ALGORITHM DEVELOPMENT

Timothy J. Barth
NASA Ames Research Center

and

Harvard Lomax
NASA Ames Research Center

Introduction

Algorithm: A set of rules for solving a problem in a finite number of steps.

Development: The progression to a more effective state.

The past decade has seen considerable activity in algorithm development for the Navier-Stokes equations. This has resulted in a wide variety of useful new techniques. It would appear, however, that there is plenty of room for further improvements. That is to say, we are far from exhausting all possible sets of rules for these problems and it is highly probable that some remaining ones will be more effective than those we have now.

It is foolish and even counterproductive to anticipate or set milestones for the detailed development of basic or even applied research. The history of science tells us that we can expect something to happen in any major field if active minds capable of original thinking are allowed to pose challenging problems and seek elegant solutions.

What we can do is look backwards and find what we are doing now in a given area of science that was not anticipated ten years ago. Some examples of this type for the numerical solution of the Navier-Stokes equations form the body of this paper. These are divided into two parts, one devoted to the incompressible Navier-Stokes equations, and the other to the compressible form. The discussion is far from being comprehensive, and, in fact, the examples for the incompressible case are strictly limited to experience at NASA Ames.

1. Incompressible Navier–Stokes Equations

In the middle and late 70s much attention was paid to the direct solution of homogeneous turbulent flows with periodic boundary conditions, see Rogallo (1981). The grids used at that time were 64^3 and the storage capacity of available computers was the limiting factor in the spatial resolution. The natural method to use for the numerical approximation of the space derivatives was the classical spectral method composed of finite Fourier series, and the algorithm used for implementation was the fast Fourier transform. The time advance was fully explicit so that all of the time and space scales could be resolved as accurately as possible. However, even with explicit schemes, time advance of a spectral method requires a minimum of two memory locations for every dependent variable at every point in the mesh. The standard third and fourth order Runge-Kutta methods both take a minimum of three locations, so the options at that time were to use an existing second order time march method or use a coarser mesh and reduce the space accuracy. This motivated Dr. Alan Wray (Ames Research Center, unpublished) to turn to the Runge-Kutta technique and search for a subset of high order methods that require a minimum (two location) amount of storage capacity. He was successful and his third-order method, used to time march the nonlinear equation,

$$\frac{du}{dt} = F(u,t) \qquad (1.1)$$

PRECEDING PAGE BLANK NOT FILMED

has the predictor-corrector form

$$\hat{u} = u_n + \alpha \Delta t F(u_n , t_n)$$
$$\dot{u} = u_n + A \Delta t F(u_n , t_n)$$
$$\ddot{u} = \hat{u} + \beta \Delta t F(\dot{u} , t_n + A \Delta t) \qquad (1.2)$$
$$\ddot{u} = \dot{u} + B \Delta t F(\ddot{u} , t_n + A \Delta t)$$
$$u_{n+1} = \dot{u} + \gamma \Delta t F(\ddot{u} , t_n + (\alpha + B) \Delta t)$$

The value of u_n is initially provided and stored. The value of $\hat{u}$ is then calculated and also stored. Then the value of $\ddot{u}$ is formed and overstores u_n which is no longer needed. The process continues through $\dot{u}$ and $\ddot{u}$, requiring at any intermediate step only two memory locations per dependent variable per mesh point. Finally, u_{n+1} overwrites $\dot{u}$, $\ddot{u}$ is discarded, and the cycle is repeated.

There are four equations for the five coefficients in eq (1.2), so we have a one-parameter family of low storage, third order Runge-Kutta methods. The four equations are

$$\alpha + \beta + \gamma = 1$$
$$(\alpha + B)\gamma + A\beta = 1/2$$
$$(\alpha + B)^2\gamma + A^2\beta = 1/3 \qquad (1.3)$$
$$AB\gamma = 1/6$$

One particular solution is given by

$$\alpha = 1/4 , \ A = 8/15 , \ \beta = 0 , \ B = 5/12 , \ \gamma = 3/4 \qquad (1.4)$$

This method is still being used to time march codes for homogeneous turbulent flows. It is a good example of an algorithm advance adding a new capability to an old concept.

A major advance in algorithms for wall-bounded turbulent simulations occurred in the early 1980's. At that time Leonard and Wray (1982) extended the concepts being used to compute homogeneous turbulent flows, to compute wall bounded turbulent flows in relatively simple geometries. Let U be the velocity vector, p the pressure, and ν the kinematic viscosity. One solves the vector equation expressing conservation of momentum,

$$U_t + U \cdot \nabla U = -\nabla p + \nu \nabla^2 U \qquad (1.5)$$

under the constraints of continuity in the domain and no slip at the walls:

$$\nabla \cdot U = 0 \ \text{in} \ D , \ U = 0 \ \text{on} \ \partial D \qquad (1.6)$$

In homogeneous flows harmonic basis functions are used and these automatically satisfy the periodic boundary conditions. Furthermore, it was easy to make the solutions solenoidal ($\nabla \cdot U = 0$) so the pressure term could be eliminated. The idea advanced by Leonard and Wray was to build the constraints (1.6) into the basis functions of a generalized spectral method for wall bounded flows, so that the constraints are automatically and exactly satisfied with each time advance, and do not need to be further enforced at each step in conjunction with (1.5). The solution is then expressed as a linear combination of global vector "basis functions" that each satisfy (1.6). Due to the constraints one needs to carry only two degrees of freedom per spectral mode while other methods usually carry four, the three velocity components and the pressure. Thus, less computer storage is needed to achieve the better resolution. For more details and further discussion see the paper by Leonard and Wray.

Where they can be formed (this can be difficult since they are geometry dependent), the choice of the generalized spectral basis functions greatly improves the numerical treatment of the spatial aspect of the problem. However, to get adequate resolution near the walls, the time integration tends to be stiff due to the eigenstructure of the viscous terms. Because of this Dr. Philippe Spalart (Ames Research Center, unpublished) devised the use of a hybrid time marching scheme which is implicit for the (linear) viscous terms and explicit for the (nonlinear) convection terms. Again because of low memory requirements he had been using existing 2nd order methods for the time march. However, he has recently extended Wray's Runge-Kutta technique and developed a hybrid method which is third order in time accuracy and still has the minimum (two location) storage requirements. Thus if we have the vector relation

$$u_t = N(u) + L \cdot u \qquad (1.7)$$

where L and N are matrix operators that are linear and nonlinear, respectively, the sequence can be made third order accurate with the proper

choice of the coefficients in

$$\tilde{u} = u_n + \Delta t[L \cdot (\alpha_1 u_n + \beta_1 \tilde{u}) + \gamma_1 N_n]$$

$$\tilde{\tilde{u}} = \tilde{u} + \Delta t[L \cdot (\alpha_2 \tilde{u} + \beta_2 \tilde{\tilde{u}}) + \gamma_2 \tilde{N} + \varsigma_1 N_n]$$

$$u_{n+1} = \tilde{\tilde{u}} + \Delta t[L \cdot (\alpha_3 \tilde{\tilde{u}} + \beta_3 u_{n+1}) + \gamma_3 \tilde{\tilde{N}} + \varsigma_2 \tilde{N}]$$

$$(1.8)$$

The treatment of the N terms is equivalent to that used in Wray's scheme. Only one solution for the coefficients is known. This is given by the conditions that

$$\gamma_1 = 0.7208762469, \quad \gamma_2 = 0.4001233399,$$

$$\gamma_3 = 0.5778221005, \quad \beta_1 = 0.3703996503,$$

$$\beta_2 = 0.0929740417, \quad \beta_3 = 0.1818702938,$$

$$\varsigma_1 = 0.4724519312, \quad \varsigma_2 = 0.2263697562$$

$$(1.9a)$$

and

$$\alpha_1 + \beta_1 = \gamma_1 \ , \ \alpha_2 + \beta_2 = \gamma_2 + \varsigma_1 \ , \ \alpha_3 + \beta_3 = \gamma_3 + \varsigma_2$$

$$(1.9b)$$

Equations (1.9b) assure that the length of each time substep is the same for both L and N. The numerical stability bounds for the model equation

$$u_t = i\lambda u - \nu u \qquad (1.10)$$

where $i\lambda u$ represents $N(u)$ and $-\nu u$ represents $L \cdot u$, are $\lambda \Delta t \leq \sqrt{3}$ and $\nu \Delta t < 47$, which were quite adequate for Spalart's purposes.

2. Compressible Navier-Stokes Algorithms

The development of compressible Navier-Stokes algorithms has also seen moments of inspiration in the last decade. We have taken several steps forward in the general development of algorithms. Some of these steps are via new concepts while most are the result of applying old concepts in a new setting. An example of a concept that was newly introduced to practical application in the field of fluid mechanics is the flux-vector splitting developed by Steger and Warming (1979). This opened up a wide new range of applications of "upwind" algorithms for the Euler and Navier-Stokes equations. Similar concepts have evolved since that time, most noticeably flux-difference splitting algorithms. Both of these methods have succeeded in removing much of the "fine tuning" of parameters which had plagued many algorithms previous to this time. A brief review of this work is given.

A concept that is probably as "new" as one can find is the total variation diminishing (TVD) theory extended to finite differencing schemes by Harten (1983). In this work, Harten extended ideas concerning total variation properties of scalar hyperbolic differential equations to discrete differencing schemes. This was an important step forward in determining the "ground rules" for designing good shock capturing methods, although it is not clear how religiously they need be followed. A complete review of this subject would be a formidable task by any measure. We chose not to do this, but rather to take some of the original underlying concepts involved and present a new perspective which hopefully will inspire new ideas.

Flux-Vector / Flux-Difference Splitting

In this section, we discuss two basic philosophies in the construction of upwind algorithms for systems of equations: flux-vector and flux-difference splitting. Each has proved to be a powerful technique in extending the upwind schemes for scalar equations to systems of equations. By the late 1970's, the theory for scalar hyperbolic equations was well established and several upwind schemes for these equations had appeared in the literature. The model nonlinear conservation equation

$$u_t + (f(u))_x = 0 \qquad (2.1)$$

had been analyzed extensively by Lax (1973) and others as an initial-value problem, yielding a fairly complete description of the equation and its solution. For smooth regions of initial data, (2.1) can be represented for a small time interval by its quasi-linear form

$$u_t + a(u)u_x = 0 \qquad a(u) = \frac{df}{du}$$

While at discontinuities, an integral form of (2.1) describes the solution behavior. The quasi-linear form has characteristic solutions for small time intervals of the form: $u(x,t) = u_0(x - at)$, i.e. the solution is constant along the characteristic lines. $\frac{dx}{dt} = a$. Upwind methods (more properly referred to as characteristic oriented methods) use this information by determining the local propagation direction, $\text{sgn}(a)$, and adapting

differencing stencils accordingly. One of the simplest upwind schemes using this strategy is the Cole-Murman scheme. This scheme can be written for the discrete mesh, $u_j^n = u(j\Delta x, n\Delta t)$, as

$$u_j^{n+1} - u_j^n + \frac{\Delta t}{\Delta x}(h_{j+\frac{1}{2}}^n - h_{j-\frac{1}{2}}^n) = 0$$

$$h_{j+\frac{1}{2}}^n = \frac{1}{2}(f_{j+1}^n + f_j^n) - \frac{1}{2}|\bar{a}|_{j+\frac{1}{2}}^n(u_{j+1}^n - u_j^n) \quad (2.2)$$

$$\text{where} \quad \bar{a}_{j+\frac{1}{2}} = \begin{cases} \frac{f_{j+1}-f_j}{u_{j+1}-u_j} & \text{if } u_j \neq u_{j+1} \\ a(u_j) & \text{otherwise} \end{cases}$$

This produces the following simple (and more recognizable) schemes for cases in which a is of uniform sign:

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}(f_j^n - f_{j-1}^n) \quad \text{if } a > 0$$

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}(f_{j+1}^n - f_j^n) \quad \text{if } a < 0$$

Obviously, if higher order accuracy is needed, then a more elaborate scheme needs to be constructed. But even for the simplest schemes (the Cole-Murman scheme for instance) one can ask the following question: *what is the simplest and most natural way to extend the scalar upwind schemes to systems of equations?* For the Euler equations, Steger and Warming (1979) and van Leer (1982) answered this question with flux-vector splitting while Roe (1981), Osher(1981) and others answered with a flux-difference splitting technique. To illustrate these methods, we consider the 1-D Euler equations

$$\mathbf{Q}_t + \partial_x \mathbf{E}(\mathbf{Q}) = 0 \quad (2.3)$$

Here $\mathbf{Q}$ is the vector of conserved variables for mass, momentum, and energy while $\mathbf{E}$ is the corresponding flux vector. Whenever needed, we assume the ideal-gas law as an equation of state.

The basic notion in flux-vector splitting is to split the flux vector into two parts

$$\mathbf{E} = \mathbf{E}^+ + \mathbf{E}^-$$

The components, $\mathbf{E}^-$ and $\mathbf{E}^+$, are to be chosen such that they can be forward and backward differenced, respectively. This choice is based on the assumption that if the individual vectors can be forward and backward differenced in a stable fashion, i.e., if

$$\mathbf{Q}_t + \mathbf{E}_x^- = 0 \quad \text{(stable for forward differencing)}$$

$$\mathbf{Q}_t + \mathbf{E}_x^+ = 0 \quad \text{(stable for backward differencing)}$$
$$(2.4a)$$

then the same differencing can be used for the full equation,

$$\mathbf{Q}_t + \mathbf{E}_x^+ + \mathbf{E}_x^- = 0 \quad (2.4b)$$

This turns out to be the case, albeit some reduction in stability characteristics may be encountered. For the van Leer splitting described below with first order explicit upwinding, van Leer (1982) mentions that this amounts to a limit of CFL ≤ 1 for (2.4a) and CFL $\leq \frac{2\gamma}{\gamma+3}$ for the full scheme, (2.4b).

Steger and Warming constructed a general class of flux-vector splittings for the Euler equations by exploiting the fact that the flux vectors are homogeneous of degree one in the conserved variables. Euler's identity then gives

$$\mathbf{E} = A\mathbf{Q} \quad \text{with} \quad A \equiv \frac{\partial \mathbf{E}}{\partial \mathbf{Q}} \quad (2.5)$$

To construct the splittings, they first diagonalized A,

$$X^{-1}AX = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix}$$

then split the diagonal matrix into nonnegative and nonpositive parts, i.e.

$$\Lambda = \Lambda^+ + \Lambda^- \quad (2.6)$$

They define the new flux vectors by using (2.5)

$$\mathbf{E}^\pm = X\Lambda^\pm X^{-1}\mathbf{Q} = A^\pm\mathbf{Q} \quad (2.7)$$

The splittings based on (2.6) are obviously not unique; Steger and Warming suggest several different splittings satisfying (2.6), of which probably the most frequently used is the " $\pm$ splitting " defined by

$$\Lambda^\pm = \frac{\Lambda \pm |\Lambda|}{2} \quad (2.8)$$

For this $\pm$ splitting they were able to determine that the resulting flux vectors had Jacobians, $\frac{\partial \mathbf{E}^+}{\partial \mathbf{Q}}$

and $\frac{\partial \mathbf{E}^-}{\partial \mathbf{Q}}$, with all positive and negative eigenvalues, respectively (at least for $1 \leq \gamma \leq \frac{5}{3}$). This is remarkable since nowhere in the development is any effort made to insure this. Unfortunately, this $\pm$ splitting leads to flux vectors that do not vary smoothly near sonic and stagnation points, even though the correct solution behaves smoothly there, and this produces "glitches" in the numerical solution. Several "fixes" have been proposed to remedy this, see Buning (1983).

Van Leer (1982) provided an alternate flux-vector splitting, which he devised using special Mach number polynomials to construct fluxes that remain smooth near stagnation and sonic points. His construction technique is quite different from that of Steger and Warming, in particular it is purely a vector construction (neither the Jacobian matrix nor its diagonalizing transforms is directly used). A reasonable question to ask is whether van Leer's flux vectors have an equivalent Steger-Warming representation via similarity transforms of A as in (2.7). We find that this is so by redefining new $\Lambda^\pm$ and direct calculation. In the case of van Leer's splitting they are given by

$$\Lambda^+ = \begin{bmatrix} \mu_1 & & \\ & \mu_2 & \\ & & \mu_3 \end{bmatrix} \quad , \quad \Lambda^- = \Lambda - \Lambda^+$$

with

$$\mu_1 = \frac{-((u-c)^2 - c^2(\gamma+1))(u+c)^2}{4c^3(\gamma+1)}$$

$$\mu_2 = \frac{(u-c)((\gamma-1)u + 2c)(u+c)^2}{4c^3(\gamma+1)}$$

$$\mu_3 = \frac{((\gamma-1)u^2 + (1-3\gamma)uc + 2(2\gamma+1)c^2)(u+c)^2}{4c^3(\gamma+1)}$$

In general, we find that these entries are of no particular uniform sign, (i.e. Λ^+ may have negative diagonal entries). This is not too surprising since the van Leer splitting only requires that the Jacobian matrices of the split fluxes, $\frac{\partial \mathbf{E}^+}{\partial \mathbf{Q}}$ and $\frac{\partial \mathbf{E}^-}{\partial \mathbf{Q}}$, have nonnegative and nonpositive eigenvalues, respectively. For illustration, we chose the state: $\rho = .9$, $u = .5$ and $c = 1.1$. In this case, the van Leer splitting gives: $\mu_1 = .5097, \mu_2 = -.2885, \mu_3 = 1.5098$, while the eigenvalues of $\frac{\partial \mathbf{E}^+}{\partial \mathbf{Q}}$ are calculated to be $0., .5795,$ and

1.6918. Thus it appears that defining splittings from (2.6) is certainly not a necessary condition. We have, in fact, considered other schemes which satisfy (2.6) yet fail to have eigenvalues of their Jacobians with signs consistent with (2.6). This is certainly an avenue for future investigation.

Flux-difference splitting has also provided a useful technique for extending scalar upwind algorithms to systems of equations. These methods use Riemann solvers to calculate the interaction of neighboring cells by the exact or approximate solution of Riemann's initial-value problem. The simplest explicit schemes for solving the Euler equations take the generic structure:

$$\frac{\mathbf{Q}_j^{n+1} - \mathbf{Q}_j^n}{\Delta t} + \frac{\mathbf{h}_{j+\frac{1}{2}} - \mathbf{h}_{j-\frac{1}{2}}}{\Delta x} = 0 \qquad (2.9)$$

where $\mathbf{h}_{j+\frac{1}{2}}$ is the numerical flux at the cell interface between the grid points j and $j+1$. The role of the local Riemann solver is to determine the numerical flux at every cell interface by examining the neighboring conditions. The best known approximate Riemann solvers are those of Roe (1981) and Osher and Solomon (1982). Roe's Riemann solver is particularly popular because of its simplicity. Roe considered the exact solution to the linearized form of (2.3),

$$\mathbf{Q}_t + A(\mathbf{Q}^L, \mathbf{Q}^R)\mathbf{Q}_x = 0 \qquad (2.10)$$

with constant left and right states specified as initial data,

$$\mathbf{Q} = \begin{cases} \mathbf{Q}^L & x < 0, \ t = 0 \\ \mathbf{Q}^R & x > 0, \ t = 0 \end{cases}$$

Here $x = 0$ corresponds to the local cell interface and A is the approximate Jacobian, obtained from a mean value linearization satisfying

$$\mathbf{E}^R - \mathbf{E}^L = A(\mathbf{Q}^L, \mathbf{Q}^R)(\mathbf{Q}^R - \mathbf{Q}^L) \qquad (2.11)$$

Equation (2.10) can be diagonalized, decoupled, solved exactly, then recoupled. This amounts to solving three linear (scalar) convection problems with step functions as initial data and constant convection velocities u, $u+c$, and $u-c$. Since the exact solution for each scalar problem is merely the translation in x of the original step function,

this results in a "shocks only" approximate Riemann solver; expansion fans, shocks, and contact discontinuities are all modelled as discontinuities. Unfortunately, this allows expansion shocks to form as solutions which must be precluded by special means (see Harten (1983) for examples).

From the local solution, the numerical flux at the cell interface can be calculated. If we construct A^+ and A^- as in (2.7) and (2.8), then the numerical flux can be written with reference to the left or right states as (details can be found in Roe (1981,1986))

$$\begin{aligned}\mathbf{h}(\mathbf{Q}^L, \mathbf{Q}^R) &= \mathbf{E}^L + A^-(\mathbf{Q}^L, \mathbf{Q}^R)(\mathbf{Q}^R - \mathbf{Q}^L) \\ &= \mathbf{E}^R - A^+(\mathbf{Q}^L, \mathbf{Q}^R)(\mathbf{Q}^R - \mathbf{Q}^L)\end{aligned}$$
$$(2.12)$$

Taking the average and applying the local solution everywhere on the discrete grid, we obtain the final form ($|A| = A^+ - A^-$)

$$\mathbf{h}_{j+\frac{1}{2}} = \frac{1}{2}(\mathbf{E}_{j+1} + \mathbf{E}_j) - \frac{1}{2}|A(\mathbf{Q}_j, \mathbf{Q}_{j+1})|(\mathbf{Q}_{j+1} - \mathbf{Q}_j)$$
$$(2.13)$$

If we again look at cases in which the local eigenvalues are of uniform sign (supersonic flow), we obtain the following conventional schemes

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n - \frac{\Delta t}{\Delta x}(\mathbf{E}_j^n - \mathbf{E}_{j-1}^n) \text{ if } [u, u-c, u+c] > 0$$

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n - \frac{\Delta t}{\Delta x}(\mathbf{E}_{j+1}^n - \mathbf{E}_j^n) \text{ if } [u, u-c, u+c] < 0$$

If we contrast this with the Cole-Murman scheme (2.2), which can also be viewed as using a "shocks only" scalar Riemann solver, we see that (2.13) is a successful extension of a scalar upwind scheme to systems.

We conclude this section by remarking that we have limited our discussion to 1-D inviscid flow. This is not really as restrictive as one might guess. Remarkable success has been attained by applying these ideas "dimension by dimension" to the two and three-dimensional Navier-Stokes equations, see Chakravarthy and Osher (1985) for some excellent examples.

Basics of TVD Schemes for Scalar Equations

In this section, we briefly mention the key elements used in the development of the TVD concept. More details as well as proofs can be found in the literature. The basic notion is to consider solutions, $u(x,t)$, of the single nonlinear conservation equation

$$u_t + (f(u))_x = 0, \qquad \frac{df}{du} = a(u) \qquad (2.14)$$

In this case, we make the usual assumption that the solutions of interest are entropy-satisfying weak solutions with convex flux functions. In the simplest case, to avoid boundary conditions, the initial value problem is considered in which the solution is specified along the x-axis, $u(x,0) = g_0(x)$, either in a periodic or compact supported fashion. This problem has been treated extensively by Lax (1973). The solution can be depicted in the $x - t$ plane by a series of converging and diverging characteristic straight lines. From the solution of (2.14), Lax provides the following observation: *the total increasing and decreasing variations of a differentiable solution between any pair of characteristics are conserved.* This means that in the absence of shocks the exact solution of (2.14) conserves the total variation of the initial data in time.

$$I(t + t_0) = I(t_0) \quad , \quad I(t) = \int_{-\infty}^{+\infty} \left| \frac{\partial u(x,t)}{\partial x} \right| dx$$
$$(2.15)$$

Moreover, in the presence of shock waves it can be shown that the total variation of the exact solution actually decreases in time (i.e. $I(t + t_0) \leq I(t_0)$). A simple heuristic argument for this decrease would be to consider solution data with a shock present, $u(x,t)$, and consider reconstructing the solution data at a previous time $u(x, t - \Delta t)$. But using characteristics, it becomes quickly obvious that this cannot be done uniquely; information (solution variation) has been irretrievably lost in the shock formation. An equally important result from Lax's observation comes from considering a monotonic solution between two nonintersecting characteristics: *between pairs of characteristics, monotonic solutions remain monotonic,* (i.e. no new extrema are created).

Although the properties described previously are those of the differential equation (2.14) and its solution, Harten developed a TVD criterion for numerical schemes by considering the discrete

form of (2.15) on a mesh $u_j^n = u(j\Delta x, n\Delta t)$, $\mathbf{u}[u_1, u_2, u_3, ...]$. The discrete total variation in this case is defined as

$$TV(\mathbf{u}) = \sum_{-\infty}^{+\infty} |u_{j+1} - u_j| \qquad (2.16)$$

with a corresponding TVD condition

$$TV(\mathbf{u}^{n+1}) \leq TV(u^n) \qquad (2.17)$$

It is not difficult to show that this TVD condition is sufficient for monotonic data with bounded total variation to remain monotonic, (to prove this, assume a new extremum is introduced and compute the new total variation). Although we will only use (2.17) to investigate conditions for constructing TVD schemes, equation (2.17), along with consistency of the scheme with the differential equation and satisfaction of the entropy inequality, is enough to guarantee convergence to the weak solution(s) (see Harten (1983)).

Equation (2.17) now provides us with an additional measure which will allow us to rule out many existing schemes which do not diminish the solution variation. More importantly, as we will see in the next section, it will be used to derive algebraic criteria which we can use to construct new TVD schemes.

Matrix Interpretation of TVD Criteria

Sufficient conditions for constructing TVD algorithms have been developed first by Harten (1983) and later in a more general form by Osher and Chakravarthy (1984), and Jameson and Lax (1984). In this section we demonstrate general sufficient conditions for TVD schemes. In developing the criteria for general explicit schemes, we independently followed a path similar to that of Jameson and Lax, although their claim of necessary and sufficient conditions is generally agreed to be in error (Harten (1986) notes that this is the danger of using their compact indicial notation). In the development of implicit schemes we depart from their strategy altogether and avoid the introduction of expansive operators. More importantly, we avoid the use of indicial notation in favor of a more compact matrix-vector notation whenever possible. As a result, the natural simplicity of constructing TVD schemes becomes evident, and we are able to give another

(and perhaps clearer) interpretation of sufficient conditions given by the previous authors.

An important step in the development of TVD schemes arises from the form chosen to express these schemes. We find it convenient to use a generalization of the form used by Osher and Chakravarthy. Since the objective is to obtain bounds on the variation of $\mathbf{u}$, the conservative difference schemes are put in a general form which uses an "apparent" $(p + q + 2)$ explicit and $(p' + q' + 2)$ implicit stencil of the solution, $\mathbf{u}$.

$$u_j^{n+1} + \sum_{i=-p'}^{q'} D(i)_{j+\frac{1}{2}} \Delta_{j+\frac{1}{2}+i} u^{n+1}$$

$$= u_j^n + \sum_{i=-p}^{q} C(i)_{j+\frac{1}{2}} \Delta_{j+\frac{1}{2}+i} u^n$$

$$(2.18)$$

where $\Delta_{j+\frac{1}{2}} u = u_{j+1} - u_j$. Because C and D are typically nonlinear functions of $\mathbf{u}$ at grid points which could be outside the apparent stencils, it should be clear that (2.18) is far from being unique. This nonuniqueness provides a large amount of freedom in designing schemes and is essentially the distinguishing feature of various schemes appearing in the literature. Although the algebraic details of putting a particular scheme into the form of (2.18) are important, we are only interested the general principles involved in the construction of TVD schemes and refer the reader to the literature for specific details.

We begin our analysis of TVD schemes by rewriting the discrete total variation in terms of the forward difference matrix, $\mathcal{D}$, (shown here for a periodic domain)

$$\mathcal{D} = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

from which $\|\mathcal{D}u\|_1 \equiv TV(u)$ and the TVD condition can be written

$$\|\mathcal{D}\mathbf{u}^{n+1}\|_1 \leq \|\mathcal{D}\mathbf{u}^n\|_1 \qquad (2.19)$$

In these expressions we are using the notation for the conventional L_1 vector norm, $\|\mathbf{v}\|_1 = \sum_j |v_j|$. Using the forward difference matrix, eq. (2.18) can be written

$$[I + \theta \widetilde{M} D]\mathbf{u}^{n+1} = [I - (1-\theta)MD]\mathbf{u}^n \quad (2.20)$$

Here M and $\widetilde{M}$ are matrix operators which can be nonlinear functions of $\mathbf{u}$. This equation, with the free parameter θ, represents various explicit and implicit forms of the evolution of $\mathbf{u}$ in time. (We chose this particular form so that if $\widetilde{M} = M$, then the scheme would be a generalized form of Harten's "linearized" implicit TVD scheme, Harten (1984).) One can also construct a scheme representing the timewise evolution of the variation, $D\mathbf{u}$. To do so multiply (2.20) from the left by D and regroup terms.

$$[I + \theta D\widetilde{M}]D\mathbf{u}^{n+1} = [I - (1-\theta)DM]D\mathbf{u}^n \quad (2.21)$$

Symbolically this can be expressed in terms of the matrix operators R and $\mathcal{L}$ as

$$\mathcal{L}\, D\mathbf{u}^{n+1} = R\, D\mathbf{u}^n \quad \text{or} \quad D\mathbf{u}^{n+1} = \mathcal{L}^{-1}R\, D\mathbf{u}^n \quad (2.22)$$

with

$$\mathcal{L} = [I + \theta D\widetilde{M}], \quad R = [I - (1-\theta)DM]$$

Next we take the L_1 vector norm of eq. (2.22) and apply the matrix-vector norm inequalities. Thus

$$\|D\mathbf{u}^{n+1}\|_1 \le \|\mathcal{L}^{-1}R\|_1 \|D\mathbf{u}^n\|_1 \quad (2.23)$$

and we find a sufficient condition for the scheme to be TVD is that $\|\mathcal{L}^{-1}R\|_1 \le 1$.

Note that for the extremely restrictive case in which $\mathcal{L}$ and R are not functions of $\mathbf{u}$, the basic definition of a matrix norm would guarantee that $\|\mathcal{L}^{-1}R\|_1 \le 1$ is both a necessary and sufficient condition for the scheme to be TVD. (Many monotone schemes would be included in this class.) Recall that the L_1 norm of a matrix is obtained by summing the absolute value of elements of individual columns of the matrix and

choosing the column whose sum is largest. Furthermore, we have the usual matrix norm inequality $\|\mathcal{L}^{-1}R\|_1 \le \|\mathcal{L}^{-1}\|_1 \|R\|_1$, so in the more general case, it is clear from (2.23) that it is sufficient to show that $\|\mathcal{L}^{-1}\|_1 \le 1$ and $\|R\|_1 \le 1$ (L_1 contractive). As we will see, these simple estimates will be enough to obtain the TVD criteria of previous investigators.

First consider the explicit operator R and multiply it from the left by the summation vector $s \equiv [1,1,1,...,1]$. It is clear that $sD = [0,0,0,...,0]$, so that that R has columns that sum to exactly unity, regardless of the particular choice of M. Because the L_1 norm of R is simply the maximum of the sum of absolute values of elements in the columns of R, it is obvious that a sufficient and necessary condition for $\|R\|_1 \le 1$ is for R to be a *nonnegative matrix*, (i.e. $R \ge 0$). Thus for explicit schemes ($\theta = 0$) to be TVD, we have the general sufficient condition that R be a nonnegative matrix. We illustrate that this leads to Harten's criteria for explicit schemes by considering his particular explicit form of (2.18), (in his notation)

$$u_j^{n+1} = u_j^n + C_{j+\frac{1}{2}}^{+}\Delta_{j+\frac{1}{2}}u^n - C_{j-\frac{1}{2}}^{-}\Delta_{j-\frac{1}{2}}u^n$$

The operator R in this case (again assuming a periodic domain) has the following banded structure

$$R = \begin{pmatrix} \ddots & \ddots & 0 & & 0 & & \ddots \\ \ddots & \ddots & C_{j+\frac{1}{2}}^{+} & & 0 & & 0 \\ 0 & C_{j-\frac{1}{2}}^{-} & 1 - C_{j+\frac{1}{2}}^{+} - C_{j+\frac{1}{2}}^{+} & C_{j+\frac{3}{2}}^{+} & 0 \\ 0 & 0 & -C_{j+\frac{1}{2}}^{-} & & & \ddots & \ddots \\ \ddots & 0 & & 0 & & \ddots & \ddots \end{pmatrix}$$

$$(2.24)$$

We need only require that this matrix be nonnegative to immediately arrive at Harten's criteria:

$$C_{j+\frac{1}{2}}^{+} \ge 0$$

$$C_{j+\frac{1}{2}}^{-} \ge 0$$

$$1 - C^+_{j+\frac{1}{2}} - C^-_{j+\frac{1}{2}} \geq 0 \qquad (2.25)$$

For the general form of $\mathcal{R}$, we can construct the matrix in the same fashion and arrive at the same conditions given by Harten, Jameson and Lax, and Osher and Chakravarthy by requiring that this resultant matrix be nonnegative.

Perhaps the more interesting use of a matrix interpretation comes when considering implicit schemes. Sufficient conditions would be to show that both $\mathcal{L}^{-1}$ and $\mathcal{R}$ are L_1 contractive. We have shown sufficient conditions for constructing $\|\mathcal{R}\|_1 \leq 1$. We now consider conditions for making $\|\mathcal{L}^{-1}\|_1 \leq 1$. From the previous development, one way to do this would be to show that $\mathcal{L}^{-1}$ is a nonnegative matrix with columns that sum to unity. At that point the development would be the same as previously discussed. This turns out to be a simple task and using some well known results from matrix theory, we can determine algebraic sufficient conditions on $\mathcal{L}$.

Note that in the following discussion, we assume that $\mathcal{L}$ is invertible, but after we have found a TVD criterion we will see that this must be so. First, we show that columns of $\mathcal{L}^{-1}$ *must* sum to unity. We use the same trick of premultiplying $\mathcal{L}$ by the summation vector, $\mathbf{s} = [1, 1, 1, ..., 1]$.

$$\mathbf{s}\mathcal{L} = \mathbf{s} \quad \rightarrow \quad \mathbf{s} = \mathbf{s}\mathcal{L}^{-1}$$

Therefore the columns of $\mathcal{L}^{-1}$ sum to unity. We need only find conditions on $\mathcal{L}$ to make its inverse nonnegative, but from matrix theory we know that a matrix whose inverse is nonnegative ($\mathcal{L}^{-1} \geq 0$) defines a *monotone matrix*. Therefore a sufficient condition would be that $\mathcal{L}$ is a monotone matrix. This is not particularly useful in itself, but a well known theorem from matrix theory allows us to develop a TVD criterion. Sufficient conditions for $\mathcal{L}$ monotone can be obtained from the theory for diagonally dominant M-matrices, a specific type of monotone matrix with positive diagonal entries and negative off-diagonal entries. To make this point clear we summarize a proof which appears in several books on matrix theory (see Lancaster, pp. 531-532 or Ortega, pp. 53-54). We begin by defining a real $n \times n$ matrix, a_{ij}, and assume that $a_{ii} > 0$ for each i and $a_{ij} \leq 0$ whenever $i \neq j$. If A is diagonally dominant, that is,

$$a_{ii} > \sum_{j=1, j \neq i}^{n} |a_{ij}|, \qquad i = 1, 2, ..., n,$$

then A is an M-matrix. To prove this, we first let $D = \text{diag}[a_{11}, a_{22}, a_{33}, ..., a_{nn}]$ and define $B = I - D^{-1}A$. Note that B has zero elements on the main diagonal and that $B \geq 0$. Also the fact that A is diagonally dominant implies that

$$\sum_{j=1}^{n} |b_{ij}| < 1, \qquad i = 1, 2, ..., n.$$

It follows immediately from Gersgorin's theorem that the maximum eigenvalue of B is less than one ($\mu_B < 1$). Now we have $D^{-1}A = I - B$, and $[I - B]^{-1}$ can be Neumann expanded into

$$[I - B]^{-1} = I + B + B^2 + B^3 + ...$$

Since $B \geq 0$, we conclude that $[I - B]^{-1} \geq 0$. It follows that $D^{-1}A$ is an M-matrix and consequently that A is an M-matrix.

Therefore, sufficient conditions for $\mathcal{L}$ to be monotone are that $\mathcal{L}$ be a diagonally dominant M-matrix, i.e. diagonally dominant with positive elements on the diagonal and negative off-diagonal elements. Also note that because of the diagonal dominance, we now can guarantee invertibility of $\mathcal{L}$ as mentioned earlier. Again, we can recover the results of other investigators from these conditions. We illustrate this using Harten's implicit form

$$u^{n+1}_j + D^+_{j+\frac{1}{2}} \Delta_{j+\frac{1}{2}} u^{n+1} - D^-_{j-\frac{1}{2}} \Delta_{j-\frac{1}{2}} u^{n+1} = u^n$$

In this case, $\mathcal{L}$ takes the general structure

$$\mathcal{L} = \begin{pmatrix} \ddots & \ddots & 0 & 0 & \ddots \\ \ddots & \ddots & -D^+_{j+\frac{1}{2}} & 0 & 0 \\ 0 & -D^-_{j-\frac{1}{2}} & 1+D^+_{j+\frac{1}{2}}+D^+_{j+\frac{1}{2}} & -D^+_{j+\frac{3}{2}} & 0 \\ 0 & 0 & -D^-_{j+\frac{1}{2}} & \ddots & \ddots \\ \ddots & 0 & 0 & \ddots & \ddots \end{pmatrix}$$

$$(2.26)$$

To obtain Harten's TVD criteria for the implicit scheme, we need only require that this be an M-matrix to obtain the following conditions, as did Harten

$$D^+_{j+\frac{1}{2}} \geq 0$$

$$D^-_{j+\frac{1}{2}} \geq 0$$

We conclude this section by noting the underlying conceptual simplicity. Once the schemes are placed in the form of (2.21), then sufficient conditions become very simple and naturally give rise to the basic concepts of nonnegative and M - matrices.

3. Concluding Remark

Looking back over the last ten years, we can see that ten years ago it would have been correct to say: "There will be considerable advances in algorithm development in the next decade. " We believe it is reasonably safe to make the same statement at this time.

4. References

Buning, P.: *Computation of Inviscid Transonic Flow Using Flux Vector Splitting in Generalized Coordinates*, PhD Thesis, Stanford University, Department of Aeronautics and Astronautics, 1983.

Chakravarthy, S. and Osher. S.: *Application of a New Class of High Accuracy TVD Schemes to the Navier-Stokes Equations*. AIAA paper 85-0165, Jan. 1985.

Harten, A.: *High Resolution Schemes for Hyperbolic Conservation Laws*, J. Comput. Phys., **49** (1983),357-393.

- - - : *On a Class of High Resolution Total - Variation - Stable Finite - Difference - Schemes*, SIAM J. Numer. Anal., **21** (1984),1-22.

- - - : *Numerical Methods for Hyperbolic Conservation Laws*, Lecture Notes: NASA Ames short course, Oct. 1-3, 1986.

Jameson, A. and Lax, P.: *Conditions for the Construction of Multi-Point Total Variation Diminishing Difference Schemes*, ICASE Report 178076, March 1986.

Lancaster, P.L.; and Tismenetsky M.: *The Theory of Matrices*, Academic Press Inc, 1985, pp. 531-532.

Lax, P.: *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM Regional Conf. Ser. Appl. Math., No. 11, Soc. Indust. Appl. Math, Philadelphia,Pa., 1973.

Van Leer, B.:*Flux-Vector Splitting for the Euler Equations*, ICASE Report 82-30, September, 1982.

Leonard, A. and Wray, A.: *A New Numerical Method for the Simulation of Three-Dimensional Flow in a Pipe*, NASA TM 84267. 1982.

Ortega, J.M.; and Rheinboldt, W.C.: *Iterative Solution of Nonlinear Equations in Several Variables*,First ed., Academic Press,Inc., 1970, pp. 53-54.

Osher S.; and Solomon, F.: *Upwind Difference Schemes Systems of Conservation Laws*, Math. of Comput.,**39**,339-374.

Osher, S.; and Chakravarthy, S.: *Very High Order Accurate TVD Schemes*, ICASE Report 84-44, September 1984.

Roe, P.L.: *Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes*, J. Comput. Phys., **43**, (1981), 357-372.

- - - : *Characteristic-Based Schemes for the Euler Equations*, Ann. Rev. Fluid Mech., **18**, (1986), 337-365.

Rogallo, Robert.: *Numerical Experiments in Homogeneous Turbulence*, NASA TM 81315, 1981.

Steger, J. and Warming R.:*Flux Vector Splitting of the Inviscid Gasdynamic Equations with Application to Finite-Difference Methods*, NASA TM 78605, 1979.