A97-32415

AIAA-97-2098

ROBUST EULER CODES

Timur Linde* and Philip L. Roe†
W. M. Keck Foundation Laboratory for Computational Fluid Dynamics
Department of Aerospace Engineering, The University of Michigan, Ann Arbor, MI 48109

Abstract

In this paper we review numerical schemes for the Euler equations that always preserve the positivity of density and pressure. We show that any one-dimensional first-order accurate positivity preserving scheme with reasonable time step restrictions can be always promoted to a multidimensional high-resolution positivity preserving scheme, also with reasonable restrictions on the time step. We further study the underlying mechanism of the loss of positivity and provide a general condition for a one-dimensional numerical scheme to be positivity preserving. This condition may become useful for analysis of complicated nonlinear schemes that prohibit analytical treatment.

I. Introduction

As computational fluid dynamics reaches maturity in a sense of being widely used in engineering design, issues such as reliability and robustness of computations become increasingly important. A robust code is one that produces reliable answers for arbitrary data with no retuning. Few codes achieve this without sacrificing accuracy. In particular, Euler codes often fail because either the density or the pressure in some cell becomes negative. This causes the sound speed in that cell to become imaginary. From a mathematical viewpoint, subsequent time steps do not correspond to a well-posed problem, and so computations must be stopped. This situation, for example, happens in situations where a flow at some initially high Mach number expands round a corner; then we expect that the pressure and density will fall to very low values, perhaps even to vacuum conditions. Small negative values may approximate the true solution within the truncation error of the scheme, but are nevertheless unacceptable. The issue is therefore distinct from that of accuracy.

*Doctoral Candidate, AIAA Member, linde@umich.edu
†Professor, AIAA Fellow, philroe@engin.umich.edu

Codes that are based on solving the equations in conservation form are particularly vulnerable, because the internal energy in a cell has to be computed as the difference between the (conserved) total energy and the kinetic energy. The latter is found from conserved values of mass and momentum. At high Mach numbers the internal energy appears as the small difference of two large quantities, and is therefore prone to large percentage errors.

Einfeldt *et al.*[1] introduced the term 'positively conservative' to refer to a conservative scheme that would, given physically meaningful data, predict positive density and pressure for all time. For one-dimensional flows they proved that the Godunov scheme[2] is positively conservative, but any Godunov-type scheme based on a linearized Riemann problem, for example Roe's scheme,[3] does not have this property. Neither does another elaborate scheme, Osher's scheme,[4] have this property since it is known to fail to compute interactions of strong shocks. On the other hand, it is not hard to prove that the Lax-Friedrichs scheme always preserves the positivity of density and pressure, but the dissipation properties of this scheme are far inferior to these of Roe's and Osher's schemes.

The apparent failure of early accurate approximate schemes to handle extreme flow conditions has added one more criterion to the design of an approximate Riemann solver - positive conservation. Today several first-order schemes qualify to be positively conservative in one space dimension. Among them are Einfeldt's modification (HLLE)[5] of the Harten-Lax-van Leer[6] (HLL) scheme, and the further modification (HLLEMR) derived in Charrier *et al.*[7] and Flandrin,[8] Liou's[9] recently reported AUSM$^+$ scheme, and Donat and Marquina's[10] combined Roe-Lax-Friedrichs scheme. Many gas-kinetic schemes have been shown to be positivity preserving (for example, see Khobalatte and Perthame,[11] Xu *et al.*,[12] Perthame and Shu,[13] Estivalezes and Villedieu,[14] and Tang and Xu[15]), some of them up to second order. However, extending a given one-dimensional first-order positively conservative scheme to more dimensions and better accuracy does not generally ap-

pear to be an obvious task. For example, Donat and Marquina[10] report that a straightforward higher order version of their scheme sometimes fails to compute strong rarefactions. Tang and Xu[15] note that an example can be constructed in which their elaborate collisional Boltzmann scheme does not preserve positivity. This conclusion is rather unfortunate since the corresponding positive collisionless Boltzmann scheme is fairly diffusive. This shows that the design of an accurate, efficient and robust scheme still remains an open issue.

It is important to stress here that positivity alone does not mean robustness. For example, the positivity of density and pressure preserves the meaning of physical entropy, but the maximum entropy principle, which is crucial for convergence studies, does not automatically follow from positivity. Moreover, it seems to be extremely difficult to prove this principle for high-resolution schemes.[11] Also, the positivity of a numerical scheme could mean its weak nonlinear stability since one could rigorously derive the maximum allowable time step. However, it was demonstrated by Quirk[16] that even entropy satisfying positive schemes do sometimes produce limited forms of instability, usually when shocks are almost aligned with the grid. Finally, the accuracy of computed solutions deserves special attention. At near-vacuum conditions, pressure is the difference of two large quantities, hence it may not be accurately computed in certain flow regions, although it is guaranteed to remain positive. This means that positive conservation does not ensure uniform accuracy. Whether this is acceptable may depend on the context. Nevertheless, it is highly desirable to preserve the physical admissibility of computed solutions, therefore positive conservation appears to be a good solid step in the quest for robust computations.

In this paper, following the analysis of Linde and Roe,[17] we will show that any one-dimensional first-order accurate positively conservative scheme can be promoted to more dimensions and formally to at least second-order accuracy. Provided that the time step limitations for the first order scheme are reasonable the time step requirements for its promoted analogue will also be reasonable. This simplifies positivity analysis to one dimension and first order. For one-dimensional schemes we will discuss the underlying mechanism of positivity loss and derive a positivity criterion that can be applied to any existing or future finite volume scheme.

II. Positive Conservation in Higher Dimensions

Consider the Euler equations in conservation form,

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{U}) = 0, \quad \mathcal{F} = (\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z), \qquad (1)$$

where

$$\mathbf{U} = \begin{pmatrix} \rho \\ m_x \\ m_y \\ m_z \\ E \end{pmatrix}, \ \mathbf{F}_x(\mathbf{U}) = \begin{pmatrix} m_x \\ m_x^2/\rho + p \\ m_x m_y/\rho \\ m_x m_z/\rho \\ m_x(E+p)/\rho \end{pmatrix}, \qquad (2)$$

and $\mathbf{F}_y(\mathbf{U})$ and $\mathbf{F}_z(\mathbf{U})$ are defined in a similar way. Here, ρ is the mass, $m_x = \rho u$, $m_y = \rho v$ and $m_z = \rho w$ are the momenta, and E is the total energy of a gas per unit volume. For ideal gases $p = (\gamma - 1)(E - |\mathbf{m}|^2/2\rho)$, where γ is the adiabatic constant.

The main tool in our work will be the analysis of dynamics of conserved gas states in the set of physically admissible states defined by

$$G = \left\{ \mathbf{U} \mid \rho > 0 \quad \text{and} \quad E - |\mathbf{m}|^2/2\rho > 0 \right\}. \qquad (3)$$

It is straightforward to show that G is an open convex cone, i.e. for any $\mathbf{U}_1, \mathbf{U}_2 \in G$ and any positive α_1 and α_2, $\alpha_1\mathbf{U}_1 + \alpha_2\mathbf{U}_2 \in G$ (note that this is more general than a convex linear combination of admissible states).

It is important to state clearly what we mean by positive conservation. We say a finite volume scheme is positively conservative if there exists a constant C, such that the scheme can update any physical initial data an arbitrary number of times with a CFL number not less than C.

We assume that a one-dimensional first-order accurate positively conservative Godunov-type scheme is available, and it is positive under a CFL-like condition $\Delta t < \Delta x/\lambda$, where λ is a scalar quantity related to the characteristic velocity which determines the allowable time step. Different schemes will have different values of λ. For example, we could take it to be equal to $\lambda(\mathbf{U}_L, \mathbf{U}_R) = \max\{|u_L| + a_L, |u_R| + a_R\}$, where a is the sound speed, which is a sensible choice of the characteristic velocity. Other than that we do no need to know any specific details about the scheme. We only make natural assumptions that the scheme is consistent and invariant with respect to coordinate transformations.

A. First-Order Schemes

Let us consider a typical first-order accurate finite volume scheme on an arbitrary computational grid. The state in cell i, which is surrounded by a set of neighbors ω_i, is updated according to

$$\mathbf{U}^i = \mathbf{U}_i - \frac{\Delta t}{V_i} \sum_{j \in \omega_i} \tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_i, \mathbf{U}_j) S_{ij}, \qquad (4)$$

where $\tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_i, \mathbf{U}_j)$ is the numerical flux normal to face ij, $\mathbf{n}_{ij}$ is the normal vector pointing from cell i to cell j, S_{ij} is the area of face ij, V_i is the volume of cell i, and Δt is the time step. Since $\sum_{j \in \omega_i} \mathbf{F}_{\mathbf{n}_{ij}}(\mathbf{U}_i) S_{ij} = 0$, then for any $\alpha_{ij}, j \in \omega_i$, which will be determined later, such that $0 < \alpha_{ij} < 1$ and $\sum_{j \in \omega_i} \alpha_{ij} = 1$ we can rewrite Equation 4 as follows,

$$\begin{aligned} \mathbf{U}^i &= \sum_{j \in \omega_i} \alpha_{ij} \mathbf{U}_i - \frac{\Delta t}{V_i} \sum_{j \in \omega_i} \Delta \mathbf{F}_{\mathbf{n}_{ij}} S_{ij} & (5) \\ &= \sum_{j \in \omega_i} \alpha_{ij} \left(\mathbf{U}_i - \frac{\Delta t S_{ij}}{\alpha_{ij} V_i} \Delta \mathbf{F}_{\mathbf{n}_{ij}} \right) & (6) \\ &= \sum_{j \in \omega_i} \alpha_{ij} T_{ij}^{-1} T_{ij} \left(\mathbf{U}_i - \frac{\Delta t S_{ij}}{\alpha_{ij} V_i} \Delta \mathbf{F}_{\mathbf{n}_{ij}} \right) & (7) \end{aligned}$$

where

$$\Delta \mathbf{F}_{\mathbf{n}_{ij}} = \tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_i, \mathbf{U}_j) - \mathbf{F}_{\mathbf{n}_{ij}}(\mathbf{U}_i). \qquad (8)$$

In the last step T_{ij} is an orthogonal matrix that defines a local rotation of state $\mathbf{U}$ from the coordinate frame $\{x, y, z\}$ to a face oriented frame $\{\mathbf{n}_{ij}, \tau_{1ij}, \tau_{2ij}\}$, i.e. $T_{ij} : (\rho, m_x, m_y, m_z, E)^T \mapsto (\rho, m_{\mathbf{n}_{ij}}, m_{\tau_{1ij}}, m_{\tau_{2ij}}, E)^T$, where τ_{1ij} and τ_{2ij} are two arbitrarily chosen tangential vectors.

Let us denote $\mathcal{U}_{ij} = T_{ij} \mathbf{U}_i$ and $\mathcal{U}_{ji} = T_{ij} \mathbf{U}_j$. Note that the orthogonality of the rotation matrix ensures that the new rotated states are physically admissible. Then using the rotational invariance of the flux function, i.e.

$$T_{ij} \tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_i, \mathbf{U}_j) = \tilde{\mathbf{F}}_x(\mathcal{U}_{ij}, \mathcal{U}_{ji}), \qquad (9)$$

we can write

$$\mathbf{U}^i = \sum_{j \in \omega_i} \alpha_{ij} T_{ij}^{-1} \mathcal{U}^{ij}. \qquad (10)$$

Therefore the new state is a superposition of states that could have arisen from some one-dimensional calculation on a grid with the equivalent cell size $\Delta x = \alpha_{ij} V_i / S_{ij}$,

$$\mathcal{U}^{ij} = \mathcal{U}_{ij} - \frac{\Delta t}{\Delta x} \left[\tilde{\mathbf{F}}_x(\mathcal{U}_{ij}, \mathcal{U}_{ji}) - \mathbf{F}_x(\mathcal{U}_{ij}) \right]. \qquad (11)$$

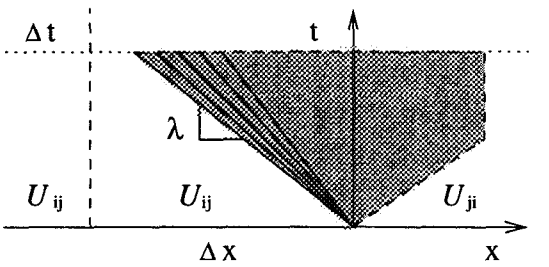


Figure 1: Equivalent one-dimensional problem.

Clearly, if all $\mathcal{U}^{ij}$ are physically admissible, then so is $\mathbf{U}^i$.

In the equivalent one-dimensional problem (see Figure 1) the left neighbor has the same state as the cell under consideration. This trivial Riemann problem that will be solved exactly by any consistent Riemann solver. Therefore $\mathcal{U}^{ij}$ will be physically admissible if

$$\Delta t < \frac{\Delta x}{\lambda(\mathcal{U}_{ij}, \mathcal{U}_{ji})}. \qquad (12)$$

If we recall the definition of λ, the above inequality becomes,

$$\Delta t \max(|u_{ni}| + a_i, |u_{nj}| + a_j) S_{ij} < \alpha_{ij} V_i. \qquad (13)$$

There are still unknown parameters α_{ij} in the above expression. To maximize the allowable time step we require that all inequalities in 13 are violated simultaneously. Then for each i we can add the inequalities to obtain

$$\Delta t < \frac{V_i}{\sum_{j \in \omega_i} \max(|u_{ni}| + a_i, |u_{nj}| + a_j) S_{ij}}, \qquad (14)$$

where we used $\sum_{j \in \omega_i} \alpha_{ij} = 1$. This gives us a CFL-like condition on the time step and proves that *given a first-order one-dimensional positively conservative scheme one can always build a first-order multidimensional positively conservative scheme for the Euler equations.*

B. Second-Order Schemes

Let us now consider a computational grid consisting of arbitrary convex cells. In order to improve both spatial and time resolution of a numerical scheme some reconstruction and time evolution techniques are needed. In the following we will consider a simple formally second-order accurate finite volume scheme.

In order to remain strictly conservative and keep the exact account of all variables we suggest to reconstruct the conservative variables. This eliminates the problem of recovery of the primitive variables from the conservative ones and vice versa discussed by van Leer.[18] From the analysis point of view, with this approach no interpolation related second-order error is introduced during the reconstruction step. Any reconstruction method (for example, see Barth[19]) may be used to compute the gradients of the conserved variables. For example, the very flexible least square approach can be used.

To prevent undesired spurious oscillations in numerical solutions and inhibit possible negative values of density and pressure that may appear due to linear reconstruction, a limiter must be applied to the reconstructed gradients. There is substantial freedom in which variables should be limited. Since extensive numerical experiments in the past have shown that limiting the primitive variables produces good results, we apply the limiter to the reconstructed *conserved* variables in such a way that the *primitive* variables are limited. We omit the details of the limiting procedure here; they can be found in Linde and Roe.[17] We only mention the main result that the positivity and nearly perfect monotonicity of the overall reconstruction step can be guaranteed by checking only the nodes of a computational cell.

As far as the time evolution step concerned, it is possible to obtain positivity results only for relatively simple convex time evolution schemes. Probably the simplest one of them is the classical Heun[20] scheme which is also known to be TVD stable.[21] This scheme can be written in the following form,

$$\mathbf{U}_i^* = \mathbf{U}_i - \Delta t \mathbf{Res}[\mathbf{U}, i] \quad (15)$$

$$\mathbf{U}^i = \frac{1}{2}\mathbf{U}_i + \frac{1}{2}(\mathbf{U}_i^* - \Delta t \mathbf{Res}[\mathbf{U}^*, i]), \quad (16)$$

where

$$\mathbf{Res}[\mathbf{U}, i] = \frac{1}{V_i}\sum_{j\in\omega_i} \tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_{ij}, \mathbf{U}_{ji}) S_{ij}, \quad (17)$$

Here, for cell i we have introduced its extrapolated face centroid states,

$$\mathbf{U}_{ij} = \mathbf{U}_i + \phi_i(\nabla\mathbf{U})_i \cdot (\mathbf{r}_{ij} - \mathbf{r}_i), \quad (18)$$

where $\mathbf{r}_i$ is the centroid of the cell, $\mathbf{r}_{ij}$ is the centroid of face ij, $(\nabla\mathbf{U})_i$ is the gradient matrix, and ϕ_i is the single scalar limiter. Since both stages of the Heun scheme contain an essentially the same operator, we only need to analyze the following generic equation,

$$\mathbf{U}^i = \mathbf{U}_i - \frac{\Delta t}{V_i}\sum_{j\in\omega_i} \tilde{\mathbf{F}}_{\mathbf{n}_{ij}}(\mathbf{U}_{ij}, \mathbf{U}_{ji}) S_{ij}. \quad (19)$$

As before, we would like to express the cell centroid state in terms of a convex linear combination of corresponding face centroid states, i.e. for each i we want to find a set of coefficients ξ_{ij}, $0 < \xi_{ij} < 1$, such that

$$\sum_{j\in\omega_i} \xi_{ij} = 1 \quad \text{and} \quad \sum_{j\in\omega_i} \xi_{ij}\mathbf{U}_{ij} = \mathbf{U}_i. \quad (20)$$

Note, since generally $\phi_i(\nabla\mathbf{U})_i \neq \mathbf{0}$, the choice of these coefficients is not arbitrary, and Equation 18 shows that we must require

$$\mathbf{r}_i = \sum_{j\in\omega_i} \xi_{ij}\mathbf{r}_{ij}. \quad (21)$$

Thus coefficients ξ_{ij} are related to the geometry of the cell. Unless the cell is a simplex, the above equation admits multiple solutions. However, Equation 21 relates the centroid of the cell to the centroids of its faces, therefore a very natural solution can be found. By dividing a two-dimensional cell into triangles with a common vertex at the centroid, or a three-dimensional cell into pyramids, we can show that all the requirements are satisfied by

$$\xi_{ij} = \delta_{dim}\frac{(\mathbf{r}_{ij} - \mathbf{r}_i)\cdot\mathbf{n}_{ij}S_{ij}}{V_i}, \quad (22)$$

where $\delta_{dim} = 1/\text{dimension}$. This formula shows why we require the convexity of the cells. If a cell is not convex, its centroid may not belong to the cell. Then some of its coefficients ξ_{ij} will be negative. However, second-order accuracy would be questionable in that cell anyway, and by setting the limiter to zero we could return to the first-order case for which the shape of the cell is irrelevant.

Now we will try to reduce the generic multidimensional problem to a set of one-dimensional ones, as it was done in the previous subsection. Let us choose some $\alpha_{ij} > 0$ and β_{ij} such that $\alpha_{ij} + \beta_{ij} = 1$ (note, we do not require $\beta_{ij} > 0$). The actual values of these coefficients will be determined later. Noticing that in general $\sum_{j\in\omega_i} \mathbf{F}_{\mathbf{n}_{ij}}(\mathbf{U}_{ij})S_{ij} \neq 0$ we can rewrite Equation 19 several times to obtain

$$\mathbf{U}^i = \sum_{j\in\omega_i} \xi_{ij}(\alpha_{ij} + \beta_{ij})\mathbf{U}_{ij} - \sum_{j\in\omega_i} \frac{\Delta t S_{ij}}{V_i}\left(\Delta\mathbf{F}_{\mathbf{n}_{ij}} + \mathbf{F}_{\mathbf{n}_{ij}}(\mathbf{U}_{ij})\right) \quad (23)$$

$$= \sum_{j\in\omega_i} \xi_{ij}\alpha_{ij}\left(\mathbf{U}_{ij} - \frac{\Delta t S_{ij}}{\xi_{ij}\alpha_{ij}V_i}\Delta\mathbf{F}_{\mathbf{n}_{ij}}\right) + \sum_{j\in\omega_i} \xi_{ij}\left(\beta_{ij}\mathbf{U}_{ij} - \frac{\Delta t S_{ij}}{\xi_{ij}V_i}\mathbf{F}_{\mathbf{n}_{ij}}(\mathbf{U}_{ij})\right), \quad (24)$$

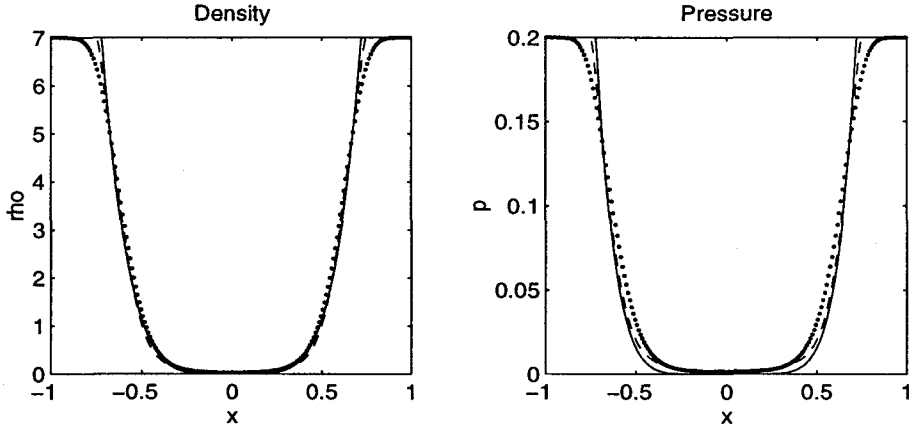


Figure 2: Double rarefaction problem. Solutions are — exact, ··· first-order and - - - second-order.

where $\Delta\mathbf{F}_{\mathbf{n}_{ij}}$ is defined as before, except now face centroid states are used in the definition. If we again introduce $\mathcal{U}_{ij} = T_{ij}\mathbf{U}_{ij}$ and $\mathcal{U}_{ji} = T_{ij}\mathbf{U}_{ji}$, we can complete the splitting,

$$\mathbf{U}^i = \sum_{j\in\omega_i} \xi_{ij}\alpha_{ij}T_{ij}^{-1}\mathcal{U}^{ij} + \sum_{j\in\omega_i} \xi_{ij}T_{ij}^{-1}\left(\beta_{ij}\mathcal{U}_{ij} - \frac{\Delta t S_{ij}}{\xi_{ij}V_i}\mathbf{F}_x(\mathcal{U}_{ij})\right), \quad (25)$$

where $\mathcal{U}^{ij}$ is again found from Equation 11 with $\Delta x = \xi_{ij}\alpha_{ij}V_i/S_{ij}$. If all terms under the summation signs in the last equation belong to the set of physically admissible states, the scheme will be positively conservative.

The first sum above contains the solutions to equivalent one-dimensional problems discussed above, therefore we immediately conclude that the time step should satisfy

$$\Delta t \max(|u_{nij}|+a_{ij}, |u_{nji}|+a_{ji})S_{ij} < \xi_{ij}\alpha_{ij}V_i. \quad (26)$$

The terms in the second sum can be computed exactly, and it is not hard to show (see the Appendix of Linde and Roe[17]) that their positivity will be preserved if

$$\Delta t\left(u_{nij} + \sqrt{\frac{\gamma-1}{2\gamma}}a_{ij}\right)S_{ij} < \xi_{ij}\beta_{ij}V_i. \quad (27)$$

To maximize the allowable time step we require that for a given face Inequalities 26 and 27 are violated simultaneously. Then we can add the inequalities to obtain

$$\Delta t < \frac{\xi_{ij}V_i}{\left(\lambda_{ij} + u_{\mathbf{n}_{ij}ij} + \sqrt{\frac{\gamma-1}{2\gamma}}a_{ij}\right)S_{ij}}, \quad (28)$$

where $\lambda_{ij} = \max(|u_{\mathbf{n}_{ij}ij}|+a_{ij}, |u_{\mathbf{n}_{ij}ji}|+a_{ji})$. At this point we no longer have any free parameters left, so, making use of Equation 22, we can finally express the time step in the following form,

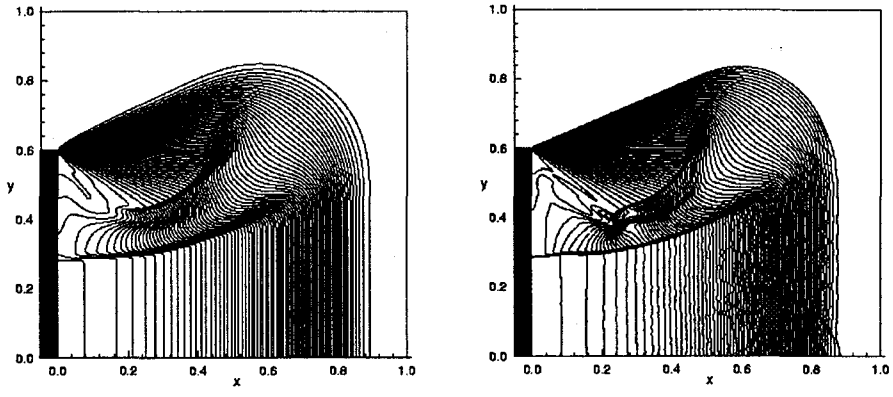
$$\Delta t < \frac{\delta_{dim}(\mathbf{r}_{ij}-\mathbf{r}_i)\cdot\mathbf{n}_{ij}}{\lambda_{ij} + u_{\mathbf{n}_{ij}ij} + \sqrt{\frac{\gamma-1}{2\gamma}}a_{ij}}. \quad (29)$$

This proves that *given a first-order one-dimensional positively conservative scheme and a computational grid consisting of convex cells we can always build a multidimensional formally second-order accurate scheme for the Euler equations.*

Similar conditions on the time steps for first and second-order schemes were obtained by Perthame and Shu.[13] On examining Equations 14 and 29 it becomes clear that the positivity requirements are not much more stringent than the requirements that come from a linear stability analysis.[19] Since the allowable time steps always remain finite, there is no limit to the time for which calculations can be continued. Thus we conclude that for a well-behaved one-dimensional positively conservative scheme multidimensional extensions do not come at significant cost.

C. Computational Results

To demonstrate the robustness of the proposed method we apply it to test problems with very extreme initial conditions. In all computations we set $\gamma = 1.4$ and use the HLLE flux function. The latter could, of course, be replaced with any other positively conservative flux function.



(a) First-Order

(b) Second-Order

Figure 3: Results for a $M = 2.5$ corner expansion at $t = 0.6$..

We first study a symmetric $M = 5$ double rarefaction wave problem. The initial distribution for this problem is $\rho_L = \rho_R = 7$, $u_L = -1$, $u_R = 1$, $p_L = p_R = 0.2$, and the y and z velocity components are set to zero. Note that the initial conditions are chosen to produce vacuum at $x = 0$. Figure 2 shows the resulting first- and second-order accurate distributions of density and pressure at $t = 0.6$ obtained on a computational grid with 200 points. Although both schemes can run at the maximum allowable time step, we find that accuracy is improved by setting the time step to half its maximum value. Both schemes produce a good approximation to the exact density distribution, but they noticeably overestimate the value of pressure in the middle of the expansion fan. This illustrates our comment that an admissible solution is not necessarily an accurate one.

We have also constructed an artificial test problem to illustrate the performance of our method in two dimensions. If an ideal gas having specific heat ratio of 1.4 diffract around a 90° corner in a Prandtl-Meyer fan at a Mach number greater than 2.56 it will form a vacuum. In our problem, an L-shaped region is initially filled with air traveling to the right at $M = 2.5$ (the initial conditions are $\rho = 7$, $u = 1$, $v = 0$, $p = 0.8$). This yields a transient flow that comes very close to cavitation. We prefer this test, despite its unrealistic initial data, to the commonly used alternative test in which a shock wave propagates along the upper branch into stationary air, because even in the limit of infinite pressure ration the flow induced behind the shock cannot be faster than $M = 1.89$. The present test is therefore more stringent. Figure 3 presents the results from the first- and the second-order schemes obtained on a 400x400 computational grid. In the calculations both pressure and density fall be three orders of magnitude and remain well-behaved. We do not believe that near the corner the percentage accuracy of pressure computation is very good, because at the corner the Mach number reaches the maximum value of only 10. This value is lower than the corresponding analytical estimate. Nevertheless, the scheme predicts correctly that pressure and density near the corner are very small, while preserving their positivity and without compromising the accuracy elsewhere.

III. Positive Conservation in One Dimension

The above discussion demonstrated that given a one-dimensional first-order positively conservative scheme it is not hard to promote the scheme to higher order and more dimensions. Thus the fundamentals of positive conservation can be studied almost entirely in one dimension. Therefore from this point on we will restrict ourselves to one-dimensional problems. For simplicity purpose only genuinely one-dimensional (no tangential slip) problems will be considered, although this simplification can be

omitted, and it is possible to extend all results to complete one-dimensional problems.

In order to perform multidimensional analysis we had to assume an underlying positively conservative scheme. However, the design of such a scheme is a difficult task by itself. Only a few one-dimensional first-order schemes have been proven to be positively conservative, and the proofs are sometimes involved and specific to a particular scheme. Some elaborate nonlinear schemes are so complicated that they simply do not admit analytical treatment. However, one would like to know their positivity properties. Hence it would be very useful to develop a general method for systematic positivity analysis. To do this one inevitably has to answer the question: *which schemes are positively conservative, and why do some fail*?

Let us recall that any finite volume scheme can be reduced to the equivalent one-dimensional problem described by Equation 11. We are going to be interested only in the dynamics of the left state. All the properties of the right state will be exactly the same due to the invariance of the scheme under coordinate transformations, in particular reflections. We can therefore treat the right state as a parameter.

Let us rewrite the equivalent one-dimensional problem in the following way,

$$\mathbf{U}^L = \mathbf{U}_L + \mathbf{H}(\mathbf{U}_L, \mathbf{U}_R, \sigma), \qquad (30)$$

where $\mathbf{U}_L$ and $\mathbf{U}_R$ are the initial left and right states, $\sigma = \Delta t/\Delta x$, and

$$\mathbf{H}(\mathbf{U}_L, \mathbf{U}_R, \sigma) = \sigma\left(\mathbf{F}(\mathbf{U}_L) - \tilde{\mathbf{F}}(\mathbf{U}_L, \mathbf{U}_R)\right). \qquad (31)$$

We can think of the scheme as of a nonlinear map: in the space of conserved states the new state is obtained from the old one by adding vector $\mathbf{H}$. To preserve positivity this map must produce a physically admissible state for any $\mathbf{U}_L \in G$, $\mathbf{U}_R \in G$ and $\sigma > 0$, i.e. the set of physically admissible states must be an invariant cone for the map.

Let us consider the set of physically admissible states in more detail. For one-dimensional flows this set is defined by $\rho > 0$ and $|m| < \sqrt{2\rho E}$. This set is shown in Figure 4. Notice that the whole boundary of the set corresponds to condition $p = 0$, while only one ray on its surface corresponds to $\rho = 0$. There is only way to produce an unphysical state - it is to cross this boundary somewhere. Since crossing it at a point where $\rho = 0$ is highly unlikely, it is clear that the point of crossing will always correspond to pressure, but usually not density, becoming negative; a conclusion which is definitely supported by numerical evidence. This shows that naming the positivity problem the problem of low densities, as it was done

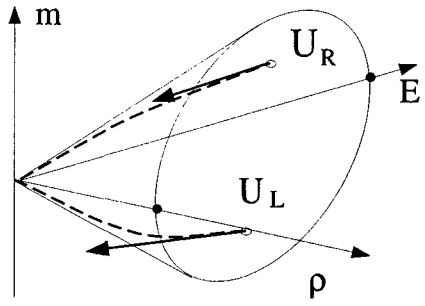


Figure 4: Dynamics of conserved states in the admissible cone. Dashed lines denote the physical trajectories, and the arrows show possible numerical motion of the states.

in Einfeldt *et al.*,[1] is a bit misleading. It would make more sense to call it the problem of low pressures (or even better temperatures).

Consider two initial states that belong to set G. For these two states one can solve the Riemann problem and find the trajectories of both states in the admissible cone. Only expansion waves can connect a physical state to vacuum (for example, see Liu and Smoller[22]), therefore only these waves are relevant to our analysis. If the initial states are such that[23]

$$u_L + \frac{2}{\gamma - 1}a_L < u_R - \frac{2}{\gamma - 1}a_R, \qquad (32)$$

the trajectories will intersect at the origin (an eventual vacuum), but do not otherwise lie on the surface unless they are initially placed on it. In a numerical calculation, the trajectory within one time step is always a straight line, and it must mimic the behavior of the exact trajectory by never crossing the boundary of the set of admissible states.

As we have mentioned above the new left state in the generic problem is obtained by adding vector $\mathbf{H}$ to the old state. If the new state happens to be in G, then no action has to be taken. It may, however, happen that the new state will lie outside of the set of physically admissible states. Then by reducing the time step, or equivalently σ, we can bring this state back into the set. This procedure can be repeated until one of the states reaches the boundary of set G. Then two possibilities exist: the state remains in G, or it tries to leave it. It the latter cases simple reduction of the time step does not help anymore, and so computations must be stopped. That this does indeed happen is known from numerical experience. For instance, for Roe's original scheme[3] applied to Sjögreen's[1] test problem there is a finite

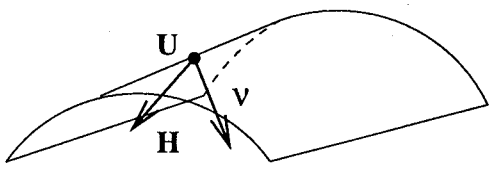


Figure 5: Whenever a conserved state reaches the boundary of the physically admissible set, it must not cross the boundary, i.e. vector H must point inward.

time until which calculations can be continued. No mater how small a time step is chosen, the scheme will fail after this time is reached. This shows that the lack of positivity for a numerical scheme is manifested in its inability to properly advance conserved states very close to the boundary of the set of admissible states.

The above argument suggests that studying the dynamics of conserved states near the boundary of set G is sufficient to deduce the positivity properties of a particular scheme. In order to do this we enlarge set G by including its boundary, i.e. we consider its closure $\overline{G}$. Let us put the left state on the boundary of $\overline{G}$ and denote this special choice of the state by $\mathbf{U}_L^0 = (\rho_L, \rho_L u_L, \frac{1}{2}\rho_L u_L^2)$. The boundary of G is a smooth surface, therefore the normal to this surface exists at every surface point except the origin. It is not hard to show that at point $\mathbf{U}_L^0$ the unit inward normal, which is basically the gradient of pressure, is given by

$$\nu(\mathbf{U}_L^0) = \left(\frac{u_L^2}{u_L^2+2}, -\frac{2u_L}{u_L^2+2}, \frac{2}{u_L^2+2}\right). \quad (33)$$

Clearly, the left state will not leave the set of physically admissible states if for any choice of the right state, $\mathbf{H}(\mathbf{U}_L^0, \mathbf{U}_R, \sigma)$ points inward (see Figure 5), that is if

$$\mathbf{H}(\mathbf{U}_L^0, \mathbf{U}_R, \sigma) \cdot \nu(\mathbf{U}_L^0) > 0. \quad (34)$$

If this condition is met, a non-zero σ can be found such that the new state does not become unphysical.

The equality sign in the above expression is also a possibility, and in that case the exact expression for $\mathbf{H}(\mathbf{U}_L^0, \mathbf{U}_R, \sigma)$ can be obtained. In fact, since $\overline{G}$ is a convex cone, the only way for $\mathbf{H}$ to be orthogonal to ν *and* at the same time belong to $\overline{G}$ is to lie along a ray connecting $\mathbf{U}_L^0$ to the vertex of the cone. Then we know that $\mathbf{H}(\mathbf{U}_L^0, \mathbf{U}_R, \sigma) = \alpha \mathbf{U}_L^0$, where α is some scalar number.

The exact flux corresponding to $\mathbf{U}_L^0$, $\mathbf{F}(\mathbf{U}_L^0) = u_L \mathbf{U}_L^0$, is orthogonal to $\nu(\mathbf{U}_L^0)$, therefore from Equation 31 it follows that the positivity condition 34 is equivalent to

$$\tilde{\mathbf{F}}(\mathbf{U}_L^0, \mathbf{U}_R) \cdot \nu(\mathbf{U}_L^0) < 0. \quad (35)$$

Let us denote $\tilde{\mathbf{F}} = (\tilde{f}_\rho, \tilde{f}_m, \tilde{f}_E)^T$. In this notation the above equation becomes

$$\frac{1}{2}u_L^2 \tilde{f}_\rho - u_L \tilde{f}_m + \tilde{f}_E < 0. \quad (36)$$

The function in this inequality depends on five parameters, and must hold for any combination of them. We can, however, use the properties of the Euler equations to reduce the dimensionality of the parameter space. Since the flux function in the Euler equations is a quasi-linear function of degree one, and the Riemann problem is self similar, any consistent numerical flux function must scale with density and velocity, i.e.

$$\frac{1}{\rho_R u_R}\left\{\tilde{f}_\rho, \frac{\tilde{f}_m}{u_R}, \frac{\tilde{f}_E}{u_R^2}\right\} = \text{fcn}\left(\frac{\rho_L}{\rho_R}, \frac{u_L}{u_R}, M_R\right). \quad (37)$$

Then it becomes clear that the positivity condition, which can be rewritten in the following form,

$$\frac{1}{2}\frac{u_L^2}{u_R^2}\frac{\tilde{f}_\rho}{\rho_R|u_R|} - \frac{u_L}{|u_R|}\frac{\tilde{f}_m}{\rho_R u_R^2} + \frac{\tilde{f}_E}{\rho_R|u_R|^3} < 0, \quad (38)$$

actually depends on only three non-dimensional parameters and the sign of u_R. Therefore without any loss of generally we can set $\rho_R = 1$, $|u_R| = 1$ and suggest the following

POSITIVITY CONDITION: Let $\tilde{\mathbf{F}}(\mathbf{U}_L, \mathbf{U}_R)$ be a one-dimensional numerical flux function for the Euler equations. Let $\mathbf{U}_L$ and $\mathbf{U}_R$ contain the following primitive data,

$$\rho_L = \rho, \quad u_L = u, \quad p_L = 0,$$
$$\rho_R = 1, \quad u_R = \pm 1, \quad p_R = p.$$

Then the corresponding Godunov-type finite volume scheme will be positively conservative if

$$\frac{1}{2}u^2 \tilde{f}_\rho - u\tilde{f}_m + \tilde{f}_E \leq 0$$

for any

$$0 < \rho < \infty,$$
$$-\infty < u < \infty,$$
$$0 < p < \infty,$$

with the equality sign holding only if

$$\tilde{\mathbf{F}}(\mathbf{U}_L, \mathbf{U}_R) = \alpha \mathbf{U}_L,$$

where α is some scalar number.

The positivity problem thus reduces to studying the sign of only one functional form defined on a three-dimensional parameter space. In some simple cases this can be done analytically. The real use for the above condition, however, comes when the flux function is complicated enough to prohibit analytical treatment. In that case finding the global maximum of the form over the parameter space will be sufficient to conclude whether the flux function produces a positively conservative scheme. Even if one has to scan the whole parameter space to find the maximum, doing so is not going to be a major difficulty with modern computers.

The parameter space may be further reduced for the class of schemes whose flux function has the following property,

$$\tilde{\mathbf{F}}(\mathbf{U},\mathbf{V}) = \lim_{\varepsilon\to 0}\left(\tilde{\mathbf{F}}(\varepsilon\mathbf{U},\mathbf{V})+\tilde{\mathbf{F}}(\mathbf{U},\varepsilon\mathbf{V})\right) \quad (39)$$

$$= \lim_{\varepsilon\to 0}\left(\tilde{\mathbf{F}}(\varepsilon\mathbf{U},\mathbf{V})+\varepsilon\tilde{\mathbf{F}}(\frac{1}{\varepsilon}\mathbf{U},\mathbf{V})\right). \quad (40)$$

For this class of numerical schemes the flux function is a positive linear combination of two special limiting cases, therefore only these two cases require analysis. This eliminates the need to scan the full range of ρ in the condition; instead it is sufficient to check only the limits $\rho \to 0$ and $\rho \to \infty$.

All flux vector splitting schemes, in particular simple gas-kinetic schemes, possess the above property. Another subclass of schemes that have this property are the schemes that can be written in conventional form,[24]

$$\tilde{\mathbf{F}}(\mathbf{U},\mathbf{V}) = \frac{\mathbf{F}(\mathbf{U})+\mathbf{F}(\mathbf{V})}{2} - \frac{Q(\mathbf{U},\mathbf{V})}{2}(\mathbf{V}-\mathbf{U}), \quad (41)$$

with the dissipation matrix Q satisfying

$$Q(\alpha\mathbf{U},\mathbf{V}) = Q(\mathbf{U},\alpha\mathbf{V}) = Q(\mathbf{U},\mathbf{V}). \quad (42)$$

For instance, the Lax-Friedrichs scheme belongs to this subclass. Finally, it is not hard to see that Donat and Marquina's[10] flux function must satisfy Equation 40 since its characteristic flux splitting is not sensitive to scaling of either conservative state. Most of these schemes are known to have a substantial amount of dissipation, and they are generally positively conservative. Also, the special form of their flux function substantially simplifies analysis, therefore analytical results exist for many of them.[11,14,15]

Unfortunately, the positivity condition does not provide us with any information about the maximum allowable time step, which in certain situations may turn out to be fairly small. One has to know the details of a particular flux function to find this information. Nevertheless, the condition is already interesting because it can provide an insight into whether such detailed analysis is necessary at all. Also, it may justify the practice of reducing the time step in some crash-prone situations.

Another interesting application of the positivity condition is to revisit some existing schemes and see whether previously unknown facts about them can be uncovered. For example, we have applied the condition to Roe's original scheme and readily found that it does not pass the test, therefore the scheme is not positively conservative. However, to our surprise we have discovered that Roe's scheme with Harten's entropy fix[24] ($\delta = \max(0, 4(a_R - a_L))$) passes the tests, and for a small enough time step it has no problem computing Sjögreen's[1] and even worse test problems in which the vacuum state appears. Thus for one-dimensional problems with no tangential slip Roe's scheme with Harten's entropy fix *is* actually positively conservative. Unfortunately, the scheme loses this property when tangential slip is allowed, although one can speculate that applying the fix to the contact wave could restore the positivity of the scheme.

To conclude this section we want to mention that an extension of this method to full one-dimensional problems is very similar. In the general case Equation 35 must be simply replaced with

$$\tilde{\mathbf{F}}(\mathbf{U}_L^0,\mathbf{U}_R)\cdot\nabla p(\mathbf{U}_L^0) < 0. \quad (43)$$

In this case set G belongs to a five-dimensional conservative variable space, and the above inequality involves a function of nine parameters. One can show that with the above scaling arguments and a suitable choice of the coordinate system the parameter space can be reduced to five dimensions (or to four if the flux function satisfies Equation 40). This space, however, is still quite large, therefore it would be useful to find a subspace of the "worst cases" in this space. More work in this direction needs to be done to determine whether this is possible.

IV. Conclusions and Future Work

In this paper we presented an extension of first-order one-dimensional positively conservative schemes for the Euler equations to more dimensions and higher order. We have proven that a multidimensional positively conservative scheme can be designed in a fairly straightforward way pro-

vided that a one-dimensional positively conservative scheme with reasonable requirements on the time step is available. By reasonable we mean that the time step remains finite for any choice of conserved states, even if some of them lie very close to the boundary of the physically admissible sets. In this case the resulting restrictions on the time step for a multidimensional scheme are also going to be reasonable and, in fact, comparable to the linear stability requirements. In this sense positive conservation is analogous to weak nonlinear stability. This conclusion is a direct consequence of the convexity of the set of physically admissible conservative states. In the constructed examples we have demonstrated that the derived schemes do indeed preserve the positivity of density and pressure, although this property may be accompanied by some local loss of accuracy of the computed solutions.

With the understanding that positive conservation can be studied in one dimension, we have begun the task of constructing tools for analyzing and ensuring the positivity of one-dimensional schemes. Based on the geometry of the Euler equations, we have proposed a general positivity condition that can be applied to any finite volume scheme, in particular the ones whose complexity prohibits analytical studies. Unfortunately, the positivity condition does not provide any information about the maximum allowable time step, which is a serious drawback from the practical point of view. We believe that additional details about a particular scheme are needed to obtain such information. Nevertheless, the condition appears very useful in uncovering the nature of the loss of positivity

Although we have formulated the positivity condition as a necessary condition, there are many reasons to believe that it also is a sufficient condition. To prove this claim one would have to show that for any sequence of iterations the time step does not converge to zero for a positively conservative scheme, i.e. that the characteristic speed λ remains finite in the limit $p \rightarrow 0$. Since expansion regions generally correspond to smooth solutions, this result could follow from the consistency of the scheme. From this it would also follow that a consistent positively conservative scheme must be able to solve the Burgers equation for velocity in the limit of vanishing pressure. Some of our numerical experiments suggest that this does indeed happen.

In one dimension we have been able to reduce analysis to the problem of global maximum over a three-dimensional parameter space. This problem can be studied using modern optimization techniques. Since flux functions are usually fairly complicated, it is hard to say *a priori* where this maximum is located for a particular flux function. Moreover, it can be shown that for some schemes the maximum can be achieved at more than one point. This shows that searching for the maximum may not generally be trivial. Therefore, to reduce the amount of work it may be worth looking for the region of the "worst cases" in the parameter space. Physical intuition tells us that this region may turn out to be universal for all sensible schemes. Clearly, more work needs to be done to determine whether our intuition serves us right.

Acknowledgments. This work was supported by a Doctoral Fellowship from the François-Xavier Bagnoud Foundation.

References

[1]Einfeldt, B., Munz, C. D., Roe, P. L., and Sjögreen, B., "On Godunov-Type Methods near Low Densities", *J. Comput. Phys.*, Vol. 92, pp. 273–295, 1991.

[2]Godunov, S. K., "A Difference Scheme for Numerical Computation of Discontinuous Solutions of Hydrodynamic Equations", *Mat. Sb.*, Vol. 47, No. 3, pp. 271–306, 1959, (in Russian).

[3]Roe, P. L., "Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes", *J. Comput. Phys.*, Vol. 43, pp. 357–372, 1981.

[4]Osher, S., and Solomon, F., "Upwind Difference Schemes for Hyperbolic Systems of Conservation Laws", *Math. Comput.*, Vol. 38, No. 158, pp. 339–374, April 1982.

[5]Einfeldt, B., "On Godunov-Type Methods for Gas Dynamics", *SIAM J. Numer. Anal.*, Vol. 25, No. 2, pp. 294–318, April 1988.

[6]Harten, A., Lax, P. D., and van Leer, B., "On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws", *SIAM Rev.*, Vol. 25, No. 1, pp. 35–61, January 1983.

[7]Charrier, P., Dubroca, B., and Flandrin, L., "Un solveur de Riemann approché pour l'étude d'écoulements hypersoniques bidimensionnels", *C.R. Acad. Sci. Paris*, Vol. 317, pp. 1083–1086, 1993.

[8]Flandrin, L., *Méthodes "cell-centered" pour l'approximation des équations d'Euler et de Navier-Stokes sur des maillages non structurés*, PhD thesis, l'Universite Bordeaux I, December 1995.

[9]Liou, M., "A Sequel to AUSM: AUSM$^+$", *J. Comput. Phys.*, Vol. 129, No. 2, pp. 364–382, December 1996.

[10]Donat, R., and Marquina, A., "Capturing Shock Reflections: An Improved Flux Formula", *J. Comput. Phys.*, Vol. 125, No. 1, pp. 42–58, April 1996.

[11]Khobalatte, B., and Perthame, B., "Maximum Principle on the Entropy and Second-Order Kinetic Schemes", *Math. Comput.*, Vol. 62, No. 205, pp. 119–131, January 1994.

[12]Xu, K., Martinelli, L., and Jameson, A., "Gas-Kinetic Finite Volume Methods, Flux-Vector Splitting, and Artificial Diffusion", *J. Comput. Phys.*, Vol. 120, No. 1, pp. 48–65, September 1995.

[13]Perthame, B., and Shu, C., "On Positive Preserving Finite Volume Schemes for Compressible Euler Equations", *Numer. Math.*, Vol. 73, pp. 119–130, 1996.

[14]Estivalezes, J. L., and Villedieu, P., "High-Order Positivity-Preserving Kinetic Schemes for the Compresible Euler Equations", *SIAM J. Numer. Anal.*, Vol. 33, No. 5, pp. 2050–2067, October 1996.

[15]Tang, T., and Xu, K., "Gas-Kinetic Schemes for the Compressible Euler Equations I: Positivity-Preserving Analysis", submitted to *SIAM J. Numer. Anal.*

[16]Quirk, J. J., "A Contribution to the Great Riemann Solver Debate", *Int. J. Numer. Methods Fluids*, Vol. 18, No. 6, pp. 555–574, 1994, Also ICASE Technical Report 92-64.

[17]Linde, T. J., and Roe, P. L., "On Multidimensional Positively Conservative High-Resolution Schemes", In *ICASE Workshop on Barriers and Challenges in Computational Fluid Dynamics*, Hampton, VA, August 1996. Proceedings to be published by Kluwer Academic Publishers.

[18]van Leer, B., "Towards the Ultimate Conservative Difference Scheme. V. A Second-Order Sequel to Godunov's Method", *J. Comput. Phys.*, Vol. 32, pp. 101–136, 1979.

[19]Barth, T. J., "On Unstructured Grids and Solvers", In *Computational Fluid Dynamics*, Lecture Series 1990-03. VKI, March 1990.

[20]Gear, C. W., *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[21]Shu, C., and Osher, S., "Efficient Implementation of Essentially Non-Oscillatory Shock-Capturing Schemes", *J. Comput. Phys.*, Vol. 77, No. 2, pp. 439–471, August 1988.

[22]Liu, T., and Smoller, J. A., "On the Vacuum State for the Isentropic Gas Dynamics Equations", *Adv. Appl. Math.*, Vol. 1, pp. 345–359, 1980.

[23]Courant, R., and Friedrichs, K. O., *Supersonic Flow and Shock Waves*, Interscience Publishers, New York, 1948.

[24]Harten, A., and Hyman, J. M., "Self-Adjusting Grid Methods for One-Dimensional Hyperbolic Conservation Laws", *J. Comput. Phys.*, Vol. 50, pp. 235–269, 1983.