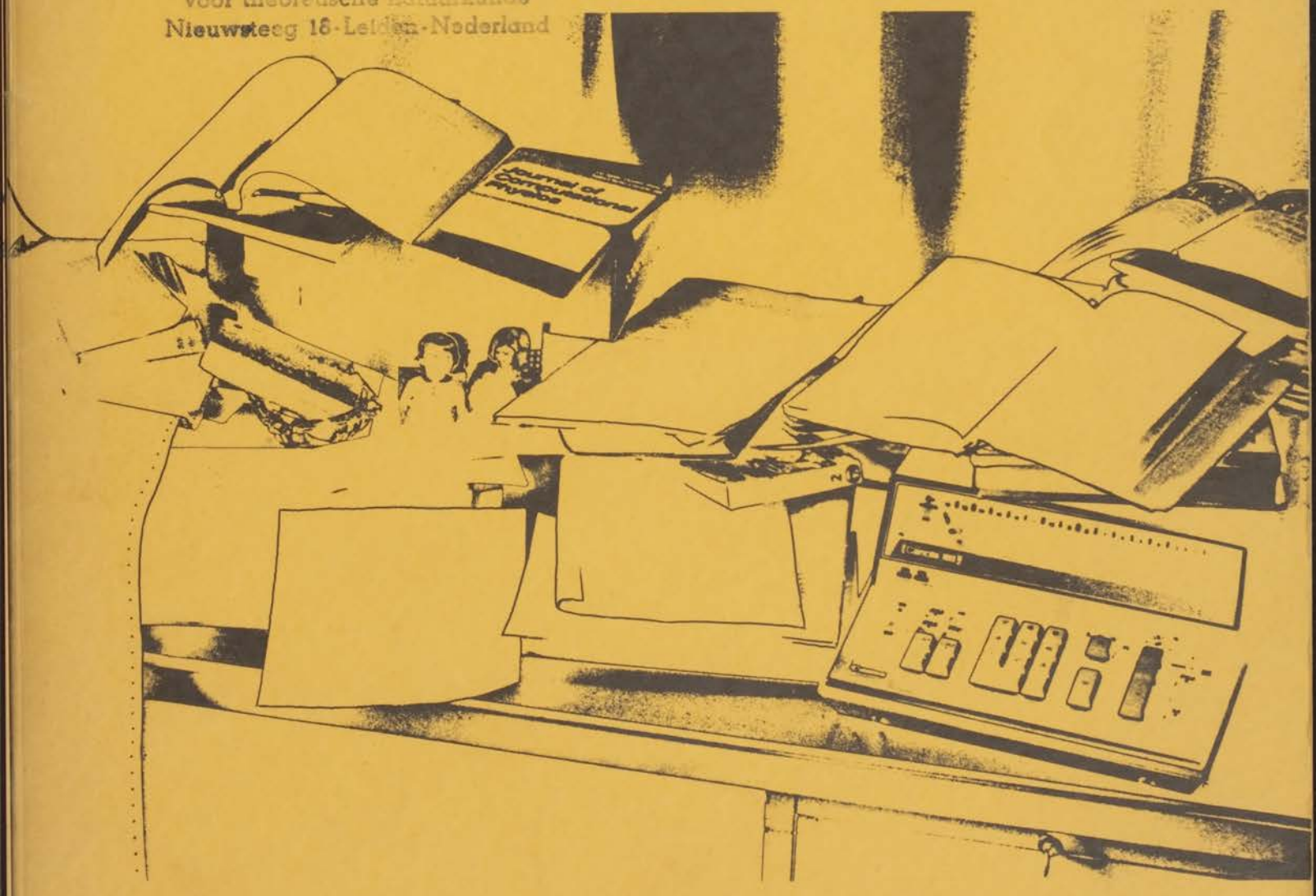
a choice of difference schemes for ideal compressible flow



bram van leer

A CHOICE OF DIFFERENCE SCHEMES FOR IDEAL COMPRESSIBLE FLOW

voor Daniël en Eva

A CHOICE OF DIFFERENCE SCHEMES FOR IDEAL COMPRESSIBLE FLOW

PROEFSCHRIFT

ter verkrijging van de graad van Doctor in de Wiskunde en Natuurwetenschappen aan de Rijksuniversiteit te Leiden, op gezag van de Rector Magnificus Dr. C. Soeteman, Hoogleraar in de Faculteit der Letteren, ten overstaan van een commissie uit de Senaat te verdedigen op woensdag 23 september 1970 te klokke 15.15 uur

door

BRAM VAN LEER

geboren te Soerabaja, Nederlands Oost Indië
(thans Indonesië), in 1942

1970

Sterrewacht Leiden

Promotor: Prof. Dr. H. C. van de Hulst

CURRICULUM VITAE ACADEMIALE

Op verzoek van de Faculteit der Wiskunde en Natuurwetenschappen volgt hier een kort overzicht van mijn akademische studie.

In september 1959 begon ik mijn studie aan de Rijksuniversiteit te Leiden, waar ik in februari 1963 het kandidaatsexamen b' aflegde (sterrekunde met wiskunde en bijvak natuurkunde). In mei 1964 werd ik aangesteld als assistent bij de Sterrewacht.

Na het kandidaatsexamen volgde ik de sterrekunde-colleges van Prof. Dr. J.H. Oort en Prof. Dr. H.C. van de Hulst, de wiskunde-colleges van Prof. Dr. C. Visser en de natuurkunde-colleges van Prof. Dr. J.A.M. Cox, Prof. Dr. P. Mazur en Prof. Dr. P.W. Kasteleyn.

In april 1965 werd mij door het Ministerie van Onderwijs en Wetenschappen een Bijzondere Toelage verleend ten behoeve van een studiereis naar de Verenigde Staten. De maanden juli en augustus 1965 bracht ik door aan het Laboratory for Astrophysics and Space Research van de University of Chicago, waar ik onder leiding van Prof. Dr. E.N. Parker de dynamika van het interstellaire gas bestudeerde.

In juni 1966 legde ik het doctoraalexamen sterrekunde met bijvak mechanika af.

Sinds juli 1966 ben ik als wetenschappelijk medewerker verbonden aan de Sterrewacht te Leiden.

CONTENTS

1. INTRODUCTION 1

2. THE PROBLEM 4

2.1 Mathematical preliminaries 4
2.2 Conservative schemes 7
2.3 Examples 12

3. THE DIFFERENCE SCHEMES 16

3.1 Selection criteria 16
3.2 Basic form 23
3.3 Stability 26
3.4 Principles of classification 31
3.5 Monotonicity 42

4. EXECUTION AND REFINEMENT 51

4.1 Two-step formulation 51
4.2 Non-basic additions 52
4.3 Uneven meshes 55

5. FLUID-DYNAMICAL INTERPRETATION 58

5.1 The equations of ideal compressible flow 58
5.2 Viscosity versus diffusion 65

6. NUMERICAL TESTS 69

6.1 Shock profiles 69
6.2 A test case from astrophysics 74

7. DESIDERATA 79

REFERENCES 82

SUMMARY 84

SAMENVATTING 86

STELLINGEN

1. Er zijn in de numerieke gasdynamika onnodig veel eerste-orde methoden in gebruik.

2. Emery {17} karakteriseert Rusanovs methode als volgt: "This improvement of Lax's method maintains the minimum artificial viscosity at each nodal point (...)". Dit is in twee opzichten misleidend.

3. De vrijwel onontkoombare kloof tussen uitvinders en gebruikers van numerieke methoden is bijzonder duidelijk waarneembaar in de kosmische gasdynamika, waar het voorkomt dat
— een methode uit een verouderd handboek wordt opgediept;
— een bepaalde methode op het verkeerde probleem wordt toegepast;
— een bestaande methode opnieuw wordt uitgevonden;
— een gelegenheidsmethode wordt verzonnen.

4. Voordat het mogelijk wordt numeriek te berekenen hoe proto-melkwegstelsels instorten zullen vele onderzoekers hetzelfde lot hebben ondergaan.

5. Bij het zoeken naar een verklaring van de hoge-snelheids wolken dient meer aandacht te worden besteed aan de hoge snelheden dan aan de precieze massaverdeling.

6. Het eerstejaars sterrekunde-praktikum is van doorslaggevende betekenis bij de werving en de selektie van sterrekunde-studenten.

7. Een niet-autoritaire opvoeding heeft het zekere voordeel dat althans de jeugd een genietbare levensfase wordt.

8. Het ritmisch begeleiden van de muziek van J. S. Bach en tijd-genoten dient op de lagere school in klasseverband te worden beoefend, zodat men er een gezonde afkeer van overhoudt.

9. Kronkel zou er beter aan doen elke tweede zaterdag griep te hebben, in plaats van zijn lezers door ouderen bedachte of gepolijste kinderhumor over godsdienst, voortplanting en dood voor te zetten.

10. De wijze waarop Behrendt in Het Parool zijn politieke mening in beeld brengt is onaanvaardbaar.

11. Het is wenselijk dat in het Nederlandse Stripschap de kritiekloze dweepzucht van met name de Tom Poeskenners plaats maakt voor de eerlijke toewijding van bijvoorbeeld de Eric de Noormanliefhebbers.

12. De redenering op grond waarvan Poe het damspel boven het schaakspel stelt is nog altijd steekhoudend.

E. A. Poe, "The Murders in the Rue Morgue".

13. De recente uitbreiding der metaalbouwdozen van Trix Vereinigte Spielwarenfabriken met tal van gespecialiseerde onderdelen is er een treffend voorbeeld van hoe oude waarden verloren raken in onze van konkurrentie bezeten maatschappij.

14. Over vijftig jaar zal men tegenover de hedendaagse vuilverbranding even vreemd staan als wij nu staan tegenover het verbranden van "gevaarlijke" lichte destillaten in de beginjaren der aardolie-industrie.

1. INTRODUCTION

In solving problems of terrestrial and cosmical gas dynamics it is often permitted to ignore the dissipative processes in the gas, i.e. viscosity and thermal conductivity. This simplification of the physical picture boils down, mathematically, to a degeneration of the partial differential equations from second-order conservation laws, the Navier-Stokes equations, to first-order conservation laws: the equations of ideal compressible flow (ICF). Even in this approximation there remains a bewildering variety of complicated flow problems.

Particularly notorious are the problems involving such a strong compression of the gas that, sooner or later, dissipation will actually dominate the flow, at least in certain regions known as shocks. In a shock all quantities characterizing the flow undergo a significant change over a distance typical of the dissipative interaction, i.e. the molecular mean free path. It is clearly impossible to infer the structure of a shock from the equations of ICF; in fact, these simply break down at the point where a shock appears.

The concept of ICF can, however, be saved and extended by representing a shock as a true discontinuity in the flow. The motion of such an idealized shock may then be derived from an integral version of the first-order conservation laws, expressing the particular conservation principle for a finite volume of fluid and a finite lapse of time. The equation thus found for the shock speed does not refer to any shock structure; this customary result is correct in so far as the assumption is valid that away from shocks dissipation can be neglected.

Nevertheless, the practical ease of dealing with lower-order equations is largely spoiled by the very fact that separate equations have to be invoked to describe the discontinuities. This circumstance generally obstructs the analytical treatment of initial-value problems in the ICF-approximation. With the rise of high-speed computers the numerical approach has become possible and popular. An anthology of numerical techniques for smooth as well as shocked ICF may be found in Richtmyer and Morton {1, Ch. 12,13}, Fox {2, Ch. 26-28} and Alder, Fernbach and Rotenberg {3}.

Especially the finite-difference methods based on artificial dissipation have proved useful. The basic consideration is that, since in ICF the effect of dissipation is ignored, it might just as well be exaggerated.

By providing a finite-difference version of the equations of ICF with sufficiently large dissipative terms it is possible to achieve that shocks, whenever they appear, possess a structure coarse enough to be resolved in the computational net-work. For an extensive justification and documentation of the use of artificial dissipation reference is made to {1, Sec. 12.10 ff}. It must be stressed that the artificial dissipation should never conflict with the physical content of any supplementary equation (cf. Goldsworthy {4}).

The first example of a difference scheme incorporating artificial dissipation was given by Von Neumann and Richtmyer {5}. They added artificial viscosity to a scheme suited for smooth ICF, in order to allow it to handle shocks as well. In their work no attention was given to the fact that the equations of ICF are conservation laws.

In the work of Lax {6} and others the concept of "conservative schemes" was developed. A conservative scheme is a difference scheme which is consistent - in a manner defined below - with the integral form of the conservation laws. It has been shown, both theoretically and in computational practice, that, in using a conservative scheme, the numerical stability of a solution containing a shock automatically guarantees the correct motion of the shock. If indeed numerical stability is to be achieved, the scheme certainly has to be dissipative. Because the integral conservation laws represent fundamental properties of the physical system, conservative schemes are less arbitrary than schemes based on other forms of the equations of ICF, like those of the Von Neumann-Richtmyer type, or on some numerical analogue of a fluid, like the Particle-In-Cell method (see Harlow in {3}).

In deriving conservative schemes for ICF, it is not necessary to refer to the detailed physical meaning of the underlying equations. This makes these schemes remarkably accessible to numerical analysis. Moreover, any other physical system that is governed by first-order conservation laws, and allows of initial-value problems, can equally well be represented by such schemes. Appropriate examples of first-order conservation laws are the idealized equations of magneto-hydrodynamics (neglecting dissipation in the gas and diffusion of the magnetic field, see Friedrichs {7}), the lowest-order shallow-water equations (see Stoker {8}) and the idealized equations of elasticity (see Broer {9}).

Although the significance of conservative schemes has been fully recognized, an exhaustive search for even the simplest possible schemes has never been made. The present work attempts to fill this gap. With "simplest" we mean in the first place that the schemes considered are based on the smallest possible set of net-points used in building the finite differences. Among these only the schemes are considered that are explicit and self-starting. The precise meaning of these terms will be explained in Sec. 2.2.

Presumably it did not seem worth-while to make such an inventory, once Lax and Wendroff {10} had worked out the most accurate scheme of the kind described above. It appears now that an important scheme, which in practice often is preferable, has so far been overlooked. This scheme is congruent, up to terms of the order of the truncation error, with Godunov's well-known method {11}; the latter however is computationally much more cumbersome. Naturally a multitude of less interesting possibilities - known and unknown - come to light as well; these merely demonstrate how the principal schemes are connected.

2. THE PROBLEM

2.1 Mathematical preliminaries

In a conservation law, the divergence of some vector field (taken over all coordinates of definition space) is equated to some source term. We shall consider a hyperbolic system of conservation laws (HSCL) of the form

$$\frac{\partial}{\partial t} w^{(k)} + \frac{\partial}{\partial x} f^{(k)}\left(w^{(1)},\ldots,w^{(n)}\right) = 0 \qquad k = 1,\ldots,n\,. \tag{1}$$

The n functions $w^{(k)}$ represent the state quantities of some physical system and depend in a yet unknown way on time t and position x. The n auxiliary quantities $f^{(k)}$ are known, well-behaved, generally nonlinear functions of all state quantities, but not of any of their derivatives. The above conservation laws hence are of the first order. Note further that in these equations

(i) no source terms are included;

(ii) the independent variables do not occur explicitly.

Extension of the present discussion beyond the scope of (i) and (ii) does not offer essentially new problems. From the numerical view-point, a larger number of independent variables can best be handled through fractional time-steps (see e.g. {1, Sec. 8.9}, Strang {12}, Gourlay and Morris {13}).

Considering the functions $w^{(k)}$ and $f^{(k)}$ as the components of column vectors w and f, we can write (1) as

$$\frac{\partial w}{\partial t} + \frac{\partial f(w)}{\partial x} = 0. \tag{2}$$

Introduction of the n×n Jacobi matrix A of f with respect to w, whose components are

$$A_{kl} = \frac{\partial f^{(k)}(w)}{\partial w^{(l)}} \qquad k,l = 1,\ldots,n\,, \tag{3}$$

permits us to write (2) in the form

$$\frac{\partial w}{\partial t} + A(w)\,\frac{\partial w}{\partial x} = 0. \tag{4}$$

Eq. (4) is hyperbolic, i.e. admits of initial-value problems, if A can be reduced to a real diagonal matrix. Hence A should have n independent real eigenvectors and corresponding real eigenvalues. The latter are called characteristic speeds and denoted by $a^{(k)}(w)$, k=1,...,n, in order of

increasing magnitude; this order is assumed to be independent of w. The functions $a^{(k)}(w)$ need not all be distinct. The largest absolute characteristic speed will be denoted by

$$a(w) \equiv \max_k |a^{(k)}(w)|. \tag{5}$$

We shall assume that $a(w)$ never vanishes. The diagonal form of A

$$A(w) = \begin{pmatrix} a^{(1)}(w) & & \emptyset \\ & \ddots & \\ \emptyset & & a^{(n)}(w) \end{pmatrix} \tag{6}$$

results from the similarity transformation

$$A(w) = P(w)\, A(w)\, P^{-1}(w), \tag{7}$$

where the columns of P^{-1} are the eigenvectors of A, suitably normalized. Eq. (4) can be reduced to normal form through left multiplication by P. With the new vector state quantity w given by

$$dw(w) = P(w)\, dw, \tag{8}$$

eq. (4) then reads

$$\frac{\partial w}{\partial t} + A(w)\, \frac{\partial w}{\partial x} = 0, \tag{9}$$

which is a short notation for the following system of convection equations:

$$\frac{\partial w^{(k)}}{\partial t} + a^{(k)}\left(w^{(1)},\ldots,w^{(n)}\right) \frac{\partial w^{(k)}}{\partial x} = 0 \qquad k = 1,\ldots,n. \tag{10}$$

Along each characteristic trajectory one component of w is conserved, a so-called Riemann invariant[1]. The normal equations therefore seem to be a favourable basis for analytical or numerical integrations.

A set-back is, however, that such integrations cannot be continued without limit. Because (9) is nonlinear, it may occur, in connection with compressive phenomena in the physical system, that characteristics of the same family run into each other. The corresponding invariant will become multi-valued, which clearly means that the differential equations cease to be valid. By physical experience we know that in w a discontinuity will appear: a shock. For any information on the motion of a shock we have to

---

[1] This term is also used for related but different quantities; see Lax {14}.

go back to an integral version of (2). According to (2), the vector field $\vec{h}$ with time component w and space component f is divergence-free in t-x space; consequently, any contour integral of $\vec{h}$ vanishes:

$$\oint_B \vec{h}(w)\cdot\vec{n}\, dB = 0. \tag{11}$$

Here $\vec{n}$ denotes the outward vector of unit length normal to the line element dB of a contour B. From (11) it follows that at a shock (2) must be replaced by the jump condition

$$U[w] = [f], \tag{12}$$

where U is the shock speed, i.e. the slope of the shock path with respect to the time axis, and the standard notation with square brackets is used to indicate the jumps.

Any function w that satisfies (11) is called a weak solution of (2); for further details see Lax {6},{14}. For a given set of initial values a weak solution need not be unique. Specifically, the past history of an initially prescribed shock can not unambiguously be inferred from (11). Likewise, the future of the reverse type of discontinuity - from which characteristics spread - is indetermined. These ambiguities can be removed by a selection criterion, the so-called entropy condition. This is usually presented as a mathematical rule of thumb, e.g.: "the number of characteristics leaving a shock must equal n-1". Its physical basis is that first-order conservation laws in fact are idealizations of second-order equations in which the inevitable dissipative processes are taken into account. The latter are responsible for the irreversibility of compression shocks and the immediate dissolution of rarefaction shocks.

Hence, in determining the proper weak solution of an initial-value problem we need eq. (9) for the continuous pieces, the jump condition (12) to connect these pieces, and the entropy condition to decide whether a discontinuity is permitted or not. The diversity of these three conditions complicates both analytical and numerical integrations.

A uniform numerical approach becomes possible if the idea of treating shocks as discontinuities is abandoned. Witness the use of first-order conservation laws, we are interested in the motion rather than the structure of

shocks. But we may very well admit some shock structure on a scale that is compatible with the fineness of numerical detail desired elsewhere. This is the principle underlying the use of dissipative difference schemes for eq. (2), i.e. finite-difference versions of (2) with a truncation error that incorporates even derivatives of the state quantities. When transported to the right-hand side of the equation, a derivative of the order 2M should have the sign $(-1)^{M+1}$ in order to yield the desired dissipative effect. The name "artificial dissipation" is particularly used when finite differences of some even order are purposely added to provide a dissipative effect of the same order.

The difference schemes considered in this paper all are dissipative, due to the inclusion of second differences. Because these differences at the same time lend overall numerical stability to the schemes, we shall call then stabilizing terms. This serves to distinguish them from artificial dissipation exclusively used to smooth compression waves, as in the Von Neumann-Richtmyer scheme for ICF. Actually, in analyzing the schemes concerned we shall hardly ever bother about shocks.

It is essential that these schemes are based on the original HSCL (2), i.e. not on any form of the equations corresponding to a different set of state variables, like (9), and not on any other form with the same state variables, like (4). As mentioned in the introduction, this adds to the uniqueness of the schemes. Accordingly, their derivation is straightforward and involves no intuition-guided guesswork.

2.2 Conservative schemes

Let us recall that, apart from being numerically stable, a finite-difference scheme first of all should be consistent with the underlying differential equations. However, schemes intended for the approximation of weak solutions of the HSCL (2) should be compared with the integral form (11), which is always valid, rather than with the differential form (2). Strictly speaking, it is required that the discrete analogue of (11), obtained by summing the particular difference version of (2) over the domain inside B, does not contain illegitimate source terms. For a uniform orthogonal computational net of points ($t^j = j\Delta t, x_m = m\Delta x$) this condition is fulfilled if the scheme can be written in the form

$$\frac{W(R_m^{j+1}) - W(R_m^j)}{\Delta t} + \frac{F(S_{m+1}^j) - F(S_m^j)}{\Delta x} = 0. \tag{13}$$

Here R denotes some fixed configuration of net points; superscript j and subscript m indicate that some cardinal point of R lies in (t^j, x_m). Further, W is a function whose arguments[2] are the values that w takes in the points of R. The set S and the function F are similarly defined. In order to make (13) consistent with (2) it is necessary and sufficient that

$$\text{if } w(t^k, x_l) = w_0 \text{ for all } (t^k, x_l) \in R_m^j \text{ then } W(R_m^j) = w_0; \tag{14}$$

$$\text{if } w(t^k, x_l) = w_0 \text{ for all } (t^k, x_l) \in S_m^j \text{ then } F(S_m^j) = f(w_0). \tag{15}$$

Difference schemes satisfying all above conditions will be called divergence-free or conservative. Appropriate changes have to be made for net-points near the boundaries of the domain of integration, and for uneven meshes; we shall come back to this in Sec. 4.3.

The above concept of conservative schemes was introduced by Lax and Wendroff {10}. They restricted themselves to the formulation of explicit two-level schemes. Such schemes involve only one net-point on the most advanced time level, while the remaining points all belong to one single lower level. This means that

$$R_m^j = (t^j, x_m) \tag{16}$$

and S becomes some sequence of net points at constant t. From (14) it follows immediately that

$$W(R_m^j) \equiv w_m^j. \tag{17}$$

The assumption that the computational net is uniform in the time direction is not essential, because in a two-level scheme only one time-difference occurs. Furthermore, two-level schemes are self-starting, i.e. the scheme itself can be used for the first time-step of a numerical integration.

The class of difference schemes that will presently be studied com-

---

[2] With a uniform net, there is no reason why the net-point values of t and x should explicitly occur in the functions W and F.

prises only the simplest of all possibilities indicated by Lax and Wendroff. These are recognized as follows. In explicit two-level schemes w_m^{j+1} is expressed solely in the values that w takes in net-points contained in the union of S_m^j and S_{m+1}^j. These points cover some space interval at $t = t^j$. The space interval which at that time actually determines the state in (t^{j+1},x_m) can be inferred from (10), in the absence of shocks. In a first approximation it reads

$$\left[\left(t^j,x_m - a^{(n)}(t^j,x_m)\cdot\Delta t\right),\left(t^j,x_m - a^{(1)}(t^j,x_m)\cdot\Delta t\right)\right]. \tag{18}$$

Meaningful numerical results can only be expected if (18) is contained in the interval occupied by $S_m^j \cup S_{m+1}^j$. This is known as the Courant-Friedrichs-Lewy (CFL) condition {15}. Because $a^{(1)}$ and $a^{(n)}$ may have opposite signs, the smallest set of net-points at t^j that seems a priori suitable for $S_m^j \cup S_{m+1}^j$, is

$$\{(t^j,x_{m-1}),(t^j,x_m),(t^j,x_{m+1})\}. \tag{19}$$

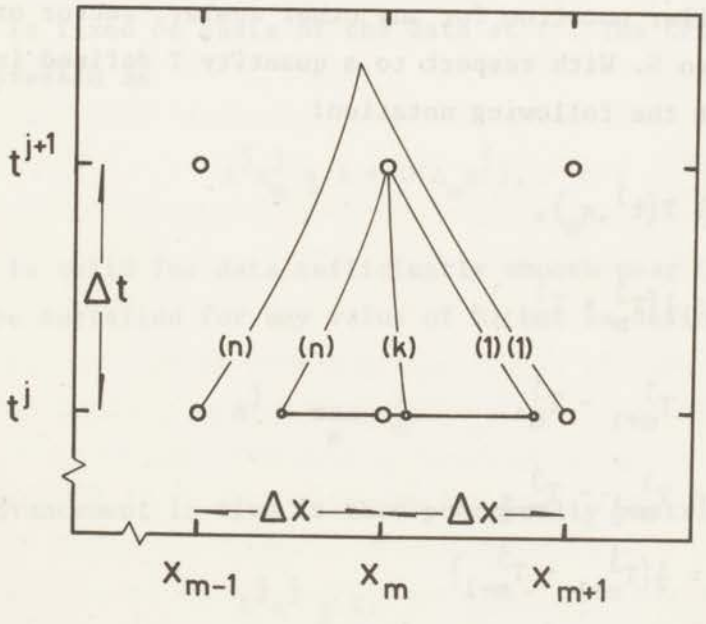


Figure 1. Net-point configuration used in explicit two-level four-point schemes. Some characteristics are drawn in order to illustrate the CFL condition; indices refer to characteristic families.

This corresponds to

$$S_m^j = \{(t^j, x_{m-1}), (t^j, x_m)\}. \tag{20}$$

The CFL condition for this set is illustrated in Fig. 1 by the condition that the characteristics emanating downwards from the upper net-point must remain within the characteristics emanating upwards from the outermost base points.

We shall devote the remainder of this paper to conservative schemes based on (16) and (20). Note that with the three net-points at t^j a second difference can be constructed; this is useful in making the schemes dissipative.

With respect to (20), let us call

$$F(S_m^j) \equiv F_{m-\frac{1}{2}}^j, \tag{21}$$

$$\Delta_m F^j \equiv F_{m+\frac{1}{2}}^j - F_{m-\frac{1}{2}}^j, \tag{22}$$

and henceforth use a similar notation for any other scalar, vector or matrix function defined on S. With respect to a quantity T defined in a single net-point we adopt the following notation:

$$T_m^j \equiv T(t^j, x_m), \tag{23}$$

$$T_{m+\frac{1}{2}}^j \equiv \tfrac{1}{2}(T_m^j + T_{m+1}^j), \tag{24}$$

$$\Delta_{m+\frac{1}{2}} T^j \equiv T_{m+1}^j - T_m^j, \tag{25}$$

$$\begin{aligned} \Delta_m T^j &\equiv T_{m+\frac{1}{2}}^j - T_{m-\frac{1}{2}}^j \\ &= \tfrac{1}{2}(T_{m+1}^j - T_{m-1}^j) \\ &= \tfrac{1}{2}(\Delta_{m+\frac{1}{2}} T^j + \Delta_{m-\frac{1}{2}} T^j), \end{aligned} \tag{26}$$

and finally

$$T_m^{j+\frac{1}{2}} \equiv \tfrac{1}{2}(T_m^j + T_m^{j+1}), \tag{27}$$

$$\Delta^{j+\frac{1}{2}} T_m \equiv T_m^{j+1} - T_m^j. \tag{28}$$

Indices will be suppressed wherever this is possible without creating confusion.

In the above notation, the explicit, two-level, four-point schemes in which we are interested take the form

$$\frac{\Delta^{j+\frac{1}{2}} w_m}{\Delta^{j+\frac{1}{2}} t} + \frac{\Delta_m F^j}{\Delta x} = 0. \tag{29}$$

To demonstrate that this is an explicit scheme we may write it as

$$w_m^{j+1} = w_m^j - \lambda^j \Delta_m F^j, \tag{30}$$

where we have introduced the mesh ratio

$$\lambda^j = \frac{\Delta^{j+\frac{1}{2}} t}{\Delta x}. \tag{31}$$

This quantity bears the superscript j because in actual computations its value is fixed on basis of the data at t^j. The CFL condition for (29) can be expressed as

$$\lambda^j a_m^j \leq 1 + O(\Delta_m a^j), \tag{32}$$

which is valid for data sufficiently smooth near (t^j, x_m). This inequality must be satisfied for any value of m. Let us define

$$a^j \equiv \max_m a_m^j; \tag{33}$$

the advancement in time is then practically restricted by

$$\lambda^j a^j \leq 1. \tag{34}$$

The expressions at the left side of (32) and (34) are called, respectively, the local Courant number

$$\sigma_m^j = \lambda^j a_m^j \tag{35}$$

and the global Courant number

$$\sigma^j = \lambda^j a^j. \tag{36}$$

We may also define a "characteristic Courant number"

$$\left(\sigma^{(k)}\right)_m^j = \lambda^j \left|a^{(k)}\right|_m^j. \tag{37}$$

From (34) it follows yet that in practice $O(\Delta t)$ and $O(\Delta x)$ are interchangeable.

2.3 Examples

The particular scheme that Lax and Wendroff discussed in detail in {10} is actually based on (20) and may serve to illustrate the general formula (29). In our notation it reads

$$w_m^{j+1} = w_m^j - \lambda^j \Delta_m f^j + \tfrac{1}{2}(\lambda^j)^2 \left(A^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} - A^j_{m-\frac{1}{2}} \Delta_{m-\frac{1}{2}}\right) f^j, \tag{38}$$

which results from inserting

$$F^j_{m+\frac{1}{2}} = f^j_{m+\frac{1}{2}} - \tfrac{1}{2}\lambda^j A^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} f^j \tag{39}$$

into (30). This is known as the Lax-Wendroff scheme. Clearly, (39) satisfies the consistency condition (15).

The first-order term at the right-hand side of (38) will be called the convection term. It represents $\Delta t.\left(\frac{\partial f}{\partial x}\right)_m^j$ as accurately as is possible with the given net-points, namely with an error of $O\left((\Delta x)^3\right)$. The second-order term is a stabilizing term; it approaches

$$\frac{(\Delta t)^2}{2}\left\{\frac{\partial}{\partial x}\left(A \frac{\partial f}{\partial x}\right)\right\}_m^j \equiv \frac{(\Delta t)^2}{2}\left\{\frac{\partial}{\partial x}\left(A^2 \frac{\partial w}{\partial x}\right)\right\}_m^j \tag{40}$$

with the same error $O\left((\Delta x)^3\right)$. In spite of its appearance, (40) does not represent the lowest-order dissipation present in the scheme, but has the effect of centering the convection term at the level $t^{j+\frac{1}{2}}$, within the margin of $O\left((\Delta x)^3\right)$. This may be seen upon comparing (38) and (40) with the

expansion

$$w_m^{j+1} = w_m^j + \Delta t\left(\frac{\partial w}{\partial t}\right)_m^j + \frac{(\Delta t)^2}{2}\left(\frac{\partial^2 w}{\partial t^2}\right)_m^j + \ldots$$

$$= w_m^j - \Delta t\left(\frac{\partial f}{\partial x}\right)_m^j + \frac{(\Delta t)^2}{2}\left\{\frac{\partial}{\partial x}\left(A\frac{\partial f}{\partial x}\right)\right\}_m^j - + \ldots \qquad (41)$$

The truncation error is thus reduced to $O\left((\Delta x)^3\right)$, which makes the Lax-Wendroff scheme the most accurate scheme admitted by (29). An expansion of (38) about the point (t^j, x_m) reveals (cf. {1, Sec. 12.14}) that this scheme is an approximation, up to terms of the magnitude $O\left((\Delta x)^4\right)$, of the equation

$$\frac{\partial w}{\partial t} + \frac{\partial f}{\partial x} + \frac{(\Delta x)^2}{6}\frac{\partial^2}{\partial x^2}\left\{(I - \lambda^2 A^2)\frac{\partial f}{\partial x}\right\} + T\left((\Delta x)^2\right) =$$

$$= -\frac{\Delta t(\Delta x)^2}{8}\frac{\partial^3}{\partial x^3}\left\{(I - \lambda^2 A^2)A^2\frac{\partial w}{\partial x}\right\} + T\left((\Delta x)^3\right). \qquad (42)$$

As compared to eq. (2), the above equation has an extra third-order convection term in the left-hand member, and a fourth-order dissipative term in the right-hand member. The notation $T\left((\Delta x)^2\right), T\left((\Delta x)^3\right)$ symbolizes terms of the magnitude $O\left((\Delta x)^2\right), O\left((\Delta x)^3\right)$ which have no systematic dissipative or convective effect, and vanish for a linear HSCL. Note that the matrix coefficients of $\frac{\partial w}{\partial x}$ and $\frac{\partial f}{\partial x}$ between the curly brackets in (42) are definite positive if the local CFL condition is fulfilled.

A linear stability analysis suggests that the Lax-Wendroff scheme is optimally stable, i.e. stable within the full range of Δt values which satisfy the CFL condition. However, because it is only weakly dissipative, the scheme is rather susceptible to nonlinear instabilities, i.e. instabilities that are not predictable from a linear analysis. These notably occur in connection with stand-off shocks; see e.g. Burstein {16}.

The lack of dissipation also appears in the poor numerical representation of shocks. In terms of a Fourier analysis, a steep transition like a shock is largely built of short-wavelength components. Dissipative terms have the effect of damping the shorter waves more strongly than the longer waves; this provides the "smoothing" of an initially sharp shock. A convective term of the order 2M+1, occurring at the left-hand side with the sign $(-1)^{M+1}$, causes a "normal" dispersion of the waves: the shorter waves move more slowly than the longer waves. Hence, if the dissipation is

insufficient, rapid oscillations will appear at the trailing end of the shock. These may persist in the post-shock region for quite a while, until they finally merge into the large-scale structure. This is a well-known feature of numerical results obtained by means of the Lax-Wendroff scheme.

Numerically generated oscillations in the state quantities often are unwanted, not only because they make the underlying smooth solution locally unrecognizable, but also because they might excite false modes of behaviour of the physical system considered.

In the oscillatory solutions produced by weakly dissipative schemes, much higher characteristic speeds occur than in smooth solutions obtained with stronger dissipation. If, in choosing a value of Δt, the anomalous characteristic speeds are simply ignored, the oscillations may become unstable. This depends on the HSCL and the problem considered, but in general the oscillations reduce the stable range of Δt. On the other hand, if Δt is chosen considerably less than the maximum value allowed by the CFL condition (34), the oscillations emerging from a shock may completely dominate the post-shock region. This certainly holds for the Lax-Wendroff scheme, where for small Δt the dissipative effect vanishes with respect to the convective error, as eq. (42) shows.

The numerical inconvenience of the Lax-Wendroff scheme can be reduced, as indicated by Lax and Wendroff themselves, by adding extra dissipative terms to the scheme. These must have the magnitude $O\left((\Delta x)^3\right)$ in order not to spoil the advantage of the scheme, namely its second-order accuracy.

The least accurate scheme of the form (29) is the scheme discussed by Lax {6}:

$$\begin{aligned} w_m^{j+1} &= w_m^j - \lambda^j \Delta_m f^j + \tfrac{1}{2}(\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}) w^j \\ &= \tfrac{1}{2}(w_{m-1}^j + w_{m+1}^j) - \lambda^j \Delta_m f^j, \end{aligned} \tag{43}$$

which follows from

$$F_{m+\frac{1}{2}}^j = f_{m+\frac{1}{2}}^j - \frac{1}{2\lambda^j} \Delta_{m+\frac{1}{2}} w^j. \tag{44}$$

It satisfies the consistency condition (15) and is optimally stable. Note that (43) is actually a three-point scheme, because (t^j, x_m) does not

contribute to it. Moreover, it is the only possible conservative three-point scheme. Lax' scheme has only first-order accuracy and is an approximation, up to terms of the magnitude $O\left((\Delta x)^3\right)$, of the equation

$$\frac{\partial w}{\partial t} + \frac{\partial f}{\partial x} - \frac{(\Delta x)^2}{3}\frac{\partial^2}{\partial x^2}\{(I - \lambda^2 A^2)\frac{\partial f}{\partial x}\} + T\left((\Delta x)^2\right) = \\ = \frac{(\Delta x)^2}{2\Delta t}\frac{\partial}{\partial x}\{(I - \lambda^2 A^2)\frac{\partial w}{\partial x}\}. \qquad (45)$$

The convective error is of the same order as in (42) but causes the reverse type of dispersion. The dissipative error is now of the second order. This provides very drastic smoothing: numerical solutions obtained with Lax' scheme are characterized by a lack of detail (cf. Emery {17}). Contrary to the Lax-Wendroff scheme, Lax' scheme never exhibits nonlinear instabilities and works very well with the maximum allowed value of Δt. It is not even advisable to employ too small a value for Δt, because this time the convection term vanishes with respect to the stabilizing term, resulting in an unpermissible loss of numerical resolution.

Between the extreme schemes (38) and (43) there is an infinity of possibilities all agreeing with (29), and all having first-order accuracy. Few of these have been spelled out and used. An example is the scheme of Rusanov {18}, which is a four-point improvement of Lax' scheme; further extensions are discussed by Van Leer {19}. The only other example is the scheme of Godunov {11}. This scheme, which we shall discuss more fully in Sec. 3.5, was especially designed in order to get around the peculiarities of the schemes of Lax and Lax-Wendroff. However, the clear improvement in the quality of the numerical results comes at the expense of a sharp increase in computing time. Though Godunov's scheme is "explicit" because w_m^{j+1} is expressed solely in quantities known at t^j, it does not contain an explicit expression for $F^j_{m+\frac{1}{2}}$. The components of $F^j_{m+\frac{1}{2}}$ follow from a set of algebraic equations involving the components of w_m^j and w_{m+1}^j; these equations have to be solved by iteration.

The question arises, how much of the benefit of Godunov's scheme is preserved if the exact solution of $F^j_{m+\frac{1}{2}}$ is replaced by, say, an approximation up to $O\left((\Delta x)^2\right)$. Instead of finding the answer to this particular question, which would be straightforward, we prefer a more general approach. In order to gain a better insight into the possibilities offered by conservative schemes we shall now investigate the complete set of schemes of the form (29).

3. THE DIFFERENCE SCHEMES

3.1 Selection criteria

In exploring the overwhelming amount of possibilities within eq. (29), we need a few criteria to sort out the basic schemes. The following four requirements are made in order to guarantee that, at least in the linear case, the differential equations share some essential properties with the difference equations.

{a} If the HSCL (2) is linearized, i.e. $\frac{\partial w}{\partial x}$ is assumed to be so small that A can be regarded as a constant matrix, then the difference scheme must become linear too.

{b} It must then be possible, through left multiplication of the scheme by P, which in this case is also constant, to obtain finite-difference versions of the normal differential equations (10).

{c} The coefficients of the finite differences in the k-th normal difference equation must depend on no other than the k-th characteristic speed.

{d} The normal difference equations must be identical, except for the value of k.

In the fifth requirement no particular reference is made to the possible linearity of the HSCL.

{e} If, in the HSCL (4), A(w) is replaced by -A(w), and the set of initial values is reflected with respect to x = 0, then the scheme should exactly reproduce the numerical solution of the original initial-value problem, apart from the reflection.

A difference scheme will be called admissible if it fulfils the above requirements. Two more requirements, which are less obvious than the first five, will arise in the course of this section.

In order to illustrate the meaning of conditions {a} through {e}, we shall assume, for the time being, that the HSCL (2) is indeed linear. By {a}, the function $F^j_{m+\frac{1}{2}}$ is now linear in w^j_m and w^j_{m+1}; it may generally be written as

$$F^j_{m+\frac{1}{2}} = f^j_{m+\frac{1}{2}} - \frac{1}{2\lambda} Q\Delta_{m+\frac{1}{2}} w^j. \tag{46}$$

Here Q is a matrix not depending on w, with the same physical dimension as λA. The general difference scheme for the linear HSCL thus becomes

$$\Delta^{j+\frac{1}{2}} w_m = - \lambda A \Delta_m w^j + \tfrac{1}{2} Q(\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}) w^j. \tag{47}$$

From {b} it follows that Q must commute with A. The diagonal form of Q will be called Q, and $q^{(k)}$ denotes the eigenvalue of Q corresponding to the same eigenvector of A as $a^{(k)}$; $q^{(k)}$ has no physical dimension. After left multiplication by P, eq. (47) reads

$$\Delta^{j+\frac{1}{2}} w_m = - \lambda A \Delta_m w^j + \tfrac{1}{2} Q(\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}) w^j, \tag{48}$$

which is a short notation for n normal difference equations of the form

$$\Delta^{j+\frac{1}{2}} \left(w^{(k)}\right)_m = - \lambda a^{(k)} \Delta_m \left(w^{(k)}\right)^j + \tfrac{1}{2} q^{(k)} (\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}) \left(w^{(k)}\right)^j. \tag{49}$$

The only physical parameter occurring in the k-th normal differential equation is $a^{(k)}$. Accordingly, in {c} it is required that $q^{(k)}$ depends only on $a^{(k)}$. Since all normal differential equations are mathematically equivalent, at least in the linear case, there is no reason to distinguish between the normal difference equations. Hence the eigenvalues $q^{(k)}$, k=1,...,n, must be values of one and the same function $q(a^{(k)})$. This is covered by {d}. As is easily checked, requirement {e} implies that $q(a^{(k)})$ is an even function of its argument, hence depends only on $(a^{(k)})$. For dimensional reasons we may conclude that $q^{(k)}$ only involves the characteristic Courant number $\sigma^{(k)}$, defined in (37).

In referring to the same computational net-work, the normal difference equations are numerically connected. If we ignore this for a moment, eq. (49) may be regarded as the general linear difference scheme for the

linear convection equation

$$\frac{\partial w^{(k)}}{\partial t} + a^{(k)} \frac{\partial w^{(k)}}{\partial x} = 0. \tag{50}$$

A scheme for this equation is sometimes called a convective difference scheme. The exact solution of an initial-value problem for eq. (50) is

$$w^{(k)}(t,x) = w^{(k)}(0,\ x - a^{(k)}t). \tag{51}$$

With respect to the computational net, we have in particular

$$w^{(k)}(t^{j+1},x_m) = w^{(k)}(t^j,\ x_m - \lambda a^{(k)}\Delta x). \tag{52}$$

The point $(t^j, x_m - \lambda a^{(k)}\Delta x)$ is not a nodal point, unless $\sigma^{(k)}$ equals zero, or one, the maximum value allowed by the CFL condition (32). In these cases a convective difference scheme ought to produce the exact solution. For any other value of $\sigma^{(k)}$ the scheme can only provide a more or less accurate interpolation.

The value $\sigma^{(k)} = 0$ occurs when $a^{(k)} = 0$ or $\lambda = 0$. In the first case the initial values of $w^{(k)}$ at once represent the solution at a later time. These should therefore remain unchanged if a difference scheme is applied. In the second case there is simply no advancement in time. Again the scheme should not affect the given data, unless it is purposely used as a "smoothing operator". This will not be considered here. As for scheme (49), we must have

$$q\left(\sigma^{(k)} = 0\right) = 0. \tag{53}$$

With $\sigma^{(k)} = 1$, eq. (52) becomes

$$\left.\begin{aligned} \left(w^{(k)}\right)_m^{j+1} &= \left(w^{(k)}\right)_{m-1}^{j} \text{ for } a^{(k)} > 0, \\ \left(w^{(k)}\right)_m^{j+1} &= \left(w^{(k)}\right)_{m+1}^{j} \text{ for } a^{(k)} < 0. \end{aligned}\right\} \tag{54}$$

Scheme (49) reduces to (54) if and only if

$$q\left(\sigma^{(k)} = 1\right) = 1. \tag{55}$$

It appears that (55) is also demanded on account of numerical stability (see Sec. 3.3), hence need not be required separately. Moreover, in dealing with a system of convection equations it is of hardly any use to require (55) for all values of k. Computationally, the convective difference schemes are coupled by the fact that in each of these the same value of λ is used. This value is related to the maximum absolute characteristic speed a by the CFL condition (32). Only the maximum characteristic Courant number, hence the local Courant number σ defined in (35), may reach unity; all other $\sigma^{(k)}$ stay below. In view of this numerical reality we may as well allow σ, which does not contain k anyway, to enter into the function $q^{(k)}$ as a parameter: $q^{(k)} \equiv q(\sigma^{(k)}, \sigma)$. This will not be considered as a violation of condition {c}. The only other dimensionless quantities admitted in $q^{(k)}$ are number constants. Condition (55) now reads, in a weaker form,

$$q\left(\sigma^{(k)} = \sigma = 1\right) = 1. \tag{56}$$

On the other hand, condition (53) remains significant with respect to any member of a system of convective schemes. It must even be supplemented because, due to the introduction of σ into $q^{(k)}$, the cases $a^{(k)} = 0$ and $\lambda = 0$ are not equivalent any more. The case $\lambda = 0$ can be separately accounted for, in demanding that

$$q\left(\sigma^{(k)} = \sigma = 0\right) = 0. \tag{57}$$

This restriction on linear admissible schemes is a special form of the sixth selection criterion given below.

{f} A difference scheme for the HSCL (2) should not yield a change in the state vector if this is not accompanied by any change in time.

In order to convert (53) into a suitable selection criterion, we must go back to the original difference scheme (47). The circumstance that one of the eigenvalues of A vanishes has the consequence that a gradient in the

corresponding Riemann invariant will not cause a gradient in f, for

$$\Delta f = P^{-1} A \Delta w. \tag{58}$$

A gradient in w however will always cause a gradient in w, as

$$\Delta w = P^{-1} \Delta w. \tag{59}$$

Hence, in this case, f(w') = f(w") does not imply that w' = w". By (53) it is now ensured that the scheme has the following property:

$$\text{if } f(w') = f(w'') = f_0 \text{ then } F(w',w'') = f_0. \tag{60}$$

Note that this is a stronger statement than the criterion for consistency (15). The unnecessary errors that arise when det A vanishes but the scheme does not agree with (60), are of the same nature as the interpolation errors inherent in any convective scheme. As implied in the notion of consistency, such errors disappear if Δt and Δx together approach zero while λ remains constant. They should however be avoided in practice whenever this is possible. It appears below that condition (60) itself is too restrictive to yield a useful selection criterion for nonlinear difference schemes.

With respect to a nonlinear HSCL (2), it may very well happen that $\Delta f = 0$ while $\Delta w \neq 0$. An important example is provided by the jumps of w and f across a standing shock. As is seen from eq. (12), f does not change in a shock with speed U = 0. The entropy condition implies that there is a family of characteristics, say, the k-th, which is absorbed by the shock. It is this particular family which is responsible for the shock, i.e. whose "breaking" must be prevented by the introduction of a discontinuity. It follows that the shock speed must lie between the pre- and post-shock values of $a^{(k)}(w)$. If the pre- and post-shock values of w are connected by a continuous sequence of states, there must be one such state for which $a^{(k)}(w)$ equals U. For a standing shock this means that $a^{(k)}(w)$ vanishes somewhere in the shock structure. This vanishing eigenvalue is reminiscent of the linear case $\Delta f = 0$, $\Delta w \neq 0$.

Imagine now that at $t = t^j$ a standing shock is given in the point $x = \frac{1}{2}\Delta x$, connecting the initial state $w_{-\infty}$, which is prescribed in all $x < \frac{1}{2}\Delta x$, with the final state $w_{+\infty}$, prescribed in all $x > \frac{1}{2}\Delta x$. The most

accurate discrete representation of this shock is of course

$$\left.\begin{aligned} w_m^j &= w_{-\infty} \quad \text{for } m \leq 0, \\ w_m^j &= w_{+\infty} \quad \text{for } m \geq 1. \end{aligned}\right\} \tag{61}$$

This situation may be preserved for all times if the scheme is consistent in the sense of (60). However, it is questionable whether this property is wanted now, because it would mean that, in the ideal numerical shock, dissipation is completely absent. In this circumstance, a slight disturbance might make the shock numerically unstable. This happens in any case for the Lax-Wendroff scheme, which indeed satisfies (60). We conclude that (60) does not make a suitable selection criterion.

A better starting point is the assumption that the HSCL (2) is linearly degenerate, i.e. that there is a family of characteristics which never gives rise to a shock; see Lax {14}. This happens when some characteristic speed, say $a^{(k)}(w)$, does not depend on $w^{(k)}$. If, in some domain of the t-x plane, the other Riemann invariants are chosen in such a way that $a^{(k)}(w)$ becomes a constant, then a linear convection equation for $w^{(k)}$ results. If in particular $a^{(k)}(w) \equiv 0$, it may again be that $\Delta f = 0$ although $\Delta w \neq 0$. It is this circumstance that should be met in the seventh selection criterion, which we now formulate as follows:

{g} A difference scheme for the HSCL (2) should not yield a change in the state vector if, due to a linear degeneration, the differential equations do not indicate a change either.

This includes the condition (53) for linear admissible schemes, because a linear HSCL is completely degenerate. Schemes that fulfil all requirements {a} through {g} will be called preferable.

To conclude this section we go back to linear schemes. Once a function q is given that agrees with {c}-{g}, the diagonal form of Q is fixed. The matrix Q itself may then be found with aid of

$$Q = P^{-1}QP. \tag{62}$$

This should be regarded as a formal rather than a practical recipe for Q. A formulation which allows of greater computational convenience, because it does not involve P, will presently be given.

Suppose that the number of distinct eigenvalues of A equals r ($\leq$ n). Then any relevant matrix Q can be written as a polynomial in λA of degree r-1, with scalar, dimensionless coefficients:

$$Q = \kappa_0 I + \kappa_1 \lambda A + \kappa_2 (\lambda A)^2 + \ldots + \kappa_{r-1} (\lambda A)^{r-1}. \tag{63}$$

In order to find the r coefficients $\kappa_0, \ldots, \kappa_{r-1}$, one must solve the following set of r linear equations:

$$\begin{pmatrix} 1 & \lambda a^{(1)} & (\lambda a^{(1)})^2 & \ldots & (\lambda a^{(1)})^{r-1} \\ \vdots & \vdots & & & \vdots \\ 1 & \lambda a^{(n)} & (\lambda a^{(n)})^2 & \ldots & (\lambda a^{(n)})^{r-1} \end{pmatrix} \begin{pmatrix} \kappa_0 \\ \vdots \\ \kappa_{r-1} \end{pmatrix} = \begin{pmatrix} q(\sigma^{(1)}) \\ \vdots \\ q(\sigma^{(n)}) \end{pmatrix}. \tag{64}$$

Here the equations for duplicate eigenvalues of A are understood to be omitted. Another way to find the coefficients is to write Q as

$$Q = \sum_{\substack{k=1 \\ a^{(k_1)} \neq a^{(k_2)}}}^{n} \left\{ q(\sigma^{(k)}) \times \prod_{\substack{i=1 \\ a^{(i)} \neq a^{(k)} \\ a^{(i_1)} \neq a^{(i_2)}}}^{n} \left(\frac{A - a^{(i)} I}{a^{(k)} - a^{(i)}} \right) \right\}. \tag{65}$$

The products contain r-1 factors, as they are taken only over those values of i corresponding to eigenvalues distinct from $a^{(k)}$ and from each other. Likewise, the summation covers multiple eigenvalues only once.

From (63) it follows that

$$Q\Delta w = \kappa_0 \Delta w + \lambda \{\kappa_1 + \kappa_2 \lambda A + \ldots + \kappa_{r-1} (\lambda A)^{r-1}\} \Delta f. \tag{66}$$

The conditions {f} and {g} imply that κ_0 must vanish if $\lambda = 0$ and if some eigenvalue of A vanishes. From (64) or (65) it can be seen that this is indeed achieved in requiring (57) and (53). In the linear case, condition {g} is also equivalent to (60). It is tempting to write $Q\Delta w$ as $Q'\lambda\Delta f$, with

$Q' = Q(\lambda A)^{-1}$; the matrix Q' may again be evaluated as a polynomial of the type (63). This of course is not always possible. It can only be done if $q(\sigma^{(k)})/\lambda a^{(k)}$ is defined for $\lambda a^{(k)} = 0$, hence for preferable schemes with

$$\frac{dq(\sigma^{(k)})}{d\sigma^{(k)}} = 0 \text{ for } \sigma^{(k)} = 0. \tag{67}$$

The distinction between preferable schemes which do or do not satisfy (67) will play an important role in the further sections of this chapter.

3.2 Basic form

Among all admissible difference schemes for the nonlinear HSCL (2), the smallest class that includes all possibilities of the linear case can be written in the form

$$\Delta^{j+\frac{1}{2}} w_m = - \lambda^j \Delta_m f^j + \tfrac{1}{2}\left(Q^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} - Q^j_{m-\frac{1}{2}} \Delta_{m-\frac{1}{2}}\right) w^j, \tag{68}$$

which results from inserting

$$F^j_{m+\frac{1}{2}} = f^j_{m+\frac{1}{2}} - \frac{1}{2\lambda^j} Q^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} w^j \tag{69}$$

into (29). Here the matrix $Q(w)$, henceforth called the stabilization matrix, depends only on <u>one</u> state vector. The subscript $m+\frac{1}{2}$ merely indicates that Q is centered in the sense of (24); this is done in view of {e}. It is further required that

—— $Q(w)$ commutes with $A(w)$, cf. {b};

—— the eigenvalue $q^{(k)}(w)$ of $Q(w)$, corresponding to the k-th eigenvector of $A(w)$, depends on w solely through the k-th characteristic Courant number $\sigma^{(k)}(w)$, the local Courant number $\sigma(w)$ and the global Courant number $\sigma(t)$, defined respectively in (37), (35) and (36), cf. {a}, {c}, {e};

—— the eigenvalues $q^{(k)}(w)$, $k=1,\ldots,n$, must be values of one and the same function $q(\sigma^{(k)}(w),\sigma(w),\sigma(t))$, cf. {d}.

Under the above restrictions, the stabilization matrix will be called admissible and the corresponding schemes will be called basic.

For a basic scheme and its stabilization matrix to be preferable (i.e. to satisfy {f} and {g}), it is required in the first place that $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ vanishes with λ^j. As in the linear case, we must have

$$q\left(\sigma^{(k)}(w) = \sigma(w) = \sigma(t) = 0\right) = 0. \tag{70}$$

Secondly, $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ must vanish if one and the same eigenvalue of A(w) vanishes in both points (t^j,x_m) and (t^j,x_{m+1}) while also $\Delta_{m+\frac{1}{2}}f^j = 0$. In order to fulfil this requirement it is necessary, but not sufficient, that

$$q\left(\sigma^{(k)}(w) = 0\right) = 0. \tag{71}$$

This does not guarantee that $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ vanishes appropriately. If we really wish to fulfil condition {g}, we must replace this term by the expression

$$(\kappa_0)^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j + \lambda^j\{\kappa_1 + \kappa_2\lambda A + \ldots + \kappa_{r-1}(\lambda A)^{r-2}\}^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^j \tag{72}$$

or by

$$(\kappa_0)^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j + \lambda^j\{(\kappa_1)^j_{m+\frac{1}{2}} + (\kappa_2)^j_{m+\frac{1}{2}}\lambda^j A^j_{m+\frac{1}{2}} + \ldots + (\kappa_{r-1})^j_{m+\frac{1}{2}}(\lambda^j A^j_{m+\frac{1}{2}})^{r-2}\}\Delta_{m+\frac{1}{2}}f^j, \tag{73}$$

which both are correct within an insignificant margin of $O\left((\Delta x)^3\right)$. The coefficients $\kappa_0(w),\ldots,\kappa_{r-1}(w)$ are functions of <u>one</u> state vector and are defined with respect to Q(w) and A(w) in the same way as in the linear case. Again, $(\kappa_i)^j_{m+\frac{1}{2}}$ merely represents the mean of $(\kappa_i)^j_m$ and $(\kappa_i)^j_{m+1}$. If, for some k, $a^{(k)}$ and $q^{(k)}$ vanish in both (t^j,x_m) and (t^j,x_{m+1}), then both $(\kappa_0)^j_m$ and $(\kappa_0)^j_{m+1}$ will vanish, as needed to satisfy {g}. Unless stipulated otherwise, it will be understood that $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ is just a short notation for a polynomial like (72) or (73).

If a preferable function q satisfies (67), it is also possible to evaluate $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ as $\lambda^j Q'^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^j$, with $Q'(w) = Q(w)\{\lambda A(w)\}^{-1}$ written as a polynomial in $\lambda A(w)$. A preferable scheme admitting of this formulation may thus be made consistent in the sense of (60). In the preceding section it was argued that this is not really wanted, in view of the danger of a

nonlinear instability. Moreover, it will be made clear in Sec. 3.3 that such a scheme, even when it is not written in the form agreeing with (60), more easily exhibits nonlinear instabilities than other preferable schemes. The Lax-Wendroff scheme becomes unstable in the computation of a standing shock, no matter whether it is executed with $Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ replaced by $(\lambda^2A^2)^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ or $\lambda^j(\lambda A)^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^j$, or as a two-step scheme (see Sec. 4.1).

Henceforth we shall concentrate on basic schemes. Though we are mainly interested in preferable basic schemes, we shall not a priori reject the remaining basic schemes. In doing so, many known schemes would be excluded, viz. all schemes indicated by Van Leer {19}. For instance, Lax' scheme (43) is basic but not preferable, because $Q(w) \equiv I$ does not agree with (70) and (71). On the other hand, the Lax-Wendroff scheme (38) clearly is a preferable basic scheme. These two examples do not yet justify the use of (68) as the starting point of a systematic search for schemes of the form (29). The proper motivation is given below.

In the first place, to any admissible scheme for the nonlinear HSCL (2) one basic scheme can be assigned. This scheme is found upon linearizing the original scheme, and then setting up a nonlinear version of the form (68). We shall call it the principal part of the original scheme.

In the second place, it seems hardly attractive to list an arbitrary number of ways in which an admissible scheme may deviate from its part. Such deviations are only of interest if they have some systematic favourable effect on the scheme or on the numerical results. A quick examination of presently known non-basic schemes reveals that four major objectives may be pursued in departing from the basic form:

(A) to elucidate the physical meaning of the scheme;

(B) to reduce computing time;

(C) to reduce numerical oscillations;

(D) to prevent nonlinear instabilities.

It may be possible to achieve several but not all of these objectives in one version of the scheme. For instance, the (preferable) scheme of Godunov might be considered physically more elegant than its principal part, but it is definitely more time-consuming. Moreover, the results obtained with the scheme and with the principal part are practically identical, as will be shown later. We conclude that objective (A) should not given much weight in choosing a scheme.

The objective (B) is important and will be taken up in Sec. 4.1, where the two-step formulation of basic schemes is discussed. Two-step schemes may also have advantages in view of (C), as found by Burstein and Rubin {20}. However, the improvement in the damping of numerical oscillations is unintentional and can not be controlled; we shall therefore not pay special attention to it.

Actually, the goals (C) and (D) can only be pursued systematically by purposely adding extra non-basic differences to the basic schemes. Such terms can be studied nearly as thoroughly as the basic schemes themselves. An outline of this subject is given in Sec. 4.2.

The deviation of an admissible scheme from its principal part necessarily disappears for a linear HSCL (cf. {a}), which means that the corresponding deviation in $F^j_{m+\frac{1}{2}}$ must vanish if $\Delta_{m+\frac{1}{2}}A^j$ vanishes. If this is a deviation in the form of extra terms such as described in Sec. 4.2, it may be contrived to vanish with any positive power of $\Delta_{m+\frac{1}{2}}w^j$. In a more regular admissible scheme, like a two-step scheme, or Godunov's scheme, the deviation can be expanded in terms of $\Delta_m w^j$, hence is at most of the magnitude $O(\Delta x)$. Such a scheme can only differ an amount $O\big((\Delta x)^2\big)$ from its principal part. If in addition the scheme is preferable, then $F^j_{m+\frac{1}{2}} - f^j_{m+\frac{1}{2}}$ occasionally vanishes with $\Delta_{m+\frac{1}{2}}f^j$; cf. {g}. Because this may happen irrespectively of the disappearance of $\Delta_{m+\frac{1}{2}}A^j$, the deviation in $F^j_{m+\frac{1}{2}}$ from the principal part must now be at most of the order $O\big((\Delta x)^2\big)$. The deviation in the scheme itself is a mere $O\big((\Delta x)^3\big)$. This falls beyond the order of magnitude of the finite differences in the principal part. Except for the Lax-Wendroff scheme, this also means that the deviation is of a higher order of magnitude than the truncation error.

3.3 Stability

The exact solution of a properly posed initial-value problem depends continuously on the initial data. From numerical solutions we may demand the same. This yields a stability criterion which for linear difference schemes is sufficient and for nonlinear schemes may also be useful.

Let us slightly perturb the set of values that w takes at t^j, which is assumed to be sufficiently smooth. The perturbation is decomposed into a Fourier spectrum of oscillations of the kind

$$\delta w_m^j(\ell) = \delta w_0^j(\ell)\, e^{2\pi i x_m/\ell}, \tag{74}$$

where ℓ is the wavelength of the Fourier component. With $\alpha = 2\pi\Delta x/\ell$, this can be written as

$$\delta w_m^j(\alpha) = \delta w_0^j(\alpha)\, e^{\alpha m}. \tag{75}$$

The local amplification matrix $G(w,\alpha,\lambda)$ is now defined by

$$\delta w_m^{j+1}(\alpha) = G\left(w_m^j,\alpha,\lambda^j\right)\delta w_m^j(\alpha) + O\left\{\left(\delta w_0^j(\alpha)\right)^2\right\}, \tag{76}$$

where $\delta w_m^{j+1}(\alpha)$ denotes the deviation of the perturbed solution from the unperturbed solution, after one time-step. The amplitude of the first variation will not grow unboundedly if the following inequality is satisfied for any complex test vector v, with conjugate v^*:

$$\left|v^* \cdot G\left(w_m^j,\alpha,\lambda^j\right)v\right| \leq \left(1 + O(\Delta t)\right)|v|^2. \tag{77}$$

This should be ensured in any place x_m and reassured at any further time level. For a given difference scheme, (77) yields an inequality in λ^j with α as a parameter; the value of λ^j must be chosen such as to satisfy this inequality for any $\alpha \in [0,2\pi]$. Condition (77) is due to Lax and Wendroff {21} and is indeed sufficient to ensure numerical stability under regular circumstances. If $G(w,\alpha,\lambda)$ has a complete set of n eigenvectors, (77) reduces to the Von Neumann condition

$$\left|g^{(k)}\left(w_m^j,\alpha,\lambda^j\right)\right| \leq 1 + O(\Delta t) \qquad k = 1,\ldots,n, \tag{78}$$

where $g^{(k)}(w_m^j,\alpha,\lambda^j)$, k=1,...,n, are the eigenvalues of $G(w_m^j,\alpha,\lambda^j)$; see {1, Sec. 4.7}.

The inequality (77) is called a linear stability criterion because, in deriving it, only the amplification of the first variation is considered. For a linear difference scheme, a linear stability analysis is of course an exact analysis. The matrix $G(\alpha,\lambda)$ is then independent of w and becomes the Fourier transform of the difference scheme itself. Accordingly, condition (77) governs the boundedness of the numerical solution itself. The term

$O(\Delta t)$ in (77) may as well be dropped, because on dimensional grounds no particular meaning can be assigned to it.

A consequence of nonlinearity is that $G_m^j(\alpha,\lambda)$ will contain terms of the magnitude $O(\Delta_m w^j)$. In connection with eq. (4) these arise from the dependence of $A(w)$ on position. Fortunately, it is not necessary to evaluate these terms exactly and draw them into the stability analysis, as such terms will fall within the margin $O(\Delta t) \equiv O(\Delta x)$ in the right-hand member of (77). The actual amplification matrix can be replaced by the main part of it, in which all coefficients refer to (t^j, x_m). This main part may be regarded as the amplification matrix of the "locally linearized" scheme, i.e. the scheme resulting when $A(w)$ is assumed to take the same value in all mesh points included in the scheme. Henceforth, when referring to the "local amplification matrix", we always mean "the main part of the local amplification matrix".

Admissible schemes for eq. (2) have an amplification matrix $G_m^j(\alpha,\lambda^j)$ which commutes with A_m^j. Any eigenvalue of $G_m^j(\alpha,\lambda^j)$ can directly be expressed in the corresponding eigenvalue of A_m^j. This makes it possible to use (78) in detecting the stable range of λ^j. For the schemes considered this is always the CFL range, as will now be shown.

The local amplification matrix associated with scheme (68) is readily found to be

$$G_m^j(\alpha,\lambda^j) = I - (1 - \cos\alpha)Q_m^j - i\lambda^j A_m^j \sin\alpha . \tag{79}$$

For convenience we shall further omit the arguments α and λ and the net-point indices, if these take only one value throughout an equation. The k-th eigenvalue of G is

$$g^{(k)} = 1 - (1 - \cos\alpha)q^{(k)} - i\lambda a^{(k)} \sin\alpha , \tag{80}$$

and its modulus, a so-called factor of growth, is given by

$$|g^{(k)}|^2 = 1 - 4\sin^2\frac{\alpha}{2} \cdot \left[q^{(k)} - \left(\sigma^{(k)}\right)^2 - \left\{\left(q^{(k)}\right)^2 - \left(\sigma^{(k)}\right)^2\right\}\sin^2\frac{\alpha}{2}\right]. \tag{81}$$

For numerical stability it is required that the expression between curly brackets is definite non-negative. This happens if and only if

$$\left(\sigma^{(k)}\right)^2 \leqq q^{(k)} \leqq 1, \tag{82}$$

which is seen to include the local CFL condition (32).

If $q^{(k)}$ is equated to the lower limit of the range permitted by (82), which means that $Q = \lambda^2 A^2$, we obtain the Lax-Wendroff scheme (38). The upper limit in (82) yields $Q = I$, hence the scheme (43) of Lax. This proves that Lax' scheme is indeed the least accurate scheme which is still stable within the scope of the CFL condition. It is furthermore seen that condition (55) is indeed a consequence of the requirement of stability, as was asserted in Sec. 3.1.

In the approximation of "local linearization", $g^{(k)}(\alpha,\lambda)$ is the Fourier transform of the k-th normal difference equation, i.e. the factor by which a Fourier component in $w^{(k)}$, with wavelength $2\pi\Delta x/\alpha$, changes upon one-time application of the scheme, with $\Delta t = \lambda\Delta x$. The modulus of this complex factor is given in eq. (81), and its argument is

$$\psi^{(k)} \equiv \arctan \frac{Im\ g^{(k)}}{Re\ g^{(k)}} = \arctan \frac{-\lambda a^{(k)} \sin\alpha}{1 - (1 - \cos\alpha) q^{(k)}}. \tag{83}$$

As follows from eq. (52), the differential equations would give rise to a factor $\exp\left(-i\lambda a^{(k)}\alpha\right)$, which is a unit vector with argument $-\lambda a^{(k)}\alpha$. The dissipative effect of the scheme may quantitatively be expressed as the amount by which $|g^{(k)}|$ is less than unity. The convective error will be given as the phase error, $\psi^{(k)} - \left(-\lambda a^{(k)}\alpha\right)$. For small values of α we have

$$1 - |g^{(k)}| = \tfrac{1}{2}\left\{q^{(k)} - \left(\sigma^{(k)}\right)^2\right\}\alpha^2 + O(\alpha^4) \tag{84}$$

$$\text{for } q^{(k)} > \left(\sigma^{(k)}\right)^2,$$

$$1 - |g^{(k)}| = \tfrac{1}{8}\left\{1 - \left(\sigma^{(k)}\right)^2\right\}\left(\sigma^{(k)}\right)^2\alpha^4 + O(\alpha^6) \tag{85}$$

$$\text{for } q^{(k)} = \left(\sigma^{(k)}\right)^2,$$

$$\psi^{(k)} - \left(-\lambda a^{(k)}\right)\alpha = -\tfrac{1}{6}\left\{1 + 2\left(\sigma^{(k)}\right)^2 - 3q^{(k)}\right\}\left(-\lambda a^{(k)}\right)\alpha^3 + O(\alpha^5). \tag{86}$$

These errors in the transformed numerical solution correspond directly to differences between the actual and the approximated differential equations,

such as given in (42) and (45) for the schemes of Lax and Lax-Wendroff. The fact that the above errors start only at $O(\alpha^2)$ reflects that the schemes considered satisfy the consistency conditions (14) and (15). Further aspects of eq. (80) will be considered in the next section.

The above stability analysis started from the assumption that $O(\Delta_m w) \equiv O(\Delta x)$. This assumption is certainly violated in shocks. These are represented by a change in w comparable to the value of w itself; this change always covers a few meshes, regardless of the value of Δx. If, at constant ratio, Δx and Δt go to zero, then $\Delta_m w/\Delta x$ becomes infinite in a shock. It is no wonder that nonlinear instabilities are often connected with shocks.

Nonlinear instabilities may arise when some factor of growth persistently is close to one for all values of α, in a mesh point where strong dissipation is actually required. It is seen from eq. (81) that $|g^{(k)}|$ approaches unity when $\sigma^{(k)}$ and $q^{(k)}$ together approach either unity or zero. The first case ($\sigma^{(k)}, q^{(k)} \to 1$) was alluded to in Sec. 2.3, in the discussion of the post-shock oscillations generated by the Lax-Wendroff scheme. The associated instabilities can simply be avoided by taking the global Courant number safely below unity. Though instabilities of this first kind do not occur for a linear HSCL, they are not usually referred to as nonlinear instabilities.

The nonlinear instabilities of the second kind (with $\sigma^{(k)}, q^{(k)} \to 0$) are the really serious ones, because these occur for any value of Δt. They have cropped up in Sec. 3.1, in connection with standing shocks. We recall that in a standing shock structure there is a point where, say, $a^{(k)}$ goes through zero; for a preferable stabilization matrix this means that $q^{(k)}$ drops to its minimum value zero. Whether this will lead to an instability depends on the flatness of the minimum, i.e. how closely $q^{(k)}$ stays near zero for small values of $|a^{(k)}|$. The danger of this happening is greatest if the graph of $q^{(k)}$, when plotted against $\sigma^{(k)}$, is tangent to the $\sigma^{(k)}$ axis in the origin. The same conclusion was reached by a different line of reasoning in connection with eq. (67).

3.4 Principles of classification

Any function $q\{\sigma^{(k)}(w),\sigma(w),\sigma(t)\}$ whose values remain between $(\sigma^{(k)})^2$ and 1, for $\sigma^{(k)}(w) \leq \sigma(w) \leq \sigma(t) \leq 1$, generates a basic scheme. In selecting one particular function for use in practice, one should of course be guided by theoretical and practical considerations. We shall furnish some of the theory in this and the next section; practice will be considered in Ch. 4 ff.

The dissipative properties of a finite-difference scheme may very well be illustrated by geometrical means. For fixed values of $\sigma^{(k)}$ and $q^{(k)}$, the eigenvalue $g^{(k)}(\alpha)$ given in (80) traces an ellipse in the complex plane when α varies from 0 to 2π:

$$\frac{\{Re\ g^{(k)} - (1 - q^{(k)})\}^2}{(q^{(k)})^2} + \frac{\{Im\ g^{(k)}\}^2}{(\sigma^{(k)})^2} = 1. \qquad (87)$$

The main axes of this ellipse are parallel to the real and imaginary axes of the complex plane; their respective half lengths are $q^{(k)}$ and $\sigma^{(k)}$. The center of the ellipse lies on the real axis, in the point $1-q^{(k)}$. For any value of $\sigma^{(k)}$ and $q^{(k)}$ the ellipse is tangent to the unit circle in the point 1 on the real axis, as required for the sake of consistency. In this point, the radius of curvature of the ellipse is given by $(\sigma^{(k)})^2/q^{(k)}$. For the Lax-Wendroff scheme, with $q^{(k)} \equiv (\sigma^{(k)})^2$, the radius equals 1. This demonstrates the high accuracy of the Lax-Wendroff scheme: the deviation of the ellipse from the unit circle is of the order α^4 for small α, cf. eq. (85). For all other schemes this deviation, i.e. the dissipation, is of the order α^2, cf. eq. (84). The ellipse corresponding to Lax' scheme, with $q^{(k)} \equiv 1$, most rapidly turns away from the unit circle, its radius of curvature being $(\sigma^{(k)})^2$ on the real axis. This ellipse is also tangent to the unit circle in the point -1, regardless of the value of $\sigma^{(k)}$. This shows that Lax' scheme is actually a three-point scheme: oscillations with $\alpha = \pi$, hence $\ell = 2\Delta x$, can not be damped at all because the scheme does not connect neighbouring net-points. The values of $\alpha > \pi$, corresponding to $\ell < 2\Delta x$, are of minor importance in determining the numerical properties of a scheme.

All ellipses corresponding to fixed $\sigma^{(k)}$ but different $q^{(k)}$ (hence different schemes) lie in the region defined by the left half of the Lax ellipse, the right half of the Lax-Wendroff ellipse, and connecting parts

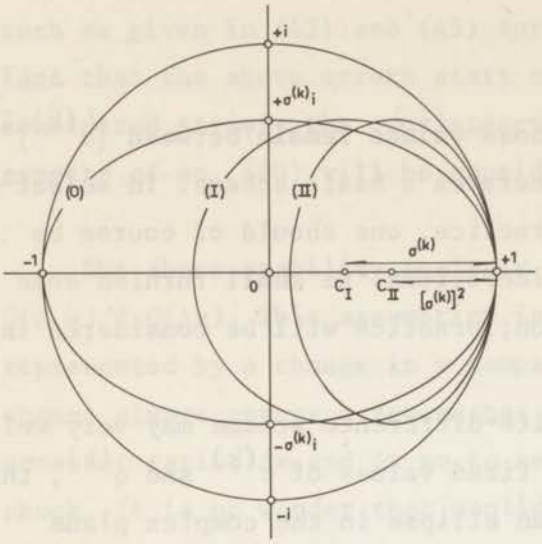


Figure 2.1. Ellipses traced by the complex vector $g(\alpha,\sigma^{(k)})$ with running α and fixed $\sigma^{(k)}$, for the schemes of Lax (0), Godunov (I) and Lax-Wendroff (II). The respective centers are 0, C_I and C_{II}. The dissipative error is smaller, the more these ellipses approach the unit circle.

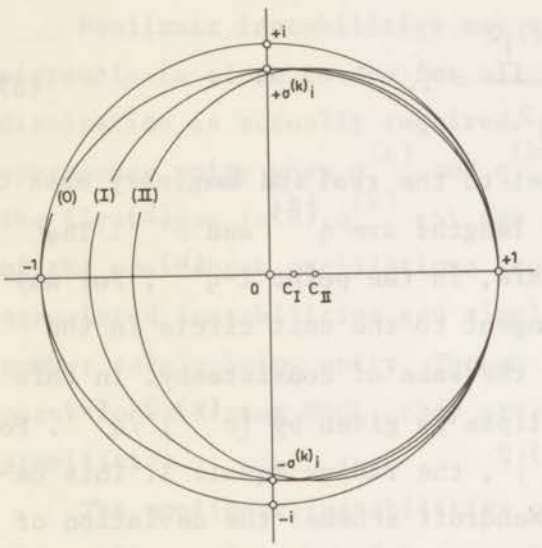


Figure 2.2. Same as Fig. 2.1, but with a value of $\sigma^{(k)}$ closer to one, yielding smaller dissipation.

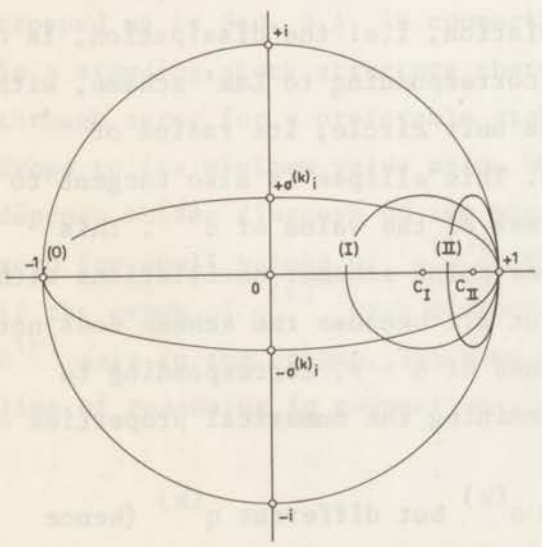


Figure 2.3. Same as Fig. 2.1, but with a value of $\sigma^{(k)}$ closer to zero. The dissipation is smaller than in Fig. 2.1 only for the preferable schemes (I) and (II).

Figure 2. Geometrical representation of dissipative errors in the complex plane.

of the lines through $\pm i\sigma^{(k)}$ parallel to the real axis; see Fig. 2.1. The centers of these ellipses lie between 0 and $1-(\sigma^{(k)})^2$. If it ever happens that $\sigma^{(k)} = 1$, then all ellipses will coincide with the unit circle; compare Fig. 2.2 with Fig. 2.1. There is no damping, and also no propagation error: the schemes yield the exact solution of the k-th normal equation (in the "locally linear" approximation). When $\sigma^{(k)} = 0$, only the preferable schemes yield the exact solution: because $q^{(k)} = 0$, the ellipses then reduce to the point 1. Lax' scheme is not preferable; damping becomes infinitely strong for all values of α except 0 and π. This is indicated in Fig. 2.3.

The ellipses corresponding to Lax' scheme are "lying" ellipses, and this holds for all other schemes with $q^{(k)} > \sigma^{(k)}$. Conversely, all schemes with $q^{(k)} < \sigma^{(k)}$, in particular the Lax-Wendroff scheme, are represented by "standing" ellipses. The scheme with $q^{(k)} \equiv \sigma^{(k)}$ is exceptional, in being represented by a <u>circle</u> for all values of $\sigma^{(k)}$. Because of this unique geometrical property, the scheme in question may be regarded as the central scheme between the extreme schemes of Lax and Lax-Wendroff. Note that it is a preferable scheme. It will be shown in the next section that this scheme is the principal part of Godunov's scheme, and has unique numerical properties. For convenience we shall henceforth refer to the three main schemes of Lax, Godunov (principal part) and Lax Wendroff as the schemes (0), (I) and (II) respectively.

The convective errors of basic difference schemes may be depicted in the same diagram as their dissipative errors. In Fig. 3 we have indicated where the angles $\psi^{(k)}$ and α occur in the geometrical representation of $g^{(k)}(\alpha)$. The convection speed $a^{(k)}$ has been taken negative; in this way the ellipses are traced counterclockwise with increasing α. We recall that there is no phase error if $\psi^{(k)} = -\lambda a^{(k)}\alpha$, which now means $\psi^{(k)} = \sigma^{(k)}\alpha$. The angle $\sigma^{(k)}\alpha$ is also shown in Fig. 3. From Fig. 3.1 it may be concluded that scheme (0) yields a lead in phase for any value of $\alpha \neq 0$ (provided that $\sigma^{(k)} < 1$). For this scheme we simply have $\tan\psi^{(k)} = \sigma^{(k)}\tan\alpha$; cf. eq. (83). For a preferable scheme like (II), the sign of the phase error is not fixed. There is a significant change in the character of the phase error when the ellipse goes through the origin, hence when $q^{(k)} = \frac{1}{2}$. For $q^{(k)} < \frac{1}{2}$ the ellipse stays at the right of the origin (see Fig. 3.2), which has the consequence that $d\psi^{(k)}/d\alpha$ changes its sign for some value of α between $\pi/2$ and π. When α goes through π, the phase angle $\psi^{(k)}$ itself becomes negative, indicating

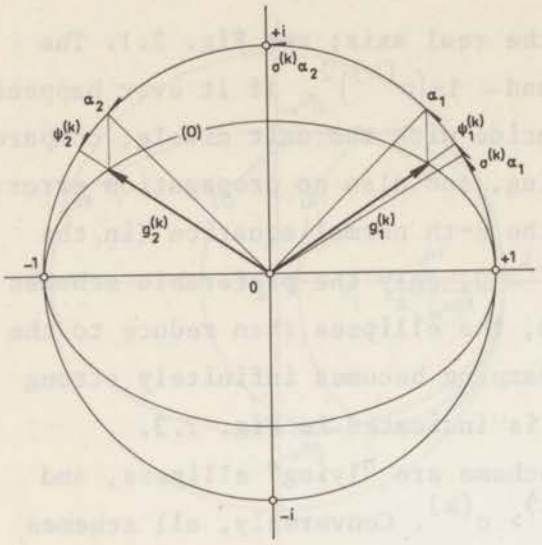


Figure 3.1. Construction of the complex vector $g^{(k)}(\alpha,\sigma^{(k)})$ for two values of α (indices 1 and 2) and fixed $\sigma^{(k)}$, for scheme (0). The angles α, $\sigma^{(k)}\alpha$ and $\psi^{(k)}$ are shown by arrows and are reckoned counter-clockwise along the unit circle, starting from the point +1 on the real axis. The convective error $\psi^{(k)} - \sigma^{(k)}\alpha$ is positive for both values of α.

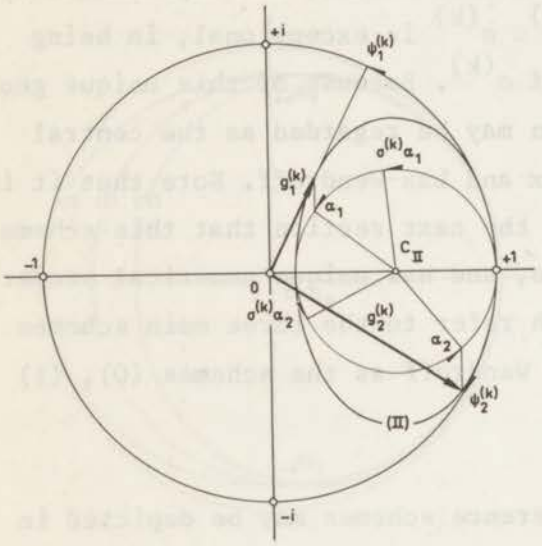


Figure 3.2. Same as in Fig. 3.1 but for scheme (II), with $(\sigma^{(k)})^2 < \frac{1}{2}$. The angles α and $\sigma^{(k)}\alpha$ are now measured along an auxiliary circle with radius $(\sigma^{(k)})^2$ centred in C_{II}. The angle α_1 is chosen such that the vector $g_1^{(k)}$ is tangent to the ellipse (II); for this value of α we have $d\psi^{(k)}/d\alpha = 0$. The phase angle $\psi^{(k)}$ is negative for $\alpha = \alpha_2$. The convective error is negative for both values of α.

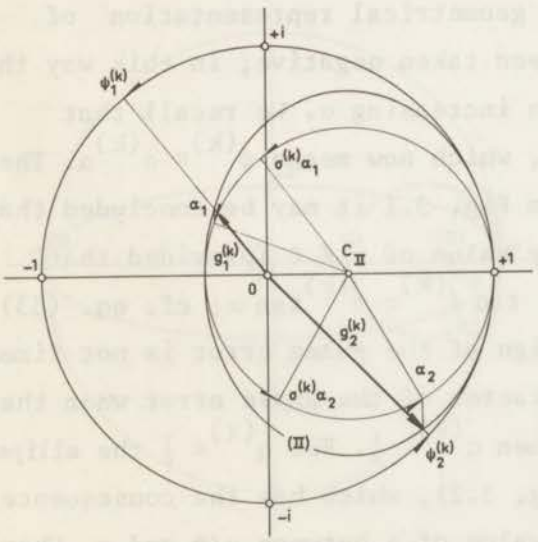


Figure 3.3. Same as Fig. 3.2, but with $(\sigma^{(k)})^2 > \frac{1}{2}$. The angle α_1 is chosen such that the convective error is zero; for $\alpha = \alpha_2$ it is positive.

Figure 3. Geometrical representation of convective errors in the complex plane.

an evident phase lag. For $q^{(k)} > \frac{1}{2}$, the ellipse crosses the imaginary axis (see Fig. 3.3); $d\psi^{(k)}/d\alpha$ is now always positive and there is a clear lead in phase when $\alpha = \pi = \psi^{(k)}$.

The dissipative and convective errors of the main schemes are summarized in the formula given below, with $-\lambda a^{(k)} = \sigma^{(k)}$.

Scheme (0): $q^{(k)} \equiv 1$;

$$\begin{aligned} |g^{(k)}|^2 &= 1 - \{1 - (\sigma^{(k)})^2\}\sin^2\alpha; \\ 1 - |g^{(k)}| &= \tfrac{1}{2}\{1 - (\sigma^{(k)})^2\}\alpha^2 + O(\alpha^4); \\ \psi^{(k)} - \sigma^{(k)}\alpha &= \tfrac{1}{3}\{1 - (\sigma^{(k)})^2\}\sigma^{(k)}\alpha^3 + O(\alpha^5). \end{aligned} \tag{88}$$

Scheme (I): $q^{(k)} \equiv \sigma^{(k)}$;

$$\begin{aligned} |g^{(k)}|^2 &= 1 - 4\sigma^{(k)}(1 - \sigma^{(k)})\sin^2\frac{\alpha}{2}; \\ 1 - |g^{(k)}| &= \tfrac{1}{2}\sigma^{(k)}(1 - \sigma^{(k)})\alpha^2 + O(\alpha^4); \\ \psi^{(k)} - \sigma^{(k)}\alpha &= -\tfrac{1}{6}(1 - 2\sigma^{(k)})(1 - \sigma^{(k)})\sigma^{(k)}\alpha^3 + O(\alpha^5). \end{aligned} \tag{89}$$

Scheme (II): $q^{(k)} \equiv (\sigma^{(k)})^2$;

$$\begin{aligned} |g^{(k)}|^2 &= 1 - 4\{1 - (\sigma^{(k)})^2\}(\sigma^{(k)})^2\sin^4\frac{\alpha}{2}; \\ 1 - |g^{(k)}| &= \tfrac{1}{8}\{1 - (\sigma^{(k)})^2\}(\sigma^{(k)})^2\alpha^4 + O(\alpha^6); \\ \psi^{(k)} - \sigma^{(k)}\alpha &= -\tfrac{1}{6}\{1 - (\sigma^{(k)})^2\}\sigma^{(k)}\alpha^3 + O(\alpha^5). \end{aligned} \tag{90}$$

Preferable schemes by definition cause no phase errors for $\sigma^{(k)} = 0$ or 1, and no dissipative error either. Only scheme (I) has the additional property that it causes no phase errors when $\sigma^{(k)} = \frac{1}{2}$, at least not for $\alpha < \pi$. This may be suspected upon considering (89) and easily be proven with aid of (83); in Fig. 4 we have shown it geometrically. The case $\sigma^{(k)} = \frac{1}{2}$ coincides for scheme (I) with the case $q^{(k)} = \frac{1}{2}$ (circle tangent to the imaginary axis), and is accompanied by a transition of the phase errors

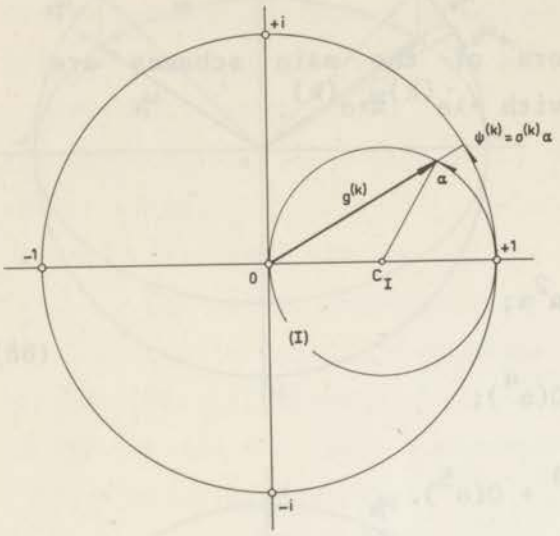


Figure 4. Construction of $g^{(k)}(\alpha,\sigma^{(k)})$ for scheme (I), in the case that $\sigma^{(k)}=\frac{1}{2}$. Here $\psi^{(k)}$ equals $\frac{1}{2}\alpha = \sigma^{(k)}\alpha$ for any value of α so that there is no convective error.

from definite negative, for $\sigma^{(k)}<\frac{1}{2}$, to definite positive for $\sigma^{(k)}>\frac{1}{2}$. From (89) it follows further that with $\sigma^{(k)}=\frac{1}{2}$ dissipation reaches its maximum for fixed α. Scheme (0) has its dissipation maximum for $\sigma^{(k)}=0$, and scheme (II) for $\sigma^{(k)}=\frac{1}{2}\sqrt{2}$.

A graphical display of the dissipative and dispersive errors of schemes (0) and (II) can be found in Vliegenthart {22}. Similar graphs referring to scheme (I) are presented in Figs. 5 and 6; these were kindly made available by Mr. Vliegenthart. In these figures are given the modulus and argument of the so-called propagation factor $T^{(k)}$, which contains the same information as $g^{(k)}$. The distinction is that, whereas $g^{(k)}$ yields the absolute dissipative and convective errors per time-step, $T^{(k)}$ yields the relative errors normalized over the time interval in which the Fourier component travels over a distance equal to its wavelength. The modulus of $T^{(k)}$ hence equals $|g^{(k)}|^{2\pi/\sigma^{(k)}\alpha}$ and its argument equals $(\psi^{(k)}-\sigma^{(k)}\alpha)\times 2\pi/\sigma^{(k)}\alpha$. Note that the normalized relative errors $1-T^{(k)}$ and $\arg T^{(k)}$ do not vanish in the limit when $\sigma^{(k)}\to 0$.

The three main schemes can be connected by a one-parameter family of schemes, in choosing

$$q^{(k)} = \left(\sigma^{(k)}\right)^{\nu}. \tag{91}$$

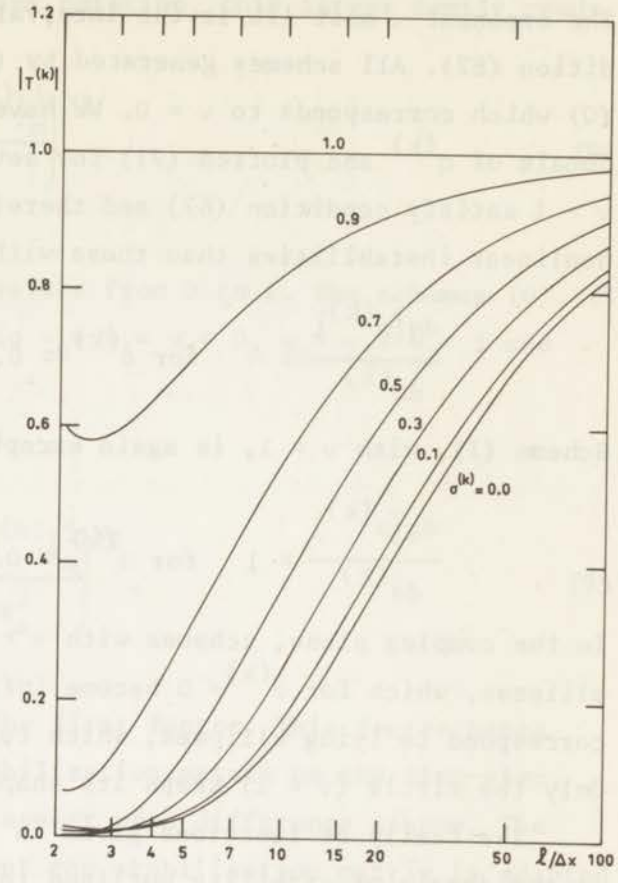


Figure 5. Dissipative properties of scheme (I). The logarithmically divided abscissa indicates the wavelength of a Fourier component; the ordinate gives the modulus of the propagation factor.

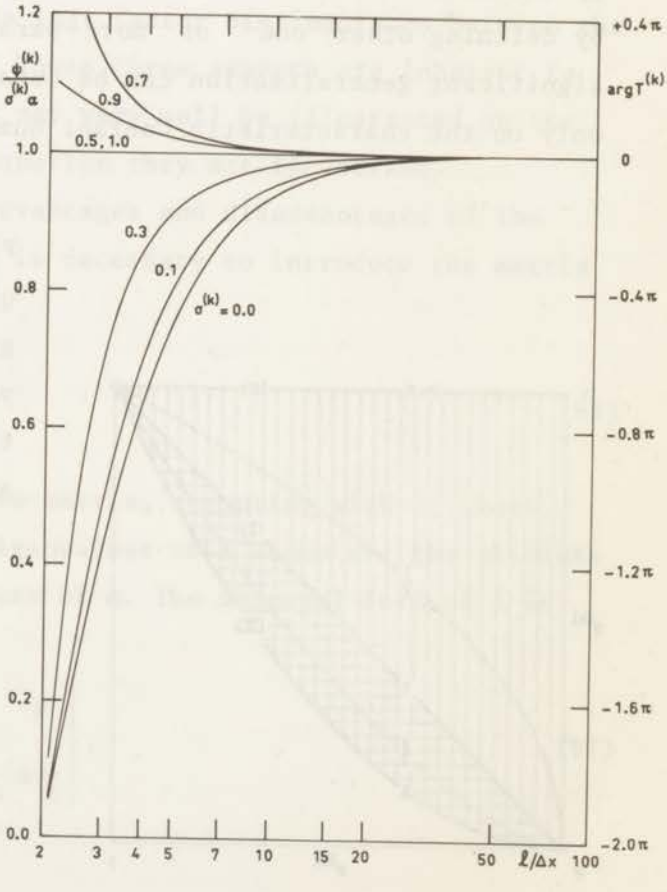


Figure 6. Convective properties of scheme (I). Abscissa same as in Fig. 5. The left vertical scale indicates the ratio between the computed and the exact velocity of a Fourier wave. The right vertical scale gives the argument of the propagation factor.

The exponent ν must lie in the interval $[0,2]$ in view of the stability condition (82). All schemes generated by (91) are preferable, except scheme (0) which corresponds to $\nu = 0$. We have indicated in Fig. 7 the stability domain of $q^{(k)}$ and plotted (91) for several values of ν. The schemes with $\nu > 1$ satisfy condition (67) and therefore are more easily susceptible to nonlinear instabilities than those with $\nu < 1$; the latter even have

$$\frac{dq(\sigma^{(k)})}{d\sigma^{(k)}} = \infty \quad \text{for } \sigma^{(k)} = 0. \tag{92}$$

Scheme (I), with $\nu = 1$, is again exceptional in having

$$\frac{dq(\sigma^{(k)})}{d\sigma^{(k)}} = 1 \quad \text{for } \sigma^{(k)} = 0. \tag{93}$$

In the complex plane, schemes with $\nu > 1$ are represented by standing ellipses, which for $\sigma^{(k)} \to 0$ become infinitely thin. Schemes with $\nu < 1$ correspond to lying ellipses, which for $\sigma^{(k)} \to 0$ become infinitely flat. Only the circle ($\nu = 1$) keeps its shape.

The family of functions given in (91) represents only one way to fill up the domain of stability outlined in Fig. 7. Nothing much would be gained by defining other one- or more-parameter families. However, a significant generalization can be obtained by permitting $q^{(k)}$ to depend not only on the characteristic Courant number but also on the local and global

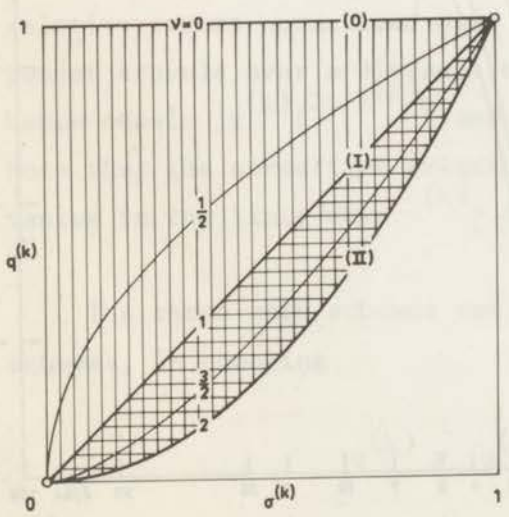
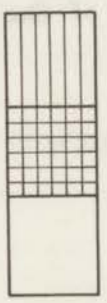


Figure 7. Stability region in the $q^{(k)}$-$\sigma^{(k)}$ plane. The function $q(\sigma^{(k)})$ given by eq. (91) is drawn for various values of the parameter ν, including those which yield schemes (0), (I) and (II).



linear stability

idem; danger of nonlinear instability for schemes arising from eq. (91)

linear instability

Courant number. In the usual net-point notation, this larger family reads

$$\left(q^{(k)}\right)^j_m = \left(\sigma^j\right)^{\nu_0}\left(\frac{\sigma^j_m}{\sigma^j}\right)^{\nu_1}\left(\frac{\left(\sigma^{(k)}\right)^j_m}{\sigma^j_m}\right)^{\nu_2}, \tag{94}$$

where all three exponents may take values from 0 to 2. The schemes (0), (I) and (II) correspond, respectively, to $\nu_0 = \nu_1 = \nu_2 = 0$, $\nu_0 = \nu_1 = \nu_2 = 1$ and $\nu_0 = \nu_1 = \nu_2 = 2$.

Eq. (94) may also be written as

$$\left(q^{(k)}\right)^j_m = \left|\lambda^j a^j\right|^{\nu_0}\left(\frac{a^j_m}{a^j}\right)^{\nu_1}\left(\frac{\left|a^{(k)}\right|^j_m}{a^j_m}\right)^{\nu_2}, \tag{95}$$

which shows that λ only appears in the first factor. This factor hence determines the dependence of the stabilization matrix on the time-step, which may be regarded as the global aspect of a difference scheme. The middle factor indicates to what extent the stabilization matrix is adapted to local needs: the local aspect. The last factor distinguishes between the normal equations: the normal aspect. These three aspects are inherent to any admissible difference scheme but may very well be illustrated on the basis of eq. (95), because in this equation they are factorized.

For a clear assessment of the advantages and disadvantages of the schemes made possible by eq. (95) it is necessary to introduce the matrix $\check{A}$, symbolically given by

$$\check{A} = \sqrt{A^2}\,. \tag{96}$$

This denotes the definite non-negative matrix, commuting with A, whose square equals the square of A. The eigenvalues of $\check{A}$ hence are the absolute values of the corresponding eigenvalues of A. The diagonal form of $\check{A}$ is

$$\check{A} = \begin{pmatrix} \left|a^{(1)}\right| & & \emptyset \\ & \ddots & \\ \emptyset & & \left|a^{(n)}\right| \end{pmatrix}, \tag{97}$$

and if we define the diagonal matrix

$$\mathcal{J} = \begin{pmatrix} \operatorname{sgn} a^{(1)} & & \emptyset \\ & \ddots & \\ \emptyset & & \operatorname{sgn} a^{(n)} \end{pmatrix}. \tag{98}$$

we can write

$$\tilde{A} = \mathcal{J}A. \tag{99}$$

Accordingly, the matrix $\tilde{A}$ can be written as

$$\tilde{A} = P^{-1}\tilde{A}P = P^{-1}\mathcal{J}AP = P^{-1}\mathcal{J}PP^{-1}AP = JA, \tag{100}$$

where J refers to the same basis as A; cf. eq. (7). With respect to scheme (I), the matrix J is the same as the matrix Q' defined in the last paragraph of Sec. 3.1. Note that $\mathcal{J}$ and J jump when some eigenvalue of A goes through zero. The fact that $\tilde{A}$ can be factorized is of no practical significance, unless the eigenvalues of A have fixed signs.

To simplify the further discussion we shall quantize the exponents ν_0, ν_1 and ν_2, replacing them by integers N_0, N_1 and N_2, which each may take the values 0, 1 or 2. The stabilization matrices corresponding to the 27 combinations of allowed values of N_0, N_1 and N_2, can be expressed in $\tilde{A}$ as follows:

$$Q_m^j = |\lambda a^j|^{N_0} \left(\frac{a_m^j}{a^j}\right)^{N_1} \left(\frac{\tilde{A}_m^j}{a_m^j}\right)^{N_2}. \tag{101}$$

In Fig. 8 we have given a three-dimensional representation of the family of 27 difference schemes generated by this equation.

The schemes corresponding to $N_2 = 0$ have been discussed by Van Leer {19}. These schemes are not preferable, and the amount of dissipation they provide is not differentiated with respect to the normal equations: it is always adjusted to the equation for the fastest characteristic. Nevertheless, these schemes have rendered good service and are very popular because of their computational simplicity (the matrix multiplication

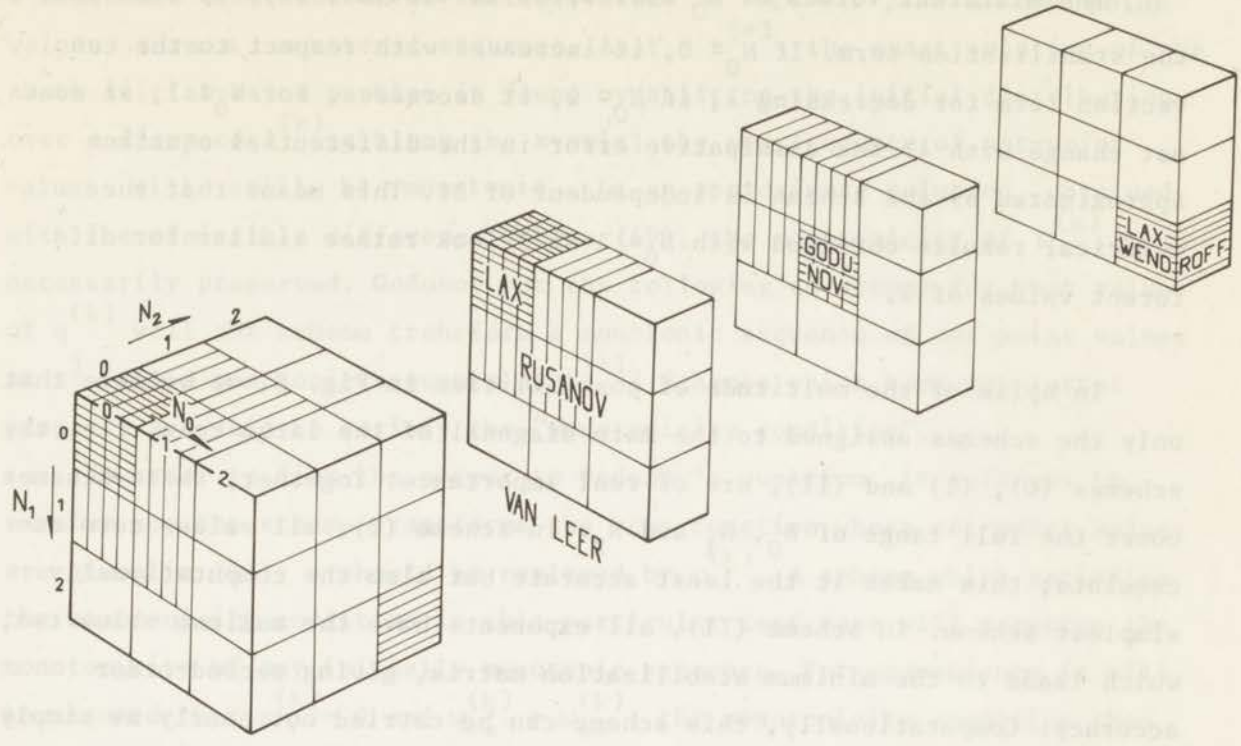


Figure 8. A choice of difference schemes. The family of 27 difference schemes arising from eq. (101) is represented by a large cube composed of 3 × 3 × 3 small cubes. The three sections correspond to different values of N_2, hence to different powers of the matrix $\tilde{A}$. The stabilization matrices of the schemes assigned to cubes of the same section differ only by a scalar factor. The cubes of schemes satisfying the monotonicity condition are lightly shaded. The heavily shaded cubes on the main diagonal correspond to the schemes (0), (I) and (II), i.e. the schemes of Lax {6}, the principal part of the scheme of Godunov {11} and the Lax-Wendroff scheme {10}. The method of Rusanov {18} occupies three cubes of the front section, which as a whole was described by Van Leer {19}.

$Q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$ simply becomes a scalar multiplication $q^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^j$) and because they do not give rise to nonlinear instabilities of the second kind. The simplest example is Lax' scheme (0). The schemes with $N_2 = 0$, $N_1 = 1$ were indicated by Rusanov {18}, who was the first one to realize that the exponents N_0 and N_1 in (101) left some play to choose schemes with other properties without requiring considerably more computing time. The introduction of the normal aspect, achieved by choosing N_2 different from zero, adds a third degree of freedom in the choice of schemes between the extremes (0) and (II).

The different values of N_0 correspond to distinct ways of behaviour of the stabilization term. If $N_0 = 0$, it increases with respect to the convection term for decreasing λ, if $N_0 = 2$, it decreases. For $N_0 = 1$, it does not change with λ: the dissipative error in the differential equation approximated by the scheme is independent of Δt. This means that the numerical results obtained with $N_0 = 1$ will look rather similar for different values of λ.

In spite of the multitude of possibilities in Fig. 8, we believe that only the schemes assigned to the main diagonal of the large cube, i.e. the schemes (0), (I) and (II), are of real importance. Together, these schemes cover the full range of N_0, N_1 and N_2. In scheme (0), all values zero accumulate; this makes it the least accurate but also the computationally simplest scheme. In scheme (II), all exponents have the maximum value two, which leads to the minimum stabilization matrix, giving second-order accuracy. Computationally, this scheme can be carried out nearly as simply as scheme (0) by employing a two-step formulation. The numerical results obtained with this scheme lack the smoothness of the results from scheme (0). In addition there is the danger of nonlinear instabilities, which can be avoided at the cost of increased complexity. Neither scheme (0) nor scheme (II) have an attractive global aspect.

Scheme (I) is represented by the central cube. Among the three schemes on the main diagonal it is the most complicated one because of occurrence of the matrix $\tilde{A}$, whose evaluation involves all r powers of A; cf. eq. (63). The stabilization matrix of (I) is the geometrical mean of the stabilization matrices of (0) and (II). Scheme (I) therefore appears to be a good all-purpose scheme: it has something of all aspects, including the favourite type of global aspect, and is not likely to exhibit persistent nonlinear instabilities.

3.5 Monotonicity

In order not to burden the previous discussion we still have left one main point out of consideration. This is the matter of monotonicity, to which Godunov {11} attached major importance in deriving his difference scheme. We shall briefly reproduce his line of reasoning.

Consider again the linear convection equation (50), supplemented with

a monotonic distribution of initial values $w^{(k)}(t = t^j, x)$. The net-point values form a monotonic sequence. At $t = t^{j+1}$, the exact solution of the above initial-value problem is found by shifting the initial distribution over a distance $a^{(k)}\Delta t$ along the x-axis; the new sequence of net-point values will still be monotonic. In an approximate solution, obtained with the admissible difference scheme (49), the monotonicity of $w^{(k)}$ is not necessarily preserved. Godunov put the following question: for what values of $q^{(k)}$ will the scheme transform a monotonic sequence of net-point values at t^j into a monotonic sequence at t^{j+1}? Schemes which have the latter property are said to satisfy the "monotonicity condition".

In order to find the answer to Godunov's question, it suffices to examine how the scheme transforms the step function whose net-point values are given in (61); w should be replaced by $w^{(k)}$. A scheme which satisfies the monotonicity condition in this particular test case will preserve the monotonicity of any initially monotonic sequence. For convenience it will be assumed that $a^{(k)} > 0$ and $w^{(k)}_{-\infty} > w^{(k)}_{+\infty}$; the monotonicity condition then reads

$$w^{(k)}_{-\infty} = \left(w^{(k)}\right)^{j+1}_{-1} \geq \left(w^{(k)}\right)^{j+1}_{0} \geq \left(w^{(k)}\right)^{j+1}_{+1} \geq \left(w^{(k)}\right)^{j+1}_{+2} = w^{(k)}_{+\infty}. \tag{102}$$

A straightforward calculation shows that these inequalities impose on $q^{(k)}$ the restriction

$$\lambda a^{(k)} \leq q^{(k)} \leq 1, \tag{103}$$

which should be compared with the requirement (82) for stability. The value $q^{(k)} = \lambda a^{(k)}$ is optimal in the sense that it differs from the value $\left(\lambda a^{(k)}\right)^2$, which assures second-order accuracy, no more than needed for the preservation of monotonicity. The corresponding scheme was therefore called by Godunov the "best" scheme.

If $q^{(k)} = \lambda a^{(k)}$ is inserted into the general formula (49), it reduces to

$$\begin{aligned}\left(w^{(k)}\right)^{j+1}_{m} &= \left(w^{(k)}\right)^{j}_{m} - \lambda a^{(k)}\left\{\left(w^{(k)}\right)^{j}_{m} - \left(w^{(k)}\right)^{j}_{m-1}\right\} \\ &= \left(1 - \lambda a^{(k)}\right)\left(w^{(k)}\right)^{j}_{m} + \lambda a^{(k)}\left(w^{(k)}\right)^{j}_{m-1}.\end{aligned} \tag{104}$$

Note that the "best" scheme is a three-point scheme. Furthermore, it is congruent with the following interpolation routine. Draw the k-th characteristic through (t^{j+1},x_m) backward in time; it will intersect the level $t = t^j$ in a point with abscissa $x_m - a^{(k)}\Delta t$. Because the characteristic speed is assumed to be positive, but smaller than $\Delta x/\Delta t$, this point of intersection lies between the net-points x_m and x_{m-1}. Find a value of $w^{(k)}$ in this point by linear interpolation between $\left(w^{(k)}\right)_m^j$ and $\left(w^{(k)}\right)_{m-1}^j$. Assign this same value to $\left(w^{(k)}\right)_m^{j+1}$.

This algorithm for a linear convection equation is also the basis of the method of Courant, Isaacson and Rees {23} for an arbitrary first-order hyperbolic system of the form (4). Because this method really starts from the normal form of the equations, it is not conservative and will not further be discussed here.

We shall now depart from Godunov's discussion and notice a peculiar property of the "best" scheme. In the matter of monotonicity it appears to be crucial that the value of $w^{(k)}$, transported along the characteristic, is the result of an interpolation rather than an extrapolation. Therefore, the choice of points used in the interpolation depends on the sign of $a^{(k)}$. For a negative value of $a^{(k)}$, the best scheme becomes

$$\begin{aligned}\left(w^{(k)}\right)_m^{j+1} &= \left(w^{(k)}\right)_m^j - \lambda a^{(k)}\Delta_{m+\frac{1}{2}}\left(w^{(k)}\right)^j \\ &= \left(1 + \lambda a^{(k)}\right)\left(w^{(k)}\right)_m^j - \lambda a^{(k)}\left(w^{(k)}\right)_{m+1}^j. \end{aligned} \tag{105}$$

The eqs.(104) and (105) can be combined in one scheme:

$$\begin{aligned}\left(w^{(k)}\right)_m^{j+1} &= \left(w^{(k)}\right)_m^j - \lambda a^{(k)}\Delta_m\left(w^{(k)}\right)^j + \tfrac{1}{2}\lambda\left|a^{(k)}\right|\left(\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}\right)\left(w^{(k)}\right)^j \\ &= \left(w^{(k)}\right)_m^j + \tfrac{1}{2}\left\{-\lambda a^{(k)}\left(\Delta_{m+\frac{1}{2}} + \Delta_{m-\frac{1}{2}}\right) + \lambda\left|a^{(k)}\right|\left(\Delta_{m+\frac{1}{2}} - \Delta_{m-\frac{1}{2}}\right)\right\}\left(w^{(k)}\right)^j \\ &= \frac{\lambda\left|a^{(k)}\right| + \lambda a^{(k)}}{2}\left(w^{(k)}\right)_{m-1}^j + \left(1 - \lambda\left|a^{(k)}\right|\right)\left(w^{(k)}\right)_m^j + \frac{\lambda\left|a^{(k)}\right| - \lambda a^{(k)}}{2}\left(w^{(k)}\right)_{m+1}^j. \end{aligned} \tag{106}$$

The modulus bars in the coefficient of the stabilizing term provide the switch by which the inappropriate net-point is eliminated.

If all linear normal difference equations satisfy the monotonicity condition with respect to the components of w, then the difference scheme for the original linear HSCL will satisfy the monotonicity condition with

respect to the components of w. If all normal difference equations are of the form (106), hence if $q^{(k)} = \sigma^{(k)}$ for all values of k, we arrive at the "best" scheme for a linear HSCL, that is: the most accurate scheme which still satisfies the monotonicity condition. This is just the (preferable) linear scheme with $Q = \lambda\tilde{A}$, and its simplest nonlinear form is the preferable basic scheme (I). It is gratifying to see that the unique position, attributed to scheme (I) in Sec. 3.4, is confirmed by the behaviour of this scheme in the matter of monotonicity.

When applied to a linear HSCL, the schemes (0) and (II) correspond to a similar interpolation procedure as scheme (I). This is illustrated in Fig. 9. As for scheme (0), the k-th normal equation may be written as

$$\left(w^{(k)}\right)_m^{j+1} = \tfrac{1}{2}\left(1 + \lambda a^{(k)}\right)\left(w^{(k)}\right)_{m-1}^{j} + \tfrac{1}{2}\left(1 - \lambda a^{(k)}\right)\left(w^{(k)}\right)_{m+1}^{j}, \tag{107}$$

which again represents a linear interpolation in the point $x_m - a^{(k)}\Delta t$, but now between the values of $w^{(k)}$ in (t^j, x_{m+1}) and (t^j, x_{m-1}). For scheme (II), the k-th normal equation becomes

$$\begin{aligned}\left(w^{(k)}\right)_m^{j+1} = {} & \tfrac{1}{2}\lambda a^{(k)}\left(1 + \lambda a^{(k)}\right)\left(w^{(k)}\right)_{m-1}^{j} + \left\{1 - \left(\lambda a^{(k)}\right)^2\right\}\left(w^{(k)}\right)_m^{j} \\ & - \tfrac{1}{2}\lambda a^{(k)}\left(1 - \lambda a^{(k)}\right)\left(w^{(k)}\right)_{m+1}^{j}.\end{aligned} \tag{108}$$

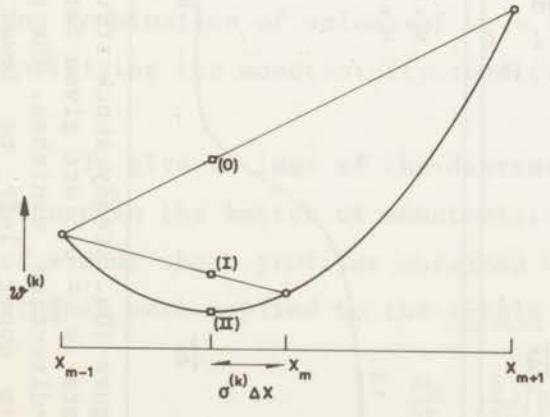


Figure 9. Difference schemes as interpolation procedures. Circles indicate net-point values of $w^{(k)}$ at t^j. The characteristic speed $a^{(k)}$ is taken negative. The squares indicate the values of $w^{(k)}$ which would arise in (t^{j+1}, x_m) upon application of schemes (0), (I) and (II). Square (II) lies on the parabola through the three circles. Scheme (II) clearly does not satisfy the monotonicity condition.

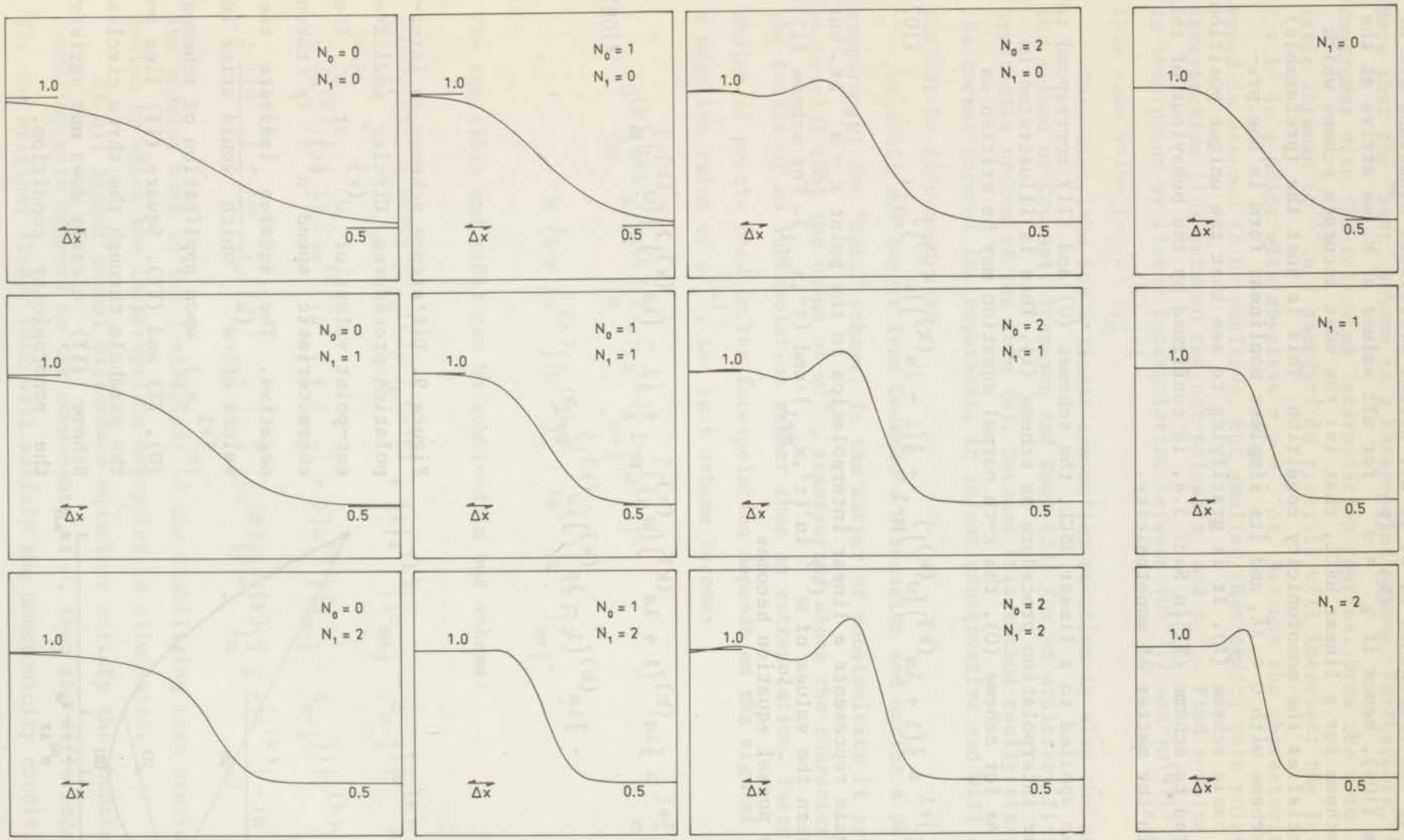


Figure 10a. Steady shock structures connecting the initial state $w_{-\infty} = 1$ with the final state $w_{+\infty} = 0.5$ are given for the schemes represented in Fig. 8. The distinction in the N_2-direction vanishes. The global Courant number σ^J equals $\frac{2}{3}$. Dissipation decreases from upper left to lower right.

Figure 10b. Same for $\sigma^J = 1$. Here also the distinction in the N_0-direction vanishes.

Figure 10. An atlas of shock profiles for the simplest nonlinear conservation law (109).

The interpolation involves all three mesh points at t^j, and is quadratic. There are no other ways of exactly linear or quadratic interpolation than the three represented by eqs. (107), (106) and (108), which confirms that (0), (I) and (II) indeed are the key schemes. Note that for schemes (0) and (II), the choice of points used in the interpolation is independent of the sign of $a^{(k)}$. Note further that in the schemes (0) and (I) the coefficients of the net-point values of $w^{(k)}$ are never negative, whereas in scheme (II) either $\left(w^{(k)}\right)^j_{m-1}$ or $\left(w^{(k)}\right)^j_{m+1}$ has a negative weight. Godunov proved that an explicit two-level difference scheme for eq. (50) will satisfy the monotonicity condition if and only if all coefficients of the net-point values employed are non-negative.

Obviously, the preservation of monotonicity with respect to a linear HSCL does not guarantee that a scheme will also preserve monotonicity for a nonlinear HSCL. To what extent the linear theory is valid in the nonlinear case can only be determined by numerical experiments. Fortunately, it appears that there is a fair, systematic agreement between the linear and nonlinear results.

For the schemes generated by eq. (94), the preservation of monotonicity is generally a local property, since condition (103) involves the values of σ^j_m and σ^j. If for just one component of w (say $w^{(k)}$) the monotonicity condition is not satisfied, due to an unfortunate combination of values of σ^j, σ^j_m and $\left(\sigma^{(k)}\right)^j_m$, it will not be satisfied for any component of w. Only for the schemes given by eq. (91), with $\nu_0 = \nu_1 = \nu_2 = \nu$, condition (103) does not vary from place to place. It simply restricts the choice of ν to $0 \leq \nu \leq 1$. This is the complement of the range of ν for which there is a virtual danger of nonlinear instabilities. This result indicates a clear correlation between numerical oscillations and nonlinear instabilities. It is further seen that the monotonicity condition is globally satisfied for any combination of values of $\nu_0, \nu_1, \nu_2 \leq 1$. In Fig. 8 the cubes of schemes satisfying the monotonicity condition are lightly shaded.

To give an idea of the distinct ways in which difference schemes may behave in the matter of monotonicity, we present in Fig. (10) a collection of steady shock profiles obtained with the schemes of eq. (101). The schemes were applied to the single conservation law

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2}w^2\right) = 0, \qquad (109)$$

where, generally, the possibilities shown in Fig. 8 reduce to 9 only, because the distinction in the N_2-direction vanishes. The initial distribution was a step function, with $w_{-\infty} = 1$ and $w_{+\infty} = \frac{1}{2}$. For the above nonlinear equation, a step function with $w_{-\infty} > w_{+\infty}$ represents a true compression shock. This shock will start running at a speed

$$U = \frac{\frac{1}{2}w_{-\infty}^2 - \frac{1}{2}w_{+\infty}^2}{w_{-\infty} - w_{+\infty}} = \frac{1}{2}(w_{-\infty} + w_{+\infty}), \tag{110}$$

which here has the value $\frac{3}{4}$.

In determining the global Courant number σ^j, we disregarded any values of $a_m^j = w_m^j$ greater than $w_{-\infty}$. The number was on this occasion defined as $\sigma^j \equiv \sigma_{-\infty}^j = \lambda^j w_{-\infty}$. This made it easier to compare the schemes that satisfy the monotonicity condition with schemes that don't. We took $\sigma^j = \frac{2}{3}$ in Fig. 10a and $\sigma^j = 1$ in Fig. 10b; for the latter value the schemes with different N_0 coincide, leaving only three possibilities. Neither value of σ^j lead to nonlinear instabilities of the first kind (see Sec. 3.3). With $\sigma^j = \frac{2}{3}$, the shock structure should move exactly one mesh in two time-steps; this provides an accurate check on the convergence of the numerical shock speed to the theoretical value. With $\sigma^j = 1$, the shock speed has to be $3\Delta x/4\Delta t$. For both values of σ^j the computations were stopped when the shock had moved $45\Delta x$. At that time the shock structures were constant in 2-4 decimals, which indicates a similar accuracy of the shock speed. Any desired accuracy could have been achieved by continuing the computations (Lax {6}).

For the simplest nonlinear conservation law (109), the choice $q_{m+\frac{1}{2}}^j = \lambda^j |w|_{m+\frac{1}{2}}^j$ exactly yields the "best" basic scheme in the sense of Godunov. This follows directly from requiring (102) with respect to eq. (109); the cases $w_{-\infty} > w_{+\infty} > 0$, $w_{-\infty} > 0 > w_{+\infty}$ and $0 < w_{-\infty} < w_{+\infty}$ must be treated separately. Indeed the shock profile for $N_0 = N_1 = 1$ exhibits no overshoot, for either value of σ^j. The overshoot for $N_0 = 1$, $N_1 = 2$ (a choice which does not satisfy the monotonicity condition) is very small for $\sigma^j = \frac{2}{3}$ (max w = 1.002) but not absent. The dependence of the shock profiles on σ^j becomes clear from a comparison between Figs. 10a and 10b; it is weakest for $N_0 = 1$. This result was anticipated in Sec. 3.4.

Godunov's way of extending the "best" linear scheme into the nonlinear case differs considerably from ours. In his derivation, the net-point values of w at the level t^j are taken to represent the step-function

$$w(t = t^j, x) = w_m^j \quad \text{for} \quad x_m - \tfrac{1}{2}\Delta x < x < x_m + \tfrac{1}{2}\Delta x. \tag{111}$$

If we were to determine the exact weak solution with these initial values, we would first have to resolve the discontinuity at the mid-point of each mesh. This notoriously complicated problem is called the Riemann initial-value problem (see Lax {14}). An arbitrary discontinuity in $x_{m+\frac{1}{2}}$ will break up into a system of r shocks, expansion waves and contact discontinuities. The latter are discontinuities occurring solely in the Riemann invariant of a linearly degenerate characteristic field; across such a discontinuity the degenerate characteristic speed is continuous.

After a time-lapse roughly equal to $\frac{1}{2}\Delta x/a^j_{m+\frac{1}{2}}$, the waves proceeding from midpoints of neighbouring meshes will start to interfere with each other. Continuation of the exact solution becomes increasingly complicated. In Godunov's difference scheme, the break up of each initial discontinuity is calculated exactly; the value of f arising in $x_{m+\frac{1}{2}}$ immediately after the break up is then employed for $F_{m+\frac{1}{2}}$.

It is easily checked that Godunov's scheme is a preferable scheme. Upon linearization of the HSCL, all waves become linear and the value of $w^{(k)}$ used in $F^j_{m+\frac{1}{2}}$ is either $\left(w^{(k)}\right)^j_m$ if $a^{(k)} > 0$ or $\left(w^{(k)}\right)^j_{m+1}$ if $a^{(k)} < 0$. This shows that the linear version of Godunov's scheme is indeed the "best" scheme (106). At the same time it means that Godunov's scheme satisfies conditions {a}-{d} from Sec. 3.1. The scheme further satisfies the symmetry requirement {e}, as $F_{m+\frac{1}{2}}$ is based on the exact solution of the Riemann problem defined by w^j_m and w^j_{m+1}. Requirement {f} is fulfilled because $F^j_{m+\frac{1}{2}}$ does not contain λ. Finally, {g} is fulfilled because of the correct treatment of linear degenerations in solving the Riemann problem.

Because Godunov's scheme is preferable, it is likely to differ only an amount $O\left((\Delta x)^3\right)$ from its principal part, scheme (I). This is indeed the case, as may be understood in the following way. The role of a characteristic speed, in the "best" linear interpolation procedure (106) is played in Godunov's scheme by either a shock speed or the mean slope of a rarefaction fan (weighted appropriately) or the speed of a contact discontinuity. Within the margin $O\left((\Delta x)^2\right)$, each of these wave speeds equals the

arithmetic mean of the proper characteristic speeds in the neighbouring net-points; cf. Lax {14}. In scheme (I) only this mean value is retained; this truncation causes the difference $O\left((\Delta x)^3\right)$ between the two schemes.

Evidently, the simpler scheme (I), which virtually gives the same numerical results (see Sec. 6.1), has great practical advantages over Godunov's original method. With this, the question posed in the last paragraph of Sec. 2.3 is finally answered.

4. EXECUTION AND REFINEMENT

4.1 Two-step formulation

From now on we shall concentrate on the case that A has three distinct eigenvalues. This particularly holds for ideal compressible flow, cf. Ch. 5. With $r = 3$, the recipe (63) for an admissible stabilization matrix involves no higher power of A than A^2. All basic schemes then allow of a two-step formulation. This means that to any basic scheme an admissible two-step scheme can be assigned whose principal part is the given basic scheme, the difference between two-step and basic scheme being only $O\big((\Delta x)^3\big)$. The computational advantage of the two-step formulation is that it involves no matrix multiplication. There are innumerable ways to construct two-step schemes starting from one single basic scheme (see e.g. Van Leer {19}); one possibility is given below.

In the first step, provisional values w are calculated by

$$w^{*}_{m+\frac{1}{2}} = w^{j}_{m+\frac{1}{2}} - \tfrac{1}{2}\lambda^{j}(\kappa_2)^{j}_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^{j}. \tag{112}$$

The subscript $m+\frac{1}{2}$ of w does not denote an average in the sense of eq. (24) but serves to indicate the proper mesh, in the sense of eq. (21). The second step is the actual difference scheme, with

$$F^{j}_{m+\frac{1}{2}} = f\big(w^{*}_{m+\frac{1}{2}}\big) - \frac{1}{2\lambda^{j}}\left\{(\kappa_0)^{j}_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}w^{j} + \lambda^{j}(\kappa_1)^{j}_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^{j}\right\}. \tag{113}$$

This yields the intended approximation of the basic scheme, because

$$f\big(w^{*}_{m+\frac{1}{2}}\big) = f^{j}_{m+\frac{1}{2}} - \tfrac{1}{2}\lambda^{j}(\kappa_2)^{j}_{m+\frac{1}{2}}A^{j}_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}}f^{j} + O\big((\Delta x)^2\big). \tag{114}$$

The finite difference with the coefficient κ_1 can equally well be accounted for in the first step, or be divided over the two steps. The linear stability condition for the two-step scheme is of course the same as for its principal part.

Suppose that the basic scheme is preferable; it may then occur that $\Delta_{m+\frac{1}{2}}f^{j} = 0$, $\Delta_{m+\frac{1}{2}}w^{j} \neq 0$, $(\kappa_0)^{j}_{m} = (\kappa_0)^{j}_{m+1} = 0$. The two-step scheme now has $F^{j}_{m+\frac{1}{2}} = f\big(w^{j}_{m+\frac{1}{2}}\big)$, which only equals $f^{j}_{m+\frac{1}{2}}$ if under these circumstances f is linear in w. In order to guarantee that the two-step version of a pre-

ferable basic scheme is again preferable, we must add the correction term

$$f^j_{m+\frac{1}{2}} - f\left(w^j_{m+\frac{1}{2}}\right) \tag{115}$$

to the right-hand side of eq. (113).

Consider the two-step versions of the schemes (0), (I) and (II), including the correction term (115). Scheme (0) is identical to its two-step version; the same is true for all other schemes with $Q \propto I$. The second-order accuracy of scheme (II) is clearly reflected in the fact that for this scheme the first step (112) is identical to scheme (0), with mesh widths reduced to $\Delta x/2$ and $\Delta t/2$. Hence we have

$$F^j_{m+\frac{1}{2}} = f\left(w^*_{m+\frac{1}{2}}\right) = f\left(w(t^{j+\frac{1}{2}}, x_{m+\frac{1}{2}})\right) + O\left((\Delta x)^2\right), \tag{116}$$

which centers the convection term at $t^{j+\frac{1}{2}}$. Note that for the two-step formulation of scheme (II), or any scheme with $Q \propto A^2$, it is not essential that $r = 3$. This means an advantage over schemes like (I), with $Q \propto A$, which involve all powers of A up to A^{r-1}. In ideal magnetohydrodynamics, with $r = 7$, it would take six steps to approximate scheme (I)! In practice, the advantage of scheme (II) is of little value: the computation of extra terms, needed to reduce numerical oscillations and to prevent nonlinear instabilities, likewise involve all powers of A. This is explained in the next section.

4.2 Non-basic additions

Extra finite differences may be added to a basic scheme in order to effect overall smoothness of the numerical results and, in particular for a preferable scheme, to remove the danger of nonlinear instabilities (points (C) and (D) in Sec. 3.2). The coefficients of these non-basic differences must vanish for a linear HSCL (cf. condition {a} in Sec. 3.1). It is logical to fight nonlinear instabilities, arising from preferable schemes, with exclusively nonlinear terms. The alternative of resorting to a non-preferable scheme is an unnecessarily drastic remedy.

Among all (preferable) basic schemes, scheme (II) is most strongly in need of an improvement in the matter of monotonicity and stability. At the same time it is also

the most gratifying scheme to be supplemented, in view of its second-order accuracy. For simplicity we shall concentrate on this scheme in the present section.

Non-basic additions to scheme (II) should not spoil the properties {e}, {f} and {g} of the scheme. The restrictions {b}, {c} and {d} have no direct consequences for the non-basic terms in the scheme but similar restrictions may be put on these terms. We shall not attempt to give accurate definitions of "admissible" and "preferable" non-basic differences, but restrict the choice of such terms in the following ways.

— The matrix coefficients of the non-basic differences must commute, in some suitable approximation, with the mesh values of A; compare {b}.

— The eigenvalues of these non-basic matrix coefficients must depend on no other than the corresponding characteristic speeds; occurrence of the local and global maximum absolute characteristic speed is allowed; compare {c}.

— All eigenvalues must be identical expressions in the corresponding characteristic speeds; compare {d}.

A difficulty in determining the effect of purely nonlinear terms is that the usual linear analysis does not yield any answer. In practice a heuristic mixture of linear and nonlinear arguments has proved satisfactory.

The quantity which is essential in constructing the k-th eigenvalue of a non-basic coefficient matrix in $F^{j}_{m+\frac{1}{2}}$, is

$$\left(\theta^{(k)}\right)^{j}_{m+\frac{1}{2}} = \frac{\frac{1}{2}\left|\Delta_{m+\frac{1}{2}}\left(a^{(k)}\right)^{j}\right|}{\left|a^{(k)}\right|^{j}_{m+\frac{1}{2}}}, \tag{117}$$

where the subscript of $\theta^{(k)}$ should be interpreted in the sense of eq. (21). It is never negative, never exceeds unity, and equals unity when $\left(a^{(k)}\right)^{j}_{m}$ and $\left(a^{(k)}\right)^{j}_{m+1}$ have opposite signs, hence when $\left(a^{(k)}\right)^{j}_{m+\frac{1}{2}}$ is close to zero. These properties can be used to make scheme (II) sufficiently dissipative in the dangerous zone of $a^{(k)}$. To achieve this result, we must add to the

mean stabilization matrix $Q^j_{m+\frac{1}{2}}$ a matrix which commutes with $A^j_{m+\frac{1}{2}}$ within the margin $O\left((\Delta x)^2\right)$ and whose k-th eigenvalue is a positive multiple of $\left(\theta^{(k)}\right)^j_{m+\frac{1}{2}}$. For scheme (II) we have $\left(q^{(k)}\right)^j_{m+\frac{1}{2}} = \left\{\left(\sigma^{(k)}\right)^2\right\}^j_{m+\frac{1}{2}}$, and we may safely raise this eigenvalue of the basic stabilization matrix by a non-basic amount

$$\left(1 - \left\{\left(\sigma^{(k)}\right)^2\right\}^j_{m+\frac{1}{2}}\right)\left(\theta^{(k)}\right)^j_{m+\frac{1}{2}}, \tag{118}$$

without coming into conflict with the linear stability criterion. If $\theta^{(k)} = 1$, the amended scheme (II) will coincide with scheme (0). The expression (118), which represents the maximum allowed change within the scope of the CFL condition, does not vanish with λ, hence does not yield a preferable scheme. A preferable scheme results, for instance, when $\left(q^{(k)}\right)^j_{m+\frac{1}{2}}$ is increased by

$$\left\{\sigma^{(k)} - \left(\sigma^{(k)}\right)^2\right\}^j_{m+\frac{1}{2}}\left(\theta^{(k)}\right)^j_{m+\frac{1}{2}}. \tag{119}$$

For $\theta^{(k)} = 1$, the scheme will now pass into scheme (I).

In practice, any change in $\left(q^{(k)}\right)^j_{m+\frac{1}{2}}$, hence in $Q^j_{m+\frac{1}{2}}$, must be accomplished through a change in the coefficients $(\kappa_i)^j_{m+\frac{1}{2}}$. For $r = 3$, the modification can again be included in a two-step procedure.

The non-basic second differences described above, which solely affect the real part of the amplification matrix of scheme (II), were discussed by Lax and Wendroff {10}. It seems to have escaped notice that another type of modification exists, which solely affects the imaginary part of the amplification matrix. For that purpose we must add to the term $f^j_{m+\frac{1}{2}}$ in $F^j_{m+\frac{1}{2}}$ some matrix times $w^j_{m+\frac{1}{2}}$; this matrix again has to satisfy the three requirements previously given. Furthermore, it must have the magnitude $O\left((\Delta x)^2\right)$, hence be proportional to $\left\{\left(\theta^{(k)}\right)^j_{m+\frac{1}{2}}\right\}^2$, in order not to swamp the truncation error $O\left((\Delta x)^3\right)$ of scheme (II). It must also be required that on any mesh this matrix equals a positive or negative matrix times $A^j_{m+\frac{1}{2}}$. The resulting non-basic first difference thus affects the modulus of each amplification factor in the same direction.

For example, the imaginary part of $g^{(k)}$, which equals $-i\lambda a^{(k)} \sin\alpha$, may safely be raised by an amount

$$2i\lambda a^{(k)}\left(\theta^{(k)}\right)^2 \sin\alpha, \tag{120}$$

without affecting the stability range of λ. The imaginary axis of the ellipse traced by $g^{(k)}$ will be reduced to

$$\sigma^{(k)}|1 - 2(\theta^{(k)})^2|. \qquad (121)$$

For $\theta^{(k)} = 1$ the original ellipse results, but traced in opposite sense.

The coefficient matrix required for non-basic first differences in the scheme should again be evaluated as a polynomial in $A^j_{m+\frac{1}{2}}$, and for $r = 3$ may be included in a two-step procedure. It generally is necessary to add a correction term of type (115) in order to keep the scheme preferable.

The positive effect of non-basic second differences on the stability and smoothness of numerical results obtained with scheme (II) was for instance discussed by Emery {17}. A proper dose of non-basic first differences will certainly be helpful in making the scheme unconditionally stable within the CFL range, and may bring the scheme still closer to satisfying the monotonicity condition. It is possible to give the non-basic differences in the scheme a very large weight, as compared to the basic differences. This however will reduce the stable range of λ (cf. Lax and Wendroff {10}) and makes the scheme rather resemble a numerical smoothing routine than an algorithm for advancing in time. To discuss this falls outside the scope of the present work.

4.3 Uneven meshes

In practice it is often convenient or even inevitable to employ uneven spatial meshes. It is therefore useful to consider the case that $\Delta_{m-\frac{1}{2}}x \neq \Delta_{m+\frac{1}{2}}x$. We shall try to reformulate the schemes (0), (I) and (II), that is, to find schemes which have the same major properties as (0), (I) and (II), and which reduce to (0), (I) and (II) for a uniform net.

It is obvious to start from the interpretation of these schemes as the interpolation procedures which were described in Sec. 3.5. Scheme (0) must remain a wide-base three-point scheme:

$$w_m^{j+1} = \tfrac{1}{2}\left(\frac{\Delta_{m+\frac{1}{2}}x}{\Delta_m x}\, w^j_{m-1} + \frac{\Delta_{m-\frac{1}{2}}x}{\Delta_m x}\, w^j_{m+1}\right) - \lambda^j \Delta_m f^j, \qquad (122)$$

with $\lambda_m^j = \Delta^{j+\frac{1}{2}}t/\Delta_m x$, etc.. Scheme (I) must remain a narrow-base three-point scheme:

$$w_m^{j+1} = w_m^j - \tfrac{1}{2}\left(\lambda^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}} + \lambda^j_{m-\frac{1}{2}}\Delta_{m-\frac{1}{2}}\right)f^j$$

$$+ \tfrac{1}{2}\left(\lambda^j_{m+\frac{1}{2}}A^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}} - \lambda^j_{m-\frac{1}{2}}A^j_{m-\frac{1}{2}}\Delta_{m-\frac{1}{2}}\right)w^j. \tag{123}$$

Scheme (II) is the unique second-order procedure:

$$w_m^{j+1} = w_m^j - \tfrac{1}{2}\lambda_m^j\left(\frac{\Delta_{m-\frac{1}{2}}x}{\Delta_{m+\frac{1}{2}}x}\Delta_{m+\frac{1}{2}} + \frac{\Delta_{m+\frac{1}{2}}x}{\Delta_{m-\frac{1}{2}}x}\Delta_{m-\frac{1}{2}}\right)f^j$$

$$+ \tfrac{1}{2}\lambda_m^j\left(\lambda^j_{m+\frac{1}{2}}A^j_{m+\frac{1}{2}}\Delta_{m+\frac{1}{2}} - \lambda^j_{m-\frac{1}{2}}A^j_{m-\frac{1}{2}}\Delta_{m-\frac{1}{2}}\right)f^j. \tag{124}$$

Because of the uneven spacing, the usual Fourier analysis presented in Sec. 3.3 can not be reproduced. A practical stability condition for these schemes is the local CFL condition

$$\max_m \lambda^j_{m+\frac{1}{2}}a^j_{m+\frac{1}{2}} \leq 1. \tag{125}$$

For a rigorous proof of the stability of (124) in a simple mesh-refinement problem we refer to Ciment {24}.

It must be stressed that the above versions of (0), (I) and (II) are in general <u>not</u> conservative. This may readily be understood by realizing that the mean mesh width $\Delta_m x$ and the mesh ratio $\Delta_{m-\frac{1}{2}}x/\Delta_{m+\frac{1}{2}}x$ are quantities assigned to the m-th nodal point, and not to any of the adjacent meshes. It is therefore not possible to write any of the schemes as

$$\frac{\Delta^{j+\frac{1}{2}}w_m}{\Delta^{j+\frac{1}{2}}t} + \frac{\Delta_m F^j}{\Delta_m X} = 0. \tag{126}$$

where $X_{m+\frac{1}{2}} \equiv X(x_m, x_{m+1})$ is a function which for a uniform net reduces to $x_{m+\frac{1}{2}}$, and $F^j_{m+\frac{1}{2}} \equiv F(w^j_m, w^j_{m+1}; x_m, x_{m+1})$ depends explicitly on the space coordinates. This can only be done if the mesh ratio does not vary from point to point. Source terms will generally arise near any point where the mesh-width does not progress geometrically. How seriously this will affect the accuracy of a numerical solution containing shocks depends on the problem

considered. In the common situation of a non-geometric progression restricted to a few refinement points in a piecewise uniform net, damage most likely will be small.

It should be mentioned that scheme (I) is the only one of the three main schemes which for uneven meshes may sometimes be conservative. For example, scheme (123) is conservative with respect to the single quadratic conservation law (109) (provided that w does not change its sign), while schemes (122) and (124) are not even conservative for a single linear conservation law. This is due to the fact that, when applied to eq. (109), scheme (I) involves only one spatial mesh at t^j. Another consequence of this fact is that scheme (I) requires no modification near a spatial boundary. To what extent these advantages are preserved with respect to an arbitrary HSCL remains to be investigated.

5. FLUID-DYNAMICAL INTERPRETATION

5.1 The equations of ideal compressible flow

The Euler equations of one-dimensional, adiabatic ICF read, in divergence-free form:

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ u \\ E \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} u \\ p + \frac{u^2}{\rho} \\ \frac{u}{\rho}(E + p) \end{pmatrix} = 0. \tag{127}$$

Here ρ, u and E denote mass, momentum and total energy per unit volume of fluid; p denotes the pressure in the fluid. If the internal energy per unit volume is called e, and the momentum, internal energy and total energy per unit mass are called u, e and E, the following relations exist

$$u = \rho u, \tag{128}$$

$$E = e + \tfrac{1}{2}u^2/\rho = \rho(e + \tfrac{1}{2}u^2) = \rho E. \tag{129}$$

The pressure is primarily regarded as a function of ρ and e:

$$p \equiv p(\rho, e) \equiv p\left(\rho, \frac{1}{\rho}(E - \tfrac{1}{2}\frac{u^2}{\rho})\right) \equiv p(\rho, u, E), \tag{130}$$

which shows that eq. (127) is indeed a HSCL of the form (1). The auxiliary quantities

$$h = e + p/\rho, \tag{131}$$

$$H = e + p/\rho + \tfrac{1}{2}u^2 = h + \tfrac{1}{2}u^2 = E + p/\rho \tag{132}$$

are called, respectively, the specific enthalpy and the stream function. The corresponding quantities per unit volume are

$$h = \rho h = e + p, \tag{133}$$

$$H = \rho H = h + \tfrac{1}{2}\frac{u^2}{\rho} = E + p. \tag{134}$$

For a fruitful discussion of the equations of ICF it is necessary to mention the specific entropy s, defined by the differential expression

$$Tds = de - \frac{p}{\rho^2}d\rho = e_p\{dp - (p_\rho + \frac{pp_e}{\rho^2})d\rho\}, \tag{135}$$

where T denotes the absolute temperature. From (135) it follows that the isentropic sound speed c, defined as the square root of $(\partial p(\rho,s)/\partial\rho)_s$, equals

$$c = \sqrt{p_\rho + \frac{pp_e}{\rho^2}}\ . \tag{136}$$

The frequently occurring dimensionless quantity p_e/ρ characterizes the difference between isentropic and isoenergetic sound speed:

$$\frac{p_e}{\rho} = \left(\frac{\partial \ln p(\rho,s)}{\partial \ln \rho}\right)_s - \left(\frac{\partial \ln p(\rho,e)}{\partial \ln \rho}\right)_e. \tag{137}$$

The Jacobian A_{Eu} arising from the Eulerian equations can be written as

$$A_{Eu} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 - \frac{p_e}{\rho}(H - u^2) & (2 - \frac{p_e}{\rho})u & \frac{p_e}{\rho} \\ -u\left(H - c^2 + \frac{p_e}{\rho}(H - u^2)\right) & H - \frac{p_e}{\rho}u^2 & (1 + \frac{p_e}{\rho})u \end{pmatrix}, \tag{138}$$

which looks fairly complicated. The evaluation of $A\Delta f$ may be somewhat simplified by recognizing that

$$df^{(3)}_{Eu} = u dH + H du, \tag{139}$$

so that $A_{Eu}\,df_{Eu} = A'_{Eu}\,df'_{Eu}$, with

$$A'_{Eu} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - (1 - \frac{p_e}{\rho})u^2 & (2 - \frac{p_e}{\rho})u & p_e u \\ u(c^2 + \frac{p_e}{\rho}u^2) & H - \frac{p_e}{\rho}u^2 & (1 + \frac{p_e}{\rho})\rho u^2 \end{pmatrix} \tag{140}$$

and

$$f'_{Eu} = \begin{pmatrix} u \\ p + \frac{u^2}{\rho} \\ H \end{pmatrix} \tag{141}$$

We hence have

$$\left(A'_{Eu}\right)^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} \left(f'_{Eu}\right)^j_{m+\frac{1}{2}} = \left(A_{Eu}\right)^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} f^j_{Eu} + O\left((\Delta x)^3\right), \tag{142}$$

and the left member is sufficiently accurate to be used in a basic scheme. When the gas considered is ideal, with $2/(\gamma-1)$ internal degrees of freedom, the pressure is given by

$$p = (\gamma - 1)\, \rho e \ , \tag{143}$$

From this it follows that

$$p_e/\rho = \gamma - 1 \ , \tag{144}$$

$$c^2 = \gamma \frac{p}{\rho} \ , \tag{145}$$

$$h = \frac{\gamma}{\gamma-1} \frac{p}{\rho} \ , \tag{146}$$

$$s \propto \ln(pV^\gamma) \ , \tag{147}$$

which considerably simplifies the computations. Still, the two-step formulation of basic schemes given in Sec. 4.1 must be preferred. In the work of Burstein {16} the full matrix multiplication was still carried out, in a two-dimensional version of scheme (II), i.e. even with 4×4 matrices.

The eigenvalues of A_{Eu} are $u-c$, u and $u+c$; the characteristics with speed u are the streamlines of the fluid. These are linearly degenerate: the Riemann invariant conserved along a streamline is the specific entropy, hence a thermodynamic quantity not depending on u. If $\partial u/\partial x$ and $\partial p/\partial x$ vanish, the HSCL (127) reduces to the single equation

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} = 0 \ , \tag{148}$$

in which u is now a constant. If $u = 0$, the prescribed values of ρ (and e,

s, T, etc.) will not change in the course of time. This is the only type of stationary solution admitted by eq. (127), apart from the standing shock discussed in Sec. 3.1.

For more-dimensional ICF, eq. (2) represents only the flow along the x axis. Though the number of equations is then greater than three, the number of distinct characteristic speeds, hence r, remains three. What merely happens is that the eigenvalue u occurs two or three times. This is fortunate, because it admits the use of two-step methods, with exactly the same coefficients κ_i as for one-dimensional ICF.

The Lagrange equations of one-dimensional adiabatic ICF follow from (127) through a transformation from the space coordinate x to the mass coordinate x of a slab of fluid. This quantity denotes the (constant) mass, in a column of unit cross-section, between the particular slab and some reference slab. The mass coordinate hence labels the slab, whose streamline may now be given as $x \equiv x(x,t)$. We have

$$dx = \rho dx \quad , \tag{149}$$

$$\left(\frac{\partial}{\partial t}\right)_x = \left(\frac{\partial}{\partial t}\right)_x + \left(\frac{\partial x}{\partial t}\right)_x \left(\frac{\partial}{\partial x}\right)_t \quad , \tag{150}$$

$$u = \left(\frac{\partial x}{\partial t}\right)_x \quad . \tag{151}$$

Inserting these relations into (127) and changing the state variables from quantities per unit volume into quantities per unit mass (exchanging the density ρ for the specific volume $V = 1/\rho$, so that p becomes $p(V,e)$), we arrive at the Lagrangean equations

$$\frac{\partial}{\partial t}\begin{pmatrix} V \\ u \\ E \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} -u \\ p \\ up \end{pmatrix} = 0 \quad . \tag{152}$$

These are again of the form (1).

The Jacobian A_{La} arising from the above equations clearly has a zero eigenvalue, since $f_{La}^{(3)}$ is the product of $f_{La}^{(1)}$ and $f_{La}^{(2)}$. The eigenvalues of A_{La} appear to be $-c$, 0 and $+c$, where c is the Lagrangean sound speed or acoustic impedance defined as the square root of $-\left(\partial p(V,s)/\partial V\right)_s$. It equals

$$c = \rho c = \sqrt{p p_e - p_V} \tag{153}$$

hence gives the mass per column fluid of unit cross-section, traversed by a sound wave in unit time. The matrix A_{La} is found to be

$$A_{La} = \begin{pmatrix} 0 & -1 & 0 \\ p_V & -p_e u & p_e \\ p_V u & p - p_e u^2 & p_e u \end{pmatrix}. \tag{154}$$

It follows that $A_{La}\, df_{La}$ can be replaced by

$$A'_{La}\, df'_{La} \equiv \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \\ -c^2 u & p \end{pmatrix} d \begin{pmatrix} -u \\ p \end{pmatrix}, \tag{155}$$

This was first shown by Lax and Wendroff {10}. If in a basic scheme for the Lagrangean equations a finite difference approximation is included of $A'_{La}\, df'_{La}$ instead of $A_{La}\, df_{La}$, the scheme can compete with its two-step version, as far as computing time and memory are concerned.

The zero eigenvalue of A_{La} corresponds to the eigenvalue u of A_{Eu}: in the Lagrangean coordinate system, the lines of constant x are also the streamlines of the fluid. If u and p exhibit no gradients, the solution of (148) is stationary, regardless of the value of u. There are no solutions with stationary shocks, because the signs of the non-degenerate characteristic speeds are fixed. It appears in practice that the ever-vanishing central characteristic speed does not lead to nonlinear instabilities of the second kind (cf. Sec. 3.3). Presumably this is due to the fact that streamlines are linearly degenerate. This argument however fails with respect to the Euler equations. Burstein {16}, when calculating the stagnation of supersonic flow by a blunt body, experienced nonlinear instabilities near the stagnation point. In this point the flow undergoes a significant change, while u vanishes. Such a singular situation is never encountered in Lagrangean flow problems, which are always one-dimensional.

The exceptional symmetry of the set of eigenvalues of A_{La} makes that the matrix A_{La} needed in scheme (I) becomes a very simple expression in A_{La}. For we have

$$A_{La} = \begin{pmatrix} -c & & \emptyset \\ & 0 & \\ \emptyset & & +c \end{pmatrix} = c \begin{pmatrix} -1 & & \emptyset \\ & 0 & \\ \emptyset & & +1 \end{pmatrix} = c\, J \quad , \tag{156}$$

hence

$$A_{La}^2 = \begin{pmatrix} c^2 & & \emptyset \\ & 0 & \\ \emptyset & & c^2 \end{pmatrix} = c \begin{pmatrix} c & & \emptyset \\ & 0 & \\ \emptyset & & c \end{pmatrix} = c\, \tilde{A}_{La} \quad , \tag{157}$$

and correspondingly:

$$\tilde{A}_{La} = \frac{1}{c} A_{La}^2 \quad . \tag{158}$$

When expanding λA_{La} in powers of λA we therefore have[3]

$$\kappa_0 = \kappa_1 = 0 \ , \quad \kappa_2 = 1/\lambda c \ . \tag{159}$$

Likewise, the expansion of non-basic coefficient matrices (see Sec. 4.2) involves only one term. For instance, the matrix $\Theta^j_{m+\frac{1}{2}}$ commuting with $A^j_{m+\frac{1}{2}}$, with eigenvalues $\left(\theta^{(k)}\right)^j_{m+\frac{1}{2}}$, k=1,...,n, can be written as

$$\Theta^j_{m+\frac{1}{2}} = \frac{\frac{1}{2}\left|\Delta_{m+\frac{1}{2}} c^j\right|}{c^j_{m+\frac{1}{2}}} \left\{ \left(\frac{A_{La}}{c} \right)^2 \right\}^j_{m+\frac{1}{2}} \tag{160}$$

It follows from eq. (158) that, for the Lagrangean equations, the schemes (I) and (II) demand roughly the same amount of computing time, no matter whether these schemes are carried out in the formulation based on eq. (155) or in the two-step formulation. For the Euler equations the situation is different. In order to evaluate $\lambda\tilde{A}$, all three powers of λA are needed. The three coefficients can be expressed in λc and the Mach number

---

[3] Note that the dimension of the Lagrangean mesh ratio $\lambda = \Delta t/\Delta x$ is velocity^{-1} × density^{-1}.

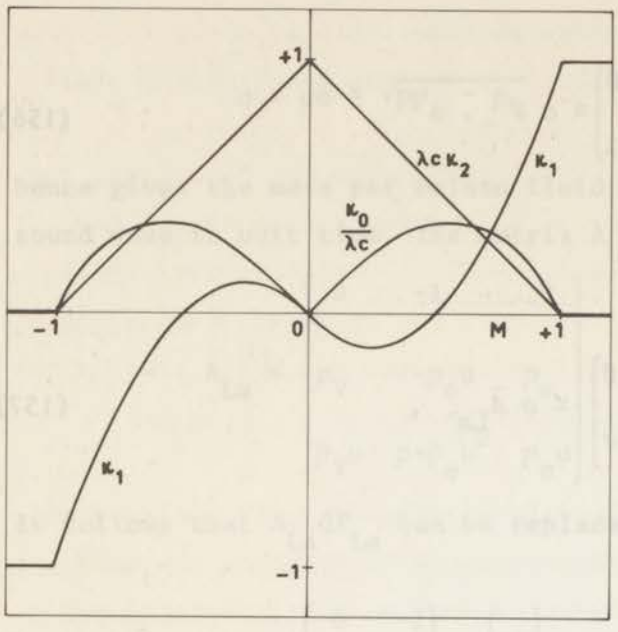


Figure 11. Dependence of the coefficients in the polynomial for $\lambda\tilde{A}_{La}$ on the Mach number M.

$$M = \frac{u}{c} . \tag{161}$$

They are found to be

$$\left.\begin{aligned} &\kappa_0 = \kappa_2 = 0 \ , \quad \kappa_1 = \operatorname{sgn} M \quad \text{for} \quad |M| > 1 \ ; \\ &\kappa_0 = \lambda c \, |M| \, (1 - M^2) \\ &\kappa_1 = M \, (2|M| - 1) \\ &\kappa_2 = (1 - |M|)/\lambda c \end{aligned}\right\} \quad \text{for} \quad |M| \leq 1 \ . \tag{162}$$

As a check, we may confirm that for supersonic flow to the right ($M > 1$, hence $u > +c$) we have $\tilde{A}_{Eu} = A_{Eu}$, and for supersonic flow to the left ($M < -1$, hence $u < -c$) we have $\tilde{A}_{Eu} = -A_{Eu}$. For $M = 0$ we have $\tilde{A}_{Eu} = \frac{1}{c} A_{Eu}^2$, analogous to eq. (158) , since for $u = 0$ the Eulerian coordinate system coincides with the Lagrangean system. In Fig. 11 the functions $\kappa_0/\lambda c$, κ_1 and $\lambda c \kappa_2$ are plotted against M. As mentioned earlier in this section, the expressions (162) are also valid for the more-dimensional Euler equations, that is, for the part describing the flow along the x axis.

With respect to two-step methods we may yet mention that the correction term (115) is not needed for the Euler equations if $e \propto 1/\rho$, and for the Lagrange equations if e is linear in V. These conditions are satisfied by the ideal gas law (143).

5.2 Viscosity versus diffusion

The Lagrangean equations of ICF give rise to one characteristic speed that vanishes everywhere in the x-t plane. These equations, therefore, are well suited to illustrate the distinction between preferable and non-preferable schemes.

A preferable scheme does not affect the net-point values of the state variables V, u and E unless there are gradients in u and p. When no such gradients are prescribed, or once they have disappeared, any given non-uniform distribution of V, hence of e and s, will remain unchanged. This agrees with the concept of ICF, which excludes thermal conductivity. Because the stabilizing term in a preferable scheme is effective only as long as there are gradients in the momentum of the fluid, the dissipation provided by this term may be regarded as an artificial viscosity. The viscous mechanism differs slightly from the viscous pressure in the scheme of Von Neumann and Richtmyer {5}. In conservative schemes there is not only an extra pressure $\propto -\frac{\partial u}{\partial x}$, decelerating the fluid, but also an extra velocity $\propto -\frac{\partial p}{\partial x}$, expanding the fluid (Lax and Wendroff {10}). Together, these provide the conversion of kinetic into internal energy, needed in a shock.

If a Lagrangean flow problem is attacked with a non-preferable scheme, any gradients in V and e will be affected, whether or not these are accompanied by gradients in u and p. The stabilization term in such a scheme provides not only momentum diffusion (viscosity) but, independently, also energy diffusion (heat conduction) and even mass diffusion (which physically doesn't make sense at all). We shall refer to this collection of dissipative effects by the term artificial diffusion (see e.g. Fox {2, Sec. 27.19}).

The inconvenience of stabilization by artificial diffusion, compared to artificial viscosity, was already noticed by Lax {6}. It is most clear in the case of a contact discontinuity, i.e. a sudden jump in s (and V and e) not accompanied by any jump in p and u. We shall assume that the ideal gas law (143) holds, and take one of the difference schemes from Sec. 3.4 with $N_2 = 0$, hence $Q = \kappa_0 I$. When such a non-preferable difference scheme is applied, V changes in the following way:

$$\Delta^{j+\frac{1}{2}} V_m = \tfrac{1}{2}\left\{\left(\kappa_0\right)^j_{m+\frac{1}{2}} \Delta_{m+\frac{1}{2}} - \left(\kappa_0\right)^j_{m-\frac{1}{2}} \Delta_{m-\frac{1}{2}}\right\} V^j \,. \qquad (163)$$

The change in e is exactly proportional to the change in V, so that p and u

are not affected. Note that (163) is a finite-difference approximation of the diffusion equation

$$\frac{\partial V}{\partial t} = \frac{\partial}{\partial x}\left(\frac{(\Delta x)^2}{2\Delta t} \kappa_0 \frac{\partial V}{\partial x} \right) . \qquad (164)$$

Accordingly, the initial discontinuity will spread so that the maximum slope in the resulting structure falls off in time as $1/\sqrt{j}$, where j denotes the number of time-steps (Lax {6}). When Δt and Δx go to zero at constant ratio λ (the usual procedure to investigate whether a scheme for eq. (2) is consistent), the diffusion coefficient in (164) will vanish too. However, this theoretical argument has little value in computational practice. During the first 100 time-steps, a contact discontinuity may easily spread to an equivalent width[4] of the order of 10 meshes. This result is very disappointing, as it completely defeats the purpose of dissipative schemes, viz. to represent a shock as a transition covering roughly two meshes. Yet it is a clearly understandable result. Due to the tendency of the characteristics to break, the width of a genuine shock in dissipative approximation remains limited to a few meshes. But a discontinuity in a field with parallel characteristics exhibits no inflow of waves to balance the dissipative spread. Such a discontinuity will gradually decay, because there is no mechanism to keep it upright.

Artificial diffusion does not just destroy contact discontinuities, but generally tries to convert any distribution of V into a linear one. The usual result is that, outside shocks, the flow tends to become isentropic. At the least, this yields a quantitative error in the numerical solution. A mesh refinement may be needed to bring the desired accuracy and fineness of detail.

More serious are those cases where the physical situation is unstable under the influence of an artificially large diffusion. The numerical errors may then lead to a qualitative change of the solution. This may not always be recognized, so that a mesh refinement will not even be tried. A large variety of unstable situations can be designed, especially when extra thermodynamic or other processes are introduced into the equations in the

---

[4] The equivalent width of a transition curve from $V_{-\infty}$ to $V_{+\infty}$ may be defined as $\int_{-\infty}^{+\infty} \left|\frac{\partial V}{\partial x}\right| dx \; / \left|\frac{\partial V}{\partial x}\right|_{max}$. Due to the modulus bars in the integration, an oscillatory region following a steep transition will increase the equivalent width. For a monotonic transition it simply equals $|V_{-\infty} - V_{+\infty}| / \left|\frac{\partial V}{\partial x}\right|_{max}$.

form of source terms. An example of a thermal instability is discussed in Sec. 6.2.

Altogether, it may be concluded that stabilization by artificial viscosity should be preferred to stabilization by artificial diffusion, in view of the undesired effects of mass and heat diffusion on the numerical results. This argument weighs the more heavily if we realize that, in artificial diffusion, essentially the artificial viscosity is responsible for the heating in a shock. The combination of heat and mass diffusion does not yield any friction, at least not when the ideal gas law is observed.

On the other hand, a less welcome effect of artificial viscosity is that rarefaction waves are distorted. Since the fluid is too viscous, the sharp head and foot of a rarefaction fan tend to be smoothed out. This type of truncation error is commonly regarded as less inconvenient than the errors caused by artificial diffusion. Viscosity tries to eliminate the second derivatives of velocity and pressure; this effect is relatively unimportant, since the nature of the flow itself is to eliminate the first derivatives.

Nevertheless, extreme care should be taken in supplementing the Lagrangean equations with source terms (and extra equations) representing microscopic physical processes that provide heating. The spurious numerical viscosity - which falsifies the ratio of kinetic and internal energy - may come into conflict with a mechanism based on genuine molecular properties. The combination may then lead to physically irrelevant solutions. An example connected with the propagation of radiative ionization fronts is discussed by Goldsworthy {4}.

It is remarkable that, although a dissipative difference scheme is apparently best for treating compressive flow, and the method of characteristics for expansive flow, no numerical techniques have been invented to combine the specific benefits of these methods. The tendency rather is in the opposite direction. In Gary {25}, the smooth part of the flow is computed with the dissipative scheme (II); a shock is regarded as a discontinuity and accounted for by a shock fitting technique.

With respect to the Lagrangean equations we may say that preferable schemes, running on pure viscosity, are indeed preferable to other admissible schemes, running on diffusion. Such a strong statement cannot be made about the Euler equations. If $\frac{\partial u}{\partial x}$ and $\frac{\partial p}{\partial x}$ vanish but $u \neq 0$, so that the ρ distribution moves through the computational net, there is diffusion even in a preferable scheme. With use of the ideal gas law, the finite-difference

versions of the three Euler equations reduce to one and the same scheme for eq. (148). The streamlines, which in the Lagrangean equations have been solved in advance, must now be traced by means of some interpolation routine (see Sec. 3.5). For all schemes, the interpolation error causes spreading of an initially steep transition. The random walk of information across streamlines (i.e. the "confusion" of streamlines) is again a diffusion process. The width of the transition will generally grow $\propto \sqrt{j}$. In the one exception of scheme (II) the spreading would be $\propto \sqrt[4]{j}$; this diffusive error however is swamped by the convective error, which causes an oscillatory region, growing $\propto \sqrt[3]{j}$.

It may be concluded that one-dimensional problems of ICF should be treated, whenever possible, in the Lagrangean formulation, in combination with a preferable scheme. For more-dimensional problems, especially of transient flow, the use of the Euler equations is practically inevitable. The problem to obtain a sufficiently detailed numerical solution with the methods considered in this paper becomes largely a matter of mesh size and, therefore, of computing time. Any mesh refinement will appear in the number of operations in a power equal to the number of independent variables, including t!

It is advisable never to trust a single integration of an initial-value problem, and always to compare numerical solutions obtained with different mesh widths but equal mesh ratios. This may give an idea of the direction in which the exact solution can be found. However, extrapolation of the numerical results does not necessarily lead to a better approximation of the exact solution, since the truncation error may not be an analytic function of the mesh width (see Gourlay and Morris {26}). Much remains to be investigated in this respect.

A class of numerical solutions for which good numerical checks on the reliability are often available, is formed by the stationary solutions obtained asymptotically with non-stationary methods (such as for the blunt-body problem). In these solutions all time derivatives vanish, and the distinction between first- and second-order accuracy in difference schemes weighs less heavily. The strongly diffusive schemes, like (0), give less detail and slower convergence of the gross features than the weakly diffusive schemes. On the other hand, a scheme like (II) starts out with abundant small-scale structure, which goes through many changes before subsiding in the stationary solution. Scheme (I) may be the best bet in giving fast convergence to a considerably detailed stationary solution.

6. NUMERICAL TESTS

6.1 Shock profiles

In Sec. 3.5 we have displayed some numerical results obtained with the schemes of Fig. 8 in a sample problem based on a single conservation law. Similar competitive sample computations might be made on the basis of the Eulerian and Lagrangean equations of ICF. We do not think this would be really worth-while. In the preceding sections enough arguments have been given in favour of scheme (I), being the best scheme of first-order accuracy, and scheme (II), the sole scheme of second-order accuracy. We believe that any further efforts should be directed at examining various non-basic additions that may be made to amend these schemes.

Scheme (II), known for a decade, has been widely used, non-basic additions included. We have already pointed out (Sec. 4.2) that, in improving on this scheme, the optimum has not yet been reached. We shall not further pursue this subject but concentrate our attention on scheme (I).

The original scheme of Godunov appears to be very popular in the Soviet Union (see e.g. Belotserkovskii and Chushkin in Holt {27}), but has practically been ignored in western countries. Its principal part, which we have designated as scheme (I), has in all probability been used only in our own work, and we treated only Lagrangean flow problems. Although we cannot claim a wide experience with this scheme, the optimistic undertone in the discussion of this scheme seems really justified. In support of this optimism, we shall now summarize the results of some simple test computations.

In a Lagrangean net running from $x = 0$ to 101 with $\Delta x = 1$, the following dimensionless values were prescribed at $t = 0$ in an ideal gas with $\gamma = 5/3$.

x	$\leqq 64$	65	$\geqq 66$
V	1	0.462475	1/3
u	1	0.663369	1/3
E	3/5	0.412525	19/45
e	1/10	0.192495	11/30
p	1/15	0.277485	11/15
c	1/3	1	$\frac{1}{3}\sqrt{33} = 1.91485$
x	x	64.731238	$65.129142 + \frac{1}{3}(x - 66)$

The state variables in $x \geqq 66$ have exactly the values that would arise if the gas at $x \leqq 64$ were shocked by a shock travelling at a Lagrangean[5] shock speed $U = 1$. This is easily checked with aid of the jump equations (12). In fluid dynamics these are usually called the Rankine-Hugoniot relations. The one-parameter family of post-shock states related to one single pre-shock state through different shock speeds defines a so-called Hugoniot curve in phase space; see Zel'dovich and Raizer {28, Sec. 1.14}. The state in the net-point $x = 65$ has (in the absence of means to anticipate the shock structure) been chosen to lie on the Hugoniot curve through the states assigned to $x \leqq 64$ and $x \geqq 66$. We particularly have taken $c = 1$ so that the characteristic speed $-c$ in this point equals the speed of the shock that would separate the states in the adjacent net-points. The above set of initial values is intended to represent two homogeneous states separated by a single shock. The reason for inserting an intermediate state is, in the first place, that it spreads the initial shock structure over two meshes, which is in better agreement with the final steady profile. In the second place, it yields the a priori most plausible definition of the shock position, viz. the point where the characteristics merge into the shock path. This definition was inspired by Lax and Wendroff {10}.

In an Eulerian coordinate system the above initial values represent a standing shock, as $u_{64}/V_{64} - u_{66}/V_{66} = 0$. Following Godunov {11}, the Eulerian coordinate of a Lagrangean net-point was calculated, not only at $t = 0$ but at any time level, by summing the specific volumina in the net-points in the following way:

$$x(t^j, x_m) = \left(\tfrac{1}{2}V_0^j + \sum_{i=1}^{m-1} V_i^j + \tfrac{1}{2}V_m^j\right) \Delta x \qquad (165)$$

The truncation error in the resulting values x_m^j has the magnitude $O\left((\Delta x)^2\right)$. This is one order lower than might be reached with a scheme of first-order accuracy and two orders lower than might be reached with scheme(II). A more accurate computation of a streamline than by (165) would require a weighted mean of the fluid velocity at different time levels. The danger in this technique is that the streamlines may cross if the velocity field is not smooth enough. Such an anomaly is of course excluded in (165).

---

[5] The Lagrangean shock speed equals the mass of fluid in a column with unit cross-section, through which the shock travels per unit time.

The values of the state variables at x = 0 and 101 were kept constant; this virtually did not affect the numerical solution near the shock. We followed the shock with aid of the one-step version of scheme (I), adopting values of λ equal to $\frac{1}{4}$, $\frac{1}{3}$, $\frac{5}{12}$, $\frac{1}{2}$ and $\frac{1}{11}\sqrt{33}$ = 0.522234. The latter value is the maximum allowed by the CFL condition; it gave a stable computation with practically the same results as for $\lambda = \frac{1}{2}$. The use of simple fractional values of λ provides a convenient check on the convergence of the shock profile; see Sec. 3.5.

In Fig. 12 we have given the pressure profiles for $\lambda = \frac{1}{4}$ and $\frac{1}{2}$. In order to draw the profiles more easily we have superimposed the set of pressure values at t = 37.5 (shock expected in x = 27.5) over the values at t = 50.0 (shock expected in x = 15.0) after a shift of -12.5 along the x axis. Within the accuracy of the drawing, these sets define the same profile, as appears clearly from the comparison of the values at t = 25.0 (shock in x = 40) with those at t = 50.0. Here the values of p in corresponding points agree up to four decimals for $\lambda = \frac{1}{2}$, and up to (at least) six decimals for $\lambda = \frac{1}{4}$.

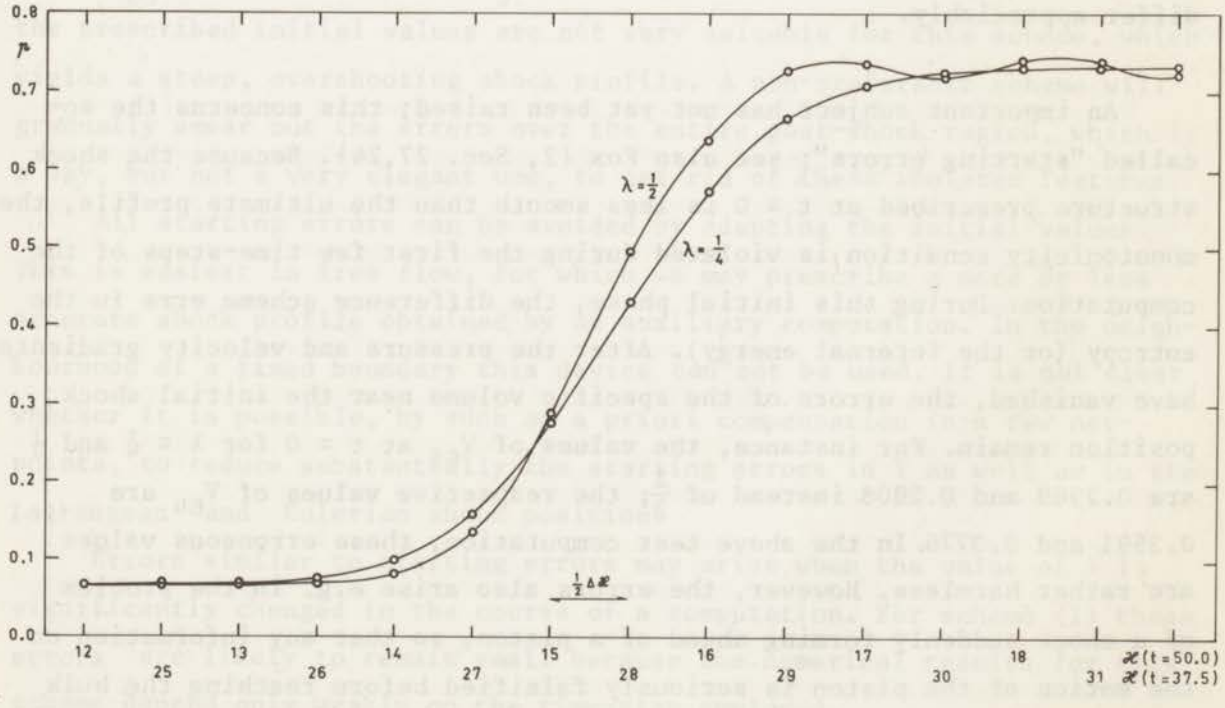


Figure 12. Stationary profiles in a Lagrangean shock, obtained with scheme (I) for different values of the mesh ratio.

The results for $\lambda = \frac{1}{2}$ exhibit small post-shock oscillations between the pressure values 0.72 and 0.74. The exact post-shock pressure, 0.733333, is monotonically approached by the values for $\lambda = \frac{1}{4}$. The results for $\lambda = \frac{1}{3}$ were also monotonic.

It is seen from Fig. 12 that the effective shock width increases roughly by a factor $\frac{4}{3}$ when λ drops by a factor 2. This is a gratifying result, but further improvement is feasible. A slight increase of the dissipation in scheme (I) by non-basic additions will make the shock profiles monotonic for all allowed values of λ, and even less dependent on λ. Profiles similar to the ones in Fig. 12, but for the two-step version of scheme (II), are given e.g. in Richtmyer and Morton {1, Figs. 12.6, 12.7}. These curves exhibit a far stronger variation with λ.

Godunov {11} gives some pressure profiles for strong shocks obtained with his scheme; unfortunately the adopted global Courant number is not mentioned. One of these (rather carelessly drawn) profiles is neatly traced in {1, Fig. 12.8}. It is practically congruent with the profile which we obtained with scheme (I) for $\lambda = \frac{5}{12}$. This curve has not been drawn in Fig. 12 because the dependence on λ was clearly linear, anyhow. We conclude that the results of Godunov's scheme and its principal part (I) do not differ appreciably.

An important subject has not yet been raised; this concerns the so-called "starting errors"; see also Fox {2, Sec. 27.24}. Because the shock structure prescribed at $t = 0$ is less smooth than the ultimate profile, the monotonicity condition is violated during the first few time-steps of the computation. During this initial phase, the difference scheme errs in the entropy (or the internal energy). After the pressure and velocity gradients have vanished, the errors of the specific volume near the initial shock position remain. For instance, the values of V_{65} at $t = 0$ for $\lambda = \frac{1}{4}$ and $\frac{1}{2}$ are 0.2969 and 0.2908 instead of $\frac{1}{3}$; the respective values of V_{64} are 0.3591 and 0.3776. In the above test computation, these erroneous values are rather harmless. However, the errors also arise e.g. in the problem of a shock suddenly forming ahead of a piston, so that any information on the motion of the piston is seriously falsified before reaching the bulk of the gas.

The starting error appears also in the position of the shock. This may be illustrated by the values obtained at $t = 50$ in $x = 15$ with scheme (I), for $\lambda = \frac{1}{4}$ and $\frac{1}{2}$.

λ	$\frac{1}{4}$	$\frac{1}{2}$
V	0.549110	0.537015
u	0.674399	0.674988
p	0.276846	0.289427
c	0.916671	0.947765
x	64.5775	64.6660

A comparison with the table on page 69 reveals that the values of u and p are remarkably close to their initial values in $x = 65$. The values of V and the dependent quatities c and $\dot{x}$ have erred. Because the sound speed differs considerably from unity, the point $x = 15$ does not designate the exact position of the shock according to our previous definition.

In addition these values may serve to illustrate that the dissipative shock structure does not consist of states lying on the same Hugoniot curve (which was not expected either).

Any dissipative difference scheme exhibits starting errors, in particular Godunov's original scheme and scheme (II). When repeating the test computations described above with scheme (II) we found for $\lambda = \frac{1}{2}$ a final value $V_{65} = 0.2267$ instead of $\frac{1}{3}$. The large error arises from the fact that the prescribed initial values are not very suitable for this scheme, which yields a steep, overshooting shock profile. A non-preferable scheme will gradually smear out the errors over the entire post-shock region, which is a way, but not a very elegant one, to get rid of these isolated features.

All starting errors can be avoided by adapting the initial values. This is easiest in free flow, for which we may prescribe a more or less accurate shock profile obtained by an auxiliary computation. In the neighbourhood of a fixed boundary this device can not be used. It is not clear whether it is possible, by such an a priori compensation in a few net-points, to reduce substantially the starting errors in V as well as in the Lagrangean and Eulerian shock position.

Errors similar to starting errors may arise when the value of λ is significantly changed in the course of a computation. For scheme (I) these errors are likely to remain small because the numerical results for this scheme depend only weakly on the time-step employed.

We repeated the above test for the Euler equations, with essentially the same shock, but running through the computational net at various speeds.

For a non-zero speed, the features of the Lagrangean results were reproduced. In the critical case of a standing shock, a mild instability occurred (mild compared to the fast instability occurring for scheme (II)). This behaviour shows that condition (67), which just excludes the existence of nonlinear instabilities in scheme (I), should not be taken too literally. We may conclude that scheme (I), when applied to the Euler equations, must be made more strongly dissipative by adding non-basic differences.

6.2 A test case from astrophysics

A particular flow problem in astrophysics was described and qualitatively solved by Savedoff, Hovenier and Van Leer {29}. They considered, in a one-dimensional simplification, the hypersonic impact of a hypothetical extragalactic wind on the galactic atmosphere. Both gases are assumed to consist mainly of hydrogen and helium, with small fractions of heavier elements. The gases have extremely low densities and, initially, also low pressures. Accordingly, both gases are heated by shocks proceeding from the contact discontinuity which is established at the moment of collision.

As the gases differ in density, the two shocks are not equally strong. Numerically this means that the post-shock Courant numbers are different. This circumstance suggests the use of scheme (I), which was shown in the preceeding section to yield shock profiles that depend only weakly on the Courant number.

A further complication is that the gases, once ionized, start to radiate, mainly through the impurities. The radiation follows upon collisional excitation, hence is, per unit volume and time, proportional to ρ^2. Because of this strong dependence of the radiative loss term on ρ, we thought it better not to use scheme (II). A radiative instability might arise from the large post-shock oscillations resulting from this scheme. A similar instability, triggered by a starting error, occurred in one of our test computations.

Since the gases are optically thin, the radiation escapes and the net effect is cooling of the gas. The solution becomes interesting at

the time when somewhere in the flow the gas starts to recombine. A layer of highly compressed neutral hydrogen results. The study cited was originally undertaken in the hope that this compressed hydrogen layer might be identified with the high-velocity gas features observed by the radio astronomers (see e.g. Hulsbosch {30}). The real situation is undoubtedly more complicated.

We have indeed arrived at a detailed solution (unpublished) of the flow problem described above, but only after considerable trouble with the numerical techniques. The most adequate results were obtained with scheme (I) in the Lagrangean formulation. To obtain a reliable solution with a non-preferable scheme, or in Euler coordinates, would have been very demanding on machine time.

In the remainder of this section we shall describe a test made on a somewhat simpler problem, to show what happens if the Eulerian formulation is employed. The simplification introduced is that both gases are assumed to have the same uniform density, so that the flow becomes symmetric with respect to the contact surface. The problem is thus reduced to the computation of one single shock followed by a radiating zone. This type of flow will tend to a stationary pattern, the properties of which can be computed a priori (see below). The purpose of the test calculation was to check if the numerical solution would indeed approach this pattern.

The time-dependent Eulerian flow equations including radiative losses are identical to eqs. (127), apart from a source term $-\rho^2 L(e)$ at the right-hand side of the energy equation. Stationary solutions will simply satisfy the equations

$$\rho u = C_1 \quad (< 0, \text{ say})$$

$$p + \rho u^2 = C_2 > 0 \tag{166}$$

$$\frac{\partial}{\partial x}\left(\rho u \left(e + \frac{p}{\rho} + \tfrac{1}{2}u^2\right)\right) = -\rho^2 L(e) < 0$$

which are easily solved analytically if an expression for $L(e)$ is given. The first two of these equations yield the combinations of ρ, u and p which occur in the cooling region; by integrating the third equation downstream

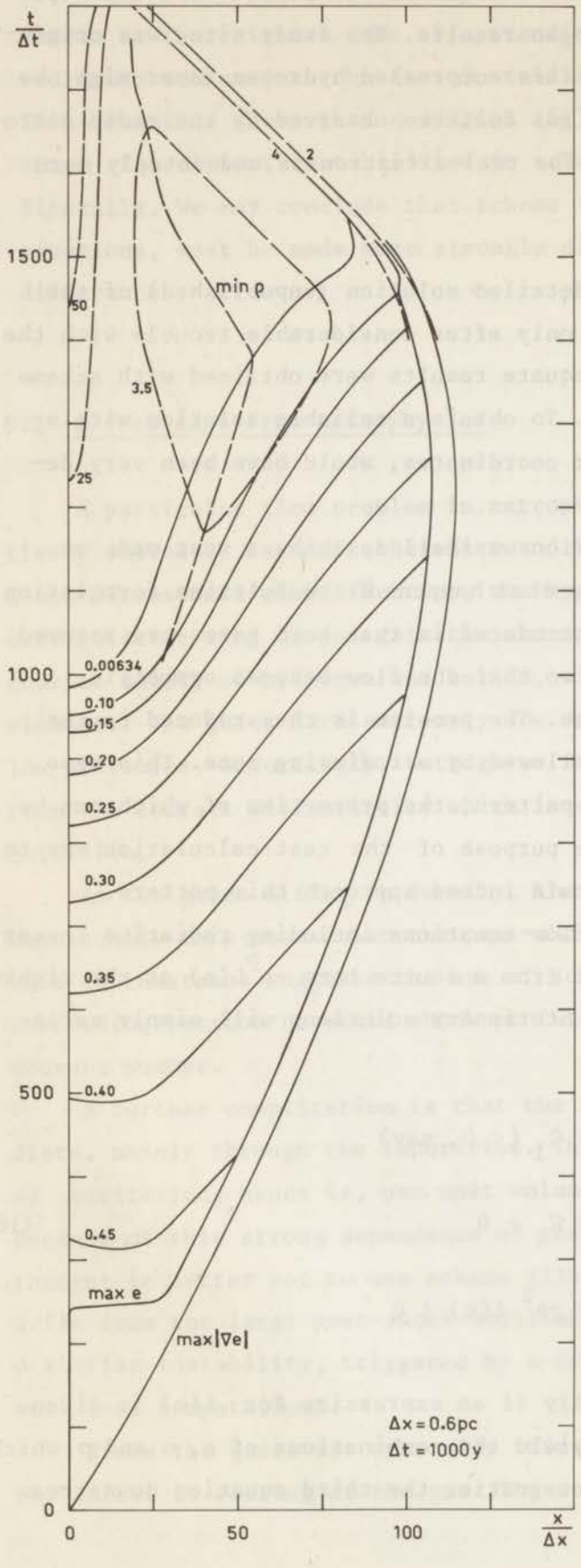


Figure 13. Collapse of radiating flow due to artificial diffusion. Description in the text.

we find the position at which a particular combination occurs. Starting from their post-shock values, u, ρ and e relatively decrease sharply, p increases slightly. At the point where e has dropped so far that radiation ceases, the flow becomes uniform, as in the pre-shock region. Examples of such stationary solutions can be found in Field, Rather, Aannestad and Orszag {31}.

The main results of the time-dependent solution of the test problem described above are shown in Fig. 13. Scheme (O) was employed. A uniform gas with dimensionless density $\rho = 1$ (corresponding to 0.05 H atoms/cm^3), coming from the right, meets its mirror image or a fixed boundary at $x = 0$. Its velocity u_∞ equals 250 km/sec, which, in astronomically more suitable units, becomes 255 parsec/10^6 year. The original internal energy of the gas equals 0.00634, expressed in the unit u_∞^2; the initial post-shock value very nearly equals $\frac{1}{2}$. No streamlines are drawn in Fig. 13 but only some isolines of e (solid) and ρ (broken) and three other curves. One of these is the solid shockpath (max$|\nabla e|$), which on this occasion was determined by the points in the shock structures where $|\frac{\partial e}{\partial x}|$ was greatest. The other solid curve (max e) connects the points at which e itself was greatest. In the initial phase this curve lingers at $x = 0$, due to an overproduction of heat (a starting error). Later the curve follows the shock. As the temperature in the post-shock region drops, the pressure drops too, the cooler gas being compressed by the hotter gas. Consequently, the shock decelerates. At $t = 0.983 \times 10^6$y, the gas at $x = 0$ has cooled down to its original temperature, and radiation stops.

What should happen next is that, in the cooling region, the flow gradually adjusts itself to the known stationary solution superimposed on a small drift velocity. In this final regime the front of cold gas advances by the drift velocity 0.00423 $|u_\infty|$; the shock advances at the same speed and all isolines have the same (small) inclination with the t axis as the cold-front and the shock. Note that the strong reduction of the flow velocity corresponds to a 236-fold compression of the gas! Before this final regime is reached, the shock is likely to overshoot the ultimate stand-off distance of about 150 meshes; then approach it from the right.

In the numerical solution, the nearly stationary state was never reached. Due to the heat conduction in the difference scheme, the internal energy of the hot gas just behind the shock diffuses into the cold gas

near $x = 0$, where it is very efficiently radiated away because of the high density at that place. The shock, after having stopped at $x \approx 110\Delta x$, is pushed back to the origin and soon meets the cold-front $e = 0.00634$, hence becomes isothermal! The position of the shock cannot be determined any more by means of $\max|\nabla e|$ or $\max e$; the lines $\rho = 2$ and $\rho = 4$ are chosen to indicate the shock path. To give an idea of the distribution of ρ along the x axis, the minima are indicated by a broken line $\min \rho$.

At $t = 1800$ the post-shock region had collapsed so far that the computation was automatically stopped due to an alarm built in the computer program. What happens after the collapse can be found from similar computations made by Van Deursen {32}, who investigated the high-velocity cloud problem with the Particle-In-Cell method (see {3}) and the Fluid-In-Cell method (see Gentry, Martin and Daly {33}). The mixed Lagrangean-Eulerian PIC method, with the feature of a quantized density, appeared to be not accurate enough to handle the sensitive radiation law. The FLIC method does not have this disadvantage but shares the common diffusion errors with all other Eulerian methods. The FLIC results exactly reproduced the flow pattern of Fig. 13. After the collapse, the region of cold gas bounded by the isothermal shock started to grow slowly. The cooling zone, however, could not be built up again.

7. DESIDERATA

It has emerged from the preceding three chapters that, among all basic schemes, the preferable schemes (I) and (II) stand out by their unique properties. The newly found scheme (I) certainly has all capacities to replace the scheme of Von Neumann and Richtmyer {5}, which in many branches of gas dynamics still seems to be the most popular first-order method. Both schemes (I) and (II) can still be put into a variety of non-basic forms; the optimal versions remain yet to be found.

Apart from the non-basic additions, there are other modifications and features which have not been thoroughly examined. We have raised in Sec. 4.2 the subject of a non-uniform computational net, in which the schemes are no longer conservative. The modification of the schemes required at a spatial boundary has hardly been mentioned. Scheme (I) simply reduces to a three-point scheme, but scheme (II), for which four points are needed to preserve the second-order accuracy, becomes very skew; see Gourlay and Morris {26}. As in uneven meshes, care must be taken not to create a spurious source term. Last but not least we recall the starting errors pointed out in Sec. 6.1.

With all these problems unsolved for even the simplest conservative schemes, it is a bit surprising to observe a tendency to look for explicit more-point, more-step, higher-order methods. We refer particularly to the schemes of third-order accuracy, which have recently cropped up in the literature; see e.g. Rusanov {34}, Burnstein and Mirin {35}. The basic philosophy is that, just as in analogue computations, a complicated scheme (circuit) of higher-order accuracy used in a coarse net still may give better results.

The complete set of all third-order schemes has numerous ramifications. The present study of the first-order methods suggests that it would be an enormous job to gain an equally clear insight into the properties of this set. Moreover, the difficulties connected with boundaries, etc., strongly increase with the number of net-points involved. Since the schemes are more accurate, they are less dissipative, hence more easily susceptible to nonlinear instabilities. We therefore do not believe there is much point in pursuing higher-order schemes until these annoying accessory problems have been adequately solved for the simpler schemes.

Another trend is to combine basic schemes, particularly scheme (II), into a new scheme in order to attain some improvement. An ingenious method is the "zero average phase error" method of Fromm {36}. He succeeded to reduce the phase errors in scheme (II) substantially by combining a forward and a backward time-step (which yield opposite phase errors).

The method of repeating the integration of an initial-value problem for various values of the mesh width, in order to extrapolate towards the true solution, has been touched upon in Sec. 5.2. Its application is already difficult in simple cases (cf. {26}) and becomes highly problematic with respect to solutions containing shocks, as may be judged from the persistence of starting errors. This provides another argument to consider the matter of starting errors quite seriously.

Finally, we should not forget to mention the development of implicit conservative difference schemes, mostly of second-order accuracy. Although the treatment of boundaries is a complicated matter in such schemes (Gourlay and Morris {13}), we believe that this is a most promising line. Implicit methods usually have higher stability than explicit methods, so that some play is left in choosing the amount of dissipation. This is certainly no luxury in designing a difference scheme for the Euler equations. As for the Lagrangean equations, we may refer to the work of Popov and Samarskii {37}, in which special care is taken to represent the balance between kinetic and internal energy accurately.

In the present work we have dealt mainly with a simple system of conservation laws without source terms. Starting from Ch. 4 we have restricted the number of equations, i.e. the number of state variables, to the practical value 3, which is just enough to treat ideal compressible flow. As a reminder we note that this assumption does not imply a restriction on the number of coordinates. We may now pose the question: is the experience thus gained at all useful in choosing an adequate difference scheme for solving a larger system of equations, such as occurs in magneto-hydrodynamics?

The answer cannot sound too optimistic. Considering first the Lagrangean equations, we essentially have to choose between the preferable schemes (I) and (II), supplemented with the optimum non-basic differences. As explained in Sec. 4.1, both schemes, when applied to the magneto-hydrodynamic equations, would require a six-step formulation, unless we would

prefer the explicit evaluation of the 7×7 stabilization matrix. In practice some simplifications will arise from the symmetry in the values of the characteristic speeds. In the Euler equations such a symmetry is not found. As in Euler coordinates diffusion cannot be avoided, we may as well use a non-preferable scheme. Note that an additional spurious diffusion now appears: diffusion of the magnetic field. In view of this fact, a good numerical solution seems hardly feasible.

We have in several places warned for the dangers connected with source terms, and given a dramatic example in Sec. 6.2. Still, such terms give rise to the most interesting flow problems. Galactic gas dynamics may be mentioned as a good example. Here the action certainly is not caused by supersonicly moving blunt bodies (although this may be assumed in simple models) but rather by radiation, gravitation and the like.

We do not know what methods will eventually be found most useful to handle these complicated flow problems. There are reasons to fear that the methods will become as numerous as the problems. In such a situation, an analysis of the fundamental properties, similar to the present work, may again be helpful to sort out the most promising methods prior to numerical tests.

REFERENCES

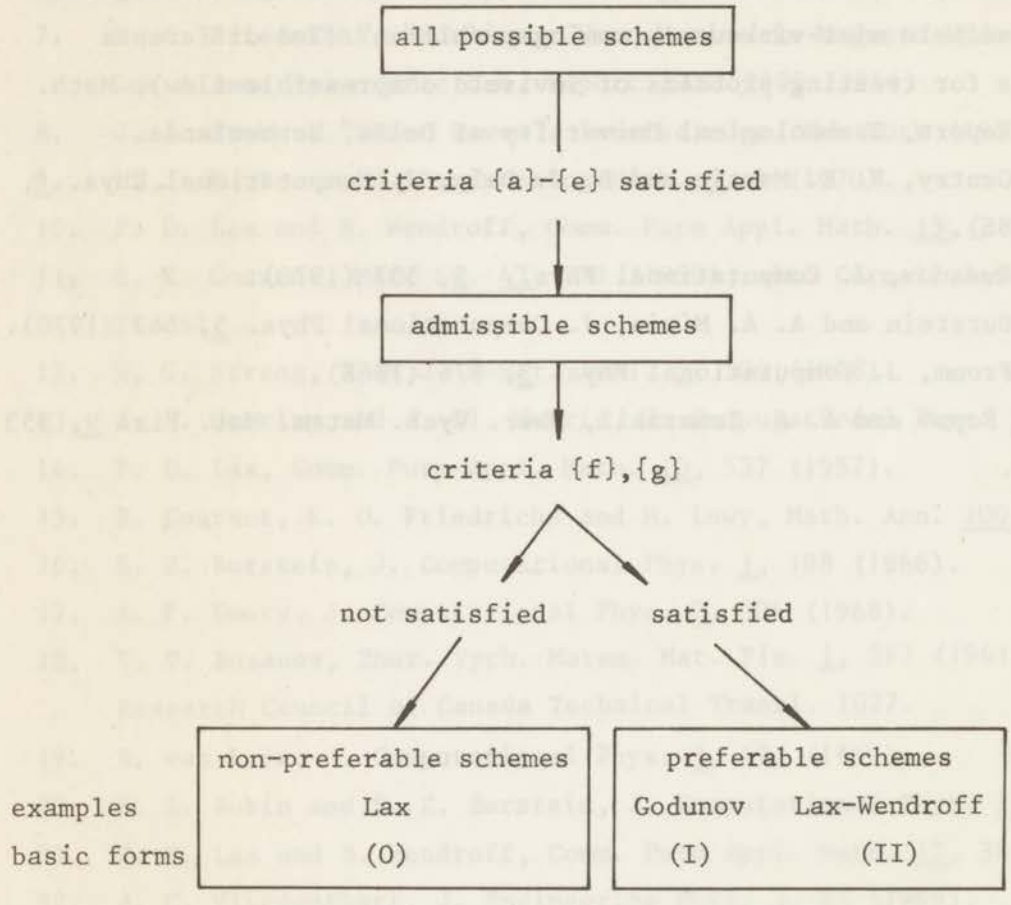
1. R. D. Richtmyer and K. W. Morton, "Difference Methods for Initial-Value Problems", 2nd edition. Interscience, New York, 1967.
2. L. Fox (ed.), "Numerical Solution of Ordinary and Partial Differential Equations". Pergamon, Oxford, 1962.
3. B. J. Alder, S. Fernbach and M. Rotenberg (eds.), "Methods in Computational Physics", Vol. 3. Academic Press, New York, 1964.
4. F. A. Goldsworthy, "The Dynamics of HII regions", to appear in Rev. Mod. Phys..
5. J. von Neumann and R. D. Richtmyer, J. Appl. Phys. 21, 232 (1950).
6. P. D. Lax, Comm. Pure Appl. Math. 7, 159 (1954).
7. K. O. Friedrichs, "Nonlinear Wave Motion in Magneto-Hydrodynamics", Los Alamos Scientific Lab. Report LAMS 2105 (1954).
8. J. J. Stoker, "Water Waves", Interscience, New York (1957).
9. L. J. F. Broer, J. Engineering Math. 4, 1 (1970).
10. P. D. Lax and B. Wendroff, Comm. Pure Appl. Math. 13, 217 (1960).
11. S. K. Godunov, Mat. Sb. 47, 271 (1959); also Cornell Aeronautical Lab. Transl..
12. W. G. Strang, SIAM J. Numer. Anal. 5, 506 (1968).
13. A. R. Gourlay and J. Ll. Morris, J. Computational Phys. 5, 229 (1970).
14. P. D. Lax, Comm. Pure Appl. Math. 10, 537 (1957).
15. R. Courant, K. O. Friedrichs and H. Lewy, Math. Ann. 100, 32 (1928).
16. S. Z. Burstein, J. Computational Phys. 1, 198 (1966).
17. A. F. Emery, J. Computational Phys. 2, 306 (1968).
18. V. V. Rusanov, Zhur. Vych. Matem. Mat. Fiz. 1, 267 (1961); also Nat. Research Council of Canada Technical Transl. 1027.
19. B. van Leer, J. Computational Phys. 3, 473 (1969).
20. E. L. Rubin and S. Z. Burstein, J. Computational Phys. 2, 178 (1967).
21. P. D. Lax and B. Wendroff, Comm. Pure Appl. Math. 17, 381 (1964).
22. A. C. Vliegenthart, J. Engineering Phys. 3, 81 (1969).
23. R. Courant, E. Isaacson and M. Rees, Comm. Pure Appl. Math. 5, 243 (1952).
24. M. Ciment, Report NYO-1480-100 (1968), Courant Inst. Math. Sci., New York University.
25. J. Gary, Report NYO-9603 (1962), Courant Inst. Math. Sci., New York University.

26. A. R. Gourlay and J. Ll. Morris, Computer J. 11, 95 (1968).
27. M. Holt (ed.), "Basic Developments in Fluid Dynamics", Academic Press, New York (1966).
28. Ya. B. Zel'dovich and Yu. P. Raizer, "Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena", Vol. 1. Academic Press, New York (1966).
29. M. P. Savedoff, J. W. Hovenier and B. van Leer, Bull. Astr. Inst. Netherlands 19, 107 (1967).
30. A. N. M. Hulsbosch, Bull. Astr. Inst. Netherlands 20, 33 (1968).
31. G. B. Field, J. D. G. Rather, P. A. Aannestad and S. A. Orszag, Ap. J. 131, 953 (1968).
32. A. van Deursen, "Twee differentie-methoden voor de behandeling van kompressibele niet-viskeuze stromingsproblemen" (Two difference methods for treating problems of inviscid compressible flow), Math. Dept. Report, Technological University of Delft, Netherlands.
33. R. A. Gentry, R. E. Martin and B. J. Daly, J. Computational Phys. 1, 87 (1966).
34. V. V. Rusanov, J. Computational Phys. 5, 507 (1970).
35. S. Z. Burstein and A. A. Mirin, J. Computational Phys. 5, 547 (1970).
36. J. E. Fromm, J. Computational Phys. 3, 176 (1968).
37. Yu. P. Popov and A. A. Samarskii, Zhur. Vych. Matem. Mat. Fiz. 9, 953 (1969).

SUMMARY

Difference schemes for a nonlinear hyperbolic system of first-order conservation laws in two independent variables are studied, with emphasis on the equations of ideal compressible flow. The schemes are conservative and dissipative, so that they can be used to handle shocks. All explicit schemes are considered that involve four net-points divided over two time levels.

Application of certain criteria of physical relevance leads to the following division (Ch. 3).



Preferable schemes, when applied to the Lagrangean flow equations, correspond to artificial viscosity, and non-preferable schemes to artificial diffusion (Sec. 5.2). In the Eulerian formulation artificial diffusion is unavoidable.

The principal part of a scheme is defined in Sec. 3.2; a scheme identical to its own principal part is called basic. A linear Fourier analysis of the dissipative and convective errors of the basic schemes is made and leads to the distinction of three key schemes.

(0) Lax' scheme: the only three-point scheme.

(I) Principal part of Godunov's scheme: optimal scheme of first-order accuracy regarding stability and smoothness of the results, and much simpler than Godunov's original scheme.

(II) Lax-Wendroff scheme: the only scheme of second-order accuracy.

Figure (8) shows a 3-parameter family of difference schemes in which the schemes (0), (I) and (II) occupy the main diagonal.

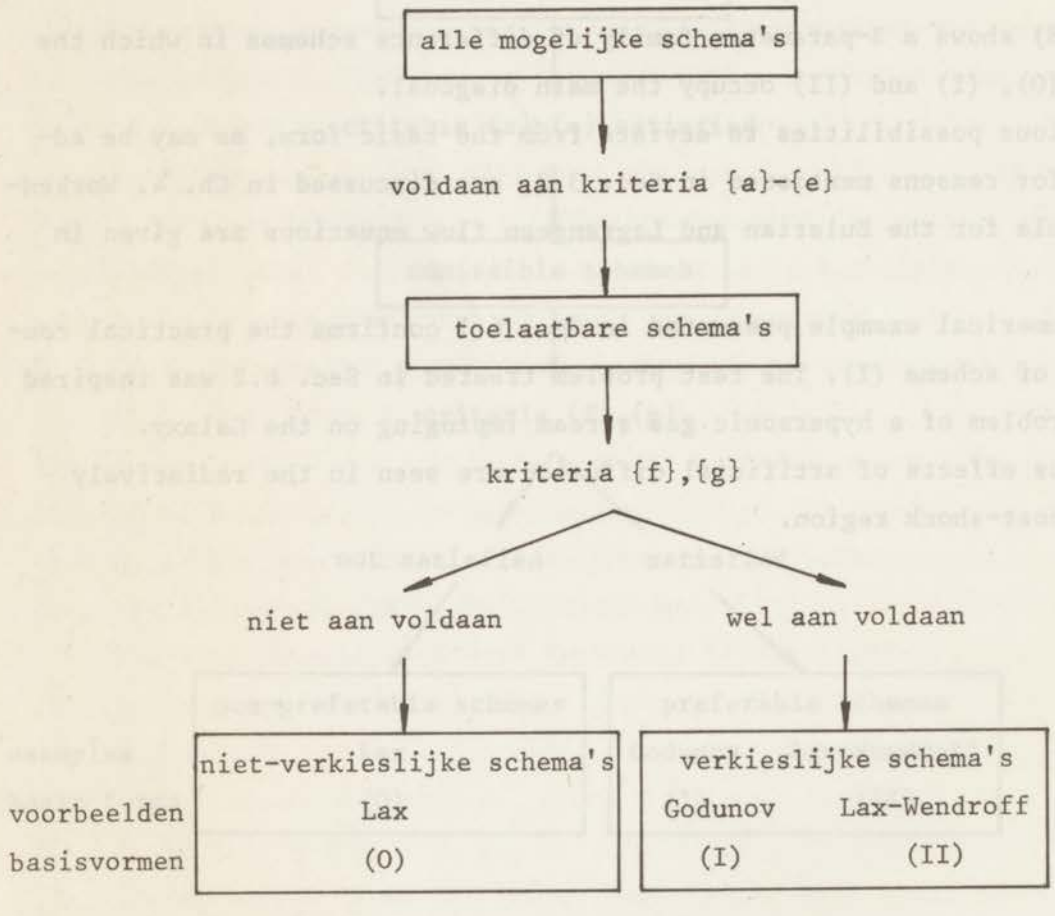
Various possibilities to deviate from the basic form, as may be advisable for reasons mentioned in Sec. 3.2, are discussed in Ch. 4. Worked-out formula for the Eulerian and Lagrangean flow equations are given in Ch. 5.

A numerical example presented in Sec. 6.1 confirms the practical convenience of scheme (I). The test problem treated in Sec. 6.2 was inspired by the problem of a hypersonic gas stream impinging on the Galaxy. Disastrous effects of artificial diffusion are seen in the radiatively cooling post-shock region.

SAMENVATTING

In dit proefschrift worden differentieschema's bestudeerd voor een niet-lineair hyperbolisch systeem van behoudswetten in twee onafhankelijke variabelen, waarbij de nadruk valt op de vergelijkingen voor ideale samendrukbare stroming. De schema's zijn konservatief en dissipatief, zodat ze kunnen worden gebruikt om schokken te behandelen. Alle expliciete schema's worden beschouwd welke gebaseerd zijn op vier roosterpunten verdeeld over twee tijdniveaus.

Toepassing van bepaalde, fysisch relevante kriteria leidt tot de volgende indeling (Hoofdstuk 3).



Wanneer een verkieslijk schema wordt gebruikt voor de stromingsvergelijkingen van Lagrange, treedt kunstmatige viskositeit op; niet-verkieslijke schema's veroorzaken kunstmatige diffusie (§ 5.2). Bij het gebruik van de vergelijkingen van Euler is kunstmatige diffusie onvermijdelijk.

Het hoofdgedeelte van een schema wordt gedefinieerd in § 3.2; een schema dat identiek is aan zijn hoofdgedeelte wordt basisschema genoemd. Een lineaire Fourier-analyse van de dissipatieve en konvektieve fouten in de basisschema's wordt gegeven; hierbij blijkt duidelijk de sleutelpositie van de volgende drie schema's.

(O) Lax' schema: het enige drie-punts schema.

(I) Het hoofdgedeelte van Godunovs schema: het optimale schema met eerste-orde nauwkeurigheid, de stabiliteit en gladheid der resultaten in aanmerking genomen; veel eenvoudiger dan Godunovs oorspronkelijke schema.

(II) Het schema van Lax en Wendroff: het enige schema met tweede-orde nauwkeurigheid.

Figuur (8) toont een 3-parameter familie van differentieschema's, waarin de schema's (O), (I) en (II) de hoofddiagonaal bezetten.

Diverse mogelijkheden om van de basisvorm af te wijken, hetgeen wenselijk kan zijn om redenen aangegeven in § 3.2, worden besproken in Hoofdstuk 4. Uitgewerkte formules voor de stromingsvergelijkingen van Euler en Lagrange zijn te vinden in Hoofdstuk 5.

In § 6.1 wordt een numeriek voorbeeld gegeven dat het praktische nut van schema (I) bevestigt. Het testprobleem dat wordt behandeld in § 6.2 is geïnspireerd op het vraagstuk van de hypersone gasstroom welke het Melkwegstelsel zou binnenvallen. De desastreuze werking van kunstmatige diffusie is te zien in het door straling afkoelende gebied achter de schok.

ERRATA

Please affix the following corrections yourself:

page 26, line 3 from top: Burstein and Rubin read Rubin and Burstein ;

page 39, eq. (97): $\tilde{A}$ read $\tilde{A}$;

page 49, line 18 from bottom: $F_{m+\frac{1}{2}}$ read $F^{j}_{m+\frac{1}{2}}$;

page 49, line 11 from bottom: $F_{m+\frac{1}{2}}$ read $F^{j}_{m+\frac{1}{2}}$;

page 63, eq. (156): J read J_{La} ;

page 63, line 6 from top: in powers of λA read in powers of λA_{La} ;

page 70, line 3 from top: $U = 1.$ read $U = -1.$;

page 79, line 11 from top: Sec. 4.2 read Sec. 4.3 .

ADDENDUM

Author's address from 1 October 1970:

Dr. B. van Leer
University of California
Leuschner Observatory
Berkeley, Cal. 94720
U.S.A.