RESIDUAL DISTRIBUTION SCHEMES FOR CONSERVATION LAWS VIA ADAPTIVE QUADRATURE

RÉMI ABGRALL* AND TIMOTHY BARTH†

Abstract. This paper considers a family of nonconservative numerical discretizations for conservation laws which retains the correct weak solution behavior in the limit of mesh refinement whenever sufficient order numerical quadrature is used. Our analysis of 2-D discretizations in nonconservative form follows the 1-D analysis of Hou and Le Floch [14]. For a specific family of nonconservative discretizations, it is shown under mild assumptions that the error arising from nonconservation is strictly smaller than the discretization error in the scheme. In the limit of mesh refinement under the same assumptions, solutions are shown to satisfy an entropy inequality. Using results from this analysis, a variant of the "N" (Narrow) residual distribution scheme of van der Weide and Deconinck [31] is developed for first-order systems of conservation laws. The modified form of the N-scheme supplants the usual exact single-state mean-value linearization of flux divergence, typically used for the Euler equations of gasdynamics, by an equivalent integral form on simplex interiors. This integral form is then numerically approximated using an adaptive quadrature procedure. This renders the scheme nonconservative in the sense described earlier so that correct weak solutions are still obtained in the limit of mesh refinement. Consequently, we then show that the modified form of the N-scheme can be easily applied to general (non-simplicial) element shapes and general systems of first-order conservation laws equipped with an entropy inequality where exact mean-value linearization of the flux divergence is not readily obtained, e.g. magnetohydrodynamics, the Euler equations with certain forms of chemistry, etc. Numerical examples of subsonic, transonic and supersonic flows containing discontinuities together with multi-level mesh refinement are provided to verify the analysis.

Key words. Residual Distribution, Fluctuation Splitting, Symmetric Hyperbolic, Entropy Symmetrization

AMS subject classifications. 35L02, 65M02, 65K02, 76N02

1. Motivations. Discrete conservation has become a standard design criteria in the development of numerical discretization techniques for conservation laws that admit discontinuous solutions. From the Lax-Wendroff theorem [21], the ingredients of consistency, stability, and discrete conservation yield convergent approximations of conservation laws in *divergence form* for both smooth and discontinuous solutions that are valid weak solutions in the sense of distribution theory. Even so, the development of stabilized numerical discretizations often also utilizes the *quasilinear form* (a.k.a. nonconservative form) of the conservation law system to approximate simple or plane wave solutions for use in upwind stabilization mechanisms. As we will illustrate later, the use of quasilinear forms is often at odds with the requirement of discrete conservation unless specialized mean-value linearized variants of the discrete quasilinear form are used. As a practical matter, obtaining simple expressions for these mean-value linearizations in closed form is often extremely complicated or even impossible. In addressing this difficulty, our goal is to develop a general framework that avoids these complications while still insuring that valid weak solutions of the conservation law system are obtained in the limit of mesh refinement.

As a motivating example, consider the scalar Cauchy problem in one space dimension and time

$$\begin{cases} u_{,t} + (f(u))_{,x} = 0 & \text{for } (x,t) \in \mathbb{R} \times \mathbb{R}^+ \\ u(x,0) = u_0(x) \end{cases} \tag{1.1}$$

with $u \in \mathbb{R}$ and $f(u) : \mathbb{R} \mapsto \mathbb{R}$. In this equation $u_0(x)$ is assumed to be periodic or compactly supported data. Let $\Delta x_{j+1/2} = x_{j+1} - x_j$ denote a general nonuniform partitioning of space so that u_j represents the numerical approximation $u(x_j, t)$. Next, consider the prototype conservative semi-discrete scheme

$$\frac{du_j}{dt} + \frac{h_{j+1/2} - h_{j-1/2}}{\Delta x_j} = 0 \tag{1.2}$$

with $h_{j\pm 1/2}$ the numerical flux. This prototype scheme is conservative in space due to the mutual telescoping of numerical fluxes. A first order accurate upwind scheme is easily obtained via the flux function

$$h_{j+1/2}(u_j, u_{j+1}) = \frac{1}{2}\left(f(u_j) + f(u_{j+1})\right) - \frac{1}{2}|a|_{j+1/2}\left(u_{j+1} - u_j\right) \tag{1.3}$$

*Mathématiques Appliquées de Bordeaux, Université Bordeaux I, 351 Cours de la Libération, 33 405 Talence cedex, France (abgrall@math.u-bordeaux.fr)

†NASA Ames Research Center, Information Sciences Directorate, Moffett Field, California, 94035-1000 USA (barth@nas.nasa.gov).

with $a_{j+1/2}$ an approximation of the flux Jacobian df/du at $x_{j+1/2}$. Observe that whenever the exact mean-value linearizations are used, e.g.

$$f(u_{j+1}) - f(u_j) = \langle a \rangle_{j+1/2} \, (u_{j+1} - u_j) \tag{1.4}$$

so that $a_{j\pm1/2} = \langle a \rangle_{j\pm1/2}$, the first order upwind scheme can be written equivalently as

$$\frac{\partial u_j}{\partial t} + \langle a \rangle^-_{j+1/2} \frac{u_{j+1} - u_j}{\Delta x_{j+1/2}} + \langle a \rangle^+_{j-1/2} \frac{u_j - u_{j-1}}{\Delta x_{j-1/2}} = 0 \ . \tag{1.5}$$

Note that this discretization is nonconservative in space unless the exact mean-value linearization (1.4) is used. Nonconservative schemes of this form are known to converge to incorrect weak solutions. More precisely, Hou and Le Floch [14] have shown (in 1-D) that if the nonconservative scheme (1.5) converges, it converges to a solution of

$$u_{,t} + (f(u))_{,x} = \mu$$

where μ is Borel measure source term that is expected to be zero in the regions where u is smooth and concentrated where u is not smooth. The construction of an exact mean-value linearization is readily accomplished in 1-D by the general path integration

$$\begin{aligned} f(u_B) - f(u_A) &= \int_{u_A}^{u_B} d\, f = \int_{u_A}^{u_B} a(u)\ du \\ &= \int_{\xi_A}^{\xi_B} a(u(\xi))\ \frac{du}{d\xi}\ d\xi \ . \end{aligned} \tag{1.6}$$

Without loss of generality, one can restrict $u(\xi)$ to the space of polynomials, P_k. A particularly convenient choice are P_1 linear polynomials since

$$\begin{aligned} f(u_B) - f(u_A) &= \int_{\xi_A}^{\xi_B} a(u(\xi))\ \frac{du}{d\xi}\ d\xi \Big|_{u(\xi)\in P_1} \\ &= \int_{\xi_A}^{\xi_B} a(u(\xi)) d\xi \Big|_{u(\xi)\in P_1} \left(\frac{u_B - u_A}{\xi_B - \xi_A} \right) \end{aligned} \tag{1.7}$$

so that the following mean-value speed is obtained

$$\langle a \rangle (u_A, u_B) = \frac{1}{\xi_B - \xi_A} \int_{\xi_A}^{\xi_B} a(u(\xi))\ d\xi \Big|_{u(\xi)\in P_1} \ . \tag{1.8}$$

In Harten, Lax, and van Leer [13] this expression is interpreted as an integration in state space parameterized along the line $\pi u(\xi) = u_A + \xi\ (u_B - u_A),\ \xi \in [0,1]$.

$$\langle a \rangle (u_A, u_B) = \int_0^1 a(\pi u(\xi)) d\xi \ .$$

When the locations A and B are not coincident, one can equivalently interpret this as an integration in physical space assuming the P_1 Lagrange interpolant

$$\pi u(x) = u_A + \frac{x - x_A}{x_B - x_A}\ (u_B - u_A), \quad x \in [x_A, x_B]$$

so that $\xi = \frac{x - x_A}{x_B - x_A}$ and

$$\langle a \rangle (u_A, u_B) = \frac{1}{x_B - x_A} \int_{x_A}^{x_B} a(\pi u(x))\ dx \ . \tag{1.9}$$

This latter interpretation is useful since it generalizes the mean-value contruction to simplices and more arbitrary regions. Next consider an approximation of Eqn. (1.9) using NQ-point numerical quadrature

$$\langle a \rangle (u_A, u_B) = \sum_{l=1}^{NQ} \omega_l \; a(\pi u(q_l)) + R_{NQ+1} \tag{1.10}$$

where ω_l are quadrature weights, q_l quadrature positions, and R_{NQ+1} is the numerical remainder term. This renders the scheme (1.5) nonconservative in space. In later sections, we derive (under suitable assumptions) the same result as Hou and LeFloch and are able to characterize more precisely the Borel measure μ. In particular, the dependency of μ with respect to the number of quadrature points is given. If an adequate number of quadrature points is taken, the error terms due to nonconservation are shown to be comparable or smaller than the discretization error of the scheme. In addition, a discrete entropy inequality is formally obtained in the limit of mesh refinement. From the practical point of view, these results are important with the following consequences

- Exact mean-value linearization is no longer needed. This is useful when solving systems of conservation laws for which exact mean-value linearizations are not known in closed form, e.g. magnetohydrodynamics, Euler equations with certain forms of chemistry, etc.
- General finite element shapes are permitted, e.g. tetrahedra, hexahedra, prisms, pyramids. Previous exact mean-value linearizations in closed form have been restricted exclusively to simplex shapes.

The new nonconservative formulation suggests an adaptative strategy, whereby the number of quadrature points depends on the local smoothness of the numerical solution. This strategy is undertaken in Sect. 3.

2. Background. In this section, we briefly review a number of well known constructions and analytical results that we utilize later in the development and analysis of our nonconservative formulations.

2.1. Conservation Laws and Symmetric Hyperbolic Forms. Consider the Cauchy problem for m coupled conservation laws in d space dimensions and time

$$\begin{cases} \mathbf{w}_{,t} + \sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w})_{,x_i} = 0 \quad \text{for } (x,t) \in \mathbb{R}^d \times \mathbb{R}^+ \\ \mathbf{w}(x,0) = \mathbf{w}_0(x) \end{cases} \tag{2.1}$$

where $\mathbf{w} \in \mathbb{R}^m$ denotes the vector of conserved variables and $\mathbf{f}(\mathbf{w}) : \mathbb{R}^m \mapsto \mathbb{R}^{m \times d}$ a flux vector. In addition, Eqn. (2.1) is assumed to be equipped with a convex entropy extension so that the additional scalar inequality holds

$$H_{,t} + \sum_{i=1}^{d} G^i_{,x_i} \leq 0 \tag{2.2}$$

with $H(\mathbf{w}) : \mathbb{R}^m \mapsto \mathbb{R}$ the convex entropy function and $G(\mathbf{w}) : \mathbb{R}^{m \times d} \mapsto \mathbb{R}^d$ the entropy flux vector for the system. Solutions of Eqn. (2.1) satisfying (2.2) are generally of two types [22]:

- (Classical Solutions) Smooth solutions satisfying the quasilinear form of Eqn. (2.1)

$$\mathbf{w}_{,t} + \sum_{i=1}^{d} A_i(\mathbf{w}) \; \mathbf{w}_{,x_i} = 0, \quad A_i(\mathbf{w}) \equiv \mathbf{f}^i_{,\mathbf{w}} \; . \tag{2.3}$$

  As part of the symmetrization theory for first-order conservation laws developed by Godunov [11], Mock [23] and others, it is known that the existence of a convex entropy extension insures that the quasilinear form (2.3) is symmetrizable via a change of variables $\mathbf{w} \mapsto \mathbf{v}$ where $\mathbf{v} = H^T_{,\mathbf{w}} \in \mathbb{R}^m$ denotes the so-called entropy variables for the system. As consequences of symmetrization theory, performing the change of variables

$$\tilde{A}_0 \mathbf{v}_{,t} + \sum_{i=1}^{d} \tilde{A}_i \mathbf{v}_{,x_i} = 0 \; , \tag{2.4}$$

yields the matrix $\tilde{A}_0 \equiv \mathbf{w}_{,\mathbf{v}} = (H_{,\mathbf{w},\mathbf{w}})^{-1}$ symmetric positive definite and the matrices $\tilde{A}_i \equiv \mathbf{f}^i_{,\mathbf{v}} = A_i \tilde{A}_0$ symmetric. For brevity, the functional dependency of the matrices A_i and $\tilde{A}_i$ has been omitted. Motivated by the energy analysis given in subsequent sections, we assume the basic solution unknowns are the entropy $\mathbf{v}$-variables so that the shorthand notation $A_i(\mathbf{w})$ should be interpreted as $A_i(\mathbf{w}(\mathbf{v}))$. It is useful for later developments to define the real-valued matrix combination $A(\mathbf{w},\omega) \equiv \omega_i A_i(\mathbf{w}), \quad \omega \in \mathbb{R}^d$ and similarly the symmetric matrix $\tilde{A}(\mathbf{w},\omega) \equiv \omega_i \tilde{A}_i(\mathbf{w})$. Observe that symmetry of $\tilde{A}(\mathbf{w},\omega)$ implies that $A(\mathbf{w},\omega)$ has m real eigenvalues, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ and m real eigenvectors $\mathbf{r}_k(\mathbf{w},\omega) \in \mathbb{R}^m$ satisfying the standard eigenvalue problem

$$A(\mathbf{w},\omega)\,\mathbf{r}_k(\mathbf{w},\omega) = \lambda_k(\mathbf{w},\omega)\,\mathbf{r}_k(\mathbf{w},\omega), \quad k = 1,2,\ldots,m$$

since the identity

$$\tilde{A}_0^{-1/2} A(\mathbf{w},\omega) \tilde{A}_0^{1/2} = \tilde{A}_0^{-1/2} \tilde{A}(\mathbf{w},\omega) \tilde{A}_0^{-1/2}$$

shows that $A(\mathbf{w},\omega)$ is similar to a real-valued symmetric matrix.

Our keen interest in the quasilinear form (2.3) comes from its use in the construction of upwind discretizations such as variants of Godunov's method [10] utilizing approximate Riemann solvers [32, 26] and the multi-dimensional fluctuation splitting scheme described in the following section. Specifically, the quasilinear form (2.3) admits nonlinear simple wave solutions of the following form for a given direction vector ω:

$$\mathbf{w}(x,t) = \sum_{k=1}^{m} \alpha_k \boldsymbol{\mathcal{W}}^k(\sigma(\omega \cdot x - \lambda_k(\boldsymbol{\mathcal{W}}^k,\omega)\ t)) \tag{2.5}$$

where $\boldsymbol{\mathcal{W}}^k(\sigma,\omega) \in \mathbb{R}^m$ satisfies the differential relation

$$\frac{d\boldsymbol{\mathcal{W}}^k}{d\sigma} = \mathbf{r}_k(\boldsymbol{\mathcal{W}}^k(\sigma,\omega),\omega) \tag{2.6}$$

for the self-similar real-valued parameter σ. In Eqn. (2.5), $\alpha_k \in \mathbb{R}$ are expansion coefficients to be determined by matching initial data. When the matrix $A(\omega)$ is assumed locally independent of $\mathbf{w}$, then local plane wave solutions are obtained. Historically, mean-value linearized variants of the quasilinear form (2.3) have been used in 1-D to construct approximate Riemann solutions [26] for eventual use in upwind discretizations. In Sect. 2.2, we consider a multi-dimensional upwinding strategy which also uses plane wave information originating from a mean-value linearized form of Eqn. (2.3).

- (Discontinuous Solutions) Weak solutions of the divergence form (2.1) satisfying a jump condition on space-time hypersurfaces, $\mathcal{S}$, with space-time normal vector $\hat{\mathbf{n}} = (n_t, \mathbf{n}^T)^T$

$$n_t[\mathbf{w}]^+_- + \sum_{i=1}^{d} \mathbf{n}_i[\mathbf{f}^i]^+_- = 0 \tag{2.7}$$

with $[\arg((x,t))]^+_- = \lim_{\epsilon \downarrow 0} (\arg((x,t)_{\mathcal{S}} + \epsilon\,\hat{\mathbf{n}}) - \arg((x,t)_{\mathcal{S}} - \epsilon\,\hat{\mathbf{n}}))$. In Sect. 3, a Lax-Wendroff-like theorem is presented which addresses the convergence to weak solutions of a family nonconservative discretizations using approximate mean-value linearization.

Note that in the remainder of the paper, the notation $\|\cdot\|$ will denote a pointwise norm over m variables unless otherwise indicated. When the argument is dimensionally comparable with the $\mathbf{v}$–variables, the natural norm is *not* the standard Euclidian norm $||x|| = \sqrt{\sum_{i=1}^m x_i^2}$ but rather the dimensionally consistent matrix norm [15]

$$||x||^2_{\tilde{A}_0} \equiv x^T \tilde{A}_0 x \tag{2.8}$$

where $\tilde{A}_0$ is the inverse of the Hessian matrix of the entropy, $\tilde{A}_0 = (H_{,\mathbf{w},\mathbf{w}})^{-1}$.

2.2. The Residual Distribution Scheme on Simplicies. In the remaining sections, we assume a triangulation $\mathcal{T}_h$ in $\mathbb{R}^d$ of a polygonal spatial domain Ω composed of nonoverlapping simplices T_i, $\Omega = \cup T_i$, $T_i \cap T_j = \emptyset$, $i \neq j$. A simplex T in $\mathbb{R}^d$ is uniquely described by $d+1$ vertices $T(M_1, M_2, \ldots, M_{d+1})$. For purposes of analysis, the triangulation is assumed to be shape regular with maximum simplex diameter h. From the triangulation $\mathcal{T}_h$, the geometric dual tessellation $\mathcal{C}_h$ is constructed by connecting gravity centers of the simplices and the mid-points of the edges as shown in Fig 2.1. In this figure, the dual cell C_i surrounds the triangulation vertex M_i. We also define piecewise linear and piecewise constant spaces on the tessellations

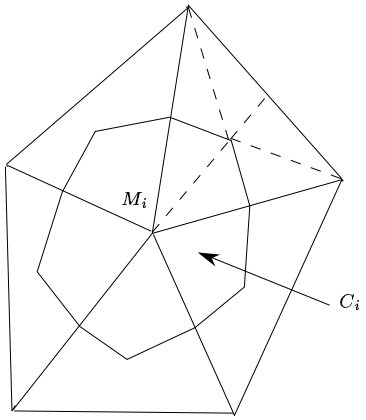


FIG. 2.1. *Dual cell C_i associated with triangulation vertex M_i in $\mathbb{R}^2$*

$\mathcal{T}_h$ and $\mathcal{C}_h$ respectively

$$\mathcal{V}_h = \{\mathbf{v}_h; \mathbf{v}_h \in C^0(\mathbb{R}^d)^m, \mathbf{v}_{h|T} \in (\mathcal{P}_1)^m, \forall\ T \in \mathcal{T}_h\}$$

$$\mathcal{X}_h = \{\mathbf{v}_h; \mathbf{v}_{h|C} \in (\mathcal{P}_0)^m, \forall\ C \in \mathcal{C}_h\}\ .$$

Let $\mathbf{V}_i \in \mathbb{R}^m$ denote the nodal degrees of freedom located at M_i which uniquely describes $\mathbf{v}_h$ in both spaces $\mathcal{V}_h$ and $\mathcal{X}_h$. For example, if $N_i(x)$ denotes the standard piecewise linear basis function for triangulation $\mathcal{T}_h$ such that $N_i(M_j) = \delta_{ij}$, then for $\mathbf{v}_h \in \mathcal{V}_h$

$$\mathbf{v}_h(x) = \sum_{M_i \in \mathcal{T}_h} N_i(x)\, \mathbf{V}_i\ .$$

Similarly, if $\chi_i(x)$ denotes the characteristic function for the dual cell $C_i \in \mathcal{C}_h$,

$$\chi_i(x) = \begin{cases} 1 & x \in C_i \\ 0 & x \notin C_i \end{cases}$$

then for $\mathbf{v}'_h \in \mathcal{X}_h$

$$\mathbf{v}'_h(x) = \sum_{M_i \in \mathcal{T}_h} \chi_i(x)\, \mathbf{V}_i\ .$$

Finally, for brevity of notation, we shall write $\mathbf{w}_h \equiv \mathbf{w}(\mathbf{v}_h)$ and $\mathbf{W}_i \equiv \mathbf{w}(\mathbf{V}_i)$ to denote the corresponding conserved variable forms.

Using these definitions, we can state the simplest prototype residual distribution scheme (explicit in time) used in discretizing (2.1).

Residual Distribution Scheme: For all $M_i \in \mathcal{T}_h$ and $n \geq 0$

$$\begin{cases} \mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \dfrac{\Delta t}{|C_i|} \displaystyle\sum_{T, M_i \in T} \mathbf{\Phi}_{i,T}^n \\ \mathbf{W}_i^0 \quad = \mathbf{w}_0(M_i) \end{cases} \tag{2.9}$$

where $\mathbf{\Phi}_T \in \mathbb{R}^m$ represents a discretization of the negated time evolution term integrated in the simplex T

$$\mathbf{\Phi}_T \equiv -\int_T (\mathbf{w}_h)_{,t}\, dx \tag{2.10}$$

with $\mathbf{\Phi}_{i,T}$ an as yet unspecified sum decomposition of $\mathbf{\Phi}_T$ among the $d+1$ vertices of the simplex T in $\mathbb{R}^d$

$$\mathbf{\Phi}_T = \mathbf{\Phi}_{1,T} + \cdots + \mathbf{\Phi}_{d+1,T} \ . \tag{2.11}$$

In the special case of Eqn. (2.1), $\mathbf{\Phi}_T$ is expressed equivalently in terms of the spatial flux divergence

$$\mathbf{\Phi}_T = \mathbf{\Phi}_{1,T} + \cdots + \mathbf{\Phi}_{d+1,T} := \int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}_h)_{,x_i} \right) dx \ . \tag{2.12}$$

The residual distribution scheme encompasses a number of well known weighted residual methods that have residual decompositions which reduce to following form for P_1 linear elements:

$$\mathbf{\Phi}_{i,T} = \left(\frac{1}{d+1} + \overline{\tau}_{i,T} \right) \mathbf{\Phi}_T \tag{2.13}$$

where $\overline{\tau} \in \mathbb{R}^{m \times m}$ is an nonsingular matrix such that $\sum_{i=1}^{d+1} \overline{\tau} = 0$. Some examples of weighted residual methods for solving (2.1) include

- The streamline diffusion method of Johnson and coworkers [17, 18].
- The streamline upwind Petrov-Galerkin (SUPG) and Galerkin least-squares finite element methods of Hughes and coworkers [15, 16].
- The cell vertex finite volume methods of Ni [25] and Morton *et al.* [24, 7]

The residual distribution formula also describes a family of monotone and positive coefficient schemes for scalar conservation laws due to Roe [27, 28] and Deconinck [8] and the system extension due to van der Weide and Deconinck [31] described later in Sect. 5. Fundamental to these residual distribution schemes is the mean-value linearization of the flux divergence formula (2.12).

$$\int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w})_{,x_i} \right) dx = \sum_{i=1}^{d} \langle A \rangle_i \int_T \mathbf{w}_{,x_i}\, dx \tag{2.14}$$

To facilitate this calculation, we follow the 1-D example of Sect. 1 by introducing an auxilliary mapping $\mathbf{z}(\mathbf{v}) : \mathbb{R}^m \mapsto \mathbb{R}^m$ and restricting $\mathbf{z}$ to the space of piecewise linear Lagrange interpolants denoted by $\pi_h \mathbf{z}$.

$$\int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}(\pi_h \mathbf{z})_{,x_i} \right) dx = \sum_{i=1}^{d} \langle A \rangle_i \int_T \mathbf{w}(\pi_h \mathbf{z})_{,x_i}\, dx \tag{2.15}$$

For the Euler equations of gas dynamics, the choice $\mathbf{z} = (\sqrt{\rho}, \sqrt{\rho}\vec{V}, \sqrt{\rho} H_t)^T$ with ρ the fluid density, $\vec{V}$ the fluid velocity, and H_t the fluid total enthalpy yields closed form expressions for the mean-value linearization of the flux divergence [30]. A striking property of this linearization is that the linearized system

$$\mathbf{w}_{,t} + \sum_{i=1}^{d} \langle A \rangle_i \mathbf{w}_{,x_i} = 0 \tag{2.16}$$

is hyperbolic. Unfortunately, no other $\mathbf{z}$ variable is known to give simple and closed form formulas leading to a hyperbolic linearized system. In addition, this approach is limited to simplices, while in many applications hexahedral brick meshes would be desirable for accuracy reasons.

From a theoretical and practical point-of-view, there is motivation to work directly with the entropy variables since the corresponding mean-value linearized form

$$\int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w})_{,x_i} \right) dx = \sum_{i=1}^{d} \langle \tilde{A} \rangle_i \int_T \mathbf{v}_{,x_i}\, dx \tag{2.17}$$

would necessarily produce a hyperbolic linearized system due to the symmetry of $\langle \tilde{A} \rangle_i$. Using symmetric forms, we also show in subsequent analysis the satisfaction of an entropy inequality in the limit of mesh refinement. Our general strategy is to utilize a piecewise linear representation of the entropy variables themselves so that $\mathbf{v}_h \in \mathcal{V}_h$ and

$$\int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}(\mathbf{v}_h))_{,x_i} \right) \, dx = \sum_{i=1}^{d} \langle \tilde{A} \rangle_i \int_T (\mathbf{v}_h)_{,x_i} \, dx = |T| \sum_{i=1}^{d} \langle \tilde{A} \rangle_i (\mathbf{v}_h)_{,x_i} \tag{2.18}$$

with

$$\langle \tilde{A} \rangle_i \equiv \frac{1}{|T|} \int_T \tilde{A}_i(\mathbf{v}_h) \, dx \ . \tag{2.19}$$

Following the 1-D motivational example of Sect. 1, Eqn. (2.19) is approximated by quadrature formula so that componentwise

$$\langle \tilde{A} \rangle_i \equiv \sum_{l=1}^{NQ} \omega_l \tilde{A}_i(\mathbf{v}_h(q_l)) + R_{NQ+1} \ . \tag{2.20}$$

Since $\mathbf{v}_h|_T \in P_1(T)$, Eqn. (2.18) has used the fact that the gradient component are constant within a simplex. Consequently, the quadrature formula used in (2.20)

$$\int_T H(x) \, dx = |T| \sum_{l=1}^{NQ} \omega_l \, H(q_l) + O(h^{k+1}) \tag{2.21}$$

should at least be exact for $H(x) \in P_k(T)$ and $k > 1$. In addition, the $O(h^{k+1})$ error is assumed to have the following behavior for use in later analysis: there exists C independent of the simplex T such that

$$O(h^{k+1}) \leq C(\mathcal{T}_h) \, \frac{h^{k+1}}{(k+1)!} \int_T ||D^{k+1} H(x)|| \, dx \tag{2.22}$$

where h is the maximum diameter of the T, $C(\mathcal{T}_h)$ is a geometrical parameter that only depends on $\mathcal{T}_h$, and

$$D^k H(x) = \left\{ \frac{\partial^\alpha H}{\partial x^\alpha}, \ |\alpha| = k \right\} \ .$$

Note that the use of numerical quadrature permits generalization of the techniques to non-simplicial elements, e.g. brick elements (Q), using the form

$$\mathbf{\Phi}_Q = |Q| \sum_{l=1}^{NQ} \omega_q \left(\sum_{i=1}^{d} \tilde{A}_i(\mathbf{v}_h(q_l))(\mathbf{v}_h(q_l))_{,x_i} \right) + R_{NQ+1} \ . \tag{2.23}$$

Hence, we are interested in residual distributive schemes that fullfil the approximate conservation relation

$$\mathbf{\Phi}_{1,T} + \cdots + \mathbf{\Phi}_{d+1,T} = |T| \sum_{l=1}^{NQ} \omega_q \left(\sum_{i=1}^{d} \tilde{A}_i(\mathbf{v}_h(q_l))(\mathbf{v}_h(q_l))_{,x_i} \right) \ . \tag{2.24}$$

In Sect. 5, a particular residual scheme known as the N-scheme is considered [30] as generalized to systems of conservation laws by van der Weide and Deconinck [31]. This system N-scheme assumes an exact mean-value linearization via the parameter vector. We then propose a variant of the system N-scheme which utilizes a piecewise linear space consisting of the entropy variables and approximates the mean-value linearization via quadrature. Analyzing this new scheme for systems of conservation laws, we show that in the limit of mesh refinement that numerical solutions satisfy an entropy inequality. We then show a similar result for the system N-scheme when the linearization is approximated via quadrature.

3. Weak–$*$ Convergence, a Lax-Wendroff Result. Consider the numerical scheme (2.9). The nodal variables $\mathbf{W}_i^n$ are assumed to map uniquely via $\mathbf{v}(\mathbf{w})$ and $\mathbf{w}(\mathbf{v})$ to and from $\mathbf{V}_i^n$ which are the degrees of freedom in the spaces $\mathcal{V}_h$ and $\mathcal{X}_h$ at time $t_n \equiv n\Delta t, n \in [0, N]$. In addition, the as yet unspecified residual decomposition $\mathbf{\Phi}_{i,T}^n$ and $\mathbf{V}_i^n$ are assumed to satisfy the following conditions :

ASSUMPTION 1 (H1). *Let $\mathcal{T}_h$ be a shape regular triangulation. For $C \in \mathbb{R}$ and any fixed n, there exists $C'(C) \in \mathbb{R}$ which depends on the triangulation $\mathcal{T}_h$ such that $\forall\ \mathbf{v}_h^n \in \mathcal{V}_h$ and $||\mathbf{v}_h^n||_{L^\infty(\mathbb{R}^d)^m} \leq C$*

$$\left|\left|\mathbf{\Phi}_{i,T}^n\right|\right| \leq C'\ h^{d-1} \sum_{M_j \in T} \left|\left|\mathbf{V}_j^n - \mathbf{V}_i^n\right|\right|, \quad \forall T \in \mathcal{T}_h \text{ and } \forall M_i \in T\ . \tag{3.1}$$

This is a continuity assumption on the residual decomposition in a simplex T in terms of the local nodal values of $\mathbf{V}_i^n, M_i \in T$. In particular, whenever $\mathbf{v}_h^n$ is constant in T we then require that $\mathbf{\Phi}_{i,T}^n = 0$.

ASSUMPTION 2 (H2). *For all $\mathbf{v}_h^n \in \mathcal{V}_h$ and fixed n*

$$\mathbf{\Phi}_T^n = \sum_{i=1}^{d+1} \mathbf{\Phi}_{i,T}^n = |T| \sum_{l=1}^{NQ} \omega_l \left(\sum_{i=1}^{d} \widetilde{A}_i(\mathbf{v}_h^n(q_l))\ (\mathbf{v}_h^n(q_l))_{,x_i} \right), \quad q_l \in T \tag{3.2}$$

where $\widetilde{A}_i = \mathbf{f}_{,\mathbf{v}}^i$ and NQ denotes the number of quadrature points. In addition, the quadrature error in the flux divergence calculation is assumed to be of the following form for all n and a given integer $k > 1$

$$\left\| \mathbf{\Phi}_T^n - \int_T \sum_{i=1}^{d} \mathbf{f}_{,x_i}^i(\mathbf{v}_h^n)\ dx \right\| \leq C(\mathcal{T}_h) \frac{h^{k+1}}{(k+1)!} \left\| D_x^{k+1} \left(\sum_{i=1}^{d} \widetilde{A}_i(\mathbf{v}_h^n)\ (\mathbf{v}_h^n)_{,x_i} \right) \right\| \tag{3.3}$$

by using a sufficient number of quadrature points.

REMARK 1.

(i) *Observe that $\mathbf{v}_h \in \mathcal{V}_h$ is C^0 continuous, consequently for neighboring simplices sharing a common spatial edge, $\Gamma_{jk} = \{x \mid \partial T_j \cap \partial T_k \neq \emptyset\}$,*

$$\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{v}_h^n(x))|_{T_j} \cdot \vec{n}_i^{jk} = \sum_{i=1}^{d} \mathbf{f}^i(\mathbf{v}_h^n(x))|_{T_k} \cdot \vec{n}_i^{jk}, \quad x \in \Gamma_{jk} \tag{3.4}$$

where $\vec{n}^{jk}$ is a directed normal on Γ_{jk}.

(ii) *For any constant C and fixed n, there exists $C'(C)$ such that $\forall \mathbf{v}^n \in \mathcal{V}_h$; $||\mathbf{v}_h^n||_{L^\infty(\mathbb{R}^d)} \leq C$. Consequently, for shape regular $T \in \mathcal{T}_h$*

$$||\mathbf{\Phi}_T^n|| \leq \frac{C'}{h} \sum_{M_i, M_j \in T} ||\mathbf{V}_j^n - \mathbf{V}_i^n|| \ . \tag{3.5}$$

(iii) *Lastly, for any sequence $(\mathbf{v}_h^n)_h$ such that $(\mathbf{v}_h^n)$ is bounded in $L^\infty(\mathbb{R}^d \times \mathbb{R}^+)^m$ independantly of h and N and converges in $L^2_{loc}(\mathbb{R}^d \times \mathbb{R}^+)^m$ to $\mathbf{v}$, we have*

$$\lim_{h \to 0} \left|\left|\mathbf{f}^i(\mathbf{v}_h) - \mathbf{f}^i(\mathbf{v})\right|\right|_{L^1_{loc}(\mathbb{R}^d \times \mathbb{R}^+)^m} = 0, \quad i = 1, 2, \ldots, d\ . \tag{3.6}$$

Our first principle result is a generalization of the Lax-Wendroff theorem to residual distribution schemes for systems of conservation laws using numerical quadrature. Note that since the mapping $H(\mathbf{w})$ is smooth and $\mathbf{w}_h$ is bounded, assumptions on $\mathbf{w}_h$ are equivalent to the same assumptions on $\mathbf{v}_h$ defined by the nodal values $\mathbf{V}_i$.

THEOREM 3.1. *Consider an initial condition $\mathbf{v}_0 \in L^\infty(\mathbb{R}^d)^m$ for time $\tau > 0$. Let $\mathbf{W}_i$ be the nodal approximation for all $M_i \in \mathcal{T}_h$ given by (2.9) from which $\mathbf{V}_i$ are obtain via $\mathbf{V}_i \equiv \mathbf{v}(\mathbf{W}_i)$. Assume that the*

scheme satisfies assumptions (H1) and (H2) and that there exists a constant C that depends only on $\mathbf{v}_0$ and functions $\mathbf{v} \in L^2(\mathbb{R}^d \times \mathbb{R}^+)^m$ and $\mathbf{v}_h$ such that for $\mathbf{v}_h \in \mathcal{V}_h$

$$\sup_h \sup_{x,t} ||\mathbf{v}_h(x,t)|| \leq C, \quad \lim_{h\to 0} ||\mathbf{v}_h - \mathbf{v}||_{L^2_{loc}(\mathbb{R}^d\times\mathbb{R}^+)^m} = 0 \ . \tag{3.7}$$

Let $\mathcal{Q} = \cup T$ be a bounded domain of $\mathbb{R}^d$ and $\tau > 0$ a bounded time. Assume that there exists a locally bounded, positive measure μ such that $||D\mathbf{v}_h||$ tends to μ in the sense of distributions as $h \to 0$. Then $\mathbf{v}(x,t)$ satisfies

$$\left|\left| \int_{\mathcal{Q}\times[0,\tau]} \left(\frac{\partial\varphi}{\partial t}\mathbf{w}(\mathbf{v}(x,t)) + \sum_{i=1}^{d} \frac{\partial\varphi}{\partial x_i}(x,t)\cdot \mathbf{f}^i(\mathbf{w}(\mathbf{v}(x,t))) \right) dx\, dt + \int_{\mathcal{Q}} \varphi(x,0)\, \mathbf{w}(\mathbf{v}_0(x))\, dx \right|\right| \leq \frac{C(\mathcal{T}_h,\mathbf{f})}{(k+1)!} \langle |\varphi|, \mu\rangle \tag{3.8}$$

where k is a integer as described in Assumption (H2) and $C(\mathcal{T}_h,\mathbf{f})$ is a constant that depends on $\mathcal{T}_h$ and $\|D^{k+1}_{\mathbf{v}}\mathbf{f}_{\mathbf{v}}\|$.

This results applies when the limit is piecewise smooth, as it is in practical applications. The proof is inspired by [20] then [2].

4. Proof of Theorem 3.1. The proof of Theorem 3.1 appeals to a sequence of lemmas that are somewhat classical but are tailored here specifically to residual distribution schemes and the use of numerical quadrature for element interior integrations. For simplicity, we assume an evolution to time τ, an N integer multiple of Δt, i.e. $\tau = N\,\Delta t$, although the generalization to arbitrary bounded values of τ is straightforward.

LEMMA 4.1. *Let $\mathcal{Q} = \cup T$ denote a bounded domain of $\mathbb{R}^d$ and $\tau > 0$ a bounded time. Further, let $(\mathbf{v}_h)_h$ denote a sequence such that $\mathbf{v}_h(\,.\,, n\Delta t) \in \mathcal{V}_h$ for any $n \leq N$. Assume there exists a constant C independant of h and N and a function $\mathbf{v} \in L^2(\mathcal{Q}\times[0,\tau])$ such that*

$$\sup_h \sup_{x,t} ||\mathbf{v}_h(x,t)|| \leq C, \quad \lim_{h\to 0} ||\mathbf{v}_h - \mathbf{v}||_{L^2_{loc}(\mathcal{Q}\times[0,\tau])^m} = 0 \ . \tag{4.1}$$

Under these assumptions, the following limits and bound are obtained

1. $\lim_{h\to 0} \sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} |T| \sum_{M_i,M_j\in T} \left|\left|\mathbf{V}_i^n - \mathbf{V}_j^n\right|\right| = 0.$
2. $\lim_{h\to 0} \sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} |T| \sum_{M_i,M_j\in T} \left|\left|\mathbf{V}_i^n - \mathbf{V}_j^n\right|\right|^2 = 0.$
3. $\lim_{h\to 0} h\ ||D_x\mathbf{v}_h||_{L^2(\mathcal{Q}\times[0,\tau])^m} = 0.$
4. *There exists C' independant of h and n such that $h||D_x\mathbf{v}_h||_{L^\infty(\mathcal{Q}\times[0,\tau])^m} \leq C'$.*

Proof. To prove this Lemma, one needs only consider real-valued functions. For any simplex T and open time interval $I^n =]t_{n-1}, t_n[$, two piecewise constant functions can be constructed which are useful in analysis, namely

$$\mathbf{v}_h(x,t)|_{T\times I^n} = \sum_{M_i\in T} \chi_{C_i\cap T}(x)\, \mathbf{V}_i^n \tag{4.2}$$

and the shifted variant

$$\widetilde{\mathbf{v}}_h(x,t)|_{T\times I^n} = \sum_{M_i\in T} \chi_{C_i\cap T}(x)\, \mathbf{V}_{\sigma(i)}^n \tag{4.3}$$

where $\sigma(i)$ denotes a cyclic permutation of the index i and $\chi_{C_i\cap T}$ is the characteristic function of $C_i \cap T$ with C_i the dual cell at node M_i. This defines two functions on $\mathcal{Q}\times[0,\tau]$ that are bounded independantly of h and N. Moreover, the following useful identity holds for these functions in a simplex T for arbitrary $p \geq 0$

$$|T| \sum_{M_i,M_j\in T} \|\mathbf{V}_i^n - \mathbf{V}_j^n\|^p = (d+1) \int_T \|\mathbf{v}_h - \widetilde{\mathbf{v}}_h\|^p\, dx \tag{4.4}$$

where the "$d+1$" factor comes from the definition of dual cells in $\mathbb{R}^d$. Integrating in time

$$\sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} |T| \sum_{M_i, M_j \in T} \|\mathbf{V}_i^n - \mathbf{V}_j^n\| = \int_0^\tau \int_{(\cup T)\in\mathcal{Q}} \|\mathbf{v}_h - \widetilde{\mathbf{v}}_h\| \, dx \, dt \ . \tag{4.5}$$

The sequence $(\mathbf{v}_h)_h$ is bounded, therefore a function $\mathbf{v}' \in L^\infty(\mathcal{Q}\times[0,\tau])^m$ exists such that $\mathbf{v}_h \to \mathbf{v}'$ for the weak–$*$ topology. From the previous assumptions, $\mathbf{v}_h \to \mathbf{v}$ in L^2_{loc} which implies $\mathbf{v} = \mathbf{v}'$ since $\mathcal{Q}\times[0,\tau]$ is bounded and $C_0^\infty(\mathcal{Q}\times[0,\tau])$ is dense in $L^1(\mathcal{Q}\times[0,\tau])$. Similarly, there exists a function $\widetilde{\mathbf{v}} \in L^\infty(\mathcal{Q}\times[0,\tau])^m$, such that $\widetilde{\mathbf{v}}_h \to \widetilde{\mathbf{v}}$ in the weak–$*$ topology. Our next task is to show that $\widetilde{\mathbf{v}} = \mathbf{v}$ and thus finally $\widetilde{\mathbf{v}} = \mathbf{v} = \mathbf{v}'$. To do so, let $\varphi(x,t) \in C_0^\infty(R^d \times \mathbb{R}^+)$, integrate $\varphi\,\mathbf{v}_h$ in $\mathcal{Q}\times[0,\tau]$, and use the definition of the shifted function $\widetilde{\mathbf{v}}$

$$\begin{aligned}
\int_0^\tau \int_{\mathcal{Q}} \varphi\, \mathbf{v}_h \, dx\, dt &= \sum_{n=0}^{n} \Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i \in T} \mathbf{V}_i^n \int_T \varphi\, \chi_{C_i\cap T} \, dx\, dt \\
&= \int_0^T \int_{\mathcal{Q}} \varphi\, \widetilde{\mathbf{v}}_h \, dx\, dt + \sum_{n=0}^{n} \Delta t \sum_{M_i\in T} \mathbf{V}_i^n \left(\int_T \varphi\, \chi_{C_i\cap T} \, dx - \int_T \varphi\, \chi_{C_{\sigma^{-1}(i)}\cap T} \, dx \right)
\end{aligned} \tag{4.6}$$

where $\sigma^{-1}(i)$ denotes the inverse cyclic index permutation such that $\sigma(\sigma^{-1}(i)) = i$. Due to use of gravity centers and edge mid-points in the definition of the dual cells C_i,

$$\int_T \chi_{C_i\cap T} \, dx = \int_T \chi_{C_{\sigma^{-1}(i)}} \, dx = |C_i \cap T|, \quad i = 1,2,\ldots,d+1 \ . \tag{4.7}$$

Using the integral mean-value theorem, points $\overline{x}_i$ and $\overline{x}'_i \in T$ can be found such that

$$\int_T \varphi\, \chi_{C_i\cap T} \, dx = |C_i \cap T|\, \varphi(\overline{x}_i), \quad \int_T \varphi\, \chi_{C_{\sigma^{-1}(i)}} \, dx = |C_i \cap T|\, \varphi(\overline{x}'_i) \ . \tag{4.8}$$

Since $\|D\varphi\|$ is bounded on $\mathcal{Q}\times[0,T]$ and V_h is bounded,

$$\left\| \int_0^T \int_{\mathcal{Q}} \varphi\, \mathbf{v}_h \, dx\, dt - \int_0^\tau \int_{\mathcal{Q}} \varphi\, \widetilde{\mathbf{v}}_h \, dx\, dt \right\| \le Ch \tag{4.9}$$

where C is independant of h and N. Hence in the limit $\widetilde{\mathbf{v}} = \mathbf{v}$ and finally $\widetilde{\mathbf{v}} = \mathbf{v} = \mathbf{v}'$.

Let v_h, $\widetilde{v}_h$, and v'_h denote scalar components of the respective vector-valued functions $\mathbf{v}_h$, $\widetilde{\mathbf{v}}_h$, and $\mathbf{v}'_h$. By the same technique, we see that the components (v_h^2) and $(\widetilde{v}_h^2)$ have the same weak–$*$ limit. It will now be shown that this limit is v^2. Once again appealing to the density of $C_0^\infty(\mathcal{Q}\times[0,\tau])$ in $L^1(\mathcal{Q}\times[0,\tau])$ and the fact that v_h^2 is bounded independantly of h and N, we will take test functions φ in $C_0^\infty(\mathcal{Q}\times[0,\tau])$. The function φ is bounded in $\mathcal{Q}\times[0,\tau]$ and $v_h \to v$ in $L^2_{loc}(\mathcal{Q}\times[0,\tau])$, thus

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, |v - v_h|^2 \, dx\, dt \to 0, \tag{4.10}$$

and consequently

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v^2 \, dx\, dt - 2\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v\, v_h \, dx\, dt + \int_{\mathcal{Q}\times[0,\tau]} \varphi\, v_h^2 \, dx\, dt \to 0 \ . \tag{4.11}$$

By the Cauchy-Schwarz inequality and $\varphi\, v \in L^1(\mathcal{Q}\times[0,\tau])$, the second term converges to

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v^2 \, dx\, dt \ . \tag{4.12}$$

Hence, $v_h^2 \to v^2$ in L^∞ weak–$*$. We are free to choose $\varphi = 1$ combined with the limit $\widetilde{v}_h^2 \to v^2$ in L^∞ weak–$*$, yielding

$$\int_{\mathcal{Q}\times[0,\tau]} |\widetilde{v}_h - v|^2 \, dx\, dt \to 0 \tag{4.13}$$

and finally

$$\int_{\mathcal{Q}\times[0,\tau]} |\widetilde{v}_h - v_h|^2 \; dx\, dt \to 0 \;\; . \tag{4.14}$$

Interpreting this equation of the form (4.4) gives the asserted limit (1) of Lemma 4.1. The limit (2) of Lemma 4.1 is then clear: $\mathcal{Q}\times[0,\tau]$ is bounded thus $L^2(\mathcal{Q}\times[0,\tau]) \subset L^1(\mathcal{Q}\times[0,\tau])$. To prove limit (3) of Lemma 4.1, consider Fig. 4.1 which shows a simplex with inward pointing normals scaled by the edge length. In $\mathbb{R}^d$, the normal $\vec{n}^i$ is the inward pointing vector pendicular to the $(d-1)$–dimensional simplex facet opposite vertex $M_i, i = 1, 2, \ldots, d+1$ scaled by the measure of this facet so that $\sum_{i=1}^{d+1} \vec{n}^i = 0$. Using this notation

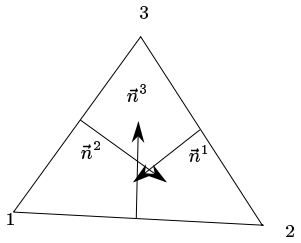


FIG. 4.1. *Depiction of inward pointing normals, $\vec{n}^i$, for a simplex in $\mathbb{R}^2$.*

$$D_x \mathbf{v}_{h|T} = \frac{1}{(d+1)\,|T|} \sum_{j=1}^{d+1} \vec{n}^j\, \mathbf{V}_j = \frac{1}{(d+1)\,|T|} \sum_{j=2}^{d+1} \vec{n}^j\, (\mathbf{V}_j - \mathbf{V}_1) \;\; . \tag{4.15}$$

Integrating in time and space

$$\sum_{n=0}^{N} \int_{I^n} \int_{\mathcal{Q}} ||D_x \mathbf{v}_h^n||^2 \; dx\, dt = \sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} |T| \; \left|\left|(D_x \mathbf{v}_h^n)_{|T}\right|\right|^2 \leq C \frac{1}{h^2} \sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i, M_j \in T} |T| \;\; \left|\left|\mathbf{V}_i^n - \mathbf{V}_j^n\right|\right|^2$$

because the the gradient is constant within a simplex and the triangulation is regular. Limit (3) of Lemma 4.1 is then obtained from the application of limit (2). To obtain the bound (4) of Lemma 4.2, we again consider Eqn. 4.15 assuming bounded $\mathbf{V}_j$

$$h\, ||D_x \mathbf{v}_h||_{L^\infty(\mathcal{Q}\times[0,\tau])^m} \leq \frac{h}{(d+1)\,|T|} \max_{1\leq j\leq d+1} |\vec{n}^j| \max_{1\leq j\leq d+1} \mathbf{V}_j \leq C \frac{h}{(d+1)\,|T|} \max_{1\leq j\leq d+1} |\vec{n}^j| \tag{4.16}$$

which is bounded from above by a constant independent of h and N for shape regular triangulations. This concludes the proof of Lemma 4.1. □

LEMMA 4.2. *Let $\varphi(x,t) \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+)$. With the assumptions of Theorem 3.1, we have*

$$\sum_{n=0}^{N} \Delta t \sum_{M_i\in\mathcal{T}_h} |C_i|\, \varphi(M_i, t_n)\, (\mathbf{W}_i^{n+1} - \mathbf{W}_i^n) + \int_{\mathbb{R}^d\times\mathbb{R}^+} \frac{\partial \varphi}{\partial t}(x,t)\, \mathbf{w}_h(x,t) \; dx\, dt + \int_{\mathbb{R}^d} \varphi(x,0)\, \mathbf{w}_0(x) \; dx \to 0 \tag{4.17}$$

when $h \to 0$. The proof is classical, see for example Kröner [19], p. 377.

LEMMA 4.3. *If $\mathbf{v}_h(x,t) \in \mathcal{V}_h$ satisfies the assumptions of Theorem 3.1, then for any bounded $\mathcal{Q}$ and smooth $\varphi(x,t)$,*

$$\lim_{h\to 0} \sup_h h^{k+1} \left| \sum_{n=0}^{N} \sum_{\forall T\in\mathcal{Q}} \pi_h \varphi(x_T, t_n) \int_T \left\| D_x^{k+1} \left(\sum_{i=1}^{d} \widetilde{A}_i(\mathbf{v}_h)\, (\mathbf{v}_h)_{,x_i} \right) \right\| \; dx \right| \leq C(\mathcal{T}_h, \mathbf{f})\, \langle |\varphi|, \mu \rangle \tag{4.18}$$

where $\pi_h\varphi(x_T,t_n)$ is the midpoint value of the linearly interpolated φ function in simplex T for constant t_n and $C(\mathcal{T}_h,\mathbf{f})$ is a bound on $\|D_{\mathbf{v}}^{k+1}\mathbf{f}_{,\mathbf{v}}(\mathbf{v}_h)\|$ for bounded $\mathbf{v}_h$.

Proof. Using the bound $\|D_{\mathbf{v}}^{k+1}\mathbf{f}_{,\mathbf{v}}\| \leq C(\mathcal{T}_h,\mathbf{f})$ together with the observation that $D_x\mathbf{v}_h$ is constant in a simplex T

$$\begin{aligned} h^{k+1}||D_x^{k+1}(\sum_{i=1}^{d}\widetilde{A}_i(\mathbf{v}_h)\,(\mathbf{v}_h)_{,x_i})|| &= h^{k+1}||D_v^{k+1}(\sum_{i=1}^{d}\widetilde{A}_i(\mathbf{v}_h))\,((\mathbf{v}_h)_{,x_i})^{k+2}|| \\ &\leq h^{k+1}||D_{\mathbf{v}}^{k+1}\widetilde{A}(\mathbf{v}_h)||\ \ ||D_x\mathbf{v}_h||^{k+2} \\ &\leq C(\mathcal{T}_h,\mathbf{f}_{,\mathbf{v}})\,||D_x\mathbf{v}_h||\ . \end{aligned} \tag{4.19}$$

Using again the fact that $D_x\mathbf{v}_h$ is constant in a simplex, it follows that

$$\begin{aligned} \left|h^{k+1}\sum_{n=0}^{N}\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\int_T||D_x^{k+1}\widetilde{A}(\mathbf{v}_h)||\ ||D_x\mathbf{v}_h||^{k+1}\ dx\right| &\leq C(\mathcal{T}_h,\mathbf{f})\sum_{n=0}^{N}\sum_{\forall T\in\mathcal{Q}}|\pi_h\varphi|(x_T,t_n)\int_T||D_x\mathbf{v}_h||\ dx \\ &= C(\mathcal{T}_h,\mathbf{f})\sum_{n=0}^{N}\sum_{\forall T\in\mathcal{Q}}\int_T|\pi_h\varphi|(x_T,t_n)||D_x\mathbf{v}_h||\ dx\ . \end{aligned} \tag{4.20}$$

The function $|\varphi|$ is bounded and continuous on a bounded domain so in the $\limsup_{h\to 0}$ limit, the right-hand-side integral in Eqn. (4.20) approaches the measure-valued function $\langle|\varphi|,\mu\rangle$ which completes Lemma 4.3. □

LEMMA 4.4. *Let $\varphi(x,t)\in C_0^1(\mathbb{R}^d\times\mathbb{R}^+)$ and assume that $\mathbf{v}_h$ satisfies the conditions of Theorem 3.1. The following measure-valued bound exists for $h\to 0$*

$$\lim_{h\to 0}\sup_h\left|\left|\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\sum_{M_i\in T}\mathbf{\Phi}_{i,T}^n+\int_{\mathbb{R}^d\times\mathbb{R}^+}\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}(x,t)\,\mathbf{f}^i(\mathbf{v}_h(x,t))\ dx\,dt\right|\right| \leq \frac{C(\mathcal{T}_h,\mathbf{f})}{(k+1)!}\,\langle|\varphi|,\mu\rangle \tag{4.21}$$

where $C(\mathcal{T}_h,\mathbf{f})$ is a constant that depends on $\mathcal{T}_h$ and $\|D_{\mathbf{v}}^{k+1}\mathbf{f}_{,\mathbf{v}}\|$.

Proof. Choose $\varphi(x,t)$ such that $\text{supp}(\varphi)\subset\mathcal{Q}\times[0,\tau]$. Recall that $\mathbf{\Phi}_T^n=\sum_{i=1}^{d+1}\mathbf{\Phi}_{i,T}^n$ represents an approximation of the flux divergence integrated in a simplex T. By direct calculation

$$\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\left(\sum_{i=1}^{d+1}\mathbf{\Phi}_{i,T}^n+\epsilon_h^{(k)}\right)=\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\pi_h\varphi(x_T,t_n)\sum_{i=1}^{d}\mathbf{f}_{,x_i}^i(\mathbf{v}_h)\ dx \tag{4.22}$$

$$\begin{aligned} &= \sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{f}_{,x_i}^i(\mathbf{v}_h)\ dx \\ &+ \sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T(\pi_h\varphi(x_T,t_n)-\pi_h\varphi(x,t_n))\sum_{i=1}^{d}\mathbf{f}_{,x_i}^i(\mathbf{v}_h)\ dx \end{aligned} \tag{4.23}$$

where $\epsilon_h^{(k)}$ is the quadrature error in calculating the flux divergence. From Assumption (H2), this quadrature error is assumed to be of the form

$$\|\epsilon_h^{(k)}\|=\left\|\int_T\sum_{i=1}^{d}\mathbf{f}_{,x_i}^i(\mathbf{v}_h^n)\,dx-\sum_{i=1}^{d+1}\mathbf{\Phi}_{i,T}^n\right\|\leq C(\mathcal{T}_h)\frac{h^{k+1}}{(k+1)!}\left\|D_x^{k+1}\left(\sum_{i=1}^{d}\widetilde{A}_i(\mathbf{v}_h^n)\,(\mathbf{v}_h^n)_{,x_i}\right)\right\| \tag{4.24}$$

consequently for $\pi_h\varphi(x,t)$ bounded by a constant absorbed into $C(\mathcal{T}_h)$

$$\left\|\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\,\epsilon_h^{(k)}\right\|\leq C(\mathcal{T}_h)\frac{h^{k+1}}{(k+1)!}\sum_{n=0}^{N}\Delta t\left\|D_x^{k+1}\left(\sum_{i=1}^{d}\widetilde{A}_i(\mathbf{v}_h^n)\,(\mathbf{v}_h^n)_{,x_i}\right)\right\|\ . \tag{4.25}$$

Combining this result with Lemma 4.3 formally bounds the quadrature error term

$$\limsup_{h\to 0}\left|\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\,\epsilon_h^{(k)}\,dx\right| \le \frac{C(\mathcal{T}_h,\mathbf{f})}{(k+1)!}\,\langle|\varphi|,\mu\rangle \ . \tag{4.26}$$

Next, apply Green's formula in each simplex to the first right-hand-side sum appearing in Eqn. (4.23)

$$\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T \pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{f}^i_{,x_i}(\mathbf{v}_h)\,dx = \sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T \pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{f}^i_{,x_i}(\mathbf{v}_h)\,dx\,dt \tag{4.27}$$

$$= -\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T\sum_{i=1}^{d}\frac{\partial\pi_h\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v}_h)\,dx\,dt + \sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_{\partial T}\pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{f}^i(\mathbf{v}_h)\cdot n_i\,dx\,dt \ . \tag{4.28}$$

Recall that $\pi_h\varphi$ and $\mathbf{f}$ are both bounded and continuous functions. Upon utilizing Remark 1 (i) and the compact support of φ, it follows that the second right-hand-side sum of Eqn. (4.28) vanishes identically. Examining the remaining right-hand-side term in Eqn. (4.28), observe that

$$\left|\left|\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T\left(\sum_{i=1}^{d}\frac{\partial\pi_h\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v}_h)-\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v})\right)dx\,dt\right|\right| \le C_0\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T||\mathbf{f}(\mathbf{v}_h)-\mathbf{f}(\mathbf{v})||\ dx\,dt + C_1\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T||D_x\pi_h\varphi-D_x\varphi||\ ||\mathbf{f}(\mathbf{v}_h)||\ dx\,dt \tag{4.29}$$

The first right-hand-side sum of Eqn. (4.29) is equal to $||\mathbf{f}(\mathbf{v}_h)-\mathbf{f}(\mathbf{v})||_{L^1(\mathcal{Q}\times[0,\tau])}$ and converges to 0 as $h\to 0$, see Remark 1 (iii). Since $\mathbf{v}_h$ stays bounded and $\mathbf{f}$ continuous, $\mathbf{f}(\mathbf{v}_h)$ stays bounded by a constant. The second right-hand-side sum of Eqn. (4.29) is bounded from above by $||D_x\pi_h\varphi-D_x\varphi||_{L^1(\mathcal{Q}\times[0,\tau])}$ which also converges to 0 as $h\to 0$. Thus, it is concluded that

$$\lim_{h\to 0}\left\|\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T\left(\sum_{i=1}^{d}\frac{\partial\pi_h\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v}_h)-\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v})\right)dx\,dt\right\| = 0 \tag{4.30}$$

and consequently

$$\lim_{h\to 0}\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{f}^i_{,x_i}(\mathbf{v}_h)\,dx = -\sum_{n=0}^{N}\int_{I^n}\sum_{\forall T\in\mathcal{Q}}\int_T\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}\mathbf{f}^i(\mathbf{v})\,dx\,dt \ . \tag{4.31}$$

Considering the second right-hand-side sum term in Eqn. (4.23), from Remark 1 (ii) it follows that

$$\left\|\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T(\pi_h\varphi(x_T,t_n)-\pi_h\varphi(x,t_n))\sum_{i=1}^{d}\mathbf{f}^i_{,x_i}(\mathbf{v}_h^n)\,dx\right\| \le \Sigma$$

where

$$\Sigma \equiv C\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\left|\frac{\pi_h\varphi(x_T,t_n)-\pi_h\varphi(x,t_n)}{h}\right|\sum_{M_i,M_j\in T}\left\|\mathbf{V}_j^n-\mathbf{V}_i^n\right\|\ dx \ . \tag{4.32}$$

Since $||D_x\pi_h\varphi||$ is assumed bounded by a constant,

$$\int_T\left|\frac{\pi_h\varphi(x_T,t_n)-\pi_h\varphi(x,t_n)}{h}\right|\ dx = \int_T\left|D_x\left(\pi_h\varphi\right)\cdot\frac{x_T-x}{h}\right|\ dx \le C\,h^d$$

where C is independant of h. Inserting this bound yields

$$\left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T |\pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n)| \sum_{i=1}^{d} \mathbf{f}^i_{,x_i}(\mathbf{v}_h^n)\, dx \right\| \leq C h^d \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i, M_j \in T} \left\| \mathbf{V}_j^n - \mathbf{V}_i^n \right\| dx \tag{4.33}$$

so that

$$\lim_{h \to 0} \left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T (\pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n)) \sum_{i=1}^{d} \mathbf{f}^i_{,x_i}(\mathbf{v}_h^n)\, dx \right\| = 0 \tag{4.34}$$

Rearrangement of the bounded terms as $h \to 0$ completes Lemma 4.4. □

Proof. (Theorem 3.1) Multiply Eqn. (2.9) by $\varphi(M_i, t_n)\,|C_i|$, where $\varphi(x,t)$ is a test function in $C_0^1\left(\mathbb{R}^d \times [0, +\infty[\right)$, such that $\mathrm{supp}(\varphi) \subset \mathcal{Q} \times [0, \tau]$. Summation on the indices n and i over time slabs and vertices, respectively, yields

$$\sum_{n=0}^{N} \sum_{M_i \in \mathcal{T}_h} |C_i|\, \varphi(M_i, t_n)\, \left(\mathbf{W}_i^{n+1} - \mathbf{W}_i^n\right) + \sum_{n=0}^{N} \Delta t \sum_{M_i \in \mathcal{T}_h} \sum_{T; M_i \in T} \varphi(M_i, t_n)\, \mathbf{\Phi}_{i,T}^n = 0 \ . \tag{4.35}$$

From Lemma 4.2,

$$\lim_{h \to 0} \sum_{n=0}^{N} \sum_{M_i \in \mathcal{T}_h} |C_i|\, \varphi(M_i, t_n)\, \left(\mathbf{W}_i^{n+1} - \mathbf{W}_i^n\right) = -\int_{\mathbb{R}^d \times \mathbb{R}^+} \frac{\partial \varphi}{\partial t}\, \mathbf{w}(\mathbf{v}_h(x,t))\, dx - \int_{\mathbb{R}^d} \varphi(x,0)\, \mathbf{w}(\mathbf{v}_0(x))\, dx \ .$$

The space term is rewritten

$$\begin{aligned} \sum_{n=0}^{N} \Delta t \sum_{M_i \in \mathcal{T}_h} \sum_{T; M_i \in T} \varphi(M_i, t_n)\, \mathbf{\Phi}_{i,T}^n &= \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \pi_h \varphi(x_T, t_n)\, \mathbf{\Phi}_{i,T}^n \\ &+ \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} (\varphi(M_i, t_n) - \pi_h \varphi(x_T, t_n))\, \mathbf{\Phi}_{i,T}^n \end{aligned} \tag{4.36}$$

where once again $\pi_h \varphi(x_T, t_n)$ denotes the midpoint value of the linearly interpolated φ function for constant t_n. Examining the first right-hand-side sum of Eqn. 4.36, recall the result of Lemma 4.4,

$$\lim_{h \to 0} \sup_h \left\{ \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \pi_h \varphi(x_T, t_n) \mathbf{\Phi}_{i,T}^n + \int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x,t)\, \mathbf{f}^i(\mathbf{v}_h(x,t))\, dx\, dt \right\} \leq \frac{C(\mathcal{T}_h, \mathbf{f})}{(k+1)!} \langle |\varphi|, \mu \rangle \ .$$

Next, examine the second right-hand-side sum of Eqn. 4.36. From boundedness of $\|D\varphi\|$ combined with Assumption (H1)

$$\left|\left| \mathbf{\Phi}_{i,T}^n \right|\right| \leq C'\ h^{d-1} \sum_{M_j \in T} \left|\left| \mathbf{V}_j^n - \mathbf{V}_i^n \right|\right| \tag{4.37}$$

yielding

$$\begin{aligned} \left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} (\varphi(M_i, t_n) - \pi_h \varphi(x_T, t_n))\, \mathbf{\Phi}_{i,T}^n \right\| &\leq C h \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \|\mathbf{\Phi}_{T,i}^n\| \\ &\leq C h^d \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i, M_j \in T} \|\mathbf{V}_j^n - \mathbf{V}_i^n\| \ . \end{aligned} \tag{4.38}$$

Consequently, from Lemma 4.1 as $h \to 0$,

$$\left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} (\varphi(M_i, t_n) - \pi_h \varphi(x_T, t_n))\, \mathbf{\Phi}_{i,T}^n \right\| \to 0 \tag{4.39}$$

which completes the proof of Theorem 3.1. □

5. The "N" Residual Distribution Scheme. An important example of a residual distribution scheme is the "N" (Narrow) scheme. It was first considered by Roe [27, 28] and Deconinck [8] for scalar equations. Here we consider the system extension due to van der Weide and Deconinck [31] and generalize their scheme to symmetrizable conservation laws

$$\mathbf{w}(\mathbf{v})_{,t} + \sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}(\mathbf{v}))_{,x_i} = 0 \ . \tag{5.1}$$

Repeating Eqn. (2.18) of Sect. 2.2, our general strategy is to utilize a piecewise linear representation of the entropy variables themselves so that $\mathbf{v}_h \in \mathcal{V}_h$. In a simplex T

$$\int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}(\mathbf{v}_h))_{,x_i} \right) \, dx = \sum_{i=1}^{d} \langle \widetilde{A}_i \rangle \int_T (\mathbf{v}_h)_{,x_i} \, dx = |T| \sum_{i=1}^{d} \langle \widetilde{A}_i \rangle_T (\mathbf{v}_h)_{,x_i,|_T} \tag{5.2}$$

with

$$\langle \widetilde{A}_i \rangle = \sum_{l=1}^{NQ} \omega_l \, \widetilde{A}_i(\mathbf{v}_h(q_l)) \ , q_l \in T \tag{5.3}$$

computed using NQ point numerical quadrature. For purposes of analysis, it is convenient to define the symmetric matrices $\widetilde{K}_{j,T} \in \mathbb{R}^{m \times m}$

$$\widetilde{K}_{j,T} \equiv \frac{1}{d+1} \sum_{i=1}^{d} \vec{n}_{i,T}^{\,j} \langle \widetilde{A}_i \rangle_T \ , \quad \forall M_j \in T \tag{5.4}$$

where $\vec{n}_T^j \in \mathbb{R}^d$ are the inward pointing normal vectors of the face of simplex T opposite vertex M_j scaled by the integral measure of the face, see for example Fig. 4.1. Also define $\widetilde{K}^{\pm} = (\widetilde{K} \pm |\widetilde{K}|)/2$ where $|\widetilde{K}|$ is calculated in the usual matrix sense via eigensystem decomposition. Due to the scaling of vector normals, $\sum_{\forall M_j \in T} \vec{n}_T^j = 0$. Consequently, we have that $\sum_{\forall M_j \in T} \widetilde{K}_{j,T} = 0$ and the identity

$$\sum_{\forall M_j \in T} \widetilde{K}^{+}_{j,T} = - \sum_{\forall M_j \in T} \widetilde{K}^{-}_{j,T} \ . \tag{5.5}$$

For the set of matrices $\{\widetilde{A}_i\}$ equal to the Jacobian matrices of the Euler equations evaluated at a single state, it is shown in [1] that $\left(\sum_{\forall M_j \in T} \widetilde{K}_j^+ \right)$ is nonsingular everywhere except when the state corresponds to a stagnation point. More generally, if we define (formally) the matrix $N \in \mathbb{R}^{m \times m}$ in a simplex T

$$N_T = \left(\sum_{\forall M_j \in T} \widetilde{K}_j^+ \right)^{-1} \ , \tag{5.6}$$

it is shown in the same paper that the matrix product $\widetilde{K}_j N$, $\forall M_j \in T$ and its appearance in the N-scheme always has meaning, even at stagnation points. Hence forward, we assume that the matrix N_T always exists in the sense just described.

Using these definitions, one can easily derive the following relationship for $\mathbf{\Phi}_T$

$$\mathbf{\Phi}_T = \int_T \left(\sum_{i=1}^{d} \mathbf{f}^i(\mathbf{w}(\mathbf{v}_h))_{,x_i} \right) \, dx = |T| \sum_{i=1}^{d} \langle \widetilde{A}_i \rangle_T (\mathbf{v}_h)_{,x_i,|_T} = \sum_{\forall M_j \in T} \widetilde{K}_{j,T} \mathbf{V}_j \ . \tag{5.7}$$

Fundamental to the N-scheme is following decomposition formula for $\mathbf{\Phi}_T$

$$\mathbf{\Phi}_{j,T} = \widetilde{K}^{+}_{j,T} \left(\mathbf{V}_j - \mathbf{V}_T^{\text{inflow}} \right) \ , \quad \forall M_j \in T \tag{5.8}$$

which is often called "upwind" because it represents a generalization to $\mathbb{R}^d$ of two-point upwind differencing for model scalar advection. Perhaps surprisingly, the requirement that $\mathbf{\Phi}_{j,T}$ represent a decomposition of $\mathbf{\Phi}_T$, i.e.

$$\mathbf{\Phi}_T = \sum_{\forall M_j \in T} \mathbf{\Phi}_{j,T} = \sum_{\forall M_j \in T} \widetilde{K}_{j,T} \mathbf{V}_j \ , \tag{5.9}$$

uniquely determines $\mathbf{V}_T^{\text{inflow}}$ when N_T exists

$$\mathbf{V}_T^{\text{inflow}} = -N_T \sum_{\forall M_i \in T} \widetilde{K}^-_{i,T} \mathbf{V}_i \ . \tag{5.10}$$

After some rearrangement, the N-scheme decomposition formula can be written in the following compact form

$$\mathbf{\Phi}_{i,T} = \widetilde{K}^+_{i,T} N_T \sum_{\forall M_j \in T} \widetilde{K}^-_{j,T} (\mathbf{V}_j - \mathbf{V}_i), \quad \forall M_i \in T \ . \tag{5.11}$$

The N-scheme then evolves the solution in time using the algorithm given earlier by Eqn. 2.9. Repeating this algorithm again while taking care to indicate the underlying dependence on $\mathbf{v}_h$ and the nodal degrees of freedom $\mathbf{V}$ that describe $\mathbf{v}_h$:

N-scheme in Symmetrization Variables: For all $M_i \in \mathcal{T}_h$, $n \geq 0$, and $\mathbf{v}_h \in \mathcal{V}_h$

$$\begin{cases} \mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \dfrac{\Delta t}{|C_i|} \displaystyle\sum_{T, M_i \in T} \mathbf{\Phi}_{i,T}(\mathbf{v}_h^n) \ , \quad \mathbf{V}_i^{n+1} = \mathbf{v}(\mathbf{W}_i^{n+1}) \\ \mathbf{W}_i^0 \quad = \mathbf{w}(\mathbf{v}_0(M_i)) \end{cases} \tag{5.12}$$

The primary interest in the N-scheme for approximating conservation laws centers about a local discrete maximum principle exhibited by the N-scheme for scalar advection equations in $\mathbb{R}^d$. To see this, let v_h, V_i, and W_i denote the scalar ($m = 1$) forms of $\mathbf{v}_h$, $\mathbf{V}_i$, and $\mathbf{W}_i$ respectively. Consider a numerical solution at steady state $v_h^n = v_h^{n+1} = v_h^*$. From Eqn. 5.11, the nodal degree of freedom at vertex M_i satisfies

$$0 = \sum_{\forall T \in \mathcal{T}_h; M_i \in T} \sum_{M_j \in T; M_j \neq M_i} -\widetilde{K}^+_{i,T} N_T \widetilde{K}^-_{j,T} (V_j^* - V_i^*) \tag{5.13}$$

$$= \sum_{\forall T \in \mathcal{T}_h; M_i \in T} \sum_{M_j \in T; M_j \neq M_i} \alpha^{ij} (V_i^* - V_j^*) \ , \quad \alpha^{ij} \geq 0 \ . \tag{5.14}$$

This latter equation implies a local discrete maximum principle. More precisely, let $\text{adj}_\alpha(M_i)$ denote the set of vertices adjacent to M_i with nonzero weights α, then $\forall M_i \in \mathcal{T}_h$

$$\min_{M_j \in \text{adj}_\alpha(M_i)} V_j^* \leq V_i^* \leq \max_{M_j \in \text{adj}_\alpha(M_i)} V_j^* \ .$$

Examining the time dependent problem in the scalar ($m = 1$) case, one easily derives a similar maximum principle result for $n > 0$

$$\min_{M_j \in \text{adj}_\alpha(M_i)} (V_j^n, V_i^n) \leq V_i^{n+1} \leq \max_{M_j \in \text{adj}_\alpha(M_i)} (V_j^n, V_i^n)$$

under the CFL-like condition at each t_n

$$\Delta t \leq \max_{\forall M_i \in \mathcal{T}_h} \frac{|C|_i}{\sum_{\forall T \in \mathcal{T}_h; M_i \in T} \sum_{M_j \in T; M_j \neq M_i} -\widetilde{K}^+_{i,T} N_T \widetilde{K}^-_{j,T}} \ .$$

6. Energy and Entropy Analysis of the System N-Scheme. In this section, an energy analysis of the system ($m \geq 1$) N-scheme is undertaken as first described in Barth [5]. As a motivational exercise, we first consider analysis of the linear (constant coefficient) form of Eqn. (2.4). We then analyze the energy and entropy consequences of the N-scheme for nonlinear systems of conservation laws. This latter analysis shows that the N-scheme using symmetrization variables and numerical quadrature satisfies an entropy inequality in the limit of mesh refinement.

6.1. Energy Analysis of the System N-Scheme: the Linear Case. In the linear system case, the numerical scheme (5.12) can be viewed abstractly as an Euler explicit integration of the semi-discrete matrix equation

$$D\,\mathbf{V}_{,t} + L\,\mathbf{V} = 0 \tag{6.1}$$

where $\mathbf{V} \in \mathbb{R}^s$ is a vector representing the nodal degrees of freedom, $D \in \mathbb{R}^{s\times s}$ a symmetric positive definite (SPD) matrix, and $L \in \mathbb{R}^{s\times s}$ a general real-valued matrix. The energy evolution is then given by

$$\frac{1}{2}(\mathbf{V}^T\,D\,\mathbf{V})_{,t} + \mathbf{V}^T\widetilde{L}\,\mathbf{V} = 0 \ , \quad \widetilde{L} = (L + L^T)/2 \ . \tag{6.2}$$

where $\widetilde{L}$ denotes the symmetric part of L. Energy boundedness is demonstrated if it can be shown that the symmetric part of L is positive semi-definite, i.e. for all $\mathbf{V}$

$$\mathbf{V}^T\widetilde{L}\,\mathbf{V} = \mathbf{V}^T L\,\mathbf{V} \geq 0 \ . \tag{6.3}$$

Now suppose that this abstract matrix equation originates from a discretization procedure such as the N-scheme. The total energy associated with the matrix L can be computed and assembled on an element-by-element basis

$$\mathbf{V}^T\widetilde{L}\,\mathbf{V} = \sum_{T\in\mathcal{T}_h} \mathbf{V}_T^T\widetilde{L}_T\,\mathbf{V}_T \tag{6.4}$$

where $\mathbf{V}_T$ and L_T denote the nodal degrees of freedom and element matrix associated with a simplex T. Consequently, to demonstrate energy boundedness of the abstract linear system it is sufficient but not necessary to show

$$\mathbf{V}_T^T\widetilde{L}_T\,\mathbf{V}_T \geq 0 \ , \quad \forall T \in \mathcal{T}_h \ . \tag{6.5}$$

Turning attention now to the N-scheme. For ease of exposition, we will show the development in two space dimensions but the generalization to $\mathbb{R}^d$ will be clear. Next, consider the linear (constant coefficient) form of Eqn. (2.4). In this linear model, the conservation and symmetrization variables are related by the constant matrix $\widetilde{A}_0$, i.e.

$$\mathbf{W}_i = \widetilde{A}_0\,\mathbf{V}_i \ , \quad \forall M_i \in \mathcal{T}_h \ . \tag{6.6}$$

The SPD matrix D appearing in (6.1) would then be block diagonal with $m \times m$ blocks corresponding to each vertex M_i of the form $|C|_i\widetilde{A}_0$. In two space dimensions, the system N-scheme decomposition (5.11) reduces to the following space discretization for a simplex T with local numbering $T(M_1, M_2, M_3)$

$$L_T\mathbf{V}_T = \begin{pmatrix}\boldsymbol{\Phi}_1\\ \boldsymbol{\Phi}_2\\ \boldsymbol{\Phi}_3\end{pmatrix} = \left[\begin{bmatrix}\widetilde{K}_1^+ & & \\ & \widetilde{K}_2^+ & \\ & & \widetilde{K}_3^+\end{bmatrix} + \begin{bmatrix}\widetilde{K}_1^+\\ \widetilde{K}_2^+\\ \widetilde{K}_3^+\end{bmatrix}[N]\begin{bmatrix}\widetilde{K}_1^-\\ \widetilde{K}_2^-\\ \widetilde{K}_3^-\end{bmatrix}^T\right]\begin{pmatrix}\mathbf{V}_1\\ \mathbf{V}_2\\ \mathbf{V}_3\end{pmatrix} \tag{6.7}$$

with $\widetilde{K}^{\pm}$ symmetric and $[N]$ a block diagonal matrix $[N] \equiv \mathrm{diag}(N, N, N)$. The symmetric part of L is given by

$$\widetilde{L}_T = \begin{bmatrix}\widetilde{K}_1^+ & & \\ & \widetilde{K}_2^+ & \\ & & \widetilde{K}_3^+\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\widetilde{K}_1^+\\ \widetilde{K}_2^+\\ \widetilde{K}_3^+\end{bmatrix}[N]\begin{bmatrix}\widetilde{K}_1^-\\ \widetilde{K}_2^-\\ \widetilde{K}_3^-\end{bmatrix}^T + \frac{1}{2}\begin{bmatrix}\widetilde{K}_1^-\\ \widetilde{K}_2^-\\ \widetilde{K}_3^-\end{bmatrix}[N]\begin{bmatrix}\widetilde{K}_1^+\\ \widetilde{K}_2^+\\ \widetilde{K}_3^+\end{bmatrix}^T \ . \tag{6.8}$$

Examining rows of L_T or $\widetilde{L}_T$, observe that the row sum is nonzero. However, we can add the following block diagonal matrix to the element matrix L

$$-\frac{1}{2}\begin{bmatrix} \widetilde{K}_1 & & \\ & \widetilde{K}_2 & \\ & & \widetilde{K}_3 \end{bmatrix} \tag{6.9}$$

so that rows and columns of the L_T sum to zero. These additional terms have no impact on the constant coefficient discretization of the Cauchy problem. These terms all vanish identically when summed for all elements sharing a mesh vertex since the geometry surrounding the vertex is closed. Hence forward, we will include these terms in our definition of L_T and $\widetilde{L}_T$ yielding

$$\widetilde{L}_T = \frac{1}{2}\begin{bmatrix} |\widetilde{K}|_1 & & \\ & |\widetilde{K}|_2 & \\ & & |\widetilde{K}|_3 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \widetilde{K}_1^+ \\ \widetilde{K}_2^+ \\ \widetilde{K}_3^+ \end{bmatrix} [N] \begin{bmatrix} \widetilde{K}_1^- \\ \widetilde{K}_2^- \\ \widetilde{K}_3^- \end{bmatrix}^T + \frac{1}{2}\begin{bmatrix} \widetilde{K}_1^- \\ \widetilde{K}_2^- \\ \widetilde{K}_3^- \end{bmatrix} [N] \begin{bmatrix} \widetilde{K}_1^+ \\ \widetilde{K}_2^+ \\ \widetilde{K}_3^+ \end{bmatrix}^T . \tag{6.10}$$

Next, rewrite off-diagonal term such as

$$\widetilde{K}_i^+ N \widetilde{K}_j^- + \widetilde{K}_i^- N \widetilde{K}_j^+$$

in the following form

$$\widetilde{K}_i^+ N \widetilde{K}_j^- + \widetilde{K}_i^- N \widetilde{K}_j^+ = \widetilde{K}_i N \widetilde{K}_j - \widetilde{K}_i^+ N \widetilde{K}_j^+ - \widetilde{K}_i^- N \widetilde{K}_j^- \ .$$

Consequently, $\widetilde{L}_T$ can be rewritten as

$$\begin{aligned} \widetilde{L}_T = & \frac{1}{2}\begin{bmatrix} \widetilde{K}_1 \\ \widetilde{K}_2 \\ \widetilde{K}_3 \end{bmatrix} [N] \begin{bmatrix} \widetilde{K}_1 \\ \widetilde{K}_2 \\ \widetilde{K}_3 \end{bmatrix}^T \\ & + \frac{1}{2}\begin{bmatrix} \widetilde{K}_1^+ & & \\ & \widetilde{K}_2^+ & \\ & & \widetilde{K}_3^+ \end{bmatrix} - \begin{bmatrix} \widetilde{K}_1^+ \\ \widetilde{K}_2^+ \\ \widetilde{K}_3^+ \end{bmatrix} [N] \begin{bmatrix} \widetilde{K}_1^+ \\ \widetilde{K}_2^+ \\ \widetilde{K}_3^+ \end{bmatrix}^T \\ & + \frac{1}{2}\begin{bmatrix} -\widetilde{K}_1^- & & \\ & -\widetilde{K}_2^- & \\ & & -\widetilde{K}_3^- \end{bmatrix} - \begin{bmatrix} -\widetilde{K}_1^- \\ -\widetilde{K}_2^- \\ -\widetilde{K}_3^- \end{bmatrix} [N] \begin{bmatrix} -\widetilde{K}_1^- \\ -\widetilde{K}_2^- \\ -\widetilde{K}_3^- \end{bmatrix}^T . \end{aligned} \tag{6.11}$$

Note that the first term appearing on the right hand side of Eqn. (6.11) gives rise to a quadratic form with positive energy so our only concern is the remaining terms on the right hand side on this equation. Before proving positive semi-definiteness of (6.11), we first review a simple result concerning the spectra of noncommuting matrices.

LEMMA 6.1. *The nonzero parts of the spectrum of AB and BA are identical for all matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$.*

Proof. Omitted, see for example Axelsson [3] p. 69. □

Next we prove positive semi-definiteness of a specialized matrix in product form.

LEMMA 6.2 (Golub[12]). *The matrix*

$$L = \begin{bmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{bmatrix} - \begin{bmatrix} A \\ B \\ C \end{bmatrix} N \begin{bmatrix} A \\ B \\ C \end{bmatrix}^T , \quad N = [A+B+C]^{-1}$$

is positive semidefinite for all $A, B, C \in \mathbb{R}^{n\times n}$ symmetric positive definite.

Proof. Let

$$Z = \begin{bmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{bmatrix}$$

and congruence transform L

$$Z^{-1/2} L Z^{-1/2} = \begin{bmatrix} I_n & & \\ & I_n & \\ & & I_n \end{bmatrix} - \begin{bmatrix} A^{1/2} \\ B^{1/2} \\ C^{1/2} \end{bmatrix} N \begin{bmatrix} A^{1/2} \\ B^{1/2} \\ C^{1/2} \end{bmatrix}^T = I_{3n} - P.$$

Next use Lemma 6.1 concerning the spectra of nonsquare matrix products. In the present case Lemma 6.1 this implies that

$$\begin{aligned} \text{Eigenvalues}\left(\begin{bmatrix} A^{1/2} \\ B^{1/2} \\ C^{1/2} \end{bmatrix} N \begin{bmatrix} A^{1/2} \\ B^{1/2} \\ C^{1/2} \end{bmatrix}^T \right) &= \text{Eigenvalues}\Big(N^{1/2}(A+B+C)N^{1/2} \Big) + 2n \text{ zeros} \\ &= \text{Eigenvalues}\Big(N(A+B+C) \Big) + 2n \text{ zeros} \\ &= \text{Eigenvalues}(I_n) + 2n \text{ zeros} \end{aligned} \tag{6.12}$$

and consequently

$$I_{3n} - P$$

is positive semidefinite. From this result it follows immediately that

$$L = Z^{1/2}(I_{3n} - P)Z^{1/2}$$

is also positive semidefinite. □
The extension to $A, B, C \geq 0$ and $(A+B+C) > 0$ follows by considering the perturbed matrices $A_\epsilon = A + \epsilon I$, $B_\epsilon = B + \epsilon I$, and $C_\epsilon = C + \epsilon I$ and letting $\epsilon \downarrow 0$.

Returning to the system N-scheme, we now can prove the main result of this section.

THEOREM 6.3. *The system N-scheme discretization of the constant coefficient form of (2.4) is energy bounded with the element energy matrix (6.11) positive semi-definite, i.e.* $\mathbf{V}^T \widetilde{L} \mathbf{V} \geq 0$.

Proof. Since $N = [\widetilde{K}_1^+ + \widetilde{K}_2^+ + \widetilde{K}_3^+]^{-1} = [-\widetilde{K}_1^- - \widetilde{K}_2^- - \widetilde{K}_3^-]^{-1}$, the result follows immediately after application of the Golub lemma to (6.11) together with Eqn. (6.4). □

6.2. Energy and Entropy Analysis of the System N-Scheme: the Nonlinear Case. In this section, an energy analysis of the N-scheme is presented for nonlinear systems of conservation laws. This energy also represents an approximation to the entropy inequality equation (2.2), see Hughes [15] or Barth [6, 4] for related entropy analysis of finite element discretizations. Specifically, we show convergence to an entropy inequality for the N-scheme with exact integration. We then show that with sufficient order numerical quadrature that the entropy inequality is retained in the limit of mesh refinement.

LEMMA 6.4. *Under the assumptions of Theorem 3.1, the limit* $\mathbf{v}$ *of* $\mathbf{v}_h$ *defined by the conservative system N-scheme satisfies*

$$\frac{d}{dt} \int_\Omega H(\mathbf{v}) + \int_{\partial\Omega} \sum_{i=1}^{d} G^i(\mathbf{v})\, \vec{n}_i \; dS \leq 0 \;\; . \tag{6.13}$$

Proof. Consider the system N-scheme decomposition (6.7). Unlike the constant coefficient linear case, the diagonal term (6.9)

$$-\frac{1}{2} \begin{bmatrix} \widetilde{K}_1 & & \\ & \widetilde{K}_2 & \\ & & \widetilde{K}_3 \end{bmatrix}$$

can not be added to the element matrix L_T in the nonlinear case without changing the discretization. Consequently, the energy associated with the simplex $T(M_1, \ldots, M_{d+1})$ must include this term, i.e.

$$\mathbf{V}_T^T \widetilde{L}_T \mathbf{V}_T = \frac{1}{2} \sum_{i=1}^{d+1} \mathbf{V}_i^T \widetilde{K}_i \mathbf{V}_i + Q(\mathbf{V}_1, \ldots, \mathbf{V}_{d+1}) \tag{6.14}$$

where the quadratic form Q is positive by Theorem 6.3. The task is to show that in the limit $h \to 0$ that the first right-hand-side term appearing in Eqn. (6.14) converges to

$$\int_{\partial\Omega} \sum_{i=1}^{d} G^i(\mathbf{v})\, \vec{n}_i \; dS \;\; , \tag{6.15}$$

the integral of the entropy flux. Recall that $\mathbf{V}$ describes the nodal degrees of freedom in the piecewise linear space $\mathbf{v}_h \in \mathcal{V}_h$. In a simplex T we have

$$\sum_{j=1}^{d+1} \mathbf{V}_j^T K_j \mathbf{V}_j = \sum_{j=2}^{d+1} \left(\mathbf{V}_j^T \widetilde{K}_j \mathbf{V}_j - \mathbf{V}_1^T \widetilde{K_j} \mathbf{V}_1 \right) \tag{6.16}$$

$$= \sum_{j=2}^{d+1} (\mathbf{V}_j + \mathbf{V}_1)^T \widetilde{K}_j (\mathbf{V}_j - \mathbf{V}_1) \tag{6.17}$$

$$= \sum_{j=2}^{d+1} (\mathbf{V}_j + \mathbf{V}_1)^T \widetilde{K}_j \, (D\mathbf{v}_h \cdot \vec{l}_{j1}) \tag{6.18}$$

where $\vec{l}_{jk}$ denotes the vector from vertex M_k to M_j. Thus we can define (by identification) the vectors $\mathbf{P}_j \in R^m$ such that

$$\sum_{j=1}^{d+1} \mathbf{V}_j^T \widetilde{K}_j \mathbf{V}_j = \sum_{j=1}^{d} \mathbf{P}_j \cdot \frac{\partial \mathbf{v}_h}{\partial x_j}.$$

Specifically,

$$\mathbf{P}_j^T = \sum_{l=2}^{d+1} \vec{l}_{1j}^{\,j} \, (\mathbf{V}_l + \mathbf{V}_1)^T \widetilde{K}_j$$

where we $\vec{l}_{1j}^{\,j}$ is the j-th component of $\vec{l}_{1j}$. Consequently, we obtain

$$\frac{1}{2} \sum_{j=1}^{d} \mathbf{V}_j^T \widetilde{K}_j \mathbf{V}_j = \int_T \mathbf{v}_h^T \left(\sum_{j=1}^{d} \widetilde{A}_j(\mathbf{V}_h) \frac{\partial \mathbf{V}_h}{\partial x_j} \right) \, dx + \epsilon_T = \int_T \sum_{i=1}^{d} G^i_{,x_i} \, dx + \epsilon_T \tag{6.19}$$

where G is the entropy flux associated with $\mathbf{v}$ and

$$\epsilon_T = \int_T \left(\left[\sum_{j=1}^{d} \frac{\mathbf{P}_j}{|T|} \cdot \frac{\partial \mathbf{v}_h}{\partial x_j} \right] - \mathbf{v}_h^T \left[\sum_{j=1}^{d} A(\mathbf{v}_h) \frac{\partial \mathbf{v}_h}{\partial x_j} \right] \right) \, dx \;\; . \tag{6.20}$$

Hence, ϵ_T can be estimated,

$$|\epsilon_T| \leq \left\{ \int_T \left(\sum_{j=1}^{d} \left[\frac{\mathbf{P}_j}{|T|} - \mathbf{v}_h^T \; \widetilde{A}_j(\mathbf{v}_h) \right] \right)^2 \, dx \right\}^{1/2} \left\{ \int_T ||D\mathbf{v}_h||^2 dx \right\}^{1/2}$$

$$= \left\{ \int_T \left(\sum_{j=1}^{d} \left[\frac{\mathbf{P}_j}{|T|} - \mathbf{v}_h^T \; \widetilde{A}_j(\mathbf{v}_h) \right] \right)^2 \, dx \right\}^{1/2} \int_T ||D\mathbf{v}_h|| \, dx$$

because $D\mathbf{v}_h$ is *constant* in each simplex T. Proceeding as in Lemma 4.1, we see that the function w defined by (see Figure 6.1 for a 2D illustration)

$$w_{|T} = \sum_{[i,j] \text{ edge of } T} (\mathbf{V}_i + \mathbf{V}_j)\, \chi_{D_{ij}}$$

where χ_D is the characteristic function of the set D converges in $L^2_{loc}$to $2\mathbf{v}$ when $h \to 0$. By definition of

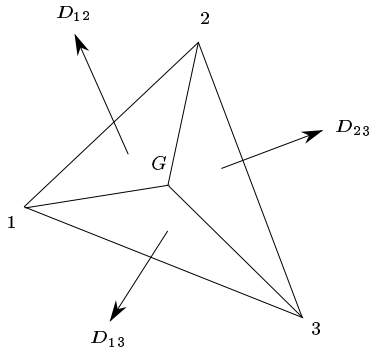


FIG. 6.1. *Geometrical elements for the definition of w*

$\widetilde{K}_j$, we have

$$\sum_{j=2}^{d+1} \widetilde{K}_j(\mathbf{V}_j - \mathbf{V}_1) = |T| \sum_{j=1}^{d} \widetilde{A}_j \frac{\partial \mathbf{v}_h}{\partial x_j},$$

thus we see that

$$\sum_{j=1}^{d} \left(\sum_{l=2}^{d+1} \vec{l}^{\,j}_{1} \widetilde{K}_j \right) \frac{\partial \mathbf{v}_h}{\partial x_j} = |T| \sum_{j=1}^{d} \widetilde{A}_j \frac{\partial \mathbf{v}_h}{\partial x_j} \ . \tag{6.21}$$

Due to the boundedness of $\mathbf{v}_h$, we can apply the dominated convergence theorem and equation (6.21) thus yielding

$$\left\{ \int_{\mathcal{Q}\times[0,\tau]} \left(\sum_{j=1}^{d} \left[\frac{\mathbf{P}_j}{|T|} - \mathbf{v}_h^T\, \widetilde{A}_j(\mathbf{v}_h) \right] \right)^2 dx \right\}^{1/2} \to 0$$

for any bounded domain $\mathcal{Q} \subset \mathbb{R}^d$. This ends the proof. □

THEOREM 6.5. *Under the assumptions of Theorem 3.1, the limit* $\mathbf{v}$ *of* $\mathbf{v}_h$ *defined by the system N-scheme satisfies the entropy enequality*

$$\int_\Omega \frac{\partial \varphi}{\partial t} H(\mathbf{v})\, dx + \int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i} G^i(\mathbf{v})\, dx + \int_{\mathbb{R}^d} \varphi(x,0)\, \mathbf{v}(x,0)\, dx \le 0 \ . \tag{6.22}$$

for any smooth $\varphi \ge 0$.

Proof. The first observation is

$$\varphi \frac{dH(\mathbf{v})}{dt} = \varphi\, \mathbf{v} \cdot \frac{d\mathbf{w}}{dt} \ . \tag{6.23}$$

The second observation is that in a simplex $T(M_1, M_2, M_3)$

$$\sum_{i=1}^{d+1} \varphi_i \mathbf{V}_i^T \mathbf{\Phi}_i = \varphi_G \sum_{i=1}^{d+1} \mathbf{V}_i^T \mathbf{\Phi}_i + \sum_{i=1}^{d+1} (\varphi_i - \varphi_G) \mathbf{V}_i^T \mathbf{\Phi}_i \tag{6.24}$$

where $\varphi_G = \frac{\varphi_1+\varphi_2+\varphi_3}{3}$. Then consequently,

$$\sum_{j=1}^{d+1} \mathbf{V}_j^T \mathbf{\Phi}_j = \frac{1}{2} \sum_{j=1}^{d+1} \mathbf{V}_j^T \widetilde{K}_j \mathbf{V}_j + Q(\mathbf{V}_1, \cdots, \mathbf{V}_{d+1}) \ , \tag{6.25}$$

where Q is positive. Thus we get, because $\varphi_j, \varphi_G \geq 0$

$$\sum_{i=1}^{d+1} \varphi_i \mathbf{V}_i^T \mathbf{\Phi}_i \geq \varphi_G \frac{1}{2} \sum_{j=1}^{d+1} \mathbf{V}_j^T \widetilde{K}_j \mathbf{V}_j + \sum_{i=1}^{d+1} (\varphi_i - \varphi_G) \mathbf{V}_i^T \mathbf{\Phi}_i \ . \tag{6.26}$$

The last observation is that in a simplex T

$$\left|\left| \sum_{i=1}^{d+1} (\varphi_i - \varphi_G) \mathbf{V}_i^T \mathbf{\Phi}_i \right|\right| \leq C \ h \, ||D\varphi||_\infty \sum_{i,j=1}^{d+1} ||\mathbf{V}_i|| \ ||C_{ij}|| \ ||\mathbf{V}_i - \mathbf{V}_j|| \tag{6.27}$$

where C is a constant depending on the mesh only and $C_{ij} = \widetilde{K}_i^+ N \widetilde{K}_j^-$. From this we conclude that

$$\left|\left| \sum_{i=1}^{d+1} (\varphi_i - \varphi_G) \mathbf{V}_i^T \mathbf{\Phi}_i \right|\right| \leq C' h^2 \, ||D\varphi||_\infty \ ||\mathbf{V}_i - \mathbf{V}_j|| \tag{6.28}$$

where C' depends on $\max_i ||\mathbf{v}_h||$ which is uniformly bounded by assumption. Since the mesh is regular, $h^2 \leq C''|T|$ for a well chose constant independant of the mesh. Lemma 4.1 shows that $B \to 0$ when $h \to 0$. Using the same arguments as in Theorem 3.1 and Lemma 6.4, we conclude that the semidiscrete scheme satisfies an entropy inequality. □ Combining the previous results of this section together, we finally conclude with the following Corollary:

COROLLARY 6.6. *Under the assumptions of theorem 3.1 and 6.5, the system N-scheme associated with the quadrature formula (2.20) satisfies in the limit $h \to 0$ an entropy inequality for any smooth $\varphi \geq 0$*

$$\int_\Omega \frac{\partial \varphi}{\partial t} H(\mathbf{v}) dx + \int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i} G^i(\mathbf{v}) \ dx + \int_{\mathbb{R}^d} \varphi(x,0) \, \mathbf{v}(x,0) \ dx \leq C \frac{1}{(k+1)!} \langle |\varphi|, \mu \rangle \ . \tag{6.29}$$

7. Numerical Results. In this section, numerical validation of Theorem 3.1 is provided via N-scheme calculation of smooth and discontinuous solutions of a scalar conservation law and system Euler equations for subsonic, transonic, and supersonic blunt body flows. Recall that Theorem 3.1 states, under classical assumptions, that numerical solutions of the N-scheme with adaptive quadrature converge to a function for which the residual

$$\int_{\mathcal{Q} \times [0,\tau]} \left(\phi_{,t} \mathbf{w}(x,t) + \sum_{i=1}^{d} \phi_{,i} \mathbf{f}_i(\mathbf{w}(x,t)) \right) dxdt + \int_{\mathcal{Q}} \phi(x,0) \mathbf{w}_0(x) dx$$

may not vanish as in the classical Lax-Wendroff theorem. Instead, the residual is bounded by a measure-valued function with strength independent of the mesh size h such that the measure strength can be made arbitrarily small by making the number of quadrature points sufficiently large. As a practical matter, as will be shown in Sect. 7.1, the convergence is very rapid when derivatives of the flux components, $\mathbf{f}^i$, are well behaved. In addition, an adaptive quadrature scheme is proposed and tested which greatly reduces the computational cost of the N-scheme with quadrature. The adaptive quadrature strategy uses a simple estimate of solution smoothness to select the number of quadrature points thus producing an overall economical discretization method since most elements need only use one interior quadrature point (even for second order accurate extensions [1]).

7.1. 1D Conservation Law. Consider the scalar Cauchy problem (1.1)

$$\begin{cases} u_{,t} + (f(u)_{,x} = 0 & \text{for } (x,t) \in \mathbb{R} \times \mathbb{R}^+ \\ u(x,0) = u_0(x) \end{cases}$$

First observe that the upwind scheme (1.5) of Sect. 1 on a uniform mesh can be rewritten as

$$\Delta x_j \frac{du_j}{dt} + \left(\Phi^+_{j-1/2} + \Phi^-_{j+1/2}\right) = 0 \tag{7.1}$$

with

$$\Phi^-_{j+1/2} = \langle a \rangle^-_{j+1/2}(u_{j+1} - u_j) \ , \quad \Phi^+_{j+1/2} = \langle a \rangle^+_{j+1/2}(u_{j+1} - u_j) \tag{7.2}$$

and from Sect. 1

$$\langle a \rangle_{j+1/2} \equiv \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} = \int_0^1 a(\pi u(\xi)))\, d\xi \ , \quad \pi u(\xi) = u_j + \xi\,(u_{j+1} - u_j) \ . \tag{7.3}$$

Note that on a nonuniform mesh, Δx_j is replaced by the lumped average $\Delta x_j = (\Delta x_{j-1/2} + \Delta x_{j+1/2})/2$ although other possible definitions are possible, e.g. $\Delta x_j = (p^+_{j-1/2}\,\Delta x_{j-1/2} + p^-_{j+1/2}\,\Delta x_{j+1/2})$, $p^\pm_{j\mp1/2} \equiv (1 \pm \mathrm{sgn}(\langle a \rangle_{j\mp1/2}))/2$. Consistent with the previous analysis, our first numerical experiment implements a variant of this residual distribution upwind scheme of the form

$$\Delta x_j \frac{du_j}{dt} + \left(\Psi^+_{j-1/2} + \Psi^-_{j+1/2}\right) = 0 \tag{7.4}$$

with residual distribution calculated via numerical quadrature

$$\Psi^-_{j+1/2} = \sum_{l=1}^{NQ} \omega_l\, a(\pi u(q_l))^-(u_{j+1} - u_j), \quad \Psi^+_{j+1/2} = \sum_{l=1}^{NQ} \omega_l\, a(\pi u(q_l))^+(u_{j+1} - u_j) \ . \tag{7.5}$$

The scheme (7.4) would be conservative if

$$\Psi^-_{j+1/2} + \Psi^+_{j+1/2} = f(u_{j+1}) - f(u_j)$$

but due to the use of numerical quadrature

$$\Psi^-_{j+1/2} + \Psi^+_{j+1/2} = \sum_{l=1}^{NQ} \omega_l\, a(\pi u(q_l))\,(u_{j+1} - u_j) \neq \int_0^1 a(\pi u(\xi))\, d\xi\,(u_{j+1} - u_j) = f(u_{j+1}) - f(u_j) \ .$$

Even so, from Theorem 3.1 we still expect convergence to weak solutions provided sufficient order numerical quadrature is employed.

7.1.1. 1D Numerical Experiment: Fixed Gauss Quadrature on Nonuniform Mesh. We first test the scheme (7.4) with Euler explicit time advancement for the smooth flux formula and initial data

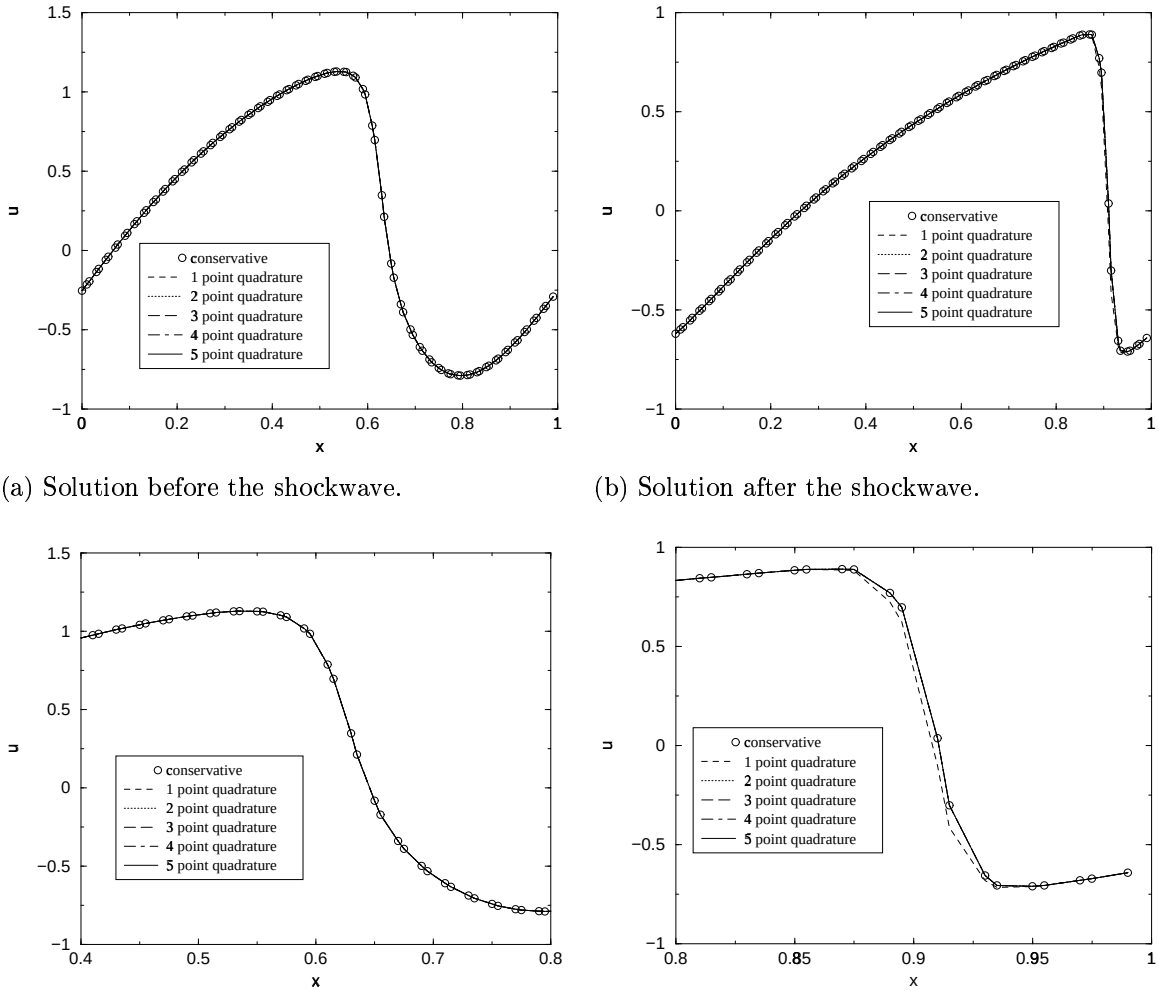
$$f(u) = e^u \ , \quad u_0(x) = \sin(2\pi x)$$

on successively refined meshes ($\Delta x = 10^{-2}$, $(1/2)\ 10^{-2}$, $(1/2)^2\ 10^{-2}$, $(1/2)^3\ 10^{-2}$ and $(1/2)^4\ 10^{-3}$). To eliminate superconvergent behavior of measured error norms due to mesh uniformity, we make the spacing between successive mesh points alternate between the values Δx and $\Delta x/2$. In evaluating the distribution formulas (7.5), NQ-point Gauss quadrature formulas are used with $1 \leq NQ \leq 3$ to validate Theorem 3.1. Selected results are given in Table 7.1 which tabulates L^1, L^2 and L^∞ norms of the difference between the numerical solution u^n_c given by standard conservative scheme (7.1) and the nonconservative u^n_{nc} provided by the scheme (7.4) on meshes with decreasing Δx at time $n\Delta t = 0.5$ (after the shockwave has appeared). Figures 7.1(a-d) graph the solutions before and after the occurence of the shockwave for the conservative scheme and the nonconservative schemes using 1, 2, 3, 4 and 5 point Gauss quadrature on a mesh containing 100 unknowns. All the solutions are virtually indistinguishable before the occurrence of the shockwave. It is only after the solution becomes discontinuous that the importance of sufficient order Gauss quadrature is visually seen and multiple quadrature points needed. The tabulated results reveal two effects addressed by the theory: (1) the L^1 error eventually stagnates when $h \to 0$ for fixed order quadrature and (2) the L^1 error decreases very rapidly with increasing NQ. In fact, upon closer inspection, this error decreases much more quickly than $(p+1)!$ typical of Gauss quadrature, see Figure 7.2.

mesh size, Δx	$L^1(u_{nc}-u_c)$	$L^2(u_{nc}-u_c)$	$L^\infty(u_{nc}-u_c)$	#quad pts, NQ
$0.100\ 10^{-1}$	0.6353110^{-2}	$0.25617\ 10^{-1}$	0.15162	1
$0.500\ 10^{-2}$	$0.67850\ 10^{-2}$	$0.38770\ 10^{-1}$	0.35783	1
$0.250\ 10^{-2}$	$0.70532\ 10^{-2}$	$0.55376\ 10^{-1}$	0.67416	1
$0.125\ 10^{-2}$	$0.72127\ 10^{-2}$	$0.73250\ 10^{-1}$	$0.10491\ 10^{1}$	1
$0.100\ 10^{-1}$	$0.12402\ 10^{-4}$	$0.50233\ 10^{-4}$	$0.30796\ 10^{-3}$	2
$0.500\ 10^{-2}$	$0.14468\ 10^{-4}$	$0.83082\ 10^{-4}$	$0.74799\ 10^{-3}$	2
$0.250\ 10^{-2}$	$0.15648\ 10^{-4}$	$0.12657\ 10^{-3}$	$0.14325\ 10^{-2}$	2
$0.125\ 10^{-2}$	$0.16296\ 10^{-4}$	$0.18732\ 10^{-3}$	$0.33085\ 10^{-2}$	2
$0.100\ 10^{-1}$	$0.28748\ 10^{-7}$	$0.42536\ 10^{-7}$	$0.23256\ 10^{-6}$	3
$0.500\ 10^{-2}$	$0.27937\ 10^{-7}$	$0.44600\ 10^{-7}$	$0.35562\ 10^{-6}$	3
$0.250\ 10^{-2}$	$0.27315\ 10^{-7}$	$0.48398\ 10^{-7}$	$0.49170\ 10^{-6}$	3
$0.125\ 10^{-2}$	$0.27017\ 10^{-7}$	$0.57222\ 10^{-7}$	$0.93983\ 10^{-6}$	3

TABLE 7.1

Numerical results for 1D Cauchy problem. Numerical error between the conservative calculation u_c and the nonconservative calculation u_{nc} using NQ point Gauss quadrature.



(a) Solution before the shockwave.

(b) Solution after the shockwave.

(c) Solution before the shockwave, zoom.

(d) Solution after the shockwave, zoom.

FIG. 7.1. *Numerical N-scheme solutions for (1.1) and $u_i^0 = \sin(2\pi ih)$. Solutions before the formation of a shockwave ((a) and (c)) and solutions after the formation of a shockwave ((b) and (d)).*

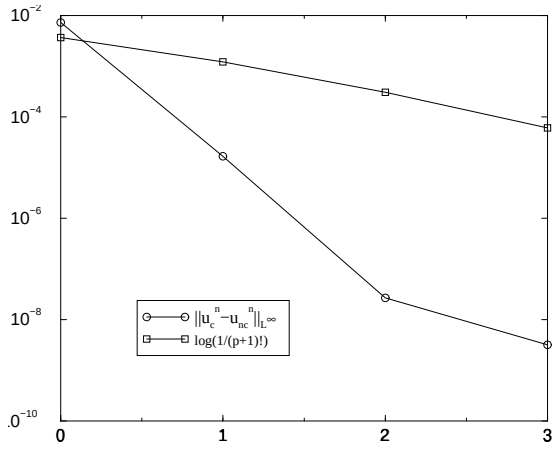


FIG. 7.2. *Comparison between* $\log\left(\frac{1}{(p+1)!}\right)$ *and* $||u_c^n - u_{nc}^n||_{L^\infty}$ *for* $h = 0.625\ 10^{-3}$.

7.1.2. 1D Numerical Experiment: Adaptive Gauss Quadrature on Nonuniform Mesh. The results of the previous 1D numerical experiment of Sect. 7.1.1 show very negligible sensitivity to the number of quadrature points whenever the solution is smooth. This observation suggests the following simple adaptive quadrature scheme which uses a nondimensional measure of solution gradient to estimate solution smoothness:

- If $\dfrac{u_{i+1} - u_i}{\max(|u_i|, |u_{i+1}|)} \leq \sqrt{\Delta x / L}$, then the solution is smooth. Compute $\psi^+_{j+1/2}$ and $\psi^-_{j+1/2}$ with NQ_{min} point quadrature.
- Else, compute $\psi^+_{j+1/2}$ and $\psi^-_{j+1/2}$ with NQ_{max} point quadrature.

Repeating the calculations of Sect. 7.1.1, Table 7.2 tabulates the corresponding numerical results using the adaptive parameters $NQ_{min} = 1$ and $NQ_{max} = 2, 3$ at the time $T = 0.5$ (after the formation of the shockwave). Note that in these calculations, nearly all cells required only $NQ_{min} = 1$ quadrature point

mesh size, Δx	$L^1(u_{nc} - u_c)$	$L^2(u_{nc} - u_c)$	$L^\infty(u_{nc} - u_c)$	NQ_{min}	NQ_{max}
$0.100\ 10^{-1}$	$0.39005\ 10^{-4}$	$0.11351\ 10^{-3}$	$0.69521\ 10^{-3}$	1	2
$0.500\ 10^{-2}$	$0.27998\ 10^{-4}$	$0.13526\ 10^{-3}$	$0.12200\ 10^{-2}$	1	2
$0.250\ 10^{-2}$	$0.21424\ 10^{-4}$	$0.16100\ 10^{-3}$	$0.18236\ 10^{-2}$	1	2
$0.125\ 10^{-2}$	$0.18526\ 10^{-4}$	$0.20768\ 10^{-3}$	$0.36673\ 10^{-2}$	1	2
$0.100\ 10^{-1}$	$0.27007\ 10^{-4}$	$0.64119\ 10^{-4}$	$0.38681\ 10^{-3}$	1	3
$0.500\ 10^{-2}$	$0.13642\ 10^{-4}$	$0.52260\ 10^{-4}$	$0.47082\ 10^{-3}$	1	3
$0.250\ 10^{-2}$	$0.57902\ 10^{-5}$	$0.34398\ 10^{-4}$	$0.38955\ 10^{-3}$	1	3
$0.125\ 10^{-2}$	$0.22238\ 10^{-5}$	$0.20259\ 10^{-4}$	$0.35797\ 10^{-3}$	1	3

TABLE 7.2

Error between the conservative scheme and the adaptive quadrature scheme at $t = 0.5$ *using 1, 2, or 3 point Gauss quadrature.*

with only 3-5 cells requiring NQ_{max} number of quadrature points. This results in a notable savings in computational resources. These numerical results indicate that the quality of the solutions is comparible with those of Table 7.1 with some reduced accuracy that would be improved by a more stringent criteria for quadrature adaptation. We have not run this case with a second order upwind scheme, but we believe that the same strategy could be used since quadrature with $NQ_{min} = 1$ points is second order accurate. In fact, it can be shown formally that to recover second order accuracy, the "exact" total residual $\Phi^-_{j+1/2} + \Phi^+_{j+1/2}$ needs only be recovered up to second order accuracy to have a second order accuracte scheme, see [1]. Finally, we note that other tests have been carried out, for example with the flux $f(u) = \exp(u^2)$ with similar results.

7.2. 2D Conservation Laws. Next, we present 2D solutions of the Euler equations of gasdynamics assuming a perfect gas relationship. Throughout the remaining 2D numerical experiments, standard quadrature formulas for simplicies are utilized with weights and quadrature point locations as given in Table 7.3, see also [29].

NQ	Accuracy	Barycentric coordinates, q_l	weights, ω_l
1	$O(h^2)$	$(1/3, 1/3, 1/3)$	1/2
3	$O(h^3)$	$(1/2, 1/2, 0)$	1/3
6	$O(h^4)$	$(0.816847572980459, 0.091576213509771, 0.091576213509771)$	0.109951743655322
		$(0.108103018168070, 0.445948490915965, 0.445948490915965)$	0.223381589678011
7	$O(h^5)$	$(1/3, 1/3, 1/3)$	0.225
		$(0.797426985353087, 0.101286507323456, 0.101286507323456)$	0.125939180544827
		$(0.470142064105115, 0.470142064105115, 0.059715871789770)$	0.132394152788506
16	$O(h^7)$	$(0.057104196114518, 0.065466994555014, 0.877428809330468)$	0.047136736386776
		$(0.276843013638124, 0.050210123211370, 0.672946863150506)$	0.070776135796160
		$(0.583590432368917, 0.028912084224389, 0.387497483406694)$	0.045168098564740
		$(0.860240135656220, 0.009703785126946, 0.130056079216834)$	0.010846451821051
		$(0.057104196114518, 0.311164552244357, 0.631731251641125)$	0.088370177044746
		$(0.276843013638124, 0.238648659731443, 0.484508326630433)$	0.132688432214078
		$(0.583590432368917, 0.137419104134574, 0.278990463496509)$	0.084679449043492
		$(0.860240135656220, 0.046122079906452, 0.093637784437328)$	0.020334519128958
		$(0.057104196114518, 0.631731251641125, 0.311164552244357)$	0.088370177044746
		$(0.276843013638124, 0.484508326630433, 0.238648659731443)$	0.132688432214078
		$(0.583590432368917, 0.278990463496509, 0.137419104134574)$	0.084679449043492
		$(0.860240135656220, 0.093637784437328, 0.046122079906452)$	0.020334519128958
		$(0.057104196114518, 0.877428809330468, 0.065466994555014)$	0.047136736386776
		$(0.276843013638124, 0.672946863150506, 0.050210123211370)$	0.070776135796160
		$(0.583590432368917, 0.387497483406694, 0.028912084224389)$	0.045168098564740
		$(0.860240135656220, 0.130056079216834, 0.009703785126946)$	0.010846451821051

TABLE 7.3

Quadrature points and weights. The missing quadrature points are obtained by cyclic permutation as needed.

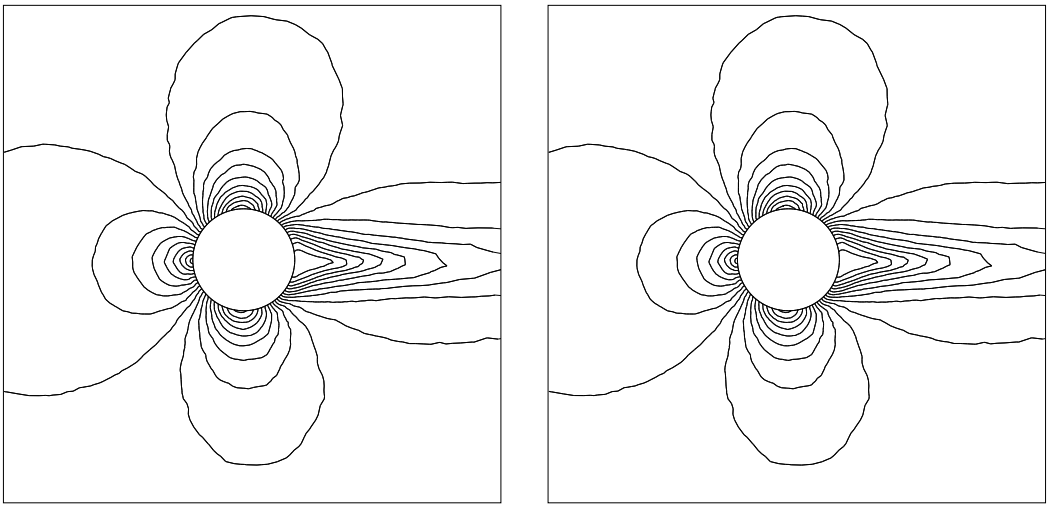
The significant computational savings obtained by adaptive numerical quadrature in 1D suggest using a similar strategy in higher space dimensions were the savings is even more dramatic. For any simplex T, a criterion must be developed which determines if the numerical solution is locally smooth or not. For efficiency reasons, this decision should ideally be made from the information available in T only. Let s_j denotes the (physical) entropy at node M_j and h_T the maximum length of the edges of T. We have implemented the following heuristic criterion for use in the adaptive quadrature strategy:

- If $\min_{M_i, M_j \in T} \left| \dfrac{s_i}{s_j} - 1 \right| \geq \sqrt{h_T / L}$, then the solution is smooth. Compute the N-scheme decomposition using NQ_{min} point quadrature.
- Else, compute the N-scheme decomposition using NQ_{max} point quadrature.

7.2.1. 2D Numerical Experiments: Euler Equations on Mesh Triangulations. We first study the effect of the loss of conservation and the influence of the number of quadrature points for the N-scheme with quadrature. To achieve this goal, we have selected three test cases that are simple yet representative of different flow regimes: a subsonic flow test case, a transonic flow case with mild shockwaves, and a supersonic flow case over a blunt body which produces a strong bow shockwave. The solutions are compared to those obtained by the reference conservative N-scheme using Z variables. Our intent is not to assess the accuracy of the solution with respect to a mesh-converged solution, but rather to see how the loss of conservation affects the structure of the solution compared with the reference solution on the same mesh. In particular, we qualitatively and quantitatively compare the overall structure of conservative and nonconservative N-scheme solutions by examining representative cross-sectional and/or boundary data plots. Additionally, we examine the behavior of numerical solutions with adaptive mesh refinement for the cases containing

discontinuities where exact discrete conservation is normally very important as it insures proper solution jump approximation. Ideally, it would be illuminating to perform *uniform* mesh refinement in evaluating the adaptive quadrature N-scheme as $h \to 0$. This was done in the 1D calculations. Unfortunately, this is prohibitively expensive in 2D so we rely on multiple levels of adaptive mesh refinement to approximate the $h \to 0$ process.

7.2.2. Subsonic Flow Case. This case is taken from Dervieux [9]. It is a flow over a cylinder with a Mach number at infinity of $M_\infty = 0.38$ computed on a relatively coarse mesh containing 3010 simplicial elements. Under these flow conditions, the flow remains subsonic and devoid of solution discontinuities. Figures 7.3(a-d) show Mach number isolines for N-scheme calculations using 1, 3, and 7 point quadrature as well as the N-scheme using the conservative Z variables. Figure 7.4(a) shows Mach number isolines for the N-scheme using the adaptive quadrature procedure described earlier with parameter values $NQ_{min} = 1$ and $NQ_{max} = 3$. Figure 7.4(b) provides a quantitative comparison of pressures on the surface of the cylinder using all the fixed and adaptive quadrature formulas as well as the conservative Z variable N-scheme. As

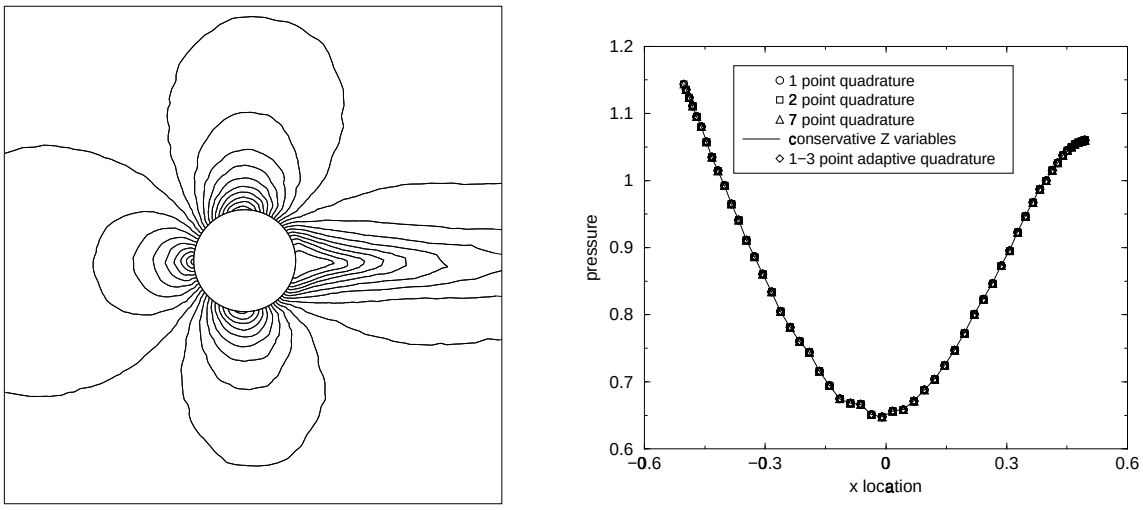


(a) 1 point quadrature (b) 3 point quadrature

(c) 7 point quadrature (d) conservative Z variables

FIG. 7.3. *(a-d) Mach number isolines for N-scheme calculations using fixed 1, 3, and 7 point quadrature and the conservative Z variables for the subsonic cylinder problem, $M_\infty = 0.38$, on a simplicial mesh containing 3010 elements.*

expected, all calculations show no discernible differences. This results confirm our analysis if we assume



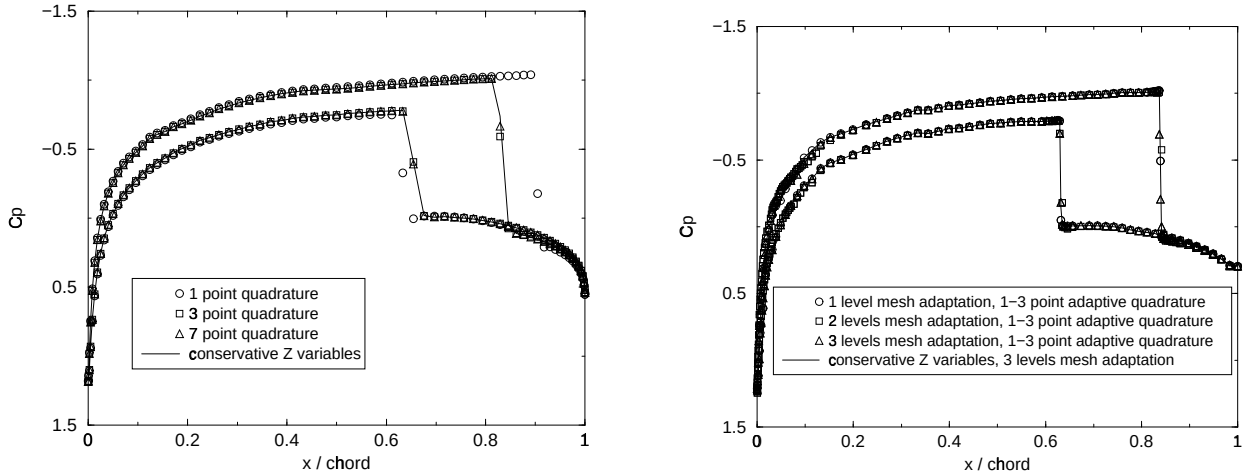
(a) 1-3 point adaptive quadrature

(b) Pressure along top-bottom line of symmetry

FIG. 7.4. *(a) Mach number isolines for N-scheme calculations using 1-3 point adaptive quadrature and (b) the resulting pressure along top-bottom line of symmetry for the subsonic cylinder problem, $M_\infty = 0.38$, on a simplicial mesh containing 3010 elements.*

that the support of the measure μ is concentrated near discontinuities in the solution. Since there are no discontinuities in this flow and our quadrature formulas are at least second order accurate using single point quadrature, a Lax Wendroff theorem is satisfied up to $O(h^2)$.

7.2.3. Transonic Flow Case. The second 2D test case consists of transonic flow, $M_\infty = 0.85$, over the NACA0012 geometry with flow incidence of 1° computed on a baseline simplicial mesh containing 5050 elements. The flow solution consists of both upper and lower surface shockwaves. Due to the 1° flow



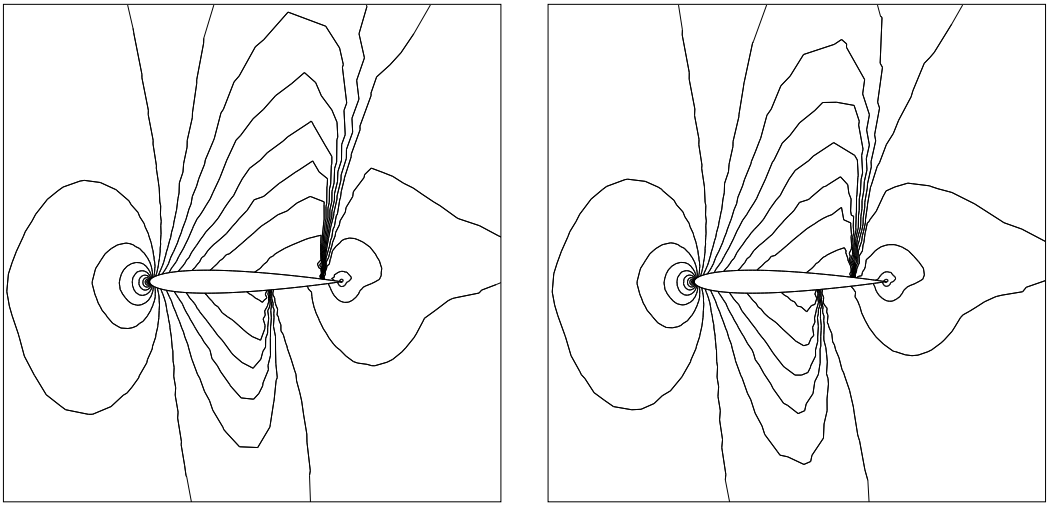
(a) Fixed 1,3, and 7 point quadrature and conservative Z variables

(b) Three levels adaptive mesh refinement with 1-3 point adaptive quadrature

FIG. 7.5. *Surface pressure coefficient distribution on the NACA0012 airfoil using the N-scheme with (a) 1, 3, and 7 point quadrature and the conservative Z variable and (b) three levels of adaptive mesh refinement together with 1-3 point adaptive quadrature.*

incidence, the upper surface shockwave is notably stronger than the lower surface shockwave. N-scheme calculations were performs using fixed 1, 3, and 7 quadrature point formulas as well as the conservative Z variables on the baseline simplicial mesh containing 5050 elements. In addition, three levels of adaptive

mesh refinement were performed and solutions computed using the adaptive quadrature N-scheme with $NQ_{min} = 1$ and $NQ_{max} = 3$. Surface pressure coefficient values are graphed in Fig. 7.5(a) and Mach number isocontours shown in Figs. 7.6(a-d) for N-scheme calculations using 1, 3, and 7 point quadrature and the conservative Z variables on fixed mesh composed of 5050 elements. Similarly, surface pressure coefficient values are graphed in Fig. 7.5(b) and Mach number isocontours in Figs. 7.7(a-d) using three levels of shockwave-adapted mesh refinement together with the 1-3 adaptive quadrature point form of the N-scheme. Although the Mach number isocontour plots look visually very similar, the Fig. 7.5(a) graph

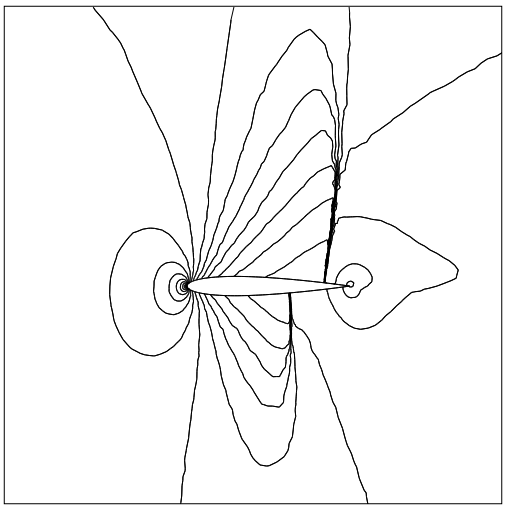


(a) 1 point quadrature (b) 3 point quadrature

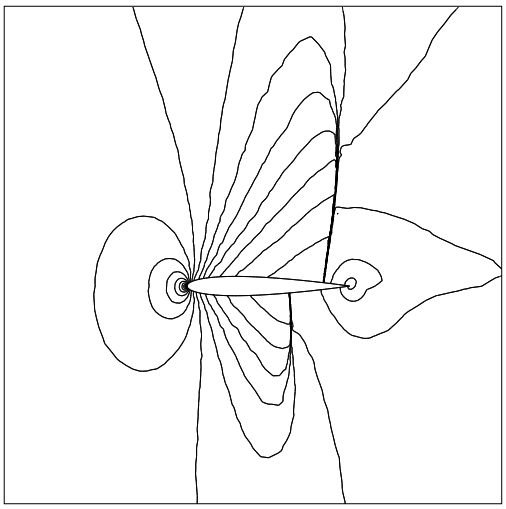
(c) 7 point quadrature (d) Conservative Z variables

FIG. 7.6. *(a-d) Mach number isolines for N-scheme calculation using fixed 1, 3, and 7 point quadrature and the conservative Z variables for the transonic NACA0012 problem, $M_\infty = 0.85$ and 1° flow incidence, on a simplicial mesh containing 5050 elements.*

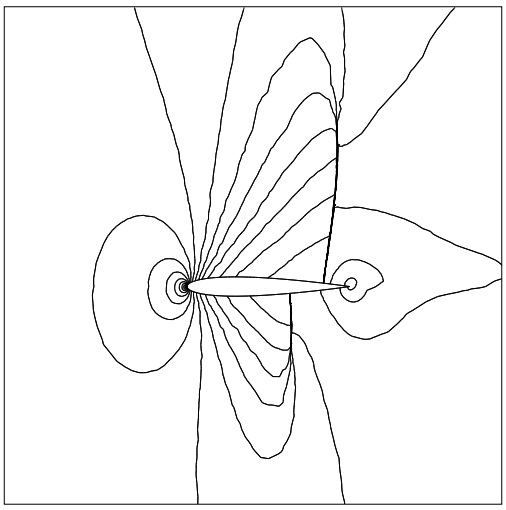
of the pressure coefficient on the body of the NACA0012 airfoil is more revealing. This graph shows that the location of the shockwaves depends on the number of quadrature points. Specifically, the use of single point quadrature leads to a significant change in shockwave location when compared to 3 and 7 point quadrature as well as the conservative Z variable scheme. For this particular flow, the effect of the measure μ is not sufficiently reduced using one quadrature point but using three or more quadrature points seems sufficient to reduce conservation error less than truncation errors present in the conservative N-scheme. The comparability of the 3 and 7 point quadrature with the conservative Z variable scheme once again suggests



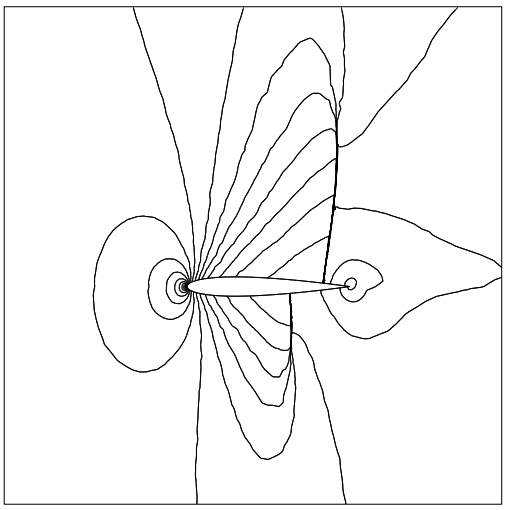
(a) 1 level adaptive mesh refinement,
1-3 point adaptive quadrature



(b) 2 levels adaptive mesh refinement,
1-3 point adaptive quadrature



(a) 3 levels adaptive mesh refinement,
1-3 point adaptive quadrature

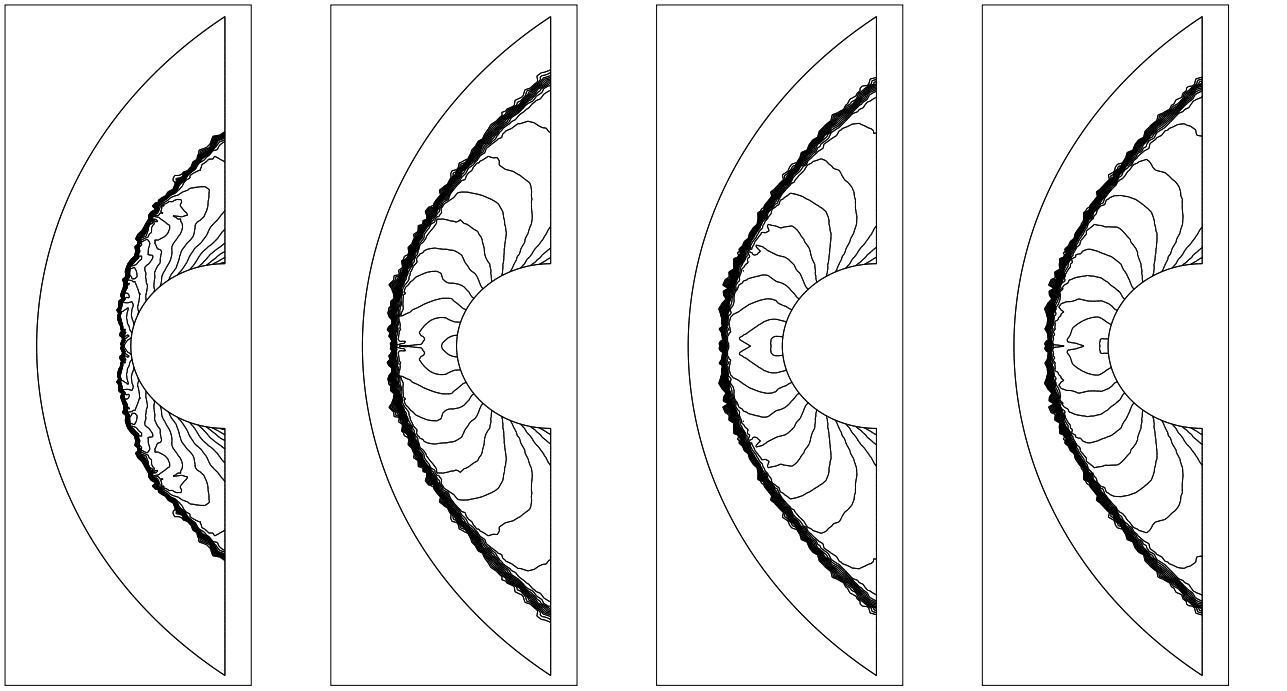


(b) 3 levels adaptive mesh refinement,
conservative Z variables

FIG. 7.7. *(a-d) Mach number isolines for N-scheme calculations using 1, 2, and 3 levels of adaptive mesh refinement and 1-3 point adaptive quadrature and the reference conservative Z variables on the 3 level refined mesh. for the transonic NACA0012 problem,* $M_\infty = 0.85$ *and* 1° *flow incidence.*

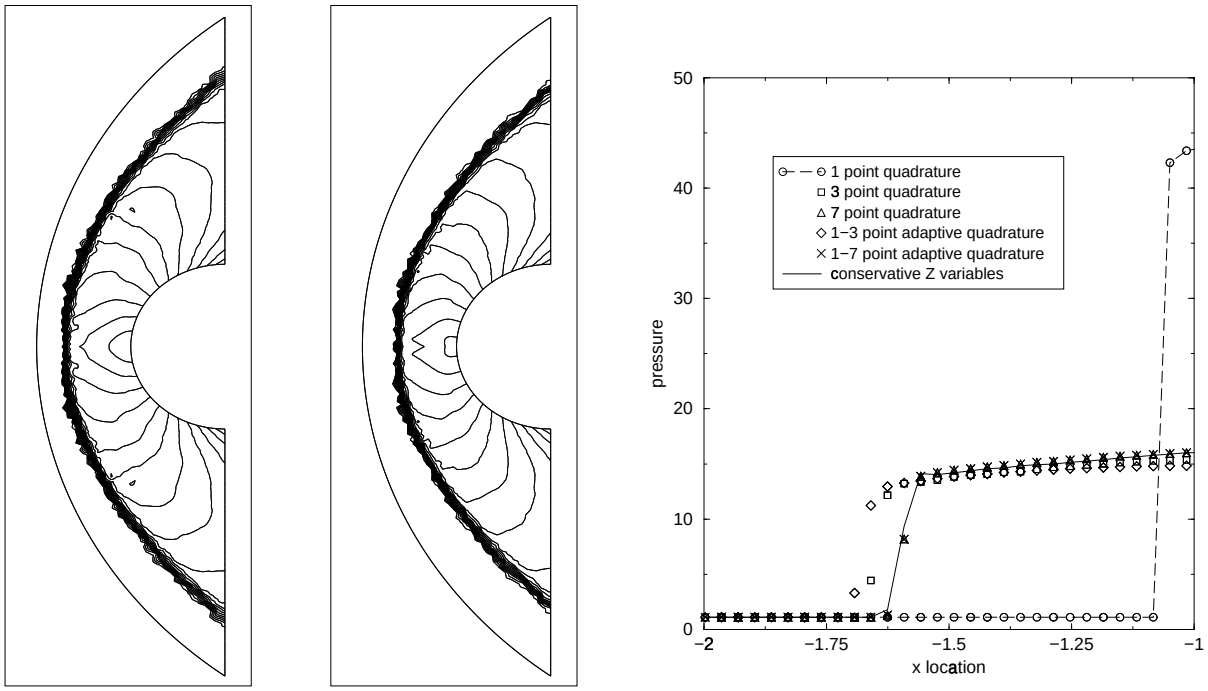
an adaptive quadrature implementation. The calculations presented in Fig. 7.5(b) are intended at checking whether the errors generated by the loss of conservation on refined meshes dominate the truncation error, even in an adaptive quadrature setting. With adaptive mesh refinement, all the computations in Fig. 7.5(b) are very comparable which further validates Theorem 3.1 and our adaptive quadrature strategy.

7.2.4. Supersonic Blunt Body Flow. The last 2D test case consists of supersonic flow, $M_\infty = 3.5$, over a circular cylinder geometry computed on a baseline simplicial mesh containing 4075 elements. The flow solution consists of strong bow shock forward of the cylinder geometry. Figures 7.8 (a-d) show Mach number isocontours for numerical solutions computed using 1, 3, and 7 point quadrature and conservative Z variable forms of the N-scheme. In addition, Mach number isocontours for 1-3 and 1-7 point adaptive quadrature N-scheme calculations are shown in Figs. 7.9 (a-b). Both fixed and adaptive quadrature calculations are compared in Fig. 7.9 for pressure data along the top-bottom line of symmetry. This latter figure shows a



(a) 1 point quadrature (b) 3 point quadrature (c) 7 point quadrature (d) conservative Z variables

FIG. 7.8. *Mach number isolines for N-scheme calculations using fixed 1, 3, and 7 point quadrature and the conservative Z variables for the supersonic cylinder problem, $M_\infty = 3.5$ on the baseline simplicial mesh containing 4075 elements.*

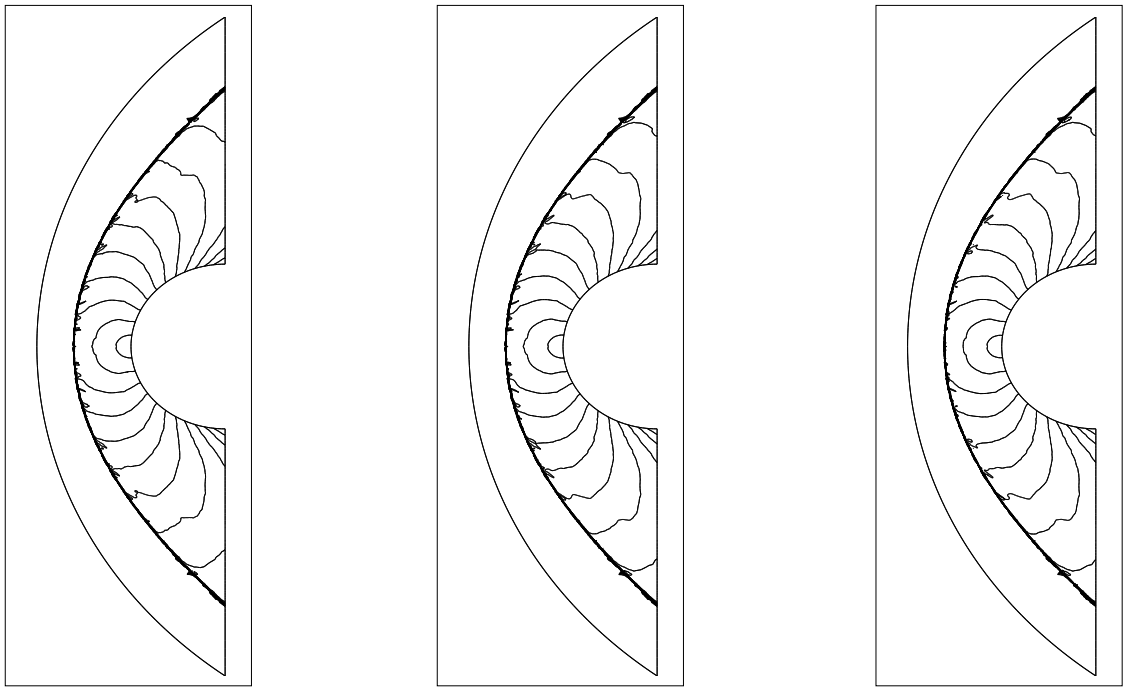


(a)1-3 point adaptive quadrature (b) 1-7 point adaptive quadrature (c) Pressure along line of top-bottom symmetry

FIG. 7.9. *(a-b) Mach number isolines for N-scheme calculations using 1-3 and 1-7 point adaptive quadrature and (c) comparison of all fixed and adaptive quadrature N-scheme calculations along the top-bottom line of symmetry for the supersonic cylinder problem, $M_\infty = 3.5$ on the baseline simplicial mesh containing 4075 elements.*

large difference between the one quadrature calculation and the other calculations. This difference is also clearly seen in the Mach number isocontour plot, Fig. 7.8(a). Perhaps more importantly, Fig. 7.9 shows that the 3 point quadrature (and 1-3 point adaptive quadrature) also produced incorrect shockwave locations on the baseline mesh, although the error is much smaller than that obtained using 1 point quadrature. Recall that for the transonic flow problem, 3 point quadrature was of sufficient order on the baseline and adaptively refined meshes for computing correct shock locations. Using 7 point fixed quadrature and 1-7 point adaptive quadrature yields solution shockwave positions in agreement with the conservative scheme. These results are also in agreement with the inequality (3.8) of Theorem 3.1, since the strength of the measure depends on the number of quadrature points but also on the supremum of a norm of higher derivatives of the flux, $D_{\mathbf{v}}^{k+1}\mathbf{f}_{,\mathbf{v}}$. Estimation of this norm is difficult, but it is reasonable that this number tends to infinity as the maximum Mach number also tends to ∞. However, since the flux $\mathbf{f}$ is analytical in $\mathbf{v}$ and the Mach number finite, the right-hand-side of (3.8) still converges to zero, albeit more slowly.

Next, we examine the effect of adaptive mesh refinement. Figures 7.10(a-c) show Mach number isocontours for the N-scheme calculations using 16 point quadrature, 1-16 point adaptive quadrature, and conservative Z variables. Figure 7.11 shows a graph of Mach number along the top-bottom symmetry line for these same schemes as well as 1-7 point adaptive quadrature. Surprisingly, Fig. 7.11 shows small differ-



(a) 16 point quadrature (b) 1-16 point adaptive quadrature (c) conservative Z variables

FIG. 7.10. *(a-b) Mach number isolines for N-scheme calculations using 16 point fixed and 1-16 point adaptive quadrature and (c) conservative Z variables for the supersonic cylinder problem, $M_\infty = 3.5$, on the baseline simplicial mesh with 3 levels of adaptive mesh refinement.*

ences in shock profile using 1-7 point adaptive quadrature for this problem with three levels of adaptive mesh refinement. It is only with 16 point fixed or adaptive quadrature that the adaptive N-scheme solutions match the conservative Z variable N-scheme. This demonstrates some slight dependency on the mesh parameter h not captured by the present analysis.

8. Concluding Remarks. A number of upwind schemes are derived in quasilinear form and discrete conservation obtained by devising specialized mean-value linearized coefficients. This approach is problematic for systems such as magnetohydrodynamics, Euler equations with certain forms of chemistry, etc. where these specialized mean-values linearizations may not exit in closed form. In the present analysis, we consider a more general construction of these upwind schemes which avoids explicitly constructing these exact mean-value linearizations. Our construction is well-tailored to systems of conservation laws with convex entropy

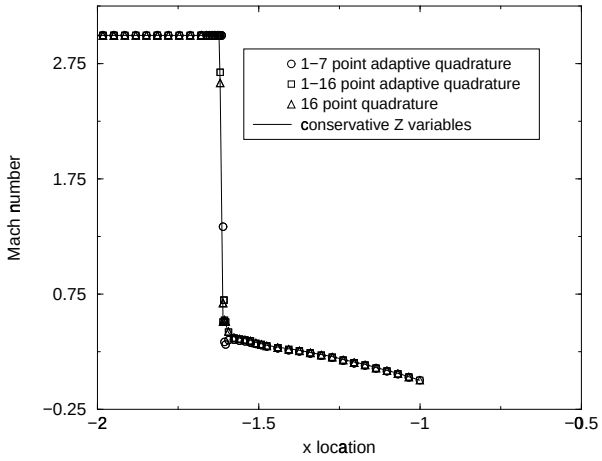


FIG. 7.11. *Comparison of the Mach number along a top-bottom symmetryline using the N-scheme with* $1-16$ *point adaptive quadrature,* 16 *point quadrature, and the conservative Z variables.*

extension. Using the tools of weak-$*$ convergence, a Lax-Wendroff theorem has been derived for this class of nonconservative schemes utilizing numerical quadrature. By using sufficient order numerical quadrature, we show that correct weak solutions of conservation laws are obtained. Numerical results confirm the basic analysis but do show some weak interdependence of the mesh space parameter h and the required order of accuracy of the numerical quadrature. This indicates that further investigation and quantification of this effect is needed.

Acknowledgements. This research was conducted while the first author was visiting RIACS, NASA Ames Research Center during 1998 and the second author during a stay at University Bordeaux I in 2000. The authors gratefully acknowledge NSF/INRIA program grant INT-0072863 and the INRIA research grant "Multiresolution Algorithms for Unstructured Meshes in Engineering". The authors also thank Katherine Mer–Nkonga of the French Atomic Commission for all the suggestions she has given to complete Sect. 3.

REFERENCES

[1] R. Abgrall. Toward the ultimate conservative scheme : following the Quest. *J. Comp. Phys*, Accepted, 2000.
[2] R. Abgrall and K. Mer. Un théorème de type Lax–Wendroff pour les schémas distributifs. Technical report, Mathématiques Appliquées de Bordeaux, 1998. Technical Report 98010.
[3] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, Cambridge, England, 1996.
[4] T. J. Barth. Simplified discontinuous Galerkin methods for systems of conservation laws with convex extension, 1999. Proceedings of the 1st International Conference on Discontinuous Galerkin Methods.
[5] T. J. Barth. Some working notes on energy analysis of the matrix N-scheme, May 1996. NASA Ames Research Center, Moffett Field, California, USA.
[6] T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In Kröner, Ohlberger, and Rohde, editors, *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, volume 5 of *Lecture Notes in Computational Science and Engineering*, pages 195–285. Springer-Verlag, Heidelberg, 1998.
[7] P.I. Crumpton, J.A. MacKenzie, and K.W. Morton. Cell vertex algorithms for the compressible Navier-Stokes equations. *J. Comp. Phys.*, 109:1–15, 1993.
[8] H. Deconinck, R. Struijs, and P. L. Roe. Compact advection schemes on unstructured grids. Technical report, VKI, 1993. VKI LS 1993-04, Computational Fluid Dynamics.
[9] A. Dervieux, B. van Leer, J. Priaux, and A. Rizzi, editors. *Numerical Simulation of Compressible Euler Flows*, volume 26 of *Note on Numerical Fluid mechanics*. Vieweg, 1989.
[10] S. K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47, 1959.
[11] S. K. Godunov. An interesting class of quasilinear systems. *Dokl. Akad. Nauk. SSSR*, 139:521–523, 1961.
[12] G. Golub, March 11, 1996. Stanford University, Private Communication.
[13] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25:35–61, 1983.
[14] T. Y. Hou and P. G. Le Floch. Why nonconservative schemes converge to wrong solutions : error analysis. *Math. Comp.*, 62(206):497–530, 1994.

[15] T. J. R. Hughes, L. P. Franca, and M. Mallet. A new finite element formulation for CFD: I. symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comp. Meth. Appl. Mech. Engrg.*, 54:223–234, 1986.

[16] T. J. R. Hughes and M. Mallet. A new finite element formulation for CFD: III. the generalized streamline operator for multidimensional advective-diffusive systems. *Comp. Meth. Appl. Mech. Engrg.*, 58:305–328, 1986.

[17] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, Cambridge, 1987.

[18] C. Johnson and A. Szepessy. Convergence of the shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws. *Math. Comp.*, 54:107–129, 1990.

[19] D. Kröner. *Numerical Schemes for Conservation Laws.* John Wiley and B.G. Teubner, New York, USA, 1997.

[20] D. Kröner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-d. *East-West J. Numer. Math.*, 4(4):279–292, 1996.

[21] P. Lax and B. Wendroff. Systems of conservation laws. *Comm. Pure Appl. Math.*, 13:217–237, 1960.

[22] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves.* SIAM, Philadelphia, Penn., 1973.

[23] M. S. Mock. Systems of conservation laws of mixed type. *J. Diff. Eqns.*, 37:70–88, 1980.

[24] K.W. Morton and E. Süli. Finite volume methods and their analysis. *IMA Journal of Numerical Analysis*, 11:241–260, 1991.

[25] R.-H. Ni. A multiple grid scheme for solving the Euler equations. *AIAA J.*, 20:1565–1571, 1981.

[26] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43, 1981.

[27] P. L. Roe. Linear advection schemes on triangular meshes. Technical Report CoA 8720, Cranfield Institute of Technology, 1987.

[28] P. L. Roe. "Optimum" upwind advection on a triangular mesh. Technical report, ICASE, NASA Langley R.C., 1990.

[29] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method.* Prentice–Hall, Englewood Cliffs, N.J., 1973.

[30] R. Struijs, H. Deconinck, and P. L. Roe. Fluctuation splitting schemes for the 2d Euler equations. Technical report, VKI, 1991. VKI LS 1991-01, Computational Fluid Dynamics.

[31] E. van der Weide and H. Deconinck. Positive matrix distribution schemes for hyperbolic systems. In *Computational Fluid Dynamics '96*, pages 747–753. Wiley, 1996.

[32] B. van Leer. Towards the ultimate conservative difference schemes V. A second order sequel to Godunov's method. *J. Comp. Phys.*, 32, 1979.