1. To advance the solution from t^n to t^{n+1},

(i) Having obtained at the previous steps the average values u_j,

compute the slopes s_j with (7.17) and the cell-boundary values $U_j^{\pm}$ with (7.19).

(ii) Compute the slopes g_j' using (7.21) and, assuming we are away from the sonic point, use (7.31) to define the numerical flux $G(U;j+1/2)$

(iii) Use (7.32) to compute the new piecewise constants u^j at time t^{n+1}.

The sonic case

We now suppose that the sonic point $\bar{u}$, with $f'(\bar{u}) = 0$, lies between U_j^- and U_{j+1}^+. We distinguish a number of cases (the reader might find it useful to draw the corresponding figures for the flux functions f, g) :

Case 1.

$$U_j^- < \bar{u} < U_j^+ \leq U_{j+1}^- \leq U_{j+1}^+ \quad \text{and} \quad f(U_j^+) > f(U_j^-) \quad \text{i.e.} \quad g_j' > 0 \,, \; g_{j+1}' > 0 \,.$$

In this case g_j' and g_{j+1}' have the same sign and we can solve (7.22) as before; the numerical flux $G(U;j+1/2)$ remains equal to

$$G_{j+1/2}^- \equiv f(U_j^+) - s_j (g_j')^2 \frac{\Delta t}{2} \,. \tag{7.31a}$$

Case 2.

$$U_j^- \leq U_j^+ \leq U_{j+1}^- < \bar{u} < U_{j+1}^+ \quad \text{and} \quad f(U_{j+1}^+) < f(U_{j+1}^-) \quad \text{i.e.} \quad g_{j+1}' < 0 \,, \; g_j' < 0 \,.$$

For the same reason nothing is changed, we have $g_j' < 0$ and $g_{j+1}' < 0$, and we take

$$G(U;j+1/2) = G_{j+1/2}^+ \equiv f(U_{j+1}^-) - s_{j+1} (g_{j+1}')^2 \frac{\Delta t}{2} \,. \tag{7.31b}$$

Case 3.

$$U_j^- < \bar{u} < U_j^+ \le U_{j+1}^- \le U_{j+1}^+ \quad \text{and} \quad f(U_j^+) < f(U_j^-) \quad \text{i.e.} \quad g_j' < 0 \ , \ g_{j+1}' > 0 \ .$$

In this case the discontinuity at $x_{j+1/2}$ for $v(x,t^n)$ is a sonic rarefaction ([67], p. 302). On the graph of $v(x,t^n)$, $\bar{u}$ is a value attained for some $x = \bar{\xi}$ with $x_j < \bar{\xi} < x_{j+1/2}$. Since $g_j' < 0$, points x from which the solution $v(x_{j+1/2},t>t^n) = v(x_{j+1/2}-a(t-t^n),t^n)$ should come from cannot be on the left of $x_{j+1/2}$ since otherwise we would then get

$$v(x,t^n) < U_j^+ \Rightarrow f(U_j^+) < g(v) < f(U_j^-) \Rightarrow g'(v) = g_j' < 0$$

so that a point $x < x_{j+1/2}$ would transmit, via $v(x,t) = v(x-a(t-t^n),t^n)$, the value of v toward the left, and not toward $x_{j+1/2}$. In the same manner, since $g_{j+1}' > 0$, we could show that the value $v(x_{j+1/2},t>t^n)$ cannot come from a point located on the right of $x_{j+1/2}$. Therefore it must come from the point $x_{j+1/2}$ itself, and since $\bar{u} < U_j^+$, we must take

$$v(x_{j+1/2},t>t^n) = U_j^+ \tag{7.32}$$

as g' is >0 for $U_j^+ < u < U_{j+1}^-$, thus forbidding the right value U_{j+1}^- due to the wave-propagation mechanism associated with (7.22) or $v_t + av_x = 0$. We then have

$$G(U;j+1/2) = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} g[v(x_{j+1/2},t)]dt = f(U_j^+) \tag{7.31c}$$

Case 4.

$$U_j^- \le U_j^+ \le U_{j+1}^- < \bar{u} < U_{j+1}^+ \quad \text{and} \quad f(U_j^+) < f(U_j^-) \quad \text{i.e.} \quad g_j' < 0 \ , \ g_{j+1}' > 0$$

(sonic rarefaction); we then take

$$v(x_{j+1/2},t>t^n) = U_{j+1}^- \qquad (7.31d)$$

$$G(U;j+1/2) = f(U_{j+1}^-) \ .$$

Case 5.

$U_j^- \le U_j^+ < \bar{u} < U_{j+1}^- \le U_{j+1}^+$ and $f(U_j^+) < f(U_j^-)$ i.e. $g_j' < 0$, $g_{j+1}' > 0$

(sonic rarefaction).

Here $x_{j+1/2}$ is a discontinuity for $v(x,t^n)$, and the sonic point $\bar{u}$ is "buried" somewhere within the jump from U_j^+ to U_{j+1}^- ; to analyze this case more closely, let us introduce an additional interpolation node $(\bar{u},f(\bar{u}))$ in the definition of $g(u)$, which becomes the piecewise linear interpolate of $f(u)$ at U_j^- , U_j^+ , $\bar{u}$, U_{j+1}^- , U_{j+1}^+ .

Defining slopes $g_{j+1/2,+}'$ and $g_{j+1/2,-}'$ with

$$g_{j+1/2,-}' = \frac{f(\bar{u})-f(U_j^+)}{\bar{u}-U_j^+} \ , \ g_{j+1/2,+}' = \frac{f(U_{j+1}^-)-f(\bar{u})}{U_{j+1}^--\bar{u}}$$

we have $g_{j+1/2,-}' < 0$ (to the left of $\bar{u}$) , $g_{j+1/2,+}' > 0$ (to the right of $\bar{u}$) and the considerations of case 3 lead us to the conclusion that the only possible choice for $v(x_{j+1/2},t>t^n)$ is $\bar{u}$, since (i) assuming for instance that $v(x_{j+1/2},t>t^n) = v(x,t^n)$ for some $x < x_{j+1/2}$ would give $f' < 0$, $g' < 0$ and

$$v(x_{j+1/2},t) = v(x_{j+1/2}-g'(v)(t-t^n),t^n) = v(\tilde{x},t^n) \quad \text{for some } \tilde{x} > x_{j+1/2} \ ,$$

a contradiction; and (ii) assuming that we would take any other choice between U_j^+ and U_{j+1}^- for $v(x_{j+1/2},t^n)$ would again lead to a value of $v(x_{j+1/2},t>t^n) = v(x_{j+1/2}-g_j'(t-t^n),t^n) = v(\hat{x},t^n)$ with $\hat{x} \neq x_{j+1/2}$ since we would then have $g_j' \neq 0$ in this formula due to our new interpolation function $g(u)$. Therefore we must take $v(x_{j+1/2},t>t^n) = \bar{u}$, $g[v(x_{j+1/2},t>t^n)] = f(\bar{u})$, and we get

$$G(U;j+1/2) = f(\bar{u}) \ . \tag{7.31e}$$

Cases 3, 4, 5 can be combined in one formula as follows ([22]):

If $g'_j < 0$ and $g'_{j+1} > 0$ then

$$G(U;j+1/2) = f(v_o) \tag{7.33a}$$

where

$$v_o \equiv \min[\max(U_j^+,\bar{u}),U_{j+1}^-] \ . \tag{7.33b}$$

A number of other cases should normally be discussed, but lack of space motivates us to refer the reader, for the discussion of these cases which correspond to the "sonic shock", to Goodman-LeVeque [22] or a forthcoming paper. All cases can be nicely put together ([22]) in the following

ALGORITHM 7.2. (Goodman and LeVeque) The new values at $t = t^{n+1}$, using Godunov's conservation scheme, are

$$u^j = u_j - \lambda[G(U;j+1/2)-G(U;j-1/2)] \tag{7.34a}$$

with the following numerical flux:

(i) If $g'_j > 0$, $g'_{j+1} < 0$ and $g'_{j+1/2}(U_{j+1}^- - U_j^+) = 0$ (7.34b)

where

$$g'_{j+1/2} = \begin{cases} \dfrac{f(U_{j+1}^-)-f(U_j^+)}{U_{j+1}^- - U_j^+} & \text{if } U_{j+1}^- \neq U_j^+ \\ \\ f'(U_j^+) & \text{if } U_{j+1}^- = U_j^+ \end{cases} \tag{7.34c}$$

then

$$G(U;j+1/2) = \begin{cases} G^-_{j+1/2} & \text{if } s_j(g'_j)^2 \geq s_{j+1}(g'_{j+1})^2 \\ G^+_{j+1/2} & \text{otherwise .} \end{cases} \tag{7.34d}$$

(ii) In all other cases,

$$G(U;j+1/2) = \begin{cases} G^-_{j+1/2} & \text{if } g'_j \geq 0 \text{ and } g'_{j+1/2} \geq 0 \quad (7.34e) \\ f(\bar{u}) & \text{if } g'_j < 0 \text{ and } g'_{j+1} > 0 \quad (7.34f) \\ G^+_{j+1/2} & \text{if } g'_{j+1/2} \leq 0 \text{ and } g'_{j+1} \leq 0 . \quad (7.34g) \end{cases}$$

In fact, as long as $g'_j \cdot g'_{j+1} > 0$ we can use the simpler form (7.31) - (7.34a) to compute the numerical flux $G(U;j+1/2)$; resorting to the detailed formulation (i) - (ii) of algorithm 7.2 is only necessary when the sonic point $\bar{u}$ is in the interval spanned by U^-_j , U^+_j , U^-_{j+1} , U^+_{j+1} . In order to show that the method is second-order accurate, one can compare the numerical flux, $G(U;j+1/2)$ in (7.31), in the nonsonic case, with the numerical flux of the Lax-Wendroff scheme, obtained from (3.3) by adding and subtracting $F(U_j)$ in the first bracket:

$$h^{LW}_{j+1/2} = \frac{1}{2}[f(u_{j+1})+f(u_j)] - \frac{\lambda}{2}[a_{j+1/2}(f(u_{j+1})-f(u_j))] \tag{7.35}$$

which can be put in the form

$$h^{LW}_{j+1/2} = \frac{1}{2}[f(u_{j+1})+f(u_j)] - \frac{\Delta t}{2}\frac{[f(u_{j+1})-f(u_j)]^2}{(u_{j+1}-u_j)^2} \cdot \frac{(u_{j+1}-u_j)}{\Delta x} . \tag{7.36}$$

A quick comparison with (7.31), using (7.21) and Taylor expansions, shows that

$$G(U;j+1/2) = h^{LW}_{j+1/2} + O(\Delta x)^2 \tag{7.37}$$

in regions of smooth flow (or for smooth solutions) and Goodman and LeVeque's

method is therefore second-order accurate. It can also be proved to be TVD (see [22]), and preliminary tests seem to be extremely promising.

7.C. *NUMERICAL EXPERIMENTS, CONCLUSION*

Due to the length of this report, we shall omit the inclusion of our own numerical experiments, and only mention that the calculations displayed in the quoted papers show that the numerical methods presented here are among the most efficient, and certainly among the most accurate for shock computations. In a forthcoming paper, we shall give more details on the extension to systems of nonlinear equations and to 2 and 3-dimensional flow computations, as well as numerical experiments in this context. Finally we would like to mention that for transonic flow computations, several other approaches are of great interest and value; among them the work of Roe [48], [49], Jameson-Schmidt-Turkel [28], [51] Angrand-Dervieux et al. [1], Lerat, Lerat-Sides [36], and the optimal control-finite element approach of Bristeau-Glowinski et al. [7].

REFERENCES

[1] F. Angrand, A. Dervieux, L. Loth and G. Vijayasundaram (1983): "Simulations of Euler transonic flows by means of explicit finite-element type schemes", Rapport INRIA No. 250, Rocquencourt, 78153 Le Chesnay, France.

[2] F. Angrand and A. Dervieux (1984): "Some explicit triangular finite element schemes for the Euler equations", Int. J. Num. Meth. in Fluids, Vol. 4, pp. 749-764.

[3] P. Arminjon and C. Beauchamp (1981): "Continuous and discontinuous finite element methods for Burgers' equation", Comp. Meth. Appl. Mech. Engng., Vol. 15, No. 1, pp. 65-84.

[4] P. Arminjon and A. Rousseau (1985): "Discontinuous finite elements and Godunov-type methods for the Eulerian equations of gas dynamics", Comp. Meth. Appl. Mech. Engng., Vol. 49, No. 1, pp. 17-36.

[5] R.M. Beam and R.F. Warming (1976): "An implicit finite-difference algorithm for hyperbolic systems in conservation-law form", J. Comp. Phys., Vol. 22, pp. 87-110.

[6] J.P. Boris and D.L. Book (1973): "Flux corrected transport, I. SHASTA, A fluid transport algorithm that works", J. Comp. Phys., Vol. 11, pp. 38-69.

[7] M. Bristeau, R. Glowinski, B. Mantel, J. Periaux, P. Perrier (1983): "Numerical Methods for the compressible Navier-Stokes equations", Proc. 6th Int. Conf. Computer Methods in Appl. Sciences and Engng, INRIA, Dec. 12-16, 1983.

[8] S.R. Chakravarthy and S. Osher (1983): "High resolution applications of the Osher upwind scheme for the Euler equations", Proc. AIAA Comp. Fluid Dynamics conference, Danvers, MA, pp. 363-372.

[9] P. Colella and P. Woodward (1982): "The piecewise parabolic method (PPM) for gas-dynamical simulations", LBL Report No. 14661. Also: J. Comp. Phys. 54 (1984), pp. 174-201.

[10] R. Courant and K.O. Friedrichs (1948): "Supersonic flow and shock waves", Interscience Publishers, New York, Reprinted (1976) by Springer Verlag.

[11] R. Courant, E. Isaacson and M. Rees (1952): "On the solution of nonlinear hyperbolic differential equations by finite differences", Comm. Pure App. Math. 5, p. 243.

[12] S.F. Davis (1984): "TVD finite difference schemes and artificial viscosity", ICASE NASA Contractor Report 172373 (June 1984).

[13] R.J. DiPerna (1983): "Convergence of approximate solutions to conservation laws", Arch. Rat. Mech. Anal. 82, pp. 27-70.

[14] B. Engquist and S. Osher (1980): "Stable and entropy condition satisfying approximations for transonic flow calculations", Math. Comp. 34, pp. 45-75.

[15] F. Fezoui (1985): "Résolution des équations d'Euler par un schema de van Leer en éléments finis", Rapport INRIA no. 358, Rocquencourt, 78153 Le Chesnay, France

[16] J.E. Fromm (1968): "A method for reducing dispersion in convective difference schemes" J. Comp. Physics vol. 3, pp. 176-189.

[17] J.E. Fromm (1971): "A Numerical study of buoyancy driven flows in room enclosures", in "Lecture Notes in Physics", Vol. 8, pp. 120-126, Springer-Verlag, Berlin.

[18] J. Glimm (1965): "Solutions in the large for nonlinear hyperbolic systems of equations", Comm. Pure Appl. Math. Vol. 18, pp. 697-715.

[19] S.K. Godunov (1959): "A difference scheme for numerical computation of discontinuous solutions of equations of fluid dynamics", Math. Sbornik, Vol. 47, pp. 271-306 (in Russian).

[20] S.K. Godunov, V.R. Riabenki (1977): Schémas aux différences, french translation from the revised russian original (1973), MIR editions, Moscow. English translation also available.

[21] S.K. Godunov, A. Zabrodin, M. Ivanov, A. Kraĭko and G. Prokopov (1979): Résolution numérique des problèmes multidimensionnels de la dynamique des gaz, translated from the Russian edition (1976), MIR, Editor, Moscow. English translation also available.

[22] J.B. Goodman and R.J. LeVeque (1984): "A geometric approach to high resolution TVD schemes", ICASE NASA contractor Report 172484 (October 1984).

[23] On Hancock's contribution: G.D. van Albada, B. van Leer and W.W. Roberts, Jr. (1982): J. Astron. Astrophys. 108, p. 76.

[24] A. Harten, J.M. Hyman and P.D. Lax (1976): "On finite-difference approximations and entropy conditions for shocks", Comm. Pure Appl. Math., 29, pp. 297-322.

[25] A. Harten (1977): "The Artificial Compression Method for computation of shocks and contact discontinuities, I. Single conservation laws", Comm. Pure Appl. Math., Vol. 30, pp. 611-638.

[26] A. Harten (1978): "The Artificial Compression Method for computation of shocks and contact discontinuities, III. Self-adjusting hybrid schemes", Comm. Pure Appl. Math., Vol. 32, pp. 363-389.

[27] A. Harten, P.D. Lax and B. van Leer (1983): "On upstream differencing and Godunov-type schemes for hyperbolic conservation laws", SIAM Review, Vol. 25, pp. 35-61.

[28] A. Jameson, W. Schmidt and E. Turkel (1981): "Numerical solution of the Euler equations by Finite Volume Methods using Runge-Kutta time stepping schemes", AIAA Paper 81-1259.

[29] P.D. Lax and B. Wendroff (1960): "Systems of conservation laws", Comm. Pure Appl. Math., Vol. 13, pp. 217-237.

[30] P.D. Lax (1973): Hyperbolic systems of conservation laws and the mathematical theory of shock waves", CBMS Regional Conference Series in Appl. Math. 11, Soc. for Industrial and Applied Mathematics, Philadelphia.

[31] B. van Leer (1972): "Towards the ultimate conservative scheme, I. The quest of monotonicity", J. Comp. Physics, Vol. 18, p. 163.

[32] B. van Leer (1974): II. "Monotonicity and conservation combined in a second order scheme", J. Comp. Phys. Vol. 14, pp. 361-370.

[33] B. van Leer (1977): III. "Upstream-centered finite difference schemes for ideal compressible flow", J. Comp. Phys. Vol. 23, pp. 263-275.

[34] B. van Leer (1977): IV. "A new approach to numerical convection", J. Comp. Phys. 23, pp. 276-299.

[35] B. van Leer (1979): V. "A second-order sequel to Godunov's method, J. Comp. Phys. 32, pp. 101-136.

[36] A. Lerat and J. Sides (1981): "Proceeding of the Conf. on Num. Meth. in Aeronautical Fluid Dynamics", Univ. of Reading, March 29 - April 1st, 1981.

[37] A. Lerat, J. Sides and V. Daru (1982): "An implicit finite volume scheme for solving the Euler equations", Proc. 8th Int. Conf. Num. Meth. in Fluid Dynamics, Aachen, 1982, edited by E. Krause, Springer-Verlag, pp. 343-349.

[38] R.W. MacCormack (1969): "The effect of viscosity in hypervelocity impact cratering", AIAA Paper 69-354.

[39] R.W. MacCormack (1982) "A numerical method for solving the equations of compressible viscous flow", AIAA Journal, Vol. 20, No. 9.

[40] J. von Neumann and R.D. Richtmyer (1950): "A method for the numerical calculations of hydrodynamical shocks", J. Appl. Phys., Vol. 21, p. 232 (1950).

[41] O. Oleinik (1957): "Discontinuous solutions of nonlinear differential equations", Uspekhi Mat. Nauk.(N.S.) 12, No. 3, pp. 3-73 (Amer. Math. Soc. Transl., Ser. 2,26, pp. 95-172).

[42] S. Osher (1984): "Riemann solvers, the entropy condition, and difference approximations", SIAM J. Numer. Anal. 21, No. 2, pp. 217-235.

[43] S. Osher (1985): "Convergence of generalized MUSCL schemes", SIAM J. Numer. Anal. 22, No. 5, pp. 947-961.

[44] S. Osher and S. Chakravarthy (1984): "Very high order accurate TVD schemes", ICASE Report. no 84-44, NASA-Langley, Virginia 23665.

[45] S. Osher and S. Chakravarthy (1984): "High resolution schemes and the entropy condition", SIAM J. Num. Anal., vol. 21, pp. 955-984.

[46] R. Peyret and T.D. Taylor (1983): "Computational Methods for Fluid Flow", Springer-Verlag, New York, Heidelberg, Berlin.

[47] R.D. Richtmyer and K.W. Morton (1967): "Difference Methods for Initial Value Problems", J. Wiley, New York.

[48] P.L. Roe (1981a): "The use of the Riemann problem in finite difference schemes", in Proc. 7th. Intern. Conf. Num. Meth. Fluid Dynamics, Stanford/NASA Ames, W.C. Reynolds and R. Mac Cormack, editors, Lecture Notes in Physics, No. 141, Springer-Verlag, New York, pp. 354-359.

[49] P.L. Roe (1981b): "Approximate Riemann solvers, parameter vectors, and difference schemes", J. Comp. Phys. 43, pp. 357-372.

[50] R. Sanders (1983): "On convergence of monotone difference schemes with variable spatial differencing", Math. Comp., Vol. 40, pp. 91-106.

[51] W. Schmidt and A. Jameson (1983): "Euler solvers as an analysis tool for aircraft aerodynamics", in: Recent advances in numerical methods in fluids, Vol. 4, W.G. Habashi (Ed.) Pineridge Press, Swansea, U.K.

[52] Y.L. Shokin (1983): "The Method of Differential Approximation", Springer-Verlag, Berlin, Heidelberg, New York.

[53] G.A. Sod (1978): "A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws", J. Comp. Phys, Vol. 27, pp. 1-31.

[54] K. Srinivas, J. Gururaja and K. Krishna Prasad (1976): "An Assessment of the Quality of Selected Finite Difference Schemes for Time Dependent Compressible Flows", J. Comp. Phys. 20, No. 2, pp. 140-159.

[55] J. Steger and R.F. Warming (1981): "Flux vector splitting of the inviscid gasdynamics equations with applications to finite-difference methods", J. Comp. Phys., Vol. 40, pp. 263-293.

[56] P.K. Sweby (1984): "High resolution schemes using flux limiters for hyperbolic conservation laws", SIAM J. Num. Anal., Vol. 21, pp. 995-1011.

[57] G. Vijayasundaram (1983): "On numerical schemes for solving the Euler equations of gas dynamics", Proc. Workshop on Num. Meth. for the Euler Equations for compressible inviscid fluids, INRIA, Versailles-Rocquencourt, Dec. 1983, R. Glowinski Editor, to appear.

[58] R.F. Warming, R.M. Beam and B.J. Hyett (1975), Math. Comp. 29, p. 1037.

[59] R.F. Warming and R.M. Beam (1976): "Upwind second-order schemes and applications in aerodynamic flow", AIAA Journal, Vol. 14, No. 9, pp. 1241-1249.

[60] R.F. Warming and R.M. Beam (1978): "On the construction and application of implicit factored schemes for conservation laws", Symposium on Computational Fluid Dynamics, New York, April 16-17, 1977, SIAM-AMS Proc. 11, p. 85.

[61] H.C. Yee, R.F. Warming and A. Harten (1983): "Implicit total variation diminishing (TVD) schemes for steady state calculations", Proc. AIAA Comp. Fluid Dynamics Conference, Danvers, MA, pp. 110-127.

[62] A.J. Chorin and J.E. Marsden (1979): "A Mathematical Introduction to Fluid Mechanics", Springer-Verlag, New York, Heidelberg, Berlin.

[63] P. Arminjon and C. Laroche (1986): "Transonic flow calculations by TVD second-order methods", in preparation.

[64] A. Harten (1983): "High resolution schemes for hyperbolic conservation laws", J. Comp. Phys. 49, pp. 357-393.

[65] A. Majda (1984): "Compressible Fluid Flow and Systems of Conservation laws in several space variables, 159 + viii p., Springer-Verlag, New York.

[66] A. Majda and S. Osher (1979): "Numerical viscosity and the entropy condition", Comm. Pure Appl. Math. 32, pp. 797-838.

[67] J. Smoller (1983): "Shock waves and the reaction-diffusion equations", Springer-Verlag, Berlin, Heidelberg, New York.

[68] E. Tadmor (1983): "Numerical viscosity and the entropy condition for conservative difference schemes", NASA Contractor Report 172141, ICASE.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique