JOURNAL OF COMPUTATIONAL PHYSICS 23, 276–299 (1977)

Towards the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection

BRAM VAN LEER

University Observatory, Leiden, The Netherlands

Received April 30, 1976; revised July 30, 1976

An approach to numerical convection is presented that exclusively yields upstream-centered schemes. It starts from a meshwise approximation of the initial-value distribution by simple basic functions, e.g., Legendre polynomials. In every mesh the integral of the distribution is conserved. The overall approximation need not be continuous. The approximate distribution is convected explicitly and then remapped meshwise in terms of the basic functions. The weights of the basic functions that approximate the initial values in a mesh may be determined by finite differencing, but the most accurate schemes are obtained by least-squares fitting. In the latter schemes, the weights of the basic functions must be regarded as independent state quantities and must be stored separately. Examples of second-order and third-order schemes are given, and the accuracy of these schemes is discussed. Several monotonicity algorithms, designed to prevent numerical oscillations, are indicated. Numerical examples are given of linear and nonlinear wave propagation, also regarding monotonicity.

1. Introduction

The approach to numerical convection described below (Sections 2–4) originated during my attempts to construct upstream-centered schemes for the conservation laws of compressible flow. Its roots lie in Godunov's numerical treatment [2] of the Lagrangean flow equations.

As explained in the previous paper [1], the common finite-difference formulation is impractical when transforming upstream convective schemes into conservative schemes for compressible flow. The convective schemes of the present paper are cast in a form that makes a better starting point for constructing such conservative schemes. The actual construction of schemes for compressible flow will be discussed in the next installment [11] of this series; a short description of the procedure can be found in [8]. The resulting schemes may be regarded as higher-order sequels to Godunov's method.

The present convection approach exclusively yields upstream-centered schemes. This is accomplished by first replacing the true initial-value distribution per mesh by a simple approximating function and then convecting the resulting distribution exactly. Besides the average value in the mesh, other parameters of the mesh functions may be integrated along as independent quantities, rather than being determined instantaneously by finite-differencing. This has a number of advantages, one of them being a potentially higher accuracy.

ISSN 0021-9991

Avoiding numerical oscillations becomes a trivial matter. The results on monotonicity from Van Leer [3] therefore are reformulated and elucidated (Section 5).

Some numerical results for linear and nonlinear convection problems are displayed in Section 6.

The notation used in this paper differs from the one used in previous installments. The reason is that, in the present approach, mesh averages play the role that in the usual finite-difference approach is assigned to nodal-point values. The new notation is compiled in Table I.

TABLE I

Notation Used in the Grid

Symbol	Definition
x_i	$x_0 + i\Delta x$, mesh boundary
$x_{i+(1/2)}$, $\overline{x}_{i+(1/2)}$	$x_0 + (i + \frac{1}{2})\Delta x$, mesh center
t^0	Initial time level
t^1	$t^0 + \Delta t$, final time level
$\overline{w}_{i+(1/2)}$	Average value of w in mesh (x_i , x_{i+1}) at t^0
$\overline{w}^{i+(1/2)}$	Average value of w in mesh (x_i , x_{i+1}) at t^1
$\langle w_i \rangle$	Average value of w at the boundary x_i during time step
$\Delta_i \overline{w}$	$\overline{w}_{i+(1/2)} - \overline{w}_{i-(1/2)}$
$\overline{\Delta}_{i+(1/2)} w/\Delta x$	Average gradient of w in mesh (x_i , x_{i+1}) at t^0
$\overline{\Delta^2}_{i+(1/2)} w/(\Delta x)^2$	Average second derivative of w in mesh (x_i , x_{i+1}) at t^0
λ	$\Delta t/\Delta x$, mesh ratio
σ	λa, signed Courant number

2. Second-order Schemes

In Godunov's first-order scheme [2] for the Lagrangean equations of ideal compressible flow, the fluid is described as a sequence of slabs rather than a sequence of point probes. When the scheme is applied to the single linear convection equation

$$\frac{\partial w}{\partial t} + a \frac{\partial w}{\partial x} = 0, \tag{1}$$

where a is a constant, it boils down to the following procedure.

Step 1. Given the complete initial-value distribution $W(t^0, x)$, determine the mesh averages

$$\overline{w}_{i+(1/2)} = \frac{1}{\Delta x} \int_{x_i}^{x_{i+1}} W(t^0, x)\, dx. \tag{2}$$

Step 2. Replace the original initial-value distribution by the piecewise constant function

$$w(t^0, x) = \overline{w}_{i+(1/2)} , \qquad x_i < x < x_{i+1} . \tag{3}$$

Colloquially speaking, we have homogenized the slabs.

Step 3. Starting from the approximate initial values (3), integrate Eq. (1) over a finite time-step Δt. This is achieved by shifting the distribution $w(t^0, x)$ over a distance $a\,\Delta t = \sigma\,\Delta x$ along the x-axis:

$$W(t^1, x) = w(t^0, x - \sigma\,\Delta x). \tag{4}$$

In view of the probable application of the scheme to systems of equations it is practical (although not necessary) to restrict σ by the usual Courant–Friedrichs–Lewy (CFL) condition

$$|\,\sigma\,| \leqslant 1. \tag{5}$$

Thus, the shift will never be greater than Δx.

Step 4 ($\equiv$ *Step* 1). Determine the new mesh averages

$$\bar{w}^{i+(1/2)} = \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} W(t^1, x)\,dx. \tag{6}$$

These steps are illustrated in Fig. 1.

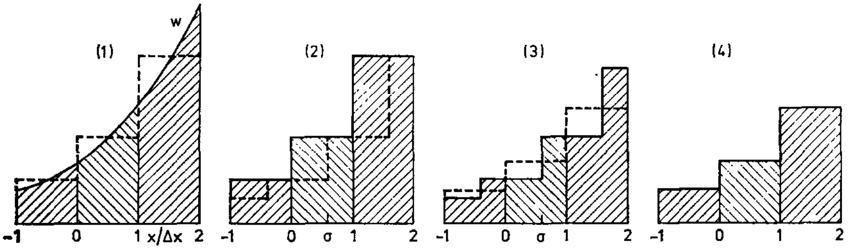

FIG. 1. The first-order upstream-centered scheme. (1) Determining the mesh averages (broken line) of the initial-value distribution (solid line). (2) The approximate initial-value distribution before (solid) and after (broken) convection over a distance $\sigma\Delta x$. (3) Determining the new mesh averages (broken) of the convected distribution (solid). (4) The initial values for the next time step.

The numerical outcome of the above procedure is

$$\bar{w}^{1/2} = \begin{cases} (1 - |\,\sigma\,|)\,\bar{w}_{1/2} + |\,\sigma\,|\,\bar{w}_{-1/2} = \bar{w}_{1/2} - \sigma\Delta_0\bar{w} & \text{if } \sigma \geqslant 0, \\ (1 - |\,\sigma\,|)\,\bar{w}_{1/2} + |\,\sigma\,|\,\bar{w}_{3/2} = \bar{w}_{1/2} - \sigma\Delta_1\bar{w} & \text{if } \sigma < 0. \end{cases} \tag{7}$$

This is precisely the upstream-centered scheme of Courant, Isaacson, and Rees (CIR) [4] applied to mesh averages of w instead of nodal-point values (cf. Van Leer [1, Eq. (10)]).

Scheme (7) can also be related to the integral version of Eq. (1) in one space-time mesh, that is, to

$$\int_{x_0}^{x_1} w(t, x)\,dx\,\Big|_{t_0}^{t^1} + \int_{t_0}^{t^1} aw(t, x)\,dt\,\Big|_{x_0}^{x_1} = 0. \tag{8}$$

This equation is equivalent to

$$(\bar{w}^{1/2} - \bar{w}_{1/2})\,\Delta x + (\langle aw_1\rangle - \langle aw_0\rangle)\,\Delta t = 0, \tag{9}$$

where the angled brackets denote averaging over the time step. Equation (9) is important in formulating the scheme for conservation laws (cf. Section 6). With the piecewise constant initial values (3) we get

$$\langle aw_i \rangle = \begin{cases} a\bar{w}_{i-(1/2)} & \text{if } a \geqslant 0, \\ a\bar{w}_{i+(1/2)} & \text{if } a < 0, \end{cases} \tag{10}$$

and the familiar form (7) results. Note that, while Eq. (9) is exact, scheme (7) is only first-order accurate since the time averages $\langle aw_i \rangle$ are derived from the crudest possible approximation of the true initial-value distribution.

Once we recognize this, extension of the scheme towards a higher order of accuracy becomes a straightforward matter. All we have to do is replace the true initial-value distribution $W(t^0, x)$ by a piecewise approximation that has a higher order of accuracy than (3). In view of Eq. (2), where the mesh average of W is defined with respect to a constant weight function, it seems natural to further approximate W piecewise in terms of Legendre polynomials:

$$w(t^0, x) = \bar{w}_{i+(1/2)} + (b_1)_{i+(1/2)} \frac{x - x_{i+(1/2)}}{\frac{1}{2}\Delta x} + (b_2)_{i+(1/2)} \left\{ \left(\frac{x - x_{i+(1/2)}}{\frac{1}{2}\Delta x} \right)^2 - \frac{1}{3} \right\} + \cdots,$$
$$x_i < x < x_{i+1}. \tag{11}$$

This ensures that integrating $w(t^0, x)$ with constant weight over any mesh (x_i, x_{i+1}) yields the proper average $\bar{w}_{i+(1/2)}$.

Let us first consider the possibilities of fitting the initial data in each mesh by a linear function. We may write

$$w(t^0, x) = \bar{w}_{i+(1/2)} + \frac{\bar{\Delta}_{i+(1/2)}w}{\Delta x}(x - x_{i+(1/2)}), \qquad x_i < x < x_{i+1}, \tag{12}$$

where

$$\frac{\bar{\Delta}_{i+(1/2)}w}{\Delta x} \equiv \overline{\left(\frac{\partial w}{\partial x} \right)}_{i+(1/2)} \tag{13}$$

is some average of the gradient of $W(t,^0 x)$ in the mesh (x_i, x_{i+1}). The evaluation of this average gradient belongs to Step 1; what sort of average is taken will be left open for the moment. Equation (12) replaces Eq. (3) in Step 2; Step 3 remains the same. The new procedure is illustrated in Fig. 2. Note that $w(t^0, x)$, although smoother than in Eq. (3), is still discontinuous. The scheme now updates $\bar{w}$ with second-order accuracy; for $\sigma \geqslant 0$ we get

$$\bar{w}^{1/2} = \bar{w}_{1/2} - \sigma\Delta_0\bar{w} - (\sigma/2)(1 - \sigma)(\bar{\Delta}_{1/2}w - \bar{\Delta}_{-1/2}w). \tag{14}$$

This, again, is an upstream scheme; a central-difference scheme such as the Lax–Wendroff [5] scheme could never result from the procedure followed above. With respect to the integral equation (9) we have, for $\sigma \geqslant 0$,

$$\langle aw_i \rangle = a\{\bar{w}_{i-(1/2)} + \tfrac{1}{2}(1 - \sigma)\,\bar{\Delta}_{i-(1/2)}w\}. \tag{15}$$

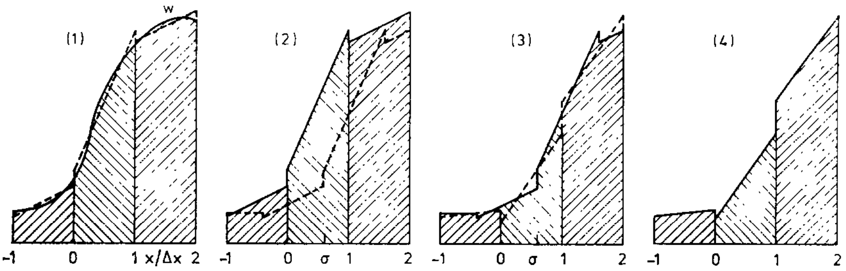

FIG. 2. The second-order upstream-centered scheme (in particular, scheme III). (1) approximating the initial-value distribution (solid line) in each slab by a linear distribution (broken line) with the same mesh integral. In this case the slopes are determined by least-squares fitting. (2) The approximate initial-value distribution before (solid) and after (broken) convection over a distance $\sigma\Delta x$. (3) Determining the new linear distributions (broken) in each mesh by least-squares fitting to the convected distribution (solid). (4) The initial values for the next time step.

The quality of scheme (14) varies considerably with the choice of $\bar{\Delta}w$. I shall demonstrate this on the basis of three examples. It is assumed everywhere that $\sigma \geqslant 0$.

SCHEME I. Determine $\bar{\Delta}w$ by central differencing of $\bar{w}$:

$$\bar{\Delta}_{i+(1/2)}w = \tfrac{1}{2}(\bar{w}_{i+(3/2)} - \bar{w}_{i-(1/2)}) = \tfrac{1}{2}(\Delta_i\bar{w} + \Delta_{i+1}\bar{w}). \tag{16}$$

Inserting this into Eq. (14) yields

$$\bar{w}^{1/2} = \bar{w}_{1/2} - \sigma\Delta_0\bar{w} - (\sigma/4)(1-\sigma)(\Delta_1\bar{w} - \Delta_{-1}\bar{w}), \tag{17}$$

which is just the finite-difference scheme of Fromm [6] applied to mesh averages of w instead of nodal-point values (cf. Van Leer [1, Eq. (34)]). Denoting a translation over $+\Delta x$ by the operator T, we may write (17) as

$$\bar{w}^{1/2} = \{1 - \sigma(1-T^{-1}) - (\sigma/4)(1-\sigma)\,T(1+T^{-1})(1-T^{-1})^2\}\,\bar{w}_{1/2}\,. \tag{18}$$

SCHEME II. Determine $\bar{\Delta}w$ by differencing $W(t^0, x)$:

$$\bar{\Delta}_{i+(1/2)}w = W(t^0, x_{i+1}) - W(t^0, x_i). \tag{19}$$

Defined this way, the quantity $\bar{\Delta}w$ is independent of the quantity $\bar{w}$ and must be integrated along, requiring a separate storage location. Unlike what we are used to in finite differencing, the scheme for updating $\bar{\Delta}w$ differs from the scheme for updating $\bar{w}$. We have

$$\bar{\Delta}^{1/2}w = \Delta_0\bar{w} + (\tfrac{1}{2} - \sigma)(\bar{\Delta}_{1/2}w - \bar{\Delta}_{-1/2}w), \tag{20}$$

and the full scheme defined in Eqs. (14) and (20) can be written as

$$\begin{pmatrix}\bar{w}\\ \bar{\Delta}w\end{pmatrix}^{1/2} = \begin{pmatrix}1-\sigma+\sigma T^{-1} & -(\sigma/2)(1-\sigma)(1-T^{-1})\\ 1-T^{-1} & (\tfrac{1}{2}-\sigma)(1-T^{-1})\end{pmatrix}\begin{pmatrix}\bar{w}\\ \bar{\Delta}w\end{pmatrix}_{1/2}. \tag{21}$$

The matrix occurring in this equation will be called G^{II}.

The above scheme has one peculiarity: for vanishingly small Courant number it yields a nonvanishing change in $\bar{\Delta}w$. This is a consequence of the discontinuities in $w(t^0, x)$ at the mesh boundaries. The scheme wants to annihilate these discontinuities by adjusting the gradients in each mesh, and will ultimately succeed if applied often enough (that is, with $\sigma = +0$). I shall come back to this property in the error analysis of the scheme in Section 3.

SCHEME III. Determine $\bar{\Delta}w$ such that in each mesh $w(t^0, x)$ has the same first moment as $W(t^0, x)$. Thus, the integrated square error in the approximation of W is minimized per mesh, with respect to a constant weight function. In formula,

$$\bar{\Delta}_{i+(1/2)}w = \frac{\int_{x_i}^{x_{i+1}} W(t^0, x) \cdot \{(x - x_{i+(1/2)})/\Delta x\}\, dx}{\int_{x_i}^{x_{i+1}} \{(x - x_{i+(1/2)})/\Delta x\}^2\, dx}$$
$$= \frac{12}{(\Delta x)^2} \int_{x_i}^{x_{i+1}} W(t^0, x) \cdot (x - x_{i+(1/2)})\, dx. \tag{22}$$

Again, $\bar{\Delta}w$ is independent of $\bar{w}$. The scheme for updating $\bar{\Delta}w$ becomes

$$\bar{\Delta}^{1/2}w = (1 - \sigma)(1 - 2\sigma - 2\sigma^2)\, \bar{\Delta}_{1/2}w - \sigma(3 - 6\sigma + 2\sigma^2)\, \bar{\Delta}_{-1/2}w + 6\sigma(1 - \sigma)\, \Delta_0\bar{w}, \tag{23}$$

and the full scheme III can be written as

$$\begin{pmatrix} \bar{w} \\ \bar{\Delta}w \end{pmatrix}^{1/2} = \begin{pmatrix} 1-\sigma+\sigma T^{-1} & -(\sigma/2)(1-\sigma)(1-T^{-1}) \\ 6\sigma(1-\sigma)(1-T^{-1}) & (1-\sigma)(1-2\sigma-2\sigma^2) - \sigma(3-6\sigma+2\sigma^2)\, T^{-1} \end{pmatrix} \begin{pmatrix} \bar{w} \\ \bar{\Delta}w \end{pmatrix}_{1/2} \tag{24}$$

The matrix occurring in this formula will be called G^{III}.

3. Accuracy of the Second-order Schemes

In order to compare the accuracy of the three sample schemes of Section 2, let us assume oscillatory initial values

$$W(t^0, x) = W_{00}e^{2\pi ix/l} = W_{00}e^{i\alpha x/\Delta x}, \tag{25}$$

with

$$\alpha = 2\pi\, \Delta x/l. \tag{26}$$

Without yet specifying the values of $\bar{w}$ and $\bar{\Delta}w$ we know that

$$T = e^{i\alpha}. \tag{27}$$

Scheme I boils down to multiplication of $\bar{w}_{1/2}$ with the amplification factor

$$g^{\mathrm{I}} = 1 - \sigma\left(\frac{1+\sigma}{2} - \frac{1-\sigma}{2}\cos\alpha\right)(1 - \cos\alpha)$$
$$- i\sigma\left(\frac{3-\sigma}{2} - \frac{1-\sigma}{2}\cos\alpha\right)\sin\alpha. \tag{28}$$

For each of the schemes II and III we find two distinct amplification factors, namely, the eigenvalues $g_1^{\mathrm{II}}, g_2^{\mathrm{II}}$ and $g_1^{\mathrm{III}}, g_2^{\mathrm{III}}$ of the matrices G^{II} and G^{III} defining those schemes. Their values are

$$\begin{aligned} g_{1,2}^{\mathrm{II}} &= e^{-i\alpha/2}[\tfrac{1}{2}\cos(\alpha/2) + i(1-2\sigma)\sin(\alpha/2) \\ &\quad \pm \tfrac{1}{2}\{\tfrac{1}{2} + 4\sigma - 4\sigma^2 + (\tfrac{1}{2} - 4\sigma + 4\sigma^2)\cos\alpha\}^{1/2}], \end{aligned} \tag{29}$$

$$\begin{aligned} g_{1,2}^{\mathrm{III}} &= e^{-i\alpha/2}[(1 - 3\sigma + 3\sigma^2)\cos(\alpha/2) + i(1-2\sigma)(1+\sigma-\sigma^2)\sin(\alpha/2) \\ &\quad \pm \sigma(1-\sigma)\{2(5+\sigma-\sigma^2) - (1+2\sigma-2\sigma^2)\cos\alpha - 3i(1-2\sigma)\sin\alpha\}^{1/2}]. \end{aligned} \tag{30}$$

For any properly upstream-centered scheme we can prove that the amplification factors satisfy the equation

$$g(1-\sigma) = e^{-i\alpha}g^*(\sigma), \tag{31}$$

in which the asterisk denotes the complex conjugate. One may check this for the factors given in (28), (29) and (30), and for the factor in the exact solution

$$W(t^1, x) = e^{-i\sigma\alpha}W(t^0, x). \tag{32}$$

Equation (31) is equivalent to the pair of symmetry relations

$$| g(1-\sigma)| = | g(\sigma)|, \tag{33}$$

$$\arg g(1-\sigma) + (1-\sigma)\alpha = -\{\arg g(\sigma) + \sigma\alpha\}. \tag{34}$$

Equation (33) implies that $| g(\sigma)|$ has an extremum for $\sigma = \frac{1}{2}$, which turns out to be a minimum. Hence, the dissipative error per time step, $1 - | g(\sigma)|$, has a maximum. On the other hand, Eq. (34) shows that the phase error per time step, $\arg g(\sigma) - (-\sigma\alpha)$, goes through zero for $\sigma = \frac{1}{2}$, regardless of the value of α. More generally, the result of a time step with $\sigma = \sigma'$ followed by a time step with $\sigma = 1 - \sigma'$ has the correct phase $-\alpha$. This is a quantitative formulation of the property Fromm [6] indicated by saying that scheme I has a "zero-average phase error."

Given this dependence of the errors on σ, it is practical to confine ourselves to comparing the dissipation of the schemes for $\sigma = \frac{1}{2}$, and their dispersion for $\sigma \rightarrow 0$.

With $\sigma = \frac{1}{2}$ we find from Eqs. (28–30) that

$$| g^{\mathrm{I}} | = |(\tfrac{5}{4} - \tfrac{1}{4}\cos\alpha)\cos(\alpha/2)|, \tag{35}$$

$$| g_{1,2}^{\mathrm{II}} | = | \tfrac{1}{2}\cos(\alpha/2) \pm \tfrac{1}{2}(\tfrac{3}{2} - \tfrac{1}{2}\cos\alpha)^{1/2} |, \tag{36}$$

$$| g_{1,2}^{\mathrm{III}} | = | \tfrac{1}{4}\cos(\alpha/2) \pm \tfrac{3}{4}(\tfrac{7}{6} - \tfrac{1}{6}\cos\alpha)^{1/2} |. \tag{37}$$

In judging the accuracy of schemes II and III, only g_1^{II} and g_1^{III} (with the plus sign) are relevant. For small α we obtain

$$| g^{\mathrm{I}} | = 1 - 3\alpha^4/128 + O(\alpha^6), \tag{38}$$

$$| g_1^{\mathrm{II}} | = 1 - \alpha^4/128 + O(\alpha^6), \tag{39}$$

$$| g_1^{\mathrm{III}} | = 1 - \alpha^4/384 + O(\alpha^6). \tag{40}$$

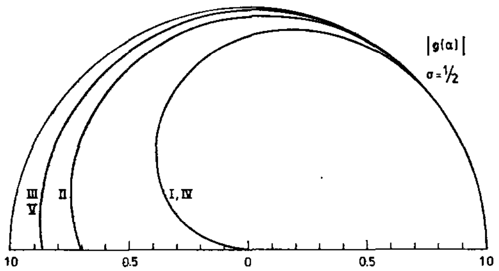

FIG. 3. Dissipation in schemes I–V. Polar plots of the damping factors per time step $|g^{\mathrm{I,IV}}|$, $|g_1^{\mathrm{II}}|$ and $|g_1^{\mathrm{III,V}}|$ as a function of the wavenumber $\alpha = 2\pi\Delta x/l$ of the wave, for Courant number $\frac{1}{2}$.

As could be expected, the error decreases from scheme I to scheme III. Polar plots of $|g^{\mathrm{I}}|$, $|g_1^{\mathrm{II}}|$, and $|g_1^{\mathrm{III}}|$ are given in Fig. 3.

In order to understand the meaning of the remaining amplification factors $g_2^{\mathrm{II,III}}$, consider for instance scheme III. Its eigenvalues g_1^{III} and g_2^{III} correspond to the eigenvectors

$$v_1^{\mathrm{III}} \equiv \begin{pmatrix} v_{11}^{\mathrm{III}} \\ v_{12}^{\mathrm{III}} \end{pmatrix} = \begin{pmatrix} \dfrac{\cos(\alpha/2) + \{7/6 - (\cos\alpha)/6\}^{1/2}}{2\cos(\alpha/4)} \\ 4i\sin(\alpha/4) \end{pmatrix}, \tag{41}$$

$$v_2^{\mathrm{III}} \equiv \begin{pmatrix} v_{21}^{\mathrm{III}} \\ v_{22}^{\mathrm{III}} \end{pmatrix} = \begin{pmatrix} 4\sin(\alpha/4) \\ -12i\dfrac{\cos(\alpha/2) + \{7/6 - (\cos\alpha)/6\}^{1/2}}{2\cos(\alpha/4)} \end{pmatrix}. \tag{42}$$

The vectors have been normalized such that they do not vanish for any value of α; otherwise the normalization is arbitrary. The eigenvectors v_1^{III} and v_2^{III} in turn define eigenfunctions V_1^{III} and V_2^{III}:

$$V_1^{\mathrm{III}} = v_{11}^{\mathrm{III}} + v_{12}^{\mathrm{III}}\frac{x - \bar{x}}{\Delta x}, \tag{43}$$

$$V_2^{\mathrm{III}} = v_{21}^{\mathrm{III}} + v_{22}^{\mathrm{III}}\frac{x - \bar{x}}{\Delta x}. \tag{44}$$

Any piecewise linear approximation of the oscillatory initial values (25) can be written, per mesh, as a combination of V_1^{III} and V_2^{III}.

The meaning of these eigenfunctions is clarified in Fig. 4. On top, an oscillation with wavelength $l = 6\,\Delta x$ (that is, $\alpha = \pi/3$) is approximated by

$$w(t^0, x) = \frac{\bar{w}_{i+(1/2)}}{v_{11}^{\mathrm{III}}} V_1^{\mathrm{III}}(x; \bar{x} = x_{i+(1/2)}), \qquad x_i < x < x_{i+1}. \tag{45}$$

The slopes of the line segments used in the meshes neatly follow the wave present in the mesh averages. Next, the same wave is approximated by

$$w(t^0, x) = \frac{\bar{w}_{i+(1/2)}}{v_{21}^{\mathrm{III}}} V_2^{\mathrm{III}}(x; \bar{x} = x_{i+(1/2)}), \qquad x_i < x < x_{i+1}. \tag{46}$$

The slopes of the line segments now are inconsistent with the mesh averages. They bear the wrong sign and, as seen from Eq. (42), remain finite when α vanishes.

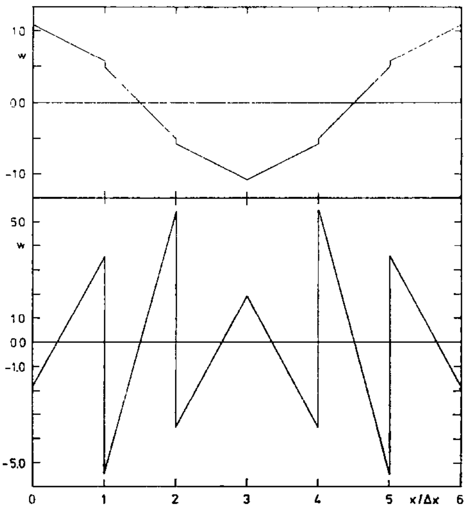

FIG. 4. Piecewise approximation of the function $\cos(\alpha x/\Delta x)$ by the eigenfunctions of scheme III for $\alpha = \pi/3$. Top: using the eigenfunction V_1^{III}. Bottom: using the eigenfunction V_2^{III}. Note the different scaling of w. Further description in Section 3.

When approximating the initial values (25) according to scheme III, we find that

$$w(t^0, x) = \left\{1 + \frac{12i}{\alpha}\left(1 - \frac{\alpha}{2}\cot\frac{\alpha}{2}\right)\frac{x - x_{i+(1/2)}}{\Delta x}\right\}\bar{w}_{i+(1/2)}, \qquad x_i < x < x_{i+1}. \tag{47}$$

The component $\sim V_2^{\mathrm{III}}$ in this function (I call it "stegosaur bias") has a weight $O(\alpha^3)$, in accordance with the second-order accuracy of scheme III. This component damps out quickly, since

$$| g_2^{\mathrm{III}} | = \tfrac{1}{2} + O(\alpha^2). \tag{48}$$

A similar story can be told about scheme II. This scheme is more dissipative than III; it damps out stegosaur bias essentially in one step, since

$$| g_2^{\mathrm{II}} | = O(\alpha^2). \tag{49}$$

Turning to the dispersive errors, let us define the velocity ratio ω:

$$\omega \equiv \frac{\text{numerical convection speed}}{\text{exact convection speed}} = \frac{\arg g - (-\sigma\alpha)}{-\sigma\alpha}. \tag{50}$$

In the limit of $\sigma \to 0$ we find

$$\omega^{\mathrm{I}} = \{(\tfrac{3}{2} - \tfrac{1}{2}\cos\alpha)\sin\alpha\}/\alpha = 1 + \alpha^2/12 + O(\alpha^4), \tag{51}$$

$$\omega_1^{\mathrm{II}} = 2\tan(\alpha/2)/\alpha = 1 + \alpha^2/12 + O(\alpha^4), \tag{52}$$

$$\begin{aligned}\omega_1^{\mathrm{III}} &= [\{(27\tfrac{1}{4} - 5\cos\alpha - 2\cos^2\alpha)^{1/2} - (-\tfrac{3}{2} + 5\cos\alpha + \cos^2\alpha)\}^{1/2} - \sin\alpha]/\alpha \\ &= 1 + \alpha^4/270 + O(\alpha^6).\end{aligned} \tag{53}$$

Polar plots of these values are given in Fig. 5.

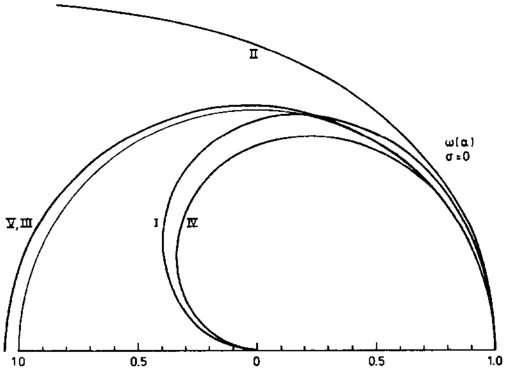

FIG. 5. Dispersion in schemes I–V. Polar plots of the ratios ω^{I}, ω_1^{II}, $\omega_1^{\text{III,V}}$, and ω^{IV} of numerical and exact convection speeds as a function of α, for vanishing Courant number.

It follows from Eq. (52) that the convection speed implied by scheme II becomes *infinite* for certain values of α. This relates to the property of scheme II to yield a finite change in $\bar{\Delta}w$ even for a vanishingly small time step. In contrast, scheme I yields a *zero* convection speed for some values of α. Scheme III is so accurate that it can stand comparison with the third-order schemes of the next section. For instance, a comparison of Eqs. (53) and (65) shows that the convective error for $\sigma \rightarrow 0$ in the finite-difference scheme IV is a factor 9 larger than in scheme III. For $\sigma = \frac{1}{2}$, scheme IV is identical to scheme I and therefore, again, less accurate than scheme III.

When examining g_1^{III} more closely, it appears that ω_1^{III} has an error $O(\alpha^4)$ for any value of σ. That is, scheme III is *third-order accurate* with respect to the eigenfunction V_1^{III}. Nevertheless, scheme III must be called a second-order scheme, since the least-squares approximation of arbitrary initial values always introduces an amount $O(\alpha^3)$ of stegosaur bias, which is convected with zero-order accuracy. Furthermore, in nonlinear convection problems, even if initially absent such bias will be created during each time step.

The above analysis, although by no means exhaustive, clearly shows that scheme III is superior to both other schemes as regards accuracy. It is undoubtedly also the most time-consuming scheme, in particular when applied to nonlinear convection equations. Scheme II is less dissipative than scheme I but equally dispersive, which makes the reduction in dissipation of questionable value. It further behaves peculiarly for $\sigma \rightarrow 0$. Its main advantage over scheme I is that it involves only one mesh in determining $\bar{\Delta}w$, just as scheme III.

However, the comparison between I and II definitely goes in favor of the latter, if these schemes are considered in conjunction with one of the monotonicity algorithms of Section 5. Any of those will bring the infinite wave speed, yielded by scheme II for $\alpha = \pi$, down to a finite value close to the correct one, while such algorithms can not raise the zero wave speed yielded by scheme I. Furthermore, they provide a strong extra dissipation where needed to prevent dispersive ripples, so the lower general level of dissipation in scheme II no longer is a disadvantage.

The convenience of updating $\bar{\Delta}_{i+(1/2)}w$ solely with information from the mesh (x_i, x_{i+1}) appears clear when boundary conditions have to be met. In the case of positive a it suffices to prescribe the value of w as a function of time at the left-hand boundary. No "virtual mesh" across the boundary need be invoked, and the scheme applies without change to the leftmost mesh. At a right-hand boundary no special values are needed.

Another favorable consequence is that a disturbance in some mesh makes itself felt only in downstream meshes. This is called the "transportive property" by Roache and Mueller [7].

Finally, note that all schemes may as well be formulated with respect to a moving grid.

4. Third-Order Schemes

It is useful to investigate how much is gained in going to the third order of accuracy. Third-order schemes result if we fit the initial values in each mesh by a quadratic polynomial. Defining some average second space derivative

$$\frac{\overline{\Delta^2}_{i+(1/2)}w}{(\Delta x)^2} = \overline{\left(\frac{\partial^2 w}{\partial x^2}\right)}_{i+(1/2)}, \tag{54}$$

we may write the approximate initial values in the mesh (x_i, x_{i+1}) as

$$w(t^0, x) = \bar{w}_{i+(1/2)} + \bar{\Delta}_{i+(1/2)}w\,\frac{x - x_{i+(1/2)}}{\Delta x} + \frac{1}{2}\,\overline{\Delta^2}_{i+(1/2)}w\left\{\left(\frac{x - x_{i+(1/2)}}{\Delta x}\right)^2 - \frac{1}{12}\right\}. \tag{55}$$

The corresponding scheme becomes

$$\begin{aligned}\bar{w}^{1/2} = {} & \bar{w}_{1/2} - \sigma\Delta_0\bar{w} - (\sigma/2)(1-\sigma)(\bar{\Delta}_{1/2}w - \bar{\Delta}_{-1/2}w)\\ & - (\sigma/12)(1-\sigma)(1-2\sigma)(\overline{\Delta^2}_{1/2}w - \overline{\Delta^2}_{-1/2}w).\end{aligned} \tag{56}$$

With respect to Eq. (9) we have

$$\langle aw_i\rangle = a\{\bar{w}_{i-(1/2)} + \tfrac{1}{2}(1-\sigma)\,\bar{\Delta}_{i-(1/2)}w + \tfrac{1}{12}(1-\sigma)(1-2\sigma)\,\overline{\Delta^2}_{i-(1/2)}w\}. \tag{57}$$

The third-order successor to scheme I is the finite-difference scheme IV, with $\bar{\Delta}w$ defined as in Eq. (16), and

$$\overline{\Delta^2}_{i+(1/2)}w = \Delta_{i+1}\bar{w} - \Delta_i\bar{w}. \tag{58}$$

It can be written as

$$\begin{aligned}\bar{w}^{1/2} = {} & \{1 - \sigma(1 - T^{-1}) - (\sigma/4)(1-\sigma)\,T(1 + T^{-1})(1 - T^{-1})^2\\ & - (\sigma/12)(1-\sigma)(1-2\sigma)\,T(1 - T^{-1})^3\}\,\bar{w}_{1/2}.\end{aligned} \tag{59}$$

A particularly attractive scheme results if we let the quadratic functions assume the values $W(t^0, x_i)$ at the slab boundaries, so that the overall approximation becomes continuous. The average first derivative again follows from Eq. (19), while the second derivative is determined by

$$\overline{\Delta^2_{i+(1\ 2)}}w = 6\{W(t^0, x_i) - 2\bar{w}_{i+(1/2)} + W(t^0, x_{i+1})\}. \tag{60}$$

Scheme V, as I shall call it, is based on two independent quantities, namely, $\bar{w}_{i+(1/2)}$ and w_i. Updating w_i is done according to

$$w^i = W(t^1, x_i) = \bar{w}_{i-(1/2)} + (\tfrac{1}{2} - \sigma)\, \bar{\Delta}_{i-(1/2)}w + \tfrac{1}{2}(\sigma^2 - \sigma + \tfrac{1}{6})\, \overline{\Delta^2_{i-(1/2)}}w. \tag{61}$$

Using Eqs. (56), (19), (60) and (61) we can write scheme V as

$$\begin{pmatrix} \bar{w}^{i+(1/2)} \\ w^i \end{pmatrix} = \begin{pmatrix} (1-\sigma)(1+\sigma-2\sigma^2) + \sigma^2(3-2\sigma)T^{-1} & -\sigma(1-\sigma)T(1-T^{-1})(1-\sigma-\sigma T^{-1}) \\ 6\sigma(1-\sigma)T^{-1} & (1-\sigma)(1-3\sigma) - \sigma(2-3\sigma)T^{-1} \end{pmatrix}$$

$$\times \begin{pmatrix} \bar{w}_{i+(1/2)} \\ w_i \end{pmatrix}. \tag{62}$$

The matrix defining this scheme will be called G^{V}.

The third-order successor to scheme III is the least-squares error scheme VI, with $\bar{\Delta}w$ as in Eq. (22), and furthermore,

$$\overline{\Delta^2_{i+(1/2)}}w = 2\,\frac{\displaystyle\int_{x_i}^{x_{i+1}} W(t^0, x)\cdot\left\{\left(\frac{x - x_{i+(1/2)}}{\Delta x}\right)^2 - \frac{1}{12}\right\} dx}{\displaystyle\int_{x_i}^{x_{i+1}} \left\{\left(\frac{x - x_{i+(1/2)}}{\Delta x}\right)^2 - \frac{1}{12}\right\}^2 dx}$$

$$= \frac{360}{\Delta x}\int_{x_i}^{x_{i+1}} W(t^0, x)\cdot\left\{\left(\frac{x - x_{i+(1/2)}}{\Delta x}\right)^2 - \frac{1}{12}\right\} dx, \tag{63}$$

a quantity independent of $\bar{w}_{i+(1/2)}$ and $\bar{\Delta}_{i+(1/2)}w$. In spite of its inherent accuracy I shall not discuss this scheme here in detail. For a single convection equation it may be profitable; however, its value for the ideal compressible flow equations at present seems doubtful. Quite probably the scheme is beyond the point of diminishing returns, because of its complexity.

The amplification factor of scheme IV is

$$g^{\mathrm{IV}} = 1 - \frac{\sigma}{3}(1 - \cos\alpha)\{1 + 3\sigma - \sigma^2 - (1 - \sigma^2)\cos\alpha\}$$

$$- i\sigma\left(\frac{4 - \sigma^2}{3} - \frac{1 - \sigma^2}{3}\cos\alpha\right)\sin\alpha. \tag{64}$$

As seen upon comparing Eqs. (59) and (18), scheme IV reduces to scheme I for $\sigma = \frac{1}{2}$. Therefore, the dissipation in that case is given by Eq. (35), with the superscript I replaced by IV. The dispersion for $\sigma \to 0$ in scheme IV is given by

$$\omega^{\mathrm{IV}} = \{(\tfrac{4}{3} - \tfrac{1}{3}\cos\alpha)\sin\alpha\}/\alpha = 1 - \alpha^4/30 + O(\alpha^6), \tag{65}$$

somewhat disappointing when compared to result (53) for scheme III. A plot of ω^{IV} is included in Fig. 5.

The amplification factors of scheme V are the eigenvalues of the matrix G^{V} occurring in Eq. (62). These turn out to be identical to the eigenvalues of G^{III}, given in Eq. (30). Scheme V therefore has exactly the same dissipation for $\sigma = \frac{1}{2}$ and dispersion for $\sigma = 0$ as scheme III. That is, they are given by Eqs. (37) and (53), with the superscript III replaced by V. Nevertheless, scheme V, being a genuine third-order scheme, is more accurate than scheme III. That is, when arbitrary initial values are approximated according to scheme V, the awkward eigenfunction V_2^{V} corresponding to g_2^{V} gets a weight of only $O(\alpha^4)$.

Scheme V, therefore, is the most accurate of schemes I–V, and by virtue of its computational simplicity, also the most economical one. This becomes even more true when the scheme is applied to ideal compressible flow problems: it takes less execution time than any of the other schemes (see [8]). It is tempting to just skip the second-order schemes and concentrate on scheme V. I have resisted this temptation because the second-order schemes are very instructive and most of my experience still lies in these schemes. Another consideration is that scheme V loses part of its simplicity when used in conjunction with one of the monotonicity algorithms of Section 5: in meshes where the monotonicity condition is enforced, the overall continuity of the initial-value approximation must temporarily be broken.

The question remains whether it is fair to compare, say, scheme V to scheme IV on the basis of the same mesh width. If computer storage space is the decisive factor, scheme V should be judged on the basis of the double mesh width, since it also uses double information per mesh. This raises its dispersive error coefficient for long waves from 1/270 to 16/270 $\approx$ 1/17, which is nearly twice the value 1/30 derived for scheme IV. However, from Fig. 5 it can be seen that for waves of length $\lesssim 6$ (single) meshes, scheme IV again has the larger dispersion.

Furthermore, to achieve the error quoted for scheme IV, four times as many meshes in time-space are used as to achieve the error quoted for scheme V. Since the schemes have about the same execution time per mesh per time step, scheme V remains the most efficient one. This applies even more strongly to the many-dimensional case.

Schemes I–VI are examples of using a polynomial of degree n to achieve a truncation error of the order $n + 1$. In satisfying ourselves with a truncation error of the order n, a degree of freedom becomes available for meeting some extra mathematical or physical requirement (e.g., an additional error bound or conservation law).

Of course, other functions than polynomials may be used as the basic approximating functions. Monotonic functions such as $(\exp cx)/c$ appear to be valuable in approximating initial values without creating oscillations (a simpler way to reach the same goal is indicated in Section 5). When dealing with a nonlinear convection equation we may select basic functions for which the integration of the equation according to Step 3 is particularly simple (see Section 6).

I do not intend to consider the use of splines, since these destroy the local character of the initial-value approximation and the explicit character of the scheme.

5. Monotonicity

The monotonicity condition says that, when a monotonic initial-value distribution is numerically convected, the resulting distribution must be monotonic again. In consequence, if $\bar{w}_{i+(1/2)}$ lies between $\bar{w}_{i-(1/2)}$ and $\bar{w}_{i+(3/2)}$, then $\bar{w}^{i+(1/2)}$ must lie between $\bar{w}^{i-(1/2)}$ and $\bar{w}^{i+(3/2)}$.

In practice it is easier to deal with the following *sufficient* requirement: if $\bar{w}_{i+(1/2)}$ lies between $\bar{w}_{i-(1/2)}$ and $\bar{w}_{i+(3/2)}$, then $\bar{w}^{i+(1/2)}$ must lie between $\bar{w}_{i-(1/2)}$ and $\bar{w}_{i+(1/2)}$ for $0 < \sigma < 1$ and between $\bar{w}_{i+(3/2)}$ and $\bar{w}_{i+(1/2)}$ for $-1 < \sigma < 0$. This has been the starting point of two preceding papers [3, 9]. In [3], Fromm's scheme was made monotonic through the inclusion of a third difference with a coefficient depending on the rate of change of the first difference, that is, on $\Delta_{i+1}\bar{w}/\Delta_i\bar{w}$. In the present approach we may think of this term as a means to reduce the value of $\bar{\Delta}_{i+(1/2)}w$ below the value given in Eq. (16).

With the help of Fig. 2 it is easily found that, regardless of how $\bar{\Delta}_{i+(1/2)}w$ is defined, its value must be limited as follows:

$$(\bar{\Delta}_{i+(1/2)}w)_{\text{mono}} = \begin{cases} \min\{2\,|\,\Delta_i\bar{w}\,|, |\,\bar{\Delta}_{i+(1/2)}w\,|, 2\,|\,\Delta_{i+1}\bar{w}\,|\}\operatorname{sgn}\bar{\Delta}_{i+(1/2)}w \\ \qquad \text{if} \quad \operatorname{sgn}\Delta_i\bar{w} = \operatorname{sgn}\Delta_{i+1}\bar{w} = \operatorname{sgn}\bar{\Delta}_{i+(1/2)}w, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{66}$$

This prescription is valid for positive as well as negative values of σ; any dependence on σ has been removed.

Equation (66) says in the first place that, in order to preserve the monotonicity of a sequence of mesh averages, *the linear function* (12) *must not take values outside the range spanned by the neighboring mesh averages.*

If $\bar{w}$ reaches an extremum in the mesh considered, that is, if $\operatorname{sgn}\Delta_i\bar{w} \neq \operatorname{sgn}\Delta_{i+1}\bar{w}$, then $\bar{\Delta}_{i+(1/2)}w$ is reduced to zero in order not to accentuate the extremum. This further guarantees that w, if initially positive, remains positive.

The use of (66) also yields extra damping of stegosaur bias, characterized by $\operatorname{sgn}\Delta_i\bar{w} = \operatorname{sgn}\Delta_{i+1}\bar{w} \neq \operatorname{sgn}\bar{\Delta}_{i+(1/2)}w$.

The limiting effect of Eq. (66) in the three distinct cases is illustrated in Fig. 6.

Equation (66) corresponds to the minimum third-difference coefficient given in [3, Eq. (27)]. The larger coefficient given in [3, Eq. (28)] yields a stronger reduction of $\bar{\Delta}w$. It applies exclusively to the Fromm scheme I, in which $\bar{\Delta}w$ is expressed solely in terms of $\Delta\bar{w}$. In the present approach we find the surprising formulation

$$(\bar{\Delta}_{i+(1/2)}w)_{\text{mono}} = \begin{cases} \dfrac{2\Delta_i\bar{w}\,\Delta_{i+1}\bar{w}}{\Delta_i\bar{w} + \Delta_{i+1}\bar{w}} & \text{if} \quad \operatorname{sgn}\Delta_i\bar{w} = \operatorname{sgn}\Delta_{i+1}\bar{w}, \\ 0 \quad \text{otherwise.} \end{cases} \tag{67}$$

In words, monotonicity may be preserved by taking for $\bar{\Delta}_{i+(1/2)}w$ the *harmonic* mean of $\Delta_i\bar{w}$ and $\Delta_{i+1}\bar{w}$ rather than the algebraic mean, as in Eq. (16). This result ought to clear up most of the mystery about monotonicity.

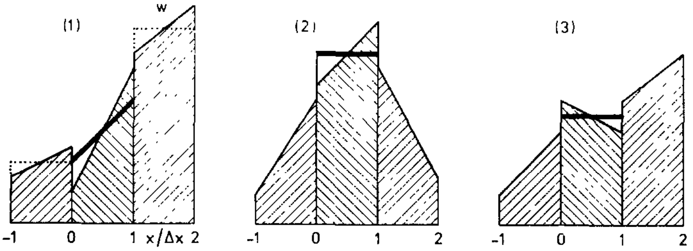

FIG. 6. The monotonicity condition (66) for the second-order scheme. (1) The slope of the linear distribution (solid line) in the mesh (x_0, x_1) is reduced (heavy solid line) so that the values in this mesh do not go beyond the average levels (dotted line) in the adjacent meshes. (2) If the mesh average reaches an extremum, the slope is reduced to zero. (3) If the slope does not agree with the trend in the mesh averages, it is also reduced to zero.

It is possible to make the limiting action of (66) about as strong as that of (67), through replacement of the factors 2 occurring in (66) by $\frac{3}{2}$.

The monotonicity algorithm (66), while derived for uniform convection through a uniform grid, remains valid for a nonuniform grid and for a broad class of convection equations with variable a. A numerical example of its use in connection with the nondissipative Burgers equation

$$\frac{\partial w}{\partial t} + \frac{\partial(\frac{1}{2}w^2)}{\partial x} = 0, \tag{68}$$

where

$$a \equiv \frac{d(\frac{1}{2}w^2)}{dw} = w, \tag{69}$$

is given in the next section.

For variable a the CFL condition (5) used in connection with Eq. (66) must be reformulated. It is most accurately expressed as: no convection path may cross more than one mesh boundary during a time step.

I have not been able to determine in general for what functions $a(t, x, w)$ the limiting effect of Eq. (66) suffices to ensure monotonicity, and for what functions it does not. The cases I have found where monotoncity is broken in spite of the use of Eq. (66) all are very contrived. In these counterexamples the deviation from monotonicity occurs in a mesh where the Courant number approaches unity; the new extremum merely sticks out a fraction $O(\Delta^2 a)$ of the differences $\Delta\bar{w}$ involved. It may be avoided through a reduction of Δt by a comparable fraction. This extra restriction on the size of the time step must not be considered too serious; in practice, the safety margin of Δt with respect to the stability limit is more likely taken to be a fraction $O(\Delta a)$.

For the third-order scheme (56), a sufficient monotonicity condition may be formulated in just the same manner as for the second-order scheme (14): the quadratic function (56) must not take values outside the range spanned by the neighboring mesh averages.

It must be mentioned here that a highly successful method to preserve monotonicity has been developed by Boris and Book [10]. The basic idea underlying their Flux-Corrected Transport (FCT) technique is the following. First, w is updated provisionally with a monotonic first-order scheme. Next, terms are added that bring second-order accuracy, but these are subjected to a limiting routine in order to preserve the monotonicity of the provisional results.

We may try out this technique, too, on our second-order scheme (14), using the embedded CIR scheme (7) as the monotonic first-order scheme. For $\sigma \geqslant 0$ we first have

$$\bar{w}^{1/2*} = \bar{w}_{1/2} - \sigma\Delta_0\bar{w} \tag{70}$$

and subsequently

$$\bar{w}^{1/2} = \bar{w}^{1/2*} - \frac{\sigma}{2}(1-\sigma)(\bar{\Delta}_{1/2}w - \bar{\Delta}_{-1/2}w)_{\text{mono}}, \tag{71}$$

where

$$(\bar{\Delta}_{i+(1/2)})_{\text{mono}} = \begin{cases} \min\left\{\dfrac{2}{\sigma(1-\sigma)}\,|\,\Delta^{i}\bar{w}^{*}\,|,\ |\,\bar{\Delta}_{i+(1/2)}w\,|,\ \dfrac{2}{\sigma(1-\sigma)}\,|\,\Delta^{i+2}\bar{w}^{*}\,|\right\}\operatorname{sgn}\bar{\Delta}_{i+(1/2)}w \\ \qquad \text{if } \operatorname{sgn}\Delta^{i}\bar{w}^{*} = \operatorname{sgn}\Delta^{i+2}\bar{w}^{*} = \operatorname{sgn}\bar{\Delta}_{i+(1/2)}w, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{72}$$

The explicit dependence on σ may be removed by adjusting the limiting effect to the minimum value of the factors $2/\{\sigma(1-\sigma)\}$:

$$(\bar{\Delta}_{i+(1/2)}w)_{\text{mono}} = \begin{cases} \min\{8\,|\,\Delta^{i}\bar{w}^{*}\,|,\ |\,\bar{\Delta}_{i+(1/2)}w\,|,\ 8\,|\,\Delta^{i+2}\bar{w}^{*}\,|\}\operatorname{sgn}\bar{\Delta}_{i+(1/2)}w \\ \qquad \text{if } \operatorname{sgn}\Delta^{i}\bar{w}^{*} = \operatorname{sgn}\Delta^{i+2}\bar{w}^{*} = \operatorname{sgn}\bar{\Delta}_{i+(1/2)}w, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{73}$$

Note that these formulas are downstream centered; for $\sigma < 0$ the differences $\Delta^{i-1}\bar{w}^{*}$ and $\Delta^{i+1}\bar{w}^{*}$ must enter.

In order to decide which of the limiters presented so far is best suited for use with scheme (14), let us compare the different formulas. A disadvantage of (72) and (73) seems that these algorithms involve data from four meshes, while (66) involves only three. On the other hand, Eqs. (72) and (73) appear to allow of much larger gradients inside the meshes than (66). The reason is that (72) and (73) take into account what value of σ will be used in the next time step; the dependence of (73) on σ is hidden in the $\bar{w}^{*}$ values.

A fairer comparison results if such dependence on σ is also included in (66). The condition then reads, for $\sigma \geqslant 0$,

$$(\bar{\Delta}_{i+(1/2)}w)_{\text{mono}} = \begin{cases} \min\left\{\dfrac{2}{\sigma}\,|\,\Delta_{i}\bar{w}\,|,\ |\,\bar{\Delta}_{i+(1/2)}w\,|,\ \dfrac{2}{1-\sigma}\,|\,\Delta_{i+1}\bar{w}\,|\right\}\operatorname{sgn}\bar{\Delta}_{i+(1/2)}w \\ \qquad \text{if } \operatorname{sgn}\Delta_{i}\bar{w} = \operatorname{sgn}\Delta_{i+1}\bar{w} = \operatorname{sgn}\bar{\Delta}_{i+(1/2)}w, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{74}$$

Note that the weights of $|\Delta_i \bar{w}|$ and $|\Delta_{i+1} \bar{w}|$ differ; for $\sigma < 0$ they become $2/(1 - |\sigma|)$ and $2/|\sigma|$, respectively. The limiting effect of (74) is now comparable to that of (72).

On the basis of their performance in the numerical examples of Section 6, I tend to prefer the upstream-centered limiters (66) and (74), in particular because they seem to introduce no phase errors of their own. However, it must yet be investigated whether they will remain the most practical when used for compressible flow problems.

6. Numerical Examples

The performance of schemes I, III, and V was investigated in a series of simple numerical experiments.

First, triangle and square waves of wavelength $l = 12\Delta x$ were convected according to the linear equation (1), using schemes I and III. Periodic boundary conditions were

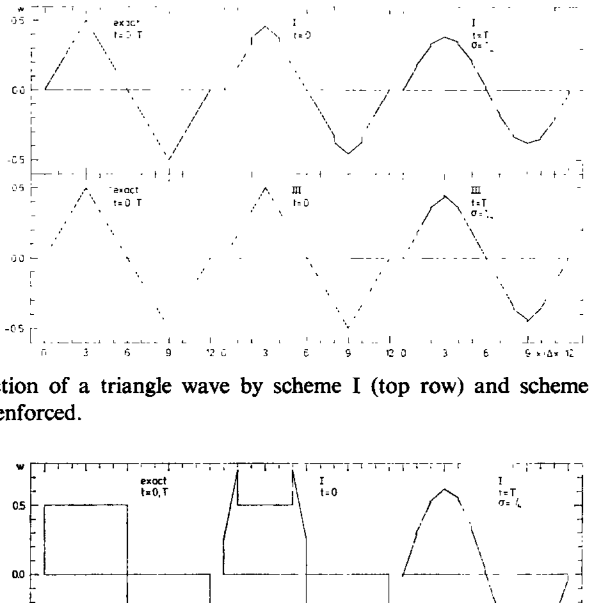

Fig. 7. Convection of a triangle wave by scheme I (top row) and scheme III (bottom row). No monotonicity enforced.

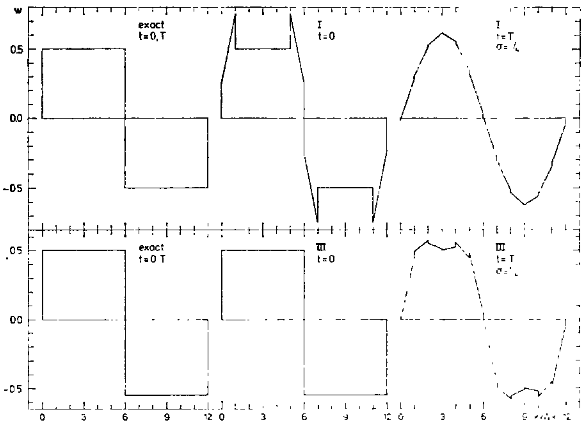

Fig. 8. Convection of a square wave by scheme I (top) and scheme III (bottom). No monotonicity enforced.

applied. The final results plotted in Figs. 7–13 are for $t = T = l/a$, when the waves should have traveled their own length once. The full distributions have been drawn; any claims to monotonicity refer to sequences of mesh *averages*. In all cases a Courant number $\frac{1}{4}$ was used, so that both dispersive and dissipative errors would show up.

Figures 7 and 8 show the superiority of scheme III over scheme I, especially regarding the preservation of the higher harmonics in the waveforms. For scheme I, the positive phase error in the fundamental wave is detectable. Note the difference in the initial values for these schemes.

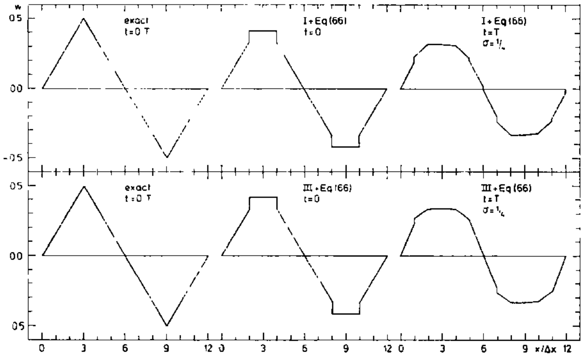

FIG. 9. Convection of a triangle wave by scheme I (top) and scheme III (bottom), using the monotonicity algorithm (66).

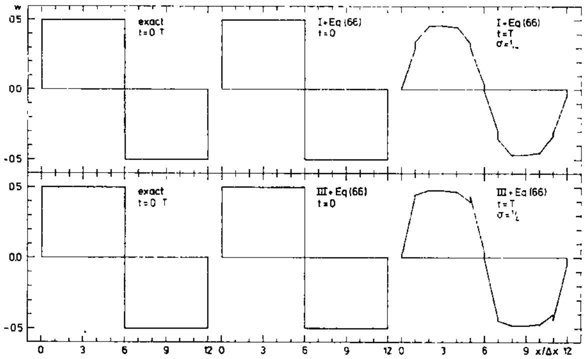

FIG. 10. Convection of a square wave by scheme I (top) and scheme III (bottom), using the monotonicity algorithm (66).

Figures 9 and 10 show the effect of the monotonicity algorithm (66) on the results of schemes I and III in the same cases as before. Note that the initial values are now the same for both schemes. The difference between the performances of I and III is largely masked by the effect of Eq. (66). The extrema of the triangle waves are strongly eroded, because those are in fact treated with the first-order scheme (7). As desired, the square waves no longer show the overshoots present in Fig. 8. For scheme I, the use of Eq. (66) appears to slow down at least the crests of the waves.

In Figs. 11–13 various monotonicity algorithms are evaluated on the basis of scheme III. In Fig. 11 the results of Eq. (66) for the triangle wave are repeated (top row) and compared with the results of Eq. (73) (FCT, bottom row). These results are very much alike, although the FCT algorithm introduces a detectable extra phase lag.

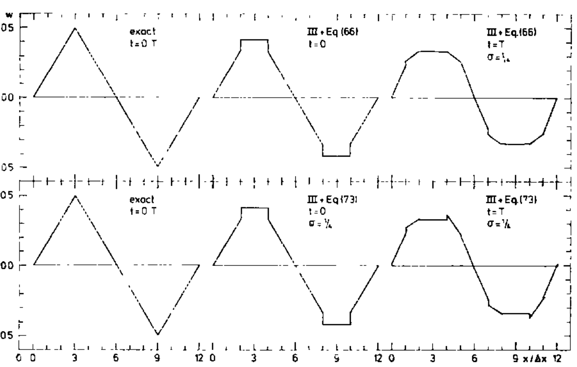

FIG. 11. Convection of a triangle wave by scheme III, using the monotonicity algorithm (66) (top) and the FCT algorithm (73) (bottom).

The internal slopes drawn for $t = T$ could be fixed according to Eq. (73) only after an extra time step was carried out with the first-order scheme (70). The implicit dependence of condition (73) on σ is easily detected from the distribution at $t = T$. In mesh (x_4, x_5) the overshoot with respect to the neighboring average level $\bar{w}_{3\frac{1}{2}}$ will not cause $\bar{w}^{4\frac{1}{2}}$ to rise above $\bar{w}^{3\frac{1}{2}}$ for $\sigma = \frac{1}{4}$ (the value used), but would do so if the next step were taken with $\sigma = \frac{3}{4}$.

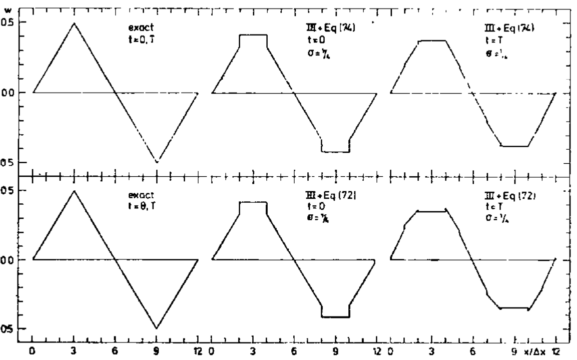

FIG. 12. Convection of a triangle wave by scheme III, using the monotonicity algorithm (74) (top) and the FCT algorithm (72) (bottom).

In Fig. 12 the results of Eq. (74) for the triangle wave are compared with those of Eq. (72) (FCT). Both formulas depend explicitly on σ. The results are clearly better than those in Fig. 11, with the FCT algorithm again yielding the largest phase error and, in this case, also the largest amplitude error.

Particularly obvious in Fig. 12 is the clipping of an extremum down to a plateau, an effect shared by all monotonicity algorithms tested. It would be extremely valuable if a more sophisticated algorithm were developed that can distinguish between a point extremum and a plateau. In the present schemes such an algorithm would involve, apart from $\Delta_i\bar{w}$ and $\Delta_{i+1}\bar{w}$, also $\bar{\Delta}_{i-(1/2)}$ and $\bar{\Delta}_{i+(3/2)}w$ in determining the largest permitted value of $\bar{\Delta}_{i+(1/2)}w$.

Figure 13 shows the monotonicity techniques at their best, that is, when applied to a square wave. In this case Eq. (73) (FCT) is the better of Eq. (66), while the results of Eq. (72) (FCT) and Eq. (74) are perfectly identical. Note that (72)–(74) do not ensure the positivity of the entire distribution inside a mesh (cf. mesh (x_{10}, x_{11})).

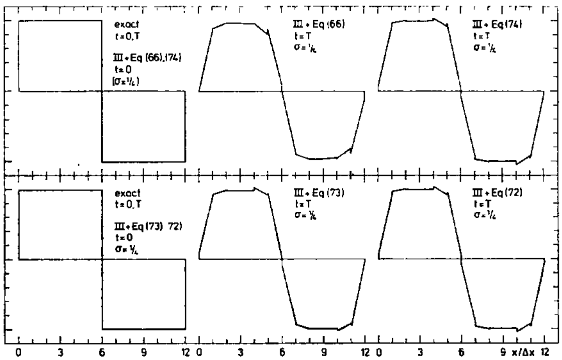

FIG. 13. Convection of a square wave by scheme III, using the monotonicity algorithms (66) and (74) (top) and the FCT algorithms (73) and (72) (bottom).

Next, some experiments were done with schemes III and V on the basis of the non-dissipative Burgers equation (68). The nonlinear versions of these schemes are given below.

Scheme III. The initial-value distribution in any mesh (x_i, x_{i+1}) is linear, with slope $\bar{\Delta}_{i+(1/2)}w/\Delta x$. Using a linear basic function is particularly convenient, since it remains linear when convected according to Eq. (68). Only, its slope becomes multiplied, per time step, by $1/(1 + \lambda\bar{\Delta}_{i+(1/2)}w)$, where λ is the mesh ratio $\Delta t/\Delta x$. Note that for

$$\lambda\bar{\Delta}w \leqslant -1 \tag{75}$$

the distribution will steepen into a shock within one time step. However, if w nowhere changes its sign, the case (75) will always be excluded by the Courant condition

$$\lambda| w | \leqslant 1. \tag{76}$$

Henceforth w is assumed to be positive.

In updating $\bar{w}$ we use the integral (conservation) form of Eq. (68):

$$\bar{w}^{i+(1/2)} = \bar{w}_{i+(1/2)} - \lambda(\langle\tfrac{1}{2}w_{i+1}^2\rangle - \langle\tfrac{1}{2}w_i{}^2\rangle), \tag{77}$$

comparable to Eq. (9) from the case of linear convection.

At the mesh boundary x_i, the initial values generally are discontinuous, with

$$w_{i-} = \bar{w}_{i-(1/2)} + \tfrac{1}{2}\bar{\Delta}_{i-(1/2)}w, \tag{78}$$

$$w_{i+} = \bar{w}_{i+(1/2)} - \tfrac{1}{2}\bar{\Delta}_{i+(1/2)}w. \tag{79}$$

Using Eq. (69) and its consequence that all convection paths are straight, it is found that

$$w^i = w_{i-}/(1 + \lambda\bar{\Delta}_{i-(1/2)}w) \tag{80}$$

and, subsequently,

$$\langle\tfrac{1}{2}{w_i}^2\rangle = \tfrac{1}{2}w_{i-}^2/(1 + \lambda\bar{\Delta}_{i-(1/2)}w), \tag{81}$$

so that step (77) can be carried out.

Updating $\bar{\Delta}w$ means that we have to determine the first moment of the final distribution $W(t^1, x)$ in the mesh (x_i, x_{i+1}). Two cases must be distinguished.

If $w_{i+} \geqslant w_{i-}$, the initial discontinuity at x_i resolves into a rarefaction fan. The final distribution is continuous and has the form

$$W(t^1, x) = \begin{cases} w^i + \dfrac{x - x_i}{\Delta x}\dfrac{\bar{\Delta}_{i-(1/2)}w}{1 + \lambda\bar{\Delta}_{i-(1/2)}w}, & x_i \leqslant x < x_i + w_{i-}\Delta t, \\ w_{i-} + \{x - (x_i + w_{i-}\Delta t)\}/\Delta t, & x_i + w_{i-}\Delta t \leqslant x \leqslant x_i + w_{i+}\Delta t, \\ w^{i+1} - \dfrac{x_{i+1} - x}{\Delta x}\dfrac{\bar{\Delta}_{i+(1/2)}w}{1 + \lambda\bar{\Delta}_{i+(1/2)}w}, & x_i + w_{i+}\Delta t < x \leqslant x_{i+1}. \end{cases} \tag{82}$$

Now $\bar{\Delta}^{i+(1/2)}w$ can be determined in the manner of Eq. (22).

If $w_{i+} < w_{i-}$, the discontinuity is a shock and remains such while moving into the mesh. Calling the final shock position x_S, we find for the discontinuous final distribution:

$$W(t^1, x) = \begin{cases} w^i + \dfrac{x - x_i}{\Delta x}\dfrac{\bar{\Delta}_{i-(1/2)}w}{1 + \lambda\bar{\Delta}_{i-(1/2)}w}, & x_i \leqslant x < x_S, \\ w^{i+1} - \dfrac{x_{i+1} - x}{\Delta x}\dfrac{\bar{\Delta}_{i+(1/2)}w}{1 + \lambda\bar{\Delta}_{i+(1/2)}w}, & x_S < x \leqslant x_{i+1}. \end{cases} \tag{83}$$

The mesh integral of this distribution is already known ($= \bar{w}^{i+(1/2)}\,\Delta x$), so that x_S can be found from the equation

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} W(t^1, x)\,dx - \bar{w}^{i+(1/2)} = 0, \tag{84}$$

which is quadratic in x_S. Once x_S is known, the first moment of $W(t^1, x)$ can be determined and $\bar{\Delta}w$ is updated.

It may appear a bit strange that the complete distribution $W(t^1, x)$ is used in determining $\bar{\Delta}^{i+(1/2)}w$, while only its first moment is needed. A direct expression for the first moment of $W(t^1, x)$ can be found by integrating the following equivalent of Eq. (68),

$$\frac{\partial\{(x - \bar{x})w\}}{\partial t} + \frac{\partial\{\frac{1}{2}(x - \bar{x})w^2\}}{\partial x} = \tfrac{1}{2}w^2, \tag{85}$$

over one space-time mesh. The source term at the right-hand side becomes

$$\int_{t^0}^{t^1} \int_{x_i}^{x_{i+1}} \tfrac{1}{2} w^2 \, dx \, dt \equiv \int_{t^0}^{t^1} \tfrac{1}{2}\overline{w^2} \, dt, \tag{86}$$

and it is tempting to evaluate this integral by using the integral form

$$\tfrac{1}{2}\overline{w^2}^{\,i+(1/2)} = \tfrac{1}{2}\overline{w}^2_{i+(1/2)} - \lambda(\langle \tfrac{1}{3} w^3_{i+1} \rangle - \langle \tfrac{1}{3} w_i{}^3 \rangle) \tag{87}$$

of another equivalent of Eq. (68),

$$\frac{\partial(\frac{1}{2}w^2)}{\partial t} + \frac{\partial(\frac{1}{3}w^3)}{\partial x} = 0. \tag{88}$$

However, the *integral* form (87) of Eq. (88) is not equivalent to the *integral* form (77) of Eq. (68) if a shock occurs in the mesh. The above procedure may therefore be used only in the case that a rarefaction wave moves into the mesh.

I must mention that the evaluation of $\bar{\Delta}^{i+(1/2)}w$ in scheme III need not be done so scrupulously as described here., Various approximations may be introduced in which no distinction is made between the rarefaction case and the shock case, and which still boil down to the least-squares formula (23) in the limit of linear convection.

Scheme V. The initial-value distribution is quadratic in each mesh and continuous at the mesh boundaries. If no shock is formed that passes x_i during the time step, we may write

$$\langle \tfrac{1}{2} w_i{}^2 \rangle = w^i [\tfrac{2}{3} w_i - \tfrac{1}{6}\{1 + \lambda(\bar{\Delta}_{i-(1/2)} w + \tfrac{1}{2}\overline{\Delta^2}_{i-(1/2)} w)\} \, w^i], \tag{89}$$

where w^i follows from the quadratic equation

$$(\tfrac{1}{2}\lambda^2 \overline{\Delta^2}_{i-(1/2)} w)(w^i)^2 - \{1 + \lambda(\bar{\Delta}_{i-(1/2)} w + \tfrac{1}{2}\overline{\Delta^2}_{i-(1/2)} w)\} \, w^i + w_i = 0. \tag{90}$$

Equations (89) and (90) supply all the information to update $\bar{w}$, $\bar{\Delta}w$, and $\overline{\Delta^2}w$.

The triangle wave of the previous experiments, superimposed onto an average level $w = 1$, was taken as the initial-value distribution for schemes III and V, again using periodic boundary conditions. The mesh ratio was $\frac{1}{4}$, so that the Courant number λw on the average was also $\frac{1}{4}$, as before. In the exact solution the negative slope in the wave steepens at $t = T/2$ into a shock, which in the course of time reduces in strength. In Fig. (14) the exact solution at $t = T$ is shown, together with the results of III and V. Monotonicity was not enforced.

The shock produced by scheme III slightly lags behind its true position, while the shock produced by scheme V is slightly ahead of it. For $\lambda = \frac{1}{2}$ (not shown here), when the average Courant number in the wave becomes $\frac{1}{2}$ too, these phase errors vanish. Both schemes then yield the shock at the right position, while the waveform is exactly antisymmetric with respect to the shock point.

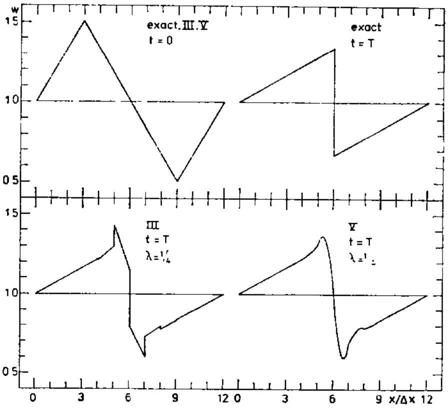

FIG. 14. Nonlinear convection of a triangle wave according to the nondissipative Burgers equation (68). Top: exact solution. Bottom: results of scheme III and scheme V. Mesh ratio $\frac{1}{4}$. No monotonicity enforced.

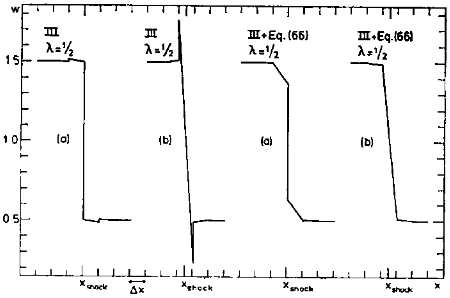

FIG. 15. Shock profiles obtained with scheme III on the basis of the nondissipative Burgers equation (68), with and without use of the monotonicity algorithm (66). Mesh ratio $\frac{1}{2}$. (a) Shock at a mesh boundary. (b) Shock in the middle of a mesh.

Some results for $\lambda = \frac{1}{2}$ are drawn in Fig. 15. This last figure shows four shock profiles obtained (as parts of long square waves) with scheme III. Note that these indeed are antisymmetric. A distinction has been made between case (a), when the shock is located at a mesh boundary, and case (b), when the shock occurs in the middle of a mesh. The adequate effect of the monotonicity algorithm (66) on the profiles is also demonstrated.

7. CONCLUSIONS

The approach to numerical convection discussed and tested in the previous sections has a number of advantages, which are summarized below.

(i) Upstream schemes emerge automatically; in fact, they seem to be the natural way to do numerical convection.

(ii) For a given order of consistency, the schemes can be made considerably more accurate than the ordinary upstream finite-difference schemes. This is achieved by using the information inside the mesh to its full potential.

(iii) In such schemes, the numerical domain of dependence does not spread with the order of consistency (as it does in finite-difference schemes). The CFL condition for this domain remains the stability condition.

(iv) Using such schemes, a disturbance in some mesh does not show up in any upstream mesh (as it generally does for higher-order finite-difference schemes).

(v) Such schemes do not change near a boundary; the exact boundary conditions may be specified.

(vi) For all schemes generated, the availability of continuous functions inside each mesh and the freedom in choosing these make it easy to introduce extra physics and to satisfy special conditions.

(vii) All schemes are fully explicit (as opposed to finite-element methods).

(viii) The schemes are well suited for use with a moving grid.

In the next installment [11] of the present series I shall show how this approach can be followed to construct schemes for the equations of ideal compressible flow. An outline of the procedure is given in Van Leer [8].

Application of the schemes to actual incompressible flow problems becomes particularly interesting if diffusion can be incorporated in the present approach. This clearly is possible, but so far has not been investigated in detail. For the time being, the easiest way to account for diffusion is to combine any diffusion terms into a separate fractional step, using a conventional finite-difference scheme.

References

1. B. van Leer, Towards the ultimate conservative difference scheme. III. Upstream-centered finite-difference schemes for ideal compressible flow, *J. Computational Phys.*, to **23** (1977), 263.
2. S. K. Godunov, *Mat. Sb.* **4** (1959), 271; also *Cornell Aeronautical Lab. Transl.*
3. B. van Leer, *J. Computational Phys.* **14** (1974), 361.
4. R. Courant, E. Isaacson, and M. Rees, *Comm. Pure Appl. Math.* **5** (1952), 243.
5. P. D. Lax and B. Wendroff, *Comm. Pure Appl. Math.* **13** (1960), 217.
6. J. E. Fromm, *J. Computational Phys.* **3** (1968), 176.
7. P. J. Roache and T. J. Mueller, *AIAA J.* **8** (1970), 530.
8. B. van Leer, MUSCL, a new approach to numerical gas dynamics, *in* "Computing in Plasmaphysics and Astrophysics. Proceedings of the Second European Conference on Computational Physics." Max-Planck-Institut für Plasmaphysik, Garching 1976.
9. B. van Leer, *in* "Lecture Notes in Physics," Vol. 18, p. 163, Springer, Berlin, 1973.
10. J. P. Boris and D. L. Book, *J. Computational Phys.* **11** (1973), 38.
11. B. van Leer, Towards the ultimate conservative difference scheme. V. A second order sequel to Godunov's method, in preparation.