# The Relaxation Schemes for Systems of Conservation Laws in Arbitrary Space Dimensions

SHI JIN
*Georgia Institute of Technology*

AND

ZHOUPING XIN
*Courant Institute*

**Abstract**

We present a class of numerical schemes (called the relaxation schemes) for systems of conservation laws in several space dimensions. The idea is to use a local relaxation approximation. We construct a linear hyperbolic system with a stiff lower order term that approximates the original system with a small dissipative correction. The new system can be solved by underresolved stable numerical discretizations without using either Riemann solvers spatially or a nonlinear system of algebraic equations solvers temporally. Numerical results for 1-D and 2-D problems are presented. The second-order schemes are shown to be total variation diminishing (TVD) in the zero relaxation limit for scalar equations. 

## 1. Introduction

In this paper we present a class of nonoscillatory numerical schemes for systems of conservation laws in several space dimensions. The basic idea is to use a local relaxation approximation. For any given system of conservation laws, we will construct a corresponding linear hyperbolic system with a stiff source term that approximates the original system with a small dissipative correction. The special structure of the new system enables one to solve it numerically with underresolved stable discretizations without using either Riemann solvers spatially or nonlinear systems of algebraic equations solvers temporally.

The idea can be illuminated by considering the Cauchy problem for the 1-D scalar conservation law of the form

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0 , \qquad (x,t) \in \mathbb{R}^1 \times \mathbb{R}_+ , \quad u \in \mathbb{R}^1 \tag{1.1}$$

with initial data

$$u(x,0) = u_0(x) . \tag{1.2}$$

We introduce a linear system with a stiff lower order term (hereafter called the *relaxation system*):

$$\begin{aligned} &\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} v = 0 , \qquad v \in \mathbb{R}^1 \\ &\frac{\partial}{\partial t} v + a \frac{\partial}{\partial x} u = -\frac{1}{\varepsilon} (v - f(u)) \end{aligned} \tag{1.3}$$

Communications on Pure and Applied Mathematics, Vol. XLVIII, 235–276 (1995)
 CCC 0010-3640/95/030235-42

where the small positive parameter $\varepsilon$ is the *relaxation rate*, and $a$ is a positive constant satisfying

$$-\sqrt{a} \le f'(u) \le \sqrt{a} \quad \text{for all} \quad u \,. \tag{1.4}$$

We choose the initial condition for (1.3) as

$$u(x,0) = u_0(x) \,, \qquad v(x,0) = v_0(x) = f(u_0(x)) \,. \tag{1.5}$$

In the small relaxation limit ($\varepsilon \to 0^+$), the relaxation system (1.3) can be approximated to leading order by

$$v = f(u) \,, \tag{1.6a}$$

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0 \,. \tag{1.6b}$$

The state satisfying (1.6a) is called the *local equilibrium*, and (1.6b) is the original conservation law (1.1). The first order approximation to (1.3) is

$$v = f(u) - \varepsilon (a - f'(u)^2) \frac{\partial}{\partial x} u \,, \tag{1.7a}$$

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = \varepsilon \frac{\partial}{\partial x} \left( (a - f'(x)^2) \frac{\partial}{\partial x} u \right) \,, \tag{1.7b}$$

which can be derived using the Chapman-Enskog expansion [2]. It is clear that (1.7b) is dissipative provided condition (1.4) is satisfied. We choose the special initial condition (1.5) for $v$ to avoid the introduction of an initial layer through the relaxation system. In doing so the state is already in the equilibrium state initially. Likewise, in boundary value problems, we impose the boundary conditions for $v$ that are consistent to the local equilibrium (this will be made more precise in Section 5). Thus we do not introduce any new boundary layers either.

By eliminating $v$ from (1.3) we can obtain the following equation:

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = \varepsilon \left( a \frac{\partial^2}{\partial x^2} u - \frac{\partial^2}{\partial t^2} u \right) \,. \tag{1.8}$$

This introduces a higher order wave perturbation of the original conservation law. Given (1.4), the characteristic $f'(u)$ of the lower order wave is bounded by the characteristics $\pm\sqrt{a}$ of the higher order wave. The linear stability of this wave hierarchy was studied by Whitham in [28]. For $\varepsilon$ small, applying the asymptotics $\frac{\partial}{\partial t} u = -\frac{\partial}{\partial x} f(u) + O(\varepsilon)$ in (1.8) again recovers the parabolic equation (1.7b).

The relaxation limit for nonlinear systems was first studied by Liu in [16]. The dissipative entropy condition was formulated for general nonlinear relaxation

systems later by Chen, Levermore and Liu in [3]. It was shown in [16] and [3] that (1.6) and (1.7) indeed govern the asymptotic behavior of the relaxation system (1.3) either as time approaches infinity or as the relaxation rate $\varepsilon$ tends to zero.

Note that the characteristic speed $\lambda = f'(u)$ for the original equation (1.1) interlaces with those $(\pm\sqrt{a})$ of relaxation system (1.3) in the sense of (1.4). This is referred to as the *subcharacteristic condition* by Liu [16].

One would expect that appropriate numerical discretizations to the relaxation system (1.3) yield accurate approximations to the original equation (1.1) when the relaxation rate $\varepsilon$ is sufficiently small. The main advantage of numerically solving the relaxation system (1.3) over the original conservation law lies in the special structure of the linear characteristic fields and the localized lower order term. The linear hyperbolic structure enables one to avoid the time-consuming Riemann solvers, which are loaded by standard high resolution methods for nonlinear hyperbolic systems; proper implicit time discretizations can be taken to overcome the stability constraints brought in by the stiffness. Moreover, since the lower order term is not fully ranked and is linear in the artificially introduced variable $v$, solving nonlinear systems of algebraic equations can be avoided.

Numerical schemes for such stiff relaxation problems in the more general nonlinear setting were studied by Jin and Levermore in [13] and by Jin in [12]. Special care must be taken to ensure that the scheme possesses the correct zero relaxation limit. A scheme for (1.3) is said to have the correct zero relaxation limit if, as $\varepsilon \to 0$, for fixed grids, it has a discrete analog of the continuous asymptotic limit that leads from (1.3) to (1.6). In [12], a second-order TVD Runge-Kutta splitting time discretization was developed by Jin for the temporal discretization of such stiff relaxation schemes which, when combined with the standard high order Godunov schemes, produces satisfactory results for these problems. By keeping the convection terms explicit and the lower order terms implicit, stable discretizations can be achieved with a time step that solely depends on the convection term and is independent of $\varepsilon$.

The main features of the relaxation schemes proposed in this paper are their simplicity and generality. General systems of higher dimensions can be treated in the same way as in the one dimension. At the relaxation level the constant linear characteristic fields allow one to apply standard upwind schemes without solving the Riemann problem, and the field-by-field decomposition becomes trivial. One can indeed design highly efficient and well vectorized algorithms. Our goal in this paper is to give a broad, general study of this approach for general systems of conservation laws in arbitrary space dimensions. We remark here that although the numerical experiments carried out in this paper are on the compressible Euler equations which are hyperbolic, the formulation of the relaxation schemes does not have that restriction. Thus they are formally valid also for problems that are not strictly hyperbolic and exhibit local linear degeneracies, and other problems of mixed types with elliptic regions.

We call the discretizations of the relaxation system (1.3) the *relaxing schemes*. The relaxing schemes depend on the artifical variable $v$ as well as the small relaxation rate $\varepsilon$. By applying the Hilbert or Chapman-Enskog expansion to the

relaxing schemes (for fixed girds and $\varepsilon \to 0$), we can also formally derive the *relaxed schemes* that are the leading order approximation of the relaxing schemes in the small $\varepsilon$ limit. These relaxed schemes are consistent and stable discretizations of the original conservation laws. They are shown to be TVD for the scalar conservation laws, thus are non-oscillatory and converge to the weak solution of the original conservation law. Here by relaxation schemes we indicate both the relaxing schemes and the relaxed schemes. When $\varepsilon$ is very small the relaxing schemes and the relaxed schemes produce essentially the same results, which has been observed in our numerical tests on the compressible Euler equations. For more general conservation laws the validity of the Chapman-Enskog expansion for the relaxation system is still unclear.

The local relaxation approximation to conservation laws is close in spirit to the description of the hydrodynamic equations by the detailed microscopic evolution of gases in kinetic theory. It is always expected that the complexity and analytic difficulties occurring at the macroscopic level can be explained and illuminated by the hydrodynamic limit process. The Boltzmann schemes introduced by Harten, Lax, and van Leer for linear systems [11] are based on this idea. Similar kinetic schemes were developed for the compressible Euler equations by, for example, Dashpande [6], Reitz [23], Pullin [22] and Perthame [20]. Our method also shares some spirit of the lattice Boltzmann schemes (see for example [7] and references therein). It seems that so far all these Boltzmann type schemes were designed specifically for the gas dynamics equations. Morawetz made a similar relaxation approach for transonic and other problems [18] using an elliptic relaxation scheme. For theoretical study of a scalar conservation law, Giga and Miyakawa [9] and Perthame and Tadmor [21] have respectively constructed artificial kinetic approximations which furnish the entropy conditions.

The rest of this paper is organized as follows. The relaxation system for general systems of conservation laws in several space variables is proposed in Section 2. The dissipative structure of the first-order correction to the original system is derived. We generalize the subcharacteristic conditions (1.4) of Liu to arbitrary systems of conservation laws with convex entropy extensions. We emphasize that the subcharacteristic condition not only guarantees the dissipative nature of the reduced system but also plays a critical role for numerical stability. As is the case for the continuous equations, the fact that the first-order correction to the original system has a dissipative structure implies that the numerical solutions to the relaxation system should also converge to the entropy solution of the original system. In Section 3, we give the details of numerical schemes solving the relaxation systems. Upwind schemes for the linear convection terms are combined with a second order TVD Runge-Kutta splitting scheme proposed by Jin [12]. These schemes are shown to have the correct zero relaxation limit as $\varepsilon \to 0$ in the sense that the limits (the relaxed schemes) are consistent and stable discretizations of the original conservation laws. We analyze the relaxed schemes in Section 4. It is shown that for scalar equations in one space variable, the second-order relaxed schemes based on the MUSCL type approximation [27] to the linear characteristic fields of the relaxation system are TVD for a large class of slope-limiter

functions. The numerical results for 1-D and 2-D gas dynamics problems are presented in Section 5. While enjoying its simplicity and efficiency, our second-order relaxation scheme seems to produce very promising results that are comparable to those produced by other high resolution methods. We conclude this paper with a few remarks in Section 6.

## 2. The Relaxation Systems

### The Relaxation Systems for 1-D Systems of Conservation Laws

Consider systems of conservation laws in one space variable

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) = 0\,, \quad (x,t) \in \mathbb{R}^1 \times \mathbb{R}\,, \quad \mathbf{u} \in \mathbb{R}^n\,, \tag{2.1}$$

where $F(\mathbf{u}) \in \mathbb{R}^n$ is a smooth vector-valued function. We introduce the *relaxation system* corresponding to (2.1) as

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{v} = 0\,, \qquad \mathbf{v} \in \mathbb{R}^n\,, \\ &\frac{\partial}{\partial t}\mathbf{v} + A\frac{\partial}{\partial x}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{v} - F(\mathbf{u}))\,, \quad \varepsilon > 0\,, \end{aligned} \tag{2.2}$$

where

$$A = \operatorname{diag}\{a_1, a_2, \ldots, a_n\} \tag{2.3}$$

is a positive diagonal matrix to be chosen. For small $\varepsilon$, applying the Chapman-Enskog expansion in the relaxation system (2.2), one can derive the following approximation for $\mathbf{u}$ as [16], [3]

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) = \varepsilon\frac{\partial}{\partial x}\left((A - F'(\mathbf{u})^2)\frac{\partial}{\partial x}\mathbf{u}\right) \tag{2.4}$$

where $F'(\mathbf{u})$ is the Jacobian matrix of the flux function $F$. Equations (2.4) govern the first-order behavior of the relaxation system (2.2). To ensure the dissipative nature of (2.4), it is necessary that

$$A - F'(\mathbf{u})^2 \geqq 0 \quad \text{for all} \quad \mathbf{u}\,. \tag{2.5}$$

It is clear that for $\mathbf{u}$ varying in a bounded domain, equation (2.5) can always be satisfied by choosing sufficiently large $A$. However, because of the CFL constraints on numerical stability, it is desirable to obtain the smallest $A$ meeting the criterion (2.5).

Assume that the original system (2.1) admits a convex entropy extension [8]. Let $(\eta(\mathbf{u}), q(\mathbf{u}))$ be an entropy-entorpy flux pair. Then

$$\eta'(\mathbf{u}) \cdot F'(\mathbf{u}) = q'(\mathbf{u})\,, \quad \text{and} \quad \eta''(\mathbf{u}) > 0\,. \tag{2.6}$$

Here the double prime denotes the Hessian matrix. It follows easily from (2.6) that both $\eta''F'$ and $F'(\eta'')^{-1}$ are symmetric matrices. Let $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$, and $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_n$ be the eigenvalues of $\eta''F'$ and $F'(\eta'')^{-1}$ respectively. They can be regarded as the generalized eigenvalues of $F'$ with respect to the matrices $\eta''$ and $(\eta'')^{-1}$ respectively, in the sense that

$$\det(F' - \lambda_i \eta'') = 0\,, \quad \det(F' - \mu_i(\eta'')^{-1}) = 0\,, \quad 1 \leqq i \leqq n\,. \tag{2.7}$$

Let

$$\lambda = \max_{1 \leqq i \leqq n} |\lambda_i|\,, \qquad \mu = \max_{1 \leqq i \leqq n} |\mu_i|\,. \tag{2.8}$$

Although for both theoretical and computational purposes it is sometimes necessary to choose $A$ to have distinct diagonal elements so as to avoid any degeneracy in the relaxation system, here for simplicity we will assume that $A$ has the special form

$$A = aI\,, \qquad a > 0 \tag{2.9}$$

where $I$ is the $n \times n$ identity matrix. Note that the relaxation system always has a complete set of eigenvectors, which are the characteristic variables $\mathbf{v} \pm A^{1/2}\mathbf{u}$, even with the choice of $A$ as in (2.9).

PROPOSITION 2.1. *The reduced system* (2.4) *is dissipative provided that*

$$\frac{\lambda\mu}{a} \leqq 1\,. \tag{2.10}$$

Proof: See Appendix.

*Remarks.*

(i) In general, instead of using (2.9), one may choose the general form of $A$ in (2.3) with distinct diagonal elements in such a way that the relaxation system (2.2) is not degenerate. In this case a sufficient condition for the dissipation condition is

$$\frac{\lambda\mu}{\min_{1 \leqq i \leqq n} a_i} \leqq 1\,. \tag{2.11}$$

(ii) Condition (2.10) is a natural generalization of the subcharacteristic condition of Liu. Since for $n = 1$, one can take $\eta(u) = \frac{1}{2}u^2$, then the generalized eigenvalues are the "genuine" eigenvalues of $f'$ so (2.10) is exactly (1.4).

(iii) In general, one cannot expect the interlace relations between the characteristic speeds for relaxation systems and those of the original system as in [3]. This

is due to the fact that the relaxation system (2.2) does not admit convex entropy extension in general.

(iv) The formulation of the dissipative condition in (2.10) by using generalized eigenvalues is aimed solely at generalizing to the multi-dimensional case. In the case of one space variable, with a slightly different definition of the dissipation, one can obtain a stronger result in terms of only the genuine eigenvalues of $F'(\mathbf{u})$. To clarify this, we say that system (2.2) is dissipative if it is strictly stable in the sense of Majda-Pego [17], i.e.,

$$L(A - F'(\mathbf{u})^2)R > 0 \; ; \tag{2.12}$$

here $L$ and $R$ are matrices such that

$$\begin{aligned} LF'(\mathbf{u})R &= \operatorname{diag}\{\lambda_1(\mathbf{u}), \ldots, \lambda_n(\mathbf{u})\} \equiv \Lambda \; , \\ LR &= I \; . \end{aligned} \tag{2.13}$$

In (2.13), $\lambda_1(\mathbf{u}) \leqq \lambda_2(\mathbf{u}) \leqq \cdots \leqq \lambda_n(\mathbf{u})$ are the genuine eigenvalues of $F'(\mathbf{u})$, and $L$ and $R$ are respectively the column and row matrices of left and right eigenvectors of $F'(\mathbf{u})$ associated with these eigenvalues. Under the assumption (2.9), it is clear that condition (2.12) is equivalent to

$$aI - \Lambda^2 > 0$$

which is satisfied provided

$$\frac{\lambda^2}{a} < 1 \; ; \tag{2.14}$$

here $\lambda = \max_{1 \leqq i \leqq n} |\lambda_i(\mathbf{u})|$.

**The Relaxation System for Multi-Dimensional Systems of Conservation Laws**

We now generalize the previous analysis to systems of conservation laws in several space variables. Consider the following $m$-dimensional $n \times n$ system

$$\frac{\partial}{\partial t}\mathbf{u} + \sum_{i=1}^{m} \frac{\partial}{\partial x_i} F_i(\mathbf{u}) = 0 \; , \qquad (x_i, t) \in \mathbb{R}^1 \times \mathbb{R}_+, \; \mathbf{u} \in \mathbb{R}^n \; , \tag{2.15}$$

where each $F_i$: $\mathbb{R}^n \to \mathbb{R}^n$ is assumed to be a smooth function ($i = 1, 2, \ldots, m$). We will introduce the relaxation system corresponding to (2.15) as the following systems of $n \cdot (m + 1)$ equations

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u} + \sum_{i=1}^{m} \frac{\partial}{\partial x_i}\mathbf{v}_i = 0 \; , \qquad \mathbf{v}_i \in \mathbb{R}^n \\ &\frac{\partial}{\partial t}\mathbf{v}_i + A_i \frac{\partial}{\partial x_i}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{v}_i - F_i(\mathbf{u})) \; , \qquad i = 1, 2, \ldots, m \; . \end{aligned} \tag{2.16}$$

As is the case for 1-D, one usually chooses $A_i$ with distinct positive diagonal elements so the relaxation system (2.16) is non-degenerate. Here for simplicity in our presentation we take

$$A_i = a_i I\,, \qquad a_i > 0\,, \quad i = 1,\dots,m\,. \tag{2.17}$$

In a similar way as for (2.2), one can show that the first-order formal asymptotic behavior of (2.16) with a small relaxation rate is governed by the following system of equations:

$$\frac{\partial}{\partial t}\mathbf{u} + \sum_{i=1}^{m} \frac{\partial}{\partial x_i} F_i(\mathbf{u}) = \varepsilon \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left( [\delta_{ij}A_i - F_i'(\mathbf{u})F_j'(\mathbf{u})]\frac{\partial}{\partial x_j}\mathbf{u}\right) \tag{2.18}$$

where $\delta_{ij}$ is the Kronecker delta. It is clear that one form of the dissipative condition for (2.18) is

$$\delta_{ij}A_i - F_i'(\mathbf{u})F_j'(\mathbf{u}) \geqq 0 \quad \text{on} \quad \mathbb{R}^{mn}\,. \tag{2.19}$$

Note that the left-hand side of (2.19) is an $(mn) \times (mn)$ matrix. The condition (2.19) is satisfied for $\mathbf{u}$ in a bounded domain by choosing a sufficiently large $A$. However, for the same reason as in the 1-D case, we want to choose the smallest $A$ satisfying (2.19). This more precise dissipative condition is formulated in terms of the entropy variable for systems of conservation laws which admit convex extensions. As observed by Friedrichs and Lax, almost all equations of classical physics of form (2.15) admit convex entropy extensions [8].

Let $(\eta(\mathbf{u}), Q(\mathbf{u}))$ be an entropy-entropy flux pair for (2.15), $Q(\mathbf{u}) = (q_1(\mathbf{u}), \dots, q_m(\mathbf{u}))$, and $\eta(\mathbf{u})$ a convex function. Then

$$\eta''(\mathbf{u}) > 0\,, \quad \eta'(\mathbf{u}) \cdot F_i'(\mathbf{u}) = q_i'(\mathbf{u})\,, \quad i = 1,\dots,m\,. \tag{2.20}$$

One deduces easily from (2.20) that both $\eta''(\mathbf{u})F_j'(\mathbf{u})$ and $F_i'(\mathbf{u})\eta''(\mathbf{u})^{-1}$ are symmetric matrices. Now multiplying (2.18) by the entropy variable

$$\mathbf{w} = \eta'(\mathbf{u}) \tag{2.21}$$

gives

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{w} + \sum_{i=1}^{m} \eta''(\mathbf{u})F_i'(\mathbf{u})\eta''(\mathbf{u})^{-1}\frac{\partial}{\partial x_i}\mathbf{w} = \varepsilon \sum_{i,j=1}^{m} \eta''(\mathbf{u}) \\ &\quad \times \frac{\partial}{\partial x_i}\left\{ \eta''(\mathbf{u})^{-1}[a_i\delta_{ij}I + \eta''(\mathbf{u})F_i'(\mathbf{u})F_j'(\mathbf{u})\eta''(\mathbf{u})^{-1}]\frac{\partial}{\partial x_j}\mathbf{w}\right\}\,. \end{aligned} \tag{2.22}$$

It follows that the system (2.22) is dissipative if

$$a_i\delta_{ij}I - (\eta''(\mathbf{u})F_i'(\mathbf{u}))(F_j'(\mathbf{u})\eta''(\mathbf{u})^{-1}) \geqq 0 \quad \text{on} \quad \mathbb{R}^{mn}\,. \tag{2.23}$$

Let $\lambda_k^{(i)}$ and $\mu_k^{(i)}$ be the eigenvalues of $F_i'(\mathbf{u})\eta''(\mathbf{u})^{-1}$ and $\eta''(\mathbf{u})F_i(\mathbf{u})$ respectively (then both $\lambda_k^{(i)}$ and $\mu_k^{(i)}$ are real), $i = 1,\dots,m$, $k = 1,\dots,n$, i.e., $\lambda_m^{(i)}$ and $\mu_k^{(j)}$ are the generalized eigenvalues of $F_i'(\mathbf{u})$ with respect to the matrices $\eta''(\mathbf{u})$ and $\eta''(\mathbf{u})^{-1}$ respectively. Setting

$$\lambda^{(i)} = \max_{k,\ell}\{|\lambda_k^{(i)}|, |\mu_\ell^{(i)}|\}\,, \tag{2.24}$$

one can obtain the following desirable dissipative condition.

PROPOSITION 2.2. *The reduced system* (2.22) ( *or* (2.18)) *is dissipative if*

$$\frac{(\lambda^{(1)})^2}{a_1} + \frac{(\lambda^{(2)})^2}{a_2} + \cdots + \frac{(\lambda^{(m)})^2}{a_m} \leqq 1 \tag{2.25}$$

Proof: See Appendix.

*Remarks.*

(i) For the scalar equation, $n = 1$, one may take $\eta(u) = \frac{1}{2}u^2$, then the condition (2.25) is reduced to

$$\frac{F_1'(u)^2}{a_1} + \frac{F_2'(u)^2}{a_2} + \cdots + \frac{F_m'(u)^2}{a_m} \leqq 1\,. \tag{2.26}$$

(ii) If, instead of requiring that there exists a convex entropy extension, one assumes that (2.16) is a symmetric system, then (2.25) can be derived quite easily with $\lambda^{(i)}$'s being genuine eigenvalues of $F_i'(\mathbf{u})$.

## 3. The Relaxation Schemes

In this section we discuss the discretizations of the relaxation systems. We call these discretizations the *relaxing schemes*, which depends on $\varepsilon$ and the artificial variable $\mathbf{v}$. We also derive the zero relaxation limit of these relaxing schemes and call the limiting schemes the *relaxed schemes*. The relaxed schemes are stable and conservative discretizations of the original conservation law, thus are independent of $\varepsilon$ and the artificial variable $\mathbf{v}$.

In our derivation we will focus on the 1-D system. Later in this section we will discuss the discretizations of the multidimensional relaxation systems, which are just the natural dimension-by-dimension extension of the 1-D case.

Consider the 1-D relaxation system

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{v} = 0\,,\\ &\frac{\partial}{\partial t}\mathbf{v} + A\frac{\partial}{\partial x}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{v} - F(\mathbf{u}))\,, \end{aligned} \tag{3.1}$$

where $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^n$, $x \in \mathbb{R}^1$, $t > 0$, and

$$A = \operatorname{diag}\{a_1, a_2, \ldots, a_n\}, \quad a_i > 0 \quad (1 \leqq i \leqq n). \tag{3.2}$$

For $\varepsilon$ sufficiently small, it is expected that by solving (3.1) properly, one can obtain good approximations to the original conservation laws

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) = 0\,. \tag{3.3}$$

Numerical schemes for nonlinear hyperbolic conservation laws with stiff source terms that share the same relaxation limit as that of (3.1) have been studied by Jin and Levermore [13] for semidiscrete schemes, by Jin [12] for implicit time discretizations, and by Caflisch, Jin and Russo for the Broadwell model of the nonlinear Boltzmann equation [1]. Adopting the approach in these works, a good numerical discretization for (3.1) should possess a discrete analogy of the continuous zero relaxation limit, in the sense that the zero relaxation limit of the numerical discretization (let $\varepsilon \to 0$ for a fixed mesh) should be a consistent and stable discretization to (3.3).

We introduce the spatial grid points $x_{j+1/2}$ with mesh width $h_j = x_{j+1/2} - x_{j-1/2}$. The discrete time level $t_n$ are spaced uniformly with space step $k = t_{n+1} - t_n$ for $n = 0, 1, 2, \ldots$ . As usual, we denote by $\mathbf{w}_j^n$ the approximate cell average of a quantity $\mathbf{w}$ in the cell $[x_{j-1/2}, x_{j+1/2}]$ at time $t_n$, and by $\mathbf{w}_{j+1/2}^n$ the approximate point value of $\mathbf{w}$ at $x = x_{j+1/2}$ and $t = t_n$.

For the relaxation system (3.1), it is convenient to treat the spatial discretization and time discretization separately. This is known as the method of lines.

**The Relaxing Schemes**

A spatial discretization to (3.1) in conservation form can be written as

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u}_j + \frac{1}{h_j}(\mathbf{v}_{j+1/2} - \mathbf{v}_{j-1/2}) = 0\,, \\ &\frac{\partial}{\partial t}\mathbf{v}_j + \frac{1}{h_j}A(\mathbf{u}_{j+1/2} - \mathbf{u}_{j-1/2}) = -\frac{1}{\varepsilon}(\mathbf{v}_j - F_j)\,, \end{aligned} \tag{3.4}$$

where the averaged quantity $F_j$ is defined by

$$\begin{aligned} F_j &= \frac{1}{h_j}\int_{x_j-1/2}^{x_j+1/2} F(\mathbf{u})\,dx = F\left(\frac{1}{h_j}\int_{x_j-1/2}^{x_j+1/2}\mathbf{u}\,dx\right) + O(h^2) \\ &= F(\mathbf{u}_j) + O(h^2) \end{aligned} \tag{3.5}$$

for $h = \max_j h_j$. Thus for sufficiently accurate spatial discretizations, we have, with an accuracy of $O(h^2)$,

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u}_j + \frac{1}{h_j}(\mathbf{v}_{j+1/2} - \mathbf{v}_{j-1/2}) = 0\,, \\ &\frac{\partial}{\partial t}\mathbf{v}_j + \frac{1}{h_j}A(\mathbf{u}_{j+1/2} - \mathbf{u}_{j-1/2}) = -\frac{1}{\varepsilon}(\mathbf{v}_j - F(\mathbf{u}_j))\,. \end{aligned} \tag{3.6}$$

Here the point value quantities $\mathbf{u}_{j+1/2}$ and $\mathbf{v}_{j+1/2}$ will be defined by upwind schemes specified below. Since the relaxation system (3.1) has two characteristic variables

$$\mathbf{v} \pm A^{1/2}\mathbf{u} \tag{3.7}$$

that travel with the *frozen* characteristic speeds $\pm A^{1/2}$ respectively, the upwind approximations will be applied to $\mathbf{v} \pm A^{1/2}\mathbf{u}$ respectively.

**The Upwind Scheme.** Applying a first-order upwind scheme to $\mathbf{v} \pm A^{1/2}\mathbf{u}$ gives respectively

$$(\mathbf{v} + A^{1/2}\mathbf{u})_{j+1/2} = (\mathbf{v} + A^{1/2}\mathbf{u})_j\,, \qquad (\mathbf{v} - A^{1/2}\mathbf{u})_{j+1/2} = (\mathbf{v} - A^{1/2}\mathbf{u})_{j+1}\,. \tag{3.8}$$

Solving (3.8) for the unknowns $\mathbf{u}_{j+1/2}$ and $\mathbf{v}_{j+1/2}$ gives

$$\begin{aligned} \mathbf{u}_{j+1/2} &= \tfrac{1}{2}(\mathbf{u}_j + \mathbf{u}_{j+1}) - \tfrac{1}{2}A^{-1/2}(\mathbf{v}_{j+1} - \mathbf{v}_j)\,, \\ \mathbf{v}_{j+1/2} &= \tfrac{1}{2}(\mathbf{v}_j + \mathbf{v}_{j+1}) - \tfrac{1}{2}A^{1/2}(\mathbf{u}_{j+1} - \mathbf{u}_j)\,. \end{aligned} \tag{3.9}$$

Plugging (3.9) into (3.6) then gives the first-order semi-discrete upwind approximation to the relaxation system (3.1):

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u}_j + \frac{1}{2h_j}(\mathbf{v}_{j+1} - \mathbf{v}_{j-1}) - \frac{1}{2h_j}A^{1/2}(\mathbf{u}_{j+1} - 2\mathbf{u}_j + \mathbf{u}_{j-1}) = 0\,, \\ &\frac{\partial}{\partial t}\mathbf{v}_j + \frac{1}{2h_j}A(\mathbf{u}_{j+1} - \mathbf{u}_{j-1}) - \frac{1}{2h_j}A^{1/2}(\mathbf{v}_{j+1} - 2\mathbf{v}_j + \mathbf{v}_{j-1}) = -\frac{1}{\varepsilon}(\mathbf{v}_j - F(\mathbf{u}_j))\,. \end{aligned} \tag{3.10}$$

**The MUSCL Scheme.** To obtain a second-order scheme we use van Leer's MUSCL scheme [27]. Instead of using the piecewise constant interpolation (3.8) which yields a first-order approximation, the MUSCL uses the piecewise linear interpolation which, applied to the $p$-th components of $\mathbf{v} \pm A^{1/2}\mathbf{u}$, gives respectively

$$\begin{aligned} (v + \sqrt{a_p}\,u)_{j+1/2} &= (v + \sqrt{a_p}\,u)_j + \tfrac{1}{2}h_j\sigma_j^+\,, \\ (v - \sqrt{a_p}\,u)_{j+1/2} &= (v - \sqrt{a_p}\,u)_{j+1} - \tfrac{1}{2}h_{j+1}\sigma_{j+1}^-\,. \end{aligned} \tag{3.11}$$

Here $u$ and $v$ are the $p$-th ($1 \leqq p \leqq n$) components of $\mathbf{u}$ and $\mathbf{v}$ respectively, $\sigma_j^{\pm}$ is the slope of $v \pm \sqrt{a_p}\,u$ on the $j$-th cell. Using Sweby's notation [26],

$$\begin{aligned} \sigma_j^{\pm} &= \frac{1}{h_j}\left(v_{j+1} \pm \sqrt{a_p}\,u_{j+1} - v_j \mp \sqrt{a_p}\,u_j\right)\phi(\theta_j^{\pm})\,, \\ \theta_j^{\pm} &= \frac{v_j \pm \sqrt{a_p}\,u_j - v_{j-1} \mp \sqrt{a_p}\,u_{j-1}}{v_{j+1} \pm \sqrt{a_p}\,u_{j+1} - v_j \mp \sqrt{a_p}\,u_j}\,. \end{aligned} \tag{3.12}$$

One simple choice of slopes is the so-called *minmod* slope,

$$\phi(\theta) = \max(0, \min(1, \theta)). \tag{3.13}$$

A sharper slope-limiter function was introduced by van Leer [27] as

$$\phi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}. \tag{3.14}$$

For scheme (3.11) to be TVD (when the lower order term is not present), a more general condition for $\phi$ is [26]

$$0 \leqq \frac{\phi(\theta)}{\theta} \leqq 2 \quad \text{and} \quad 0 \leqq \phi(\theta) \leqq 2. \tag{3.15}$$

Solving (3.11) for $u_{j+1/2}$ and $v_{j+1/2}$ gives

$$\begin{aligned} u_{j+1/2} &= \tfrac{1}{2}(u_j + u_{j+1}) - \frac{1}{2\sqrt{a_p}}(v_{j+1} - v_j) + \frac{1}{4\sqrt{a_p}}(h_j\sigma_j^+ + h_{j+1}\sigma_{j+1}^-), \\ v_{j+1/2} &= \tfrac{1}{2}(v_j + v_{j+1}) - \frac{\sqrt{a_p}}{2}(u_{j+1} - u_j) + \tfrac{1}{4}(h_j\sigma_j^+ - h_{j+1}\sigma_{j+1}^-). \end{aligned} \tag{3.16}$$

Applying (3.16) in (3.6), one gets the MUSCL for the relaxation system (3.1) componentwise as

$$\begin{aligned} &\frac{\partial}{\partial t}u + \frac{1}{2h_j}(v_{j+1} - v_{j-1}) - \frac{\sqrt{a_p}}{2h_j}(u_{j+1} - 2u_j + u_{j-1}) \\ &\quad - \frac{1}{4h_j}(h_{j+1}\sigma_{j+1}^- - h_j(\sigma_j^+ + \sigma_{j-1}^-) + h_{j-1}\sigma_{j-1}^+) = 0, \\ &\frac{\partial}{\partial t}v + \frac{1}{2h_j}(u_{j+1} - u_{j-1}) - \frac{\sqrt{a_p}}{2h_j}(v_{j+1} - 2v_j + v_{j-1}) \\ &\quad + \frac{\sqrt{a_p}}{4h_j}(h_{j+1}\sigma_{j+1}^- + h_j(\sigma_j^+ - \sigma_{j-1}^-) - h_{j-1}\sigma_{j-1}^+) = -\frac{1}{\varepsilon}(v_j - F^{(p)}(\mathbf{u}_j)) \end{aligned} \tag{3.17}$$

where $F^{(p)}$ is the $p$-th component of $F$.

*Remarks.*

(i) By setting the slope $\sigma^{\pm} = 0$, the MUSCL (3.17) reduces to the first-order upwind scheme (3.10).

(ii) At the relaxation level the field-by-field decomposition can be done trivially since the relaxation system always has a complete set of eigenvectors and is easily diagonalizable.

**The Time Discretization.** A second-order TVD Runge-Kutta splitting scheme was introduced by Jin [12] that works very effectively for such problems. This

splitting scheme takes two implicit stiff source steps and two explicit convection steps alternatively. Denote

$$D_+\mathbf{w}_j = \frac{1}{h_j}(\mathbf{w}_{j+1/2} - \mathbf{w}_{j-1/2}) . \tag{3.18}$$

Then applying the second-order TVD Runge-Kutta splitting scheme to (3.1) gives

$$\mathbf{u}^* = \mathbf{u}^n , \tag{3.19a}$$

$$\mathbf{v}^* = \mathbf{v}^n + \frac{k}{\varepsilon}(\mathbf{v}^* - F(\mathbf{u}^*)) ; \tag{3.19b}$$

$$\mathbf{u}^{(1)} = \mathbf{u}^* - kD_+\mathbf{v}^* , \tag{3.19c}$$

$$\mathbf{v}^{(1)} = \mathbf{v}^* - kAD_+\mathbf{u}^* ; \tag{3.19d}$$

$$\mathbf{u}^{**} = \mathbf{u}^{(1)} , \tag{3.19e}$$

$$\mathbf{v}^{**} = \mathbf{v}^{(1)} - \frac{k}{\varepsilon}(\mathbf{v}^{**} - F(\mathbf{u}^{**})) - 2\frac{k}{\varepsilon}(\mathbf{v}^* - F(\mathbf{u}^*)) ; \tag{3.19f}$$

$$\mathbf{u}^{(2)} = \mathbf{u}^{**} - kD_+\mathbf{v}^{**} , \tag{3.19g}$$

$$\mathbf{v}^{(2)} = \mathbf{v}^{**} - kAD_+\mathbf{u}^{**} ; \tag{3.19h}$$

$$\mathbf{u}^{n+1} = \tfrac{1}{2}(\mathbf{u}^n + \mathbf{u}^{(2)}) , \tag{3.19i}$$

$$\mathbf{v}^{n+1} = \tfrac{1}{2}(\mathbf{v}^n + \mathbf{v}^{(2)}) . \tag{3.19j}$$

Here without ambiguity we have omitted the subscript $j$.

The idea of keeping the convection terms explicit has great advantages here. First, one does not need to solve systems of linear algebraic equations that will arise if the convection terms are implicit. Secondly and more importantly, due to the special structure of the source term, one does not need to solve any systems of nonlinear algebraic equations, in spite of the implicit nonlinear source terms. As a matter of fact, one can always update $\mathbf{u}$ from the first equation, applying it to $F(\mathbf{u})$, then solving the second equation for $\mathbf{v}$ exactly, since $\mathbf{v}$ is linear in the relaxation system.

Since the source terms are treated implicitly, this time discretization is stable independent of $\varepsilon$, given that the CFL condition from the convection is satisfied.

*Remark.* Note that (3.19b) is inconsistent to the relaxation systems due to a negative parameter. It will break down if $k = O(\varepsilon)$. However here we are developing underresolved numerical methods in the regime that $k \gg \varepsilon$, thus (3.19) will never break down in our case.

**The Relaxed Schemes**

It is known that not all stable discretizations to the relaxation system (3.1) yield the correct solutions when the small relaxation rate is not well resolved [12], [1]. A good numerical discretization should possess the correct zero relaxation limit in the sense that the zero relaxation limit ($\varepsilon \to 0^+$) is a consistent and stable discretization of (3.3).

As was demonstrated in [13] and [12], the discretizations above indeed have the correct zero relaxation limit. Consider (3.19). Assume $\varepsilon \ll 1$ and $\frac{\varepsilon}{h} \ll 1$, $\frac{\varepsilon}{k} \ll 1$ (this is the underresolved regime and such a mesh is called a coarse mesh). If the initial state is the local equilibrium, namely $\mathbf{v}(x,0) = F(\mathbf{u}(x,0))$ (this is exactly how we define the initial condition for $\mathbf{v}$ to avoid an initial layer), then it was shown in [12] that

$$\mathbf{v}^* = F(\mathbf{u}^*) + O\left(\frac{\varepsilon}{k}\right), \qquad \mathbf{v}^{**} = F(\mathbf{u}^{**}) + O\left(\frac{\varepsilon}{k}\right). \tag{3.20}$$

This means that at the two intermediate time steps, the solutions are local equilibria up to an error of $O(\frac{\varepsilon}{k})$. Thus for a second-order scheme a suitable choice of $\varepsilon$ will be

$$\varepsilon = o(k^3).$$

Applying (3.20) in (3.19c) and (3.19g) respectively one arrives at the *relaxed schemes* after ignoring the $O(\frac{\varepsilon}{k})$ terms:

$$\begin{aligned} \mathbf{u}^{(1)} &= \mathbf{u}^n - k D_+ \mathbf{v}^n \big|_{\mathbf{v}^n = F(\mathbf{u}^n)}, \\ \mathbf{u}^{(2)} &= \mathbf{u}^{(1)} - k D_+ \mathbf{v}^{(1)} \big|_{\mathbf{v}^{(1)} = F(\mathbf{u}^{(1)})}, \\ \mathbf{u}^{n+1} &= \tfrac{1}{2}(\mathbf{u}^n + \mathbf{u}^{(2)}). \end{aligned} \tag{3.21}$$

This is the second-order TVD Runge-Kutta time discretization [24] for (3.3). The relaxed spatial discretization can be worked out easily [13] by projecting the numerical flux $D_+\mathbf{v}$ into the local equilibrium $\mathbf{v} = F(\mathbf{u})$. By (3.9) the relaxed upwind scheme is

$$F_{j+1/2} = \mathbf{v}_{j+1/2}\big|_{\mathbf{v}=F(\mathbf{u})} = \tfrac{1}{2}(F(\mathbf{u}_j) + F(\mathbf{u}_{j+1})) - \tfrac{1}{2}A^{1/2}(\mathbf{u}_{j+1} - \mathbf{u}_j); \tag{3.22}$$

thus

$$\begin{aligned} D_+F_j = D_+\mathbf{v}_j &= \frac{1}{h_j}(\mathbf{v}_{j+1/2} - \mathbf{v}_{j-1/2}) \\ &= \frac{1}{2h_j}(F(\mathbf{u}_{j+1}) - F(\mathbf{u}_{j-1})) - \frac{1}{2h_j}A^{1/2}(\mathbf{u}_{j+1} - 2\mathbf{u}_j + \mathbf{u}_{j-1}). \end{aligned} \tag{3.23}$$

This is a Lax-Friedrichs type scheme which is a first-order consistent discretization to (3.3). Combining (3.21) with (3.23) gives our first-order relaxed scheme for (3.3).

Similarly, the relaxed MUSCL flux can be derived from (3.16) by applying the local equilibrium $\mathbf{v} = F(\mathbf{u})$ to (3.16),

$$\begin{aligned} F^{(p)}_{j+1/2} = v_{j+1/2}\big|_{v=F^{(p)}(\mathbf{u})} &= \tfrac{1}{2}(F^{(p)}(\mathbf{u}_j) + F^{(p)}(\mathbf{u}_{j+1})) - \tfrac{1}{2}\sqrt{a_p}(u_{j+1} - u_j) \\ &\quad + \tfrac{1}{4}(h_j\sigma^+_j - h_{j+1}\sigma^-_{j+1})\big|_{v=F^{(p)}(\mathbf{u})}, \end{aligned} \tag{3.24}$$

where, from (3.12),

$$
\begin{aligned}
\sigma_j^{\pm}\big|_{v=F^{(p)}(\mathbf{u})} &= \frac{1}{h_j}\left(F^{(p)}(\mathbf{u}_{j+1}) \pm \sqrt{a_p}\,u_{j+1} - F^{(p)}(\mathbf{u}_j) \mp \sqrt{a_p}\,u_j\right)\phi(\theta_j^{\pm}) , \\
\theta_j^{\pm} &= \frac{F^{(p)}(\mathbf{u}_j) \pm \sqrt{a_p}\,u_j - F^{(p)}(\mathbf{u}_{j-1}) \mp \sqrt{a_p}\,u_{j-1}}{F^{(p)}(\mathbf{u}_{j+1}) \pm \sqrt{a_p}\,u_{j+1} - F^{(p)}(\mathbf{u}_j) \mp \sqrt{a_p}\,u_j} .
\end{aligned}
\tag{3.25}
$$

Clearly the relaxed MUSCL is a second-order consistent discretization to (3.3). The combination of (3.21), (3.18), (3.24) and (3.25) gives our second-order relaxed scheme.

In conclusion, the relaxing schemes we discussed here have the correct zero relaxation limit. In this 1-D case the limiting relaxed flux (3.24) bears some similarity with the second-order non-oscillatory central differencing scheme developed by Nessyahu and Tadmor [19], in the sense that they both are not Riemann-solver based schemes, and the reconstruction (3.25) is on the primitive variables instead of the local characteristic variables based on the nonlinear field-by-field decomposition. Moreover, in the next section we will show that, for the scalar case, the first-order relaxed scheme (3.22) is monotone and the second-order relaxed scheme (3.24) and (3.25) is TVD, as long as the CFL condition and the subcharacteristic condition (1.4) are satisfied.

**The Higher Dimensional Case**

In the case of multidimensions, similar MUSCL discretization can be applied to each space dimension. Consider the multidimensional relaxation system

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{u} + \sum_{i=1}^{m}\frac{\partial}{\partial x_i}\mathbf{v}_i = 0 , \\
&\frac{\partial}{\partial t}\mathbf{v}_i + A_i\frac{\partial}{\partial x_i}\mathbf{u} = -\frac{1}{\varepsilon}\left(\mathbf{v}_i - F_i(\mathbf{u})\right) , \quad i = 1,2,\ldots,m .
\end{aligned}
\tag{3.26}
$$

Then the MUSCL discretization (3.11) and (3.12) can be applied to the characteristic variables $\mathbf{u} \pm A_i\mathbf{v}_i$ in the $i$-th space dimension for $i = 1,2,\ldots,m$. The second-order Runge-Kutta splitting scheme (3.19) applies here, with the only difference being that $D_+$ now becomes a multidimensional discrete derivative operator.

For reader's convenience we work out explicitly the 2-D scheme here. It can be easily generated to higher dimensional case. Consider

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{v} + \frac{\partial}{\partial y}\mathbf{w} = 0 , \\
&\frac{\partial}{\partial t}\mathbf{v} + A\frac{\partial}{\partial x}\mathbf{u} = -\frac{1}{\varepsilon}\left(\mathbf{v} - F(\mathbf{u})\right) , \\
&\frac{\partial}{\partial t}\mathbf{w} + B\frac{\partial}{\partial y}\mathbf{u} = -\frac{1}{\varepsilon}\left(\mathbf{w} - G(\mathbf{u})\right) .
\end{aligned}
\tag{3.27}
$$

Here $A = \operatorname{diag}\{a_1, \ldots, a_n\}$ and $B = \operatorname{diag}\{b_1, \ldots, b_n\}$ with $a_i > 0, b_i > 0$ for $1 \leqq i \leqq n$.

Suppose the spatial grid points are located at $(x_{i+1/2}, y_{j+1/2})$, the grid widths are $h_i^x = x_{i+1/2} - x_{i-1/2}$ and $h_j^y = y_{j+1/2} - y_{j-1/2}$ respectively. For a vector $\mathbf{U}(x, y)$, the point value is $\mathbf{U}_{i+1/2,j+1/2} = \mathbf{U}(x_{i+1/2}, y_{j+1/2})$, the cell average value is defined as

$$\mathbf{U}_{ij} = \frac{1}{h_i^x h_j^y} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{y_{j-1/2}}^{y_{j+1/2}} \mathbf{U}(x, y)\, dx\, dy .$$

A conservative, semi-discrete differencing to (3.27) is

$$\frac{\partial}{\partial t}\mathbf{u}_{ij} + \frac{1}{h_i^x}(\mathbf{v}_{i+1/2,j} - \mathbf{v}_{i-1/2,j}) + \frac{1}{h_j^y}(\mathbf{w}_{i,j+1/2} - \mathbf{w}_{i,j-1/2}) = 0 ,$$

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{v}_{ij} + A\frac{1}{h_i^x}(\mathbf{u}_{i+1/2,j} - \mathbf{u}_{i-1/2,j}) = -\frac{1}{\varepsilon}(\mathbf{v}_{ij} - F(\mathbf{u}_{ij})) , \\ &\frac{\partial}{\partial t}\mathbf{w}_{ij} + B\frac{1}{h_j^y}(\mathbf{u}_{i,j+1/2} - \mathbf{u}_{i,j-1/2}) = -\frac{1}{\varepsilon}(\mathbf{w}_{ij} - G(\mathbf{u}_{ij})) . \end{aligned} \tag{3.28}$$

Applying the MUSCL discretization to $\mathbf{v} \pm A^{1/2}\mathbf{u}$ in the x-direction and $\mathbf{w} \pm B^{1/2}\mathbf{u}$ in the y-direction respectively, then similar to (3.16), for the $p$-th component,

$$\begin{aligned} u_{i+1/2,j} &= \tfrac{1}{2}(u_{ij} + u_{i+1,j}) - \frac{1}{2\sqrt{a_p}}(v_{i+1,j} - v_{ij}) \\ &\quad + \frac{1}{4\sqrt{a_p}}(h_i^x \sigma_{ij}^{x,+} + h_{i+1}^x \sigma_{i+1,j}^{x,-}) , \\ v_{i+1/2,j} &= \tfrac{1}{2}(v_{ij} + v_{i+1,j}) - \frac{\sqrt{a_p}}{2}(u_{i+1,j} - u_{ij}) \\ &\quad + \tfrac{1}{4}(h_i^x \sigma_{ij}^{x,+} - h_{i+1}^x \sigma_{i+1,j}^{x,-}) ; \end{aligned} \tag{3.29a}$$

$$\begin{aligned} u_{i,j+1/2} &= \tfrac{1}{2}(u_{ij} + u_{i,j+1}) - \frac{1}{2\sqrt{b_p}}(w_{i,j+1} - w_{ij}) \\ &\quad + \frac{1}{4\sqrt{b_p}}(h_j^y \sigma_{ij}^{y,+} + h_{j+1}^y \sigma_{i,j+1}^{y,-}) , \\ w_{i,j+1/2} &= \tfrac{1}{2}(w_{ij} + w_{i,j+1}) - \frac{\sqrt{b_p}}{2}(u_{i,j+1} - u_{ij}) \\ &\quad + \tfrac{1}{4}(h_j^y \sigma_{ij}^{y,+} - h_{j+1}^y \sigma_{i,j+1}^{y,-}) . \end{aligned} \tag{3.29b}$$

The slope limiters are defined as

$$
\begin{aligned}
\sigma_{ij}^{x,\pm} &= \frac{1}{h_i^x}\left(v_{i+1,j} \pm \sqrt{a_p}\,u_{i+1,j} - v_{ij} \mp \sqrt{a_p}\,u_{ij}\right)\phi(\theta_{ij}^{x,\pm}) , \\
\theta_{ij}^{x,\pm} &= \frac{v_{ij} \pm \sqrt{a_p}\,u_{ij} - v_{i-1,j} \mp \sqrt{a_p}\,u_{i-1,j}}{v_{i+1,j} \pm \sqrt{a_p}\,u_{i+1,j} - v_{ij} \mp \sqrt{a_p}\,u_{ij}} ; \\
\sigma_{ij}^{y,\pm} &= \frac{1}{h_j^y}\left(w_{i,j+1} \pm \sqrt{b_p}\,u_{i,j+1} - w_{ij} \mp \sqrt{b_p}\,u_{ij}\right)\phi(\theta_{ij}^{y,\pm}) , \\
\theta_{ij}^{y,\pm} &= \frac{w_{ij} \pm \sqrt{b_p}\,u_{ij} - w_{i,j-1} \mp \sqrt{b_p}\,u_{i,j-1}}{w_{i,j+1} \pm \sqrt{b_p}\,u_{i,j+1} - w_{ij} \mp \sqrt{b_p}\,u_{ij}} .
\end{aligned} \tag{3.30}
$$

For the time discretization one can just use (3.19).

By passing to the $\varepsilon \to 0$ limit we obtain the 2-D second-order relaxed scheme for the system of 2-D conservation laws

$$
\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) + \frac{\partial}{\partial y}G(\mathbf{u}) = O_t , \tag{3.31}
$$

with

$$
\frac{\partial}{\partial t}\mathbf{u}_{ij} + \frac{1}{h_i^x}(F_{i+1/2,j} - F_{i-1/2,j}) + \frac{1}{h_j^y}(G_{i,j+1/2} - G_{i,j-1/2}) = 0 , \tag{3.32}
$$

where the $p$-the components of the numerical fluxes are given by (similar to (3.24)–(3.25))

$$
\begin{aligned}
F_{i+1/2,j}^{(p)} &= v_{i+1/2,j}\big|_{v=F^{(p)}(\mathbf{u})} \\
&= \tfrac{1}{2}(F^{(p)}(\mathbf{u}_{ij}) + F^{(p)}(\mathbf{u}_{i+1,j})) - \tfrac{1}{2}\sqrt{a_p}\,(u_{i+1,j} - u_{ij}) \\
&\quad + \tfrac{1}{4}\,(h_i^x\sigma_{ij}^{x,+} - h_{i+1}^x\sigma_{i+1,j}^{x,-})\big|_{v=F^{(p)}(\mathbf{u})} , \\
G_{i,j+1/2}^{(p)} &= w_{i,j+1/2}\big|_{w=G^{(p)}(\mathbf{u})} \\
&= \tfrac{1}{2}(G^{(p)}(\mathbf{u}_{ij}) + G^{(p)}(\mathbf{u}_{i,j+1})) - \tfrac{1}{2}\sqrt{b_p}\,(u_{i,j+1} - u_{ij}) \\
&\quad + \tfrac{1}{4}\,(h_j^y\sigma_{ij}^{y,+} - h_{j+1}^y\sigma_{i,j+1}^{y,-})\big|_{w=G^{(p)}(\mathbf{u})} ,
\end{aligned} \tag{3.33}
$$

with the slope limiters

$$
\begin{aligned}
&\sigma_{ij}^{x,\pm}\big|_{v=F^{(p)}(\mathbf{u})} \\
&\quad = \frac{1}{h_i^x}(F^{(p)}(\mathbf{u}_{i+1,j}) \pm \sqrt{a_p}\,u_{i+1,j} - F^{(p)}(\mathbf{u}_{ij}) \mp \sqrt{a_p}\,u_{ij})\,\phi(\theta_{ij}^{x,\pm})\,, \\
&\theta_{ij}^{x,\pm} = \frac{F^{(p)}(\mathbf{u}_{ij}) \pm \sqrt{a_p}\,u_{ij} - F^{(p)}(\mathbf{u}_{i-1,j}) \mp \sqrt{a_p}\,u_{i-1,j}}{F^{(p)}(\mathbf{u}_{i+1,j}) \pm \sqrt{a_p}\,u_{i+1,j} - F^{(p)}(\mathbf{u}_{ij}) \mp \sqrt{a_p}\,u_{ij}}\,; \\
&\sigma_{ij}^{y,\pm}\big|_{w=G^{(p)}(\mathbf{u})} \\
&\quad = \frac{1}{h_j^y}(G^{(p)}(\mathbf{u}_{i,j+1}) \pm \sqrt{b_p}\,u_{i,j+1} - G^{(p)}(\mathbf{u}_{ij}) \mp \sqrt{b_p}\,u_{ij})\,\phi(\theta_{ij}^{y,\pm})\,, \\
&\theta_{ij}^{y,\pm} = \frac{G^{(p)}(\mathbf{u}_{ij}) \pm \sqrt{b_p}\,u_{ij} - G^{(p)}(\mathbf{u}_{i,j-1}) \mp \sqrt{b_p}\,u_{i,j-1}}{G^{(p)}(\mathbf{u}_{i,j+1}) \pm \sqrt{b_p}\,u_{i,j+1} - G^{(p)}(\mathbf{u}_{ij}) \mp \sqrt{b_p}\,u_{ij}}\,,
\end{aligned}
\tag{3.34}
$$

with $\phi(\theta)$ given in (3.13) or (3.14). The time discretization is just the second-order Runge-Kutta method (3.21).

Since the formulation of the relaxation system is multidimensional, and the structure of the multidimensional relaxation system is similar to the 1-D system, the numerical implementations for higher dimensional problems based on the dimension-by-dimension discretizations are formally not much harder than for the 1-D problems. Again, all the multidimensional algorithms here are highly efficient and well vectorized. In fact, the only IF command in the entire algorithm appears in the slope limiter function (3.13) or (3.14).

## 4. TVD Properties of the Relaxed Schemes

Since the leading behavior of the relaxing schemes is governed by the relaxed schemes when $\varepsilon$ is sufficiently small, it is important to study the behavior of the relaxed scheme. In this section, we will prove the TVD property of the relaxed schemes for the 1-D scalar conservation law (1.1).

The simple one-step conservative scheme for the relaxation system (1.3) with uniform grids can be written in the form

$$
\begin{aligned}
&\frac{u_j^{n+1} - u_j^n}{k} + \frac{v_{j+1/2}^n - v_{j-1/2}^n}{h} = 0\,, \\
&\frac{v_j^{n+1} - v_j^n}{k} + a\,\frac{u_{j+1/2}^n - u_{j-1/2}^n}{h} = -\frac{1}{\varepsilon}\,(v_j^{n+1} - f(u_j^{n+1}))\,.
\end{aligned}
\tag{4.1}
$$

The point value quantities $u_{j+1/2}$ and $v_{j+1/2}$ are defined by upwind schemes in Section 3. Here we assume the uniform grids just for simplicity. We will study first-order and second-order upwind schemes separately.

**The Upwind Scheme**

As in Section 3, the first-order upwind relaxing flux to the relaxation system (1.3) is (see (3.8) and (3.9))

$$
\begin{aligned}
u^n_{j+1/2} &= \frac{1}{2}(u^n_j + u^n_{j+1}) - \frac{1}{2\sqrt{a}}(v^n_{j+1} - v^n_j)\,, \\
v^n_{j+1/2} &= \frac{1}{2}(v^n_j + v^n_{j+1}) + \frac{\sqrt{a}}{2}(u^n_{j+1} - u^n_j)\,.
\end{aligned} \tag{4.2}
$$

Applying these fluxes in (4.1) gives

$$
\begin{aligned}
&\frac{u^{n+1}_j - u^n_j}{k} + \frac{1}{2h}(v^n_{j+1} - v^n_{j-1}) - \frac{\sqrt{a}}{2h}(u^n_{j+1} - 2u^n_j + u^n_{j-1}) = 0\,, \\
&\frac{v^{n+1}_j - v^n_j}{k} + \frac{a}{2h}(u^n_{j+1} - u^n_{j-1}) - \frac{1}{2\sqrt{a}\,h}(v^n_{j+1} - 2v^n_j + v^n_{j-1}) \\
&\qquad = -\frac{1}{\varepsilon}(v^n_j - f(u^n_j))\,.
\end{aligned} \tag{4.3}
$$

Using a Hilbert expansion gives the leading order equations (as $\varepsilon \to 0^+$),

$$
\begin{aligned}
v^n_j &= f(u^n_j)\,, \\
u^{n+1}_j &= u^n_j - \frac{\lambda}{2}(f(u^n_{j+1}) - f(u^n_{j-1})) + \frac{\sqrt{a}\lambda}{2}(u^n_{j+1} - 2u^n_j + u^n_{j-1})\,,
\end{aligned} \tag{4.5}
$$

where

$$
\lambda = \frac{k}{h}\,. \tag{4.6}
$$

The scheme in (4.5) is a first-order relaxed scheme for (1.1) which is a Lax-Friedrichs type scheme.

A numerical scheme

$$
U^{n+1}_j = \mathbf{H}(U^n; j)
$$

is called a monotone scheme if

$$
\frac{\partial}{\partial U^n_i}\mathbf{H}(U^n; j) \geqq 0 \quad \text{for all} \quad i, j, U^n\,.
$$

It can be checked easily that (4.5) is a monotone scheme provided the standard CFL condition arising from the discrete convection terms

$$
\sqrt{a}\,\lambda \leqq 1 \tag{4.7}
$$

and the subcharacteristic condition (1.4) are satisfied. Thus we have the following theorem.

THEOREM 4.1. *Under the CFL condition* (4.7) *and the subcharacteristic condition* (1.4), *the relaxed scheme* (4.5) *is a monotone scheme.*

*Remarks.*

(i) As a consistent monotone scheme, the first-order relaxed scheme (4.5) with fixed $\lambda$ converges to the entropy solution as $k \to 0$ [5].

(ii) Similar conclusion holds for scalar equations in several space variables. We omit the details.

(iii) Since the convection part in the relaxation system (1.3) is linear with constant coefficients, one might think that it would be more accurate to use the characteristic method for the characteristic fields, since it is exact for the linear convection. However, this is not the case here. To see this, notice that the characteristic method corresponds to taking $\sqrt{a}\lambda = 1$ in (4.3), so that the relaxed scheme (4.5) becomes

$$u_j^{n+1} = u_j^n - \tfrac{1}{2}\lambda(f(u_{j+1}^n) - f(u_{j-1}^n)) + \tfrac{1}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) .$$

This is the original Lax-Friedrichs scheme that is of maximum numerical dissipation among all first-order schemes.

**The Second-Order Relaxed Schemes**

A second-order relaxing schemes for the relaxation system (1.3) is obtained by applying the MUSCL scheme to the linear characteristic fields as in Section 3 (see (3.11)–(3.17)). The resulting relaxing flux can be written as

$$\begin{aligned} u_{j+1/2} &= \frac{1}{2}(u_j + u_{j+1}) - \frac{1}{2\sqrt{a}}(v_{j+1} - v_j) + \frac{h}{4\sqrt{a}}(1-\beta)(\sigma_j^+ + \sigma_{j+1}^-), \\ v_{j+1/2} &= \frac{1}{2}(v_j + v_{j+1}) - \frac{\sqrt{a}}{2}(u_{j+1} - u_j) + \frac{h}{4}(1-\beta)(\sigma_j^+ - \sigma_{j+1}^-), \end{aligned} \tag{4.8}$$

where the slopes are defined by

$$\begin{aligned} \sigma_j^{\pm} &= \frac{1}{h}(v_{j+1} \pm \sqrt{a}\,u_{j+1} - v_j \mp \sqrt{a}\,u_j)\,\phi(\theta_j^+) , \\ \theta_j^{\pm} &= \frac{v_j \pm \sqrt{a}\,u_j - v_{j-1} \mp \sqrt{a}\,u_{j-1}}{v_{j+1} \pm \sqrt{a}\,u_{j+1} - v_j \mp \sqrt{a}\,u_j} , \end{aligned} \tag{4.9}$$

and $\beta$ is a positive constant such that

$$\beta = \begin{cases} \sqrt{a}\lambda \equiv \nu \leqq 1 & \text{for forward Euler time discretization.} \\ \beta = 0 & \text{for method of lines.} \end{cases} \tag{4.10}$$

We will only discuss the case $\beta = \nu$ (the forward Euler time discretization), and remark on the other cases at the end of our analysis. Without ambiguity we omit

the index $n$ for all quantities at time $t = t_n$. The relaxed scheme takes the form (see (3.24) and (3.25))

$$
\text{(4.11)}\qquad \begin{cases} v_j = f(u_j)\,, \\ \dfrac{u_j^{n+1} - u_j}{k} = -\dfrac{1}{2h}(v_{j+1} - v_{j-1}) + \dfrac{\sqrt{a}}{2h}(u_{j+1} - 2u_j + u_{j-1}) \\ \quad + \dfrac{1-\beta}{4}(\sigma_{j+1}^- - \sigma_j^- + \sigma_{j-1}^+ - \sigma_j^+)\,. \end{cases}
$$

A scheme is called TVD if its total variation

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

decreases in time,

$$
TV(u^{n+1}) \leqq TV(u^n)\,, \qquad \forall n \geqq 0\,.
$$

Setting

$$
\text{(4.12)}\qquad \begin{aligned} C_j &= \tfrac{1}{2}\lambda\left(\sqrt{a} + \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j}\right) \geqq 0\,, \\ D_j &= \tfrac{1}{2}\lambda\left(\sqrt{a} - \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j}\right) \geqq 0\,. \end{aligned}
$$

Here the positivity of $C_j$ and $D_j$ is due to the subcharacteristic condition (1.4). One can write the relaxed scheme (4.11) in the following convenient form

$$
\text{(4.13)}\quad u_j^{n+1} = u_j - C_{j-1}(u_j - u_{j-1}) + D_j(u_{j+1} - u_j) + \frac{k}{4}(1-\beta)(\sigma_{j+1}^- - \sigma_j^- + \sigma_{j-1}^+ - \sigma_j^+)\,.
$$

It follows from (4.13) that

$$
\begin{aligned} u_{j+1}^{n+1} - u_j^{n+1} = {} & (1 - C_j - D_j)(u_{j+1} - u_j) + C_{j-1}(u_j - u_{j-1}) \\ & + D_{j+1}(u_{j+2} - u_{j+1}) + E_{j+1/2}\,, \end{aligned}
$$

where

$$
\text{(4.14)}\qquad \begin{aligned} E_{j+\frac{1}{2}} = {} & \frac{k}{4}(1-\beta)\left[((\sigma_{j+2}^- - \sigma_{j+1}^-) - (\sigma_{j+1}^- - \sigma_j^-))\right] \\ & - \frac{k}{4}(1-\beta)\left[((\sigma_{j+1}^+ - \sigma_j^+) - (\sigma_j^+ - \sigma_{j-1}^+))\right]\,. \end{aligned}
$$

To continue the analysis, one needs to estimate the last term on the right-hand side of (4.14). Noting that from (4.9) and (4.12),

$$
\text{(4.15)}\qquad \sigma_j^+ = \frac{2}{\lambda h} C_j \phi_j^+(u_{j+1} - u_j)\,, \qquad \sigma_j^- = -\frac{2}{\lambda h} D_j \phi_j^-(u_{j+1} - u_j)\,,
$$

one arrives at

$$
\begin{aligned}
E_{j+\frac{1}{2}} &= \tfrac{1}{2}(1-\beta)\Big[-\phi^-_{j+2}D_{j+2}(u_{j+3}-u_{j+2}) + 2\phi^-_{j+1}D_{j+1}(u_{j+2}-u_{j+1}) \\
&\quad - \phi^-_j D_j(u_{j+1}-u_j) + 2\phi^+_j C_j(u_{j+1}-u_j) \\
&\quad -\phi^+_{j+1}C_{j+1}(u_{j+2}-u_{j+1}) - \phi^+_{j-1}C_{j-1}(u_j-u_{j-1})\Big] \\
&= \tfrac{1}{2}(1-\beta)\Big[-\phi^-_{j+2}D_{j+2}(u_{j+3}-u_{j+2}) + (2\phi^-_{j+1}D_{j+1} - \phi^+_{j+1}C_{j+1}) \\
&\quad \times (u_{j+2}-u_{j+1}) + (2\phi^+_j C_j - \phi^-_j D_j)(u^n_{j+1}-u_j) \\
&\quad -\phi^+_{j-1}C_{j-1}(u_j-u_{j-1})\Big] .
\end{aligned}
$$

This yields

$$
\begin{aligned}
u^{n+1}_{j+1} - u^{n+1}_j &= \Big(1 - [1-(1-\beta)\phi^+_j]C_j \\
&\quad - [1+\tfrac{1}{2}(1-\beta)\phi^-_j]D_j\Big)(u_{j+1}-u_j) \\
&\quad + [1-\tfrac{1}{2}(1-\beta)\phi^+_{j-1}]C_{j-1}(u_j-u_{j-1}) \\
&\quad + \Big([1+(1-\beta)\phi^-_{j+1}]D_{j+1} - \tfrac{1}{2}(1-\beta)\phi^+_{j+1}C_{j+1}\Big) \\
&\quad \times (u_{j+2}-u_{j+1}) - \tfrac{1}{2}(1-\beta)\phi^-_{j+2}D_{j+2}(u_{j+3}-u_{j+2}) .
\end{aligned} \tag{4.16}
$$

Observing that

$$
\begin{aligned}
\theta^+_j C_j(u_{j+1}-u_j) &= C_{j-1}(u_j-u_{j-1}) , \\
\theta^-_j D_j(u_{j+1}-u_j) &= D_{j-1}(u_j-u_{j-1}) ,
\end{aligned} \tag{4.17}
$$

which follows easily from (4.9) and (4.12), we obtain

$$
\begin{aligned}
u^{n+1}_{j+1} - u_j &= \left(1 - \left[1-(1-\beta)\phi^+_j + \tfrac{1}{2}(1-\beta)\frac{\phi^+_{j+1}}{\theta^+_{j+1}}\right]C_j\right. \\
&\quad \left. - [1+\tfrac{1}{2}(1-\beta)\phi^-_j]D_j\right)\cdot(u_{j+1}-u_j) \\
&\quad + [1-\tfrac{1}{2}(1-\beta)\phi^+_{j-1}]C_{j-1}(u_j-u_{j-1}) \\
&\quad + \left[1+(1-\beta)\phi^-_{j+1} - \tfrac{1}{2}(1-\beta)\frac{\phi^-_{j+2}}{\theta^-_{j+2}}\right]D_{j+1}(u_{j+2}-u_{j+1}) .
\end{aligned} \tag{4.18}
$$

We now claim that the standard restriction on the limiters $\phi$, (3.15), and the CFL condition (4.7) imply the three coefficients on the right-hand side of (4.18) are nonnegative, i.e.,

$$
1 - \left[1-(1-\beta)\phi^+_j + \tfrac{1}{2}(1-\beta)\frac{\phi^+_{j+1}}{\theta^+_{j+1}}\right]C_j - [1+\tfrac{1}{2}(1-\beta)\phi^-_j]D_j \geqq 0 , \tag{4.19a}
$$

(4.19b) $$1 - \tfrac{1}{2}(1-\beta)\phi^+_{j-1} \geqq 0 ,$$

(4.19c) $$1 + (1-\beta)\phi^-_{j+1} - \tfrac{1}{2}(1-\beta)\frac{\phi^-_{j+2}}{\theta^-_{j+2}} \geqq 0 .$$

In (4.19b, c) we have used the positivity of $C_j$ and $D_j$. To prove this claim, we first note that by (3.15) and the CFL condition $\beta \leqq 1$, one has

$$1 - \tfrac{1}{2}(1-\beta)\phi^+_{j-1} \geqq 1 - (1-\beta) = \beta \geqq 0 ;$$

and

$$\begin{aligned} &1 + (1-\beta)\phi^-_{j+1} - \tfrac{1}{2}(1-\beta)\frac{\phi^-_{j+2}}{\theta^-_{j+2}} \\ &= 1 + \tfrac{1}{2}(1-\beta)\phi^-_{j+1} - \tfrac{1}{2}(1-\beta)\left(\phi^-_{j+1} - \frac{\phi^-_{j+2}}{\theta^-_{j+2}}\right) \\ &\geqq 1 + \tfrac{1}{2}(1-\beta)\phi^-_{j+1} - (1-\beta) = \beta + \tfrac{1}{2}(1-\beta)\phi^-_{j+1} \geqq 0 , \end{aligned}$$

where one has used the fact that

(4.20) $$\left|\phi(\theta_j) - \frac{\phi(\theta_{j+1})}{\theta_{j+1}}\right| \leqq 2 \quad \text{for all} \quad \theta_j \text{ and } \theta_{j+1} ,$$

which is a direct consequence of (3.15). This proves (4.19b, c). To verify (4.19a), one first rewrites the left-hand side of the inequality as

$$1 + \tfrac{1}{2}(1-\beta)\phi^+_j C_j - (C_j + D_j) - \tfrac{1}{2}(1-\beta)\left[\left(\frac{\phi^+_{j+1}}{\theta^+_{j+1}} - \phi^+_j\right)C_j + \phi_j D_j\right] .$$

By the definition of $C_j$ and $D_j$ in (4.12) one obtains

$$C_j + D_j = \lambda\sqrt{a} \equiv \nu \leqq 1 ,$$

$$\tfrac{1}{2}(1-\beta)\left(\left[\frac{\phi^+_{j+1}}{\theta^+_{j+1}} - \phi^+_j\right]C_j + \phi^-_j D_j\right) \leqq (1-\beta)(C_j + D_j) = (1-\nu)\nu ,$$

where (3.15), (4.7) and (4.10) have been used. Collecting these facts gives

$$\begin{aligned} &1 - \left[1 - (1-\beta)\phi^+_j + \tfrac{1}{2}(1-\beta)\frac{\phi^+_{j+1}}{\theta^+_{j+1}}\right]C_j - [1 + \tfrac{1}{2}(1-\beta)\phi^-_j]D_j \\ &\geqq 1 + \tfrac{1}{2}(1-\beta)\phi^+_j C_j - (C_j + D_j) - (1-\beta)(C_j + D_j) \\ &= 1 + \tfrac{1}{2}(1-\beta)\phi^+_j C_j - (2-\nu)\nu \\ &\geqq 1 + \tfrac{1}{2}(1-\beta)\phi^+_j C_j - \max_{0 \leqq \nu \leqq 1}(2-\nu)\nu = \tfrac{1}{2}(1-\beta)\phi^+_j C_j \geqq 0 , \end{aligned}$$

which verifies our claim (4.19). As an immediate consequence, one obtains

$$
\begin{aligned}
&|u_{j+1}^{n+1} - u_j^{n+1}| \\
&\quad \leqq \left(1 - \left[1 - (1-\beta)\phi_j^+ + \tfrac{1}{2}(1-\beta)\frac{\phi_{j+1}^+}{\theta_{j+1}^+}\right] C_j - [1 + \tfrac{1}{2}(1-\beta)\phi_j^-] D_j\right) \\
&\quad \cdot |u_{j+1} - u_j| + [1 - \tfrac{1}{2}(1-\nu)\phi_{j-1}^+] C_{j-1} |u_j - u_{j-1}| \\
&\quad + \left[1 + (1-\beta)\phi_{j+1}^- - \tfrac{1}{2}(1-\beta)\frac{\phi_{j+2}^-}{\theta_{j+2}^-}\right] D_{j+1} |u_{j+2} - u_{j+1}| \,.
\end{aligned} \tag{4.21}
$$

We want to show that (4.21) implies that the second-order relaxed schemes are TVD. This can be achieved now quite easily by noting from (3.15) and (4.17) that

$$
\begin{aligned}
&\frac{\phi_j^+}{\theta_j^+} C_{j-1} |u_j - u_{j-1}| = \phi_j^+ C_j |u_{j+1} - u_j| \,, \\
&\frac{\phi_j^-}{\theta_j^-} D_{j-1} |u_j - u_{j-1}| = \phi_j^- D_j |u_{j+1} - u_j| \,,
\end{aligned} \tag{4.22}
$$

where the fact that $\phi(\theta) = 0$, for all $\theta \leqq 0$ has been used. Combining (4.21) with (4.22) gives

$$
\begin{aligned}
&|u_{j+1}^{n+1} - u_j^{n+1}| \\
&\quad \leqq \left(1 - [1 - \tfrac{1}{2}(1-\beta)\phi_j^+] C_j - [1 + \tfrac{1}{2}(1-\beta)\phi_j^-] D_j\right) |u_{j+1} - u_j| \\
&\qquad + [1 - \tfrac{1}{2}(1-\beta)\phi_{j-1}^+] C_{j-1} |u_j - u_{j-1}| \\
&\qquad + [1 + \tfrac{1}{2}(1-\beta)\phi_{j+1}^-] D_{j+1} |u_{j+1} - u_j| \\
&\qquad + \tfrac{1}{2}(1-\beta)\phi_j^+ C_j |u_{j+1} - u_j| - \tfrac{1}{2}(1-\beta)\phi_{j+1}^+ C_{j+1} |u_{j+2} - u_{j+1}| \\
&\qquad + \tfrac{1}{2}(1-\beta)\phi_{j+1}^- D_{j+1} |u_{j+2} - u_{j+1}| \\
&\qquad - \tfrac{1}{2}(1-\beta)\phi_{j+2}^- D_{j+2} |u_{j+3} - u_{j+2}| \,.
\end{aligned} \tag{4.23}
$$

Summing (4.23) up over all $j$ yields immediately the desired estimate,

$$
TV(u^{n+1}) \leqq TV(u^n) \,, \qquad \forall n \geqq 0 \,. \tag{4.24}
$$

We summarize the above analysis by the following theorem.

THEOREM 4.2. *For the slope-limiters satisfying* (3.15), *the corresponding second-order relaxed schemes* (4.11) *for conservation law* (1.1) *are TVD provided that the subcharacteristic conditions* (1.4) *and CFL condition* (4.7) *are satisfied.*

*Remarks.*

(i) The consistency and the TVD property of our second-order relaxed scheme implies that it is convergent, and converges to the weak solution of the original conservation law (1.1) [15].

(ii) For the method of the lines ($\beta = 0$), the same analysis as before with only a slight modification shows that the second-order relaxed schemes are TVD under the additional restriction on CFL,

$$\sqrt{a}\lambda \leqq \tfrac{1}{2} . \tag{4.26}$$

(iii) Combining the previous results with those of Shu and Osher [24] shows that our second-order Runge-Kutta relaxed schemes, (3.21) and (3.14), are also TVD.

## 5. Numerical Results

In this section we present numerical results that demonstrate the performance of the relaxation schemes for the compressible Euler equations in both 1-D and 2-D cases.

### Boundary Conditions

In the boundary value problems we impose the boundary conditions for $v$ that are consistent to the local equilibrium $v = f(u)$. Let $\Omega$ be the integration domain and $\partial\Omega$ its boundary. If $u|_{\partial\Omega}$ is given then we set $v|_{\partial\Omega} = f(u|_{\partial\Omega})$. If $u$ satisfies the Neumann boundary condition $\frac{\partial}{\partial x} u = 0$ then, since $\frac{\partial}{\partial x} v = f'(u) \frac{\partial}{\partial x} u$ we obtain $\frac{\partial}{\partial x} v = 0$ on $\partial\Omega$. Other boundary conditions can be similarly imposed. For a system of conservation laws such a determination of the boundary condition for $\mathbf{v}$ is done componentwise.

In our numerical examples we use the boundary conditions for $\mathbf{v}$ determined as discussed above. No new artificial boundary layer behavior is numerically observed. Two possibilities may exist. First, the choice of the boundary condition that is consistent to the local equilibrium as defined above does not introduce a new boundary layer at all. Second, if the artificial boundary layer does exist, it should be of $O(\varepsilon)$, which may have been ignored by the underresolved numerical methods with $k, h >> \varepsilon$.

We would like to point out that these assumptions are not theoretically supported and further theoretic investigation will be interesting. Such an investigation is not only important for the development of the relaxing schemes. As a matter of fact we can always use the relaxed schemes which do not depend on $\varepsilon$ at all. Its importance also lies in that it will shed light on the understanding of the boundary layer behavior for more general hyperbolic systems with stiff relaxations in the sense of Liu [16], among which our relaxation system is the simplest.

**1-D Tests**

We consider the approximate solution of the 1-D Euler equations of gas dynamics,

$$\begin{aligned} &\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x} m = 0 , \\ &\frac{\partial}{\partial t} m + \frac{\partial}{\partial x}(\rho u^2 + p) = 0 , \\ &\frac{\partial}{\partial t} E + \frac{\partial}{\partial x}(u(E + p)) = 0 . \end{aligned} \tag{5.1}$$

Here $\rho$, $u$, $m = \rho u$, $p$ and $E$ are respectively the density, velocity, momentum, pressure and total energy. For a polytropic gas, the equation of state is given by

$$p = (\gamma - 1)(E - \tfrac{1}{2}\rho u^2) . \tag{5.2}$$

We take $\gamma = 1.4$. Let

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} , \qquad F(\mathbf{u}) = \begin{pmatrix} m \\ \rho u^2 + p \\ u(E + p) \end{pmatrix} , \tag{5.3}$$

then (5.1) can be written as

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x} F(\mathbf{u}) = 0 . \tag{5.4}$$

The relaxation system for (5.4) is

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{v} = 0 , \\ &\frac{\partial}{\partial t}\mathbf{v} + A\frac{\partial}{\partial x}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{v} - F(\mathbf{u})) . \end{aligned} \tag{5.5}$$

Here $A = \operatorname{diag}\{a_1, a_2, a_3\}$ with $a_i > 0$ for $1 \leqq i \leqq 3$. Since the 1-D Euler equations have three eigenvalues $u, u \pm c$ where $c = \sqrt{\gamma p/\rho}$ is the sound speed, in our test we either take $a_1 = \sup|u - c|, a_2 = \sup|u|$ and $a_3 = \sup|u + c|$, which satisfies the characteristic interlace relation defined in [3], or simply $a_1 = a_2 = a_3 = \max(\sup|u - c|, \sup|u|, \sup|u + c|)$ for all $t$ and $x$. We solve (5.5) by the relaxing schemes described in Section 3. The CFL number is defined as

$$\text{CFL number} = \max_i a_i \frac{k}{h} . \tag{5.6}$$

We carried out the following standard 1-D tests.

(a) *Sod Shock Tube* [25]. We take the initial data

$$\mathbf{u}_L = \begin{pmatrix} 1 \\ 0 \\ 2.5 \end{pmatrix}, \quad \text{if} \quad 0 \leqq x < 0.5; \quad \mathbf{u}_R = \begin{pmatrix} 0.125 \\ 0 \\ 0.25 \end{pmatrix}, \quad \text{if} \quad 0.5 \leqq x \leqq 1 .$$

We first choose $a_1 = 1, a_2 = 1.68, a_3 = 5.045$ so the relaxation system (5.5) has distinct eigenvalues. We take $h = 0.005$ (200 zones) and CFL = 0.75. The output time is $t = 0.1644$. In Figure 1(a), we display the result of the first-order (upwind) relaxing scheme with $\varepsilon = 10^{-8}$. The behavior is similar to that of the Lax-Friedrichs scheme, as predicted by the first-order relaxed scheme. In Figure 1(b) we exhibit the result by the second-order relaxing scheme with the van Leer slope limiter and $\varepsilon = 10^{-8}$. In Figure 1(c) we repeat the test in Figure 1(b) but with $\varepsilon = 10^{-4}$. Here $\varepsilon$ plays the role of viscosity coefficient so the undershots in Figure 1(b) are flattened by a relatively larger $\varepsilon$. In Figure 1(d) we take $a_1 = a_2 = a_3 = 5.045$ so the relaxation system (5.5) has only two distinct eigenvalues $\pm a_1$, and $\varepsilon = 10^{-8}$. By comparing it with Figure 1(b) one sees that this set of parameters seems to introduce more numerical viscosity than the set used in Figure 1(a). As a consequence the contact discontinuity is smeared out by two more grid points.

(b) *Lax Shock Tube* [14]. We take the initial data

$$\mathbf{u}_L = \begin{pmatrix} 0.445 \\ 0.311 \\ 8.928 \end{pmatrix}, \quad \text{if} \quad 0 \leqq x < 0.5;$$

$$\mathbf{u}_R = \begin{pmatrix} 0.5 \\ 0 \\ 0.4275 \end{pmatrix}, \quad \text{if} \quad 0.5 \leqq x \leqq 1 .$$

We choose $a_1 = 2.4025, a_2 = 11, a_3 = 22.2056$. In both tests in Figure 2 we use $h = 0.005$ (200 zones), CFL = 0.5, $\varepsilon = 10^{-8}$, and output the results at $t = 0.16$. In Figure 2(a) the result was obtained by the second-order relaxing scheme with the van Leer slope limiter. In Figure 2(b) it is rerun with the minmod slope limiter, which is more dissipative.

(c) *Interacting Blast Wave* [29]. The initial data are

$$\rho = 1, \quad u = 0 \quad \text{for} \quad 0 \leqq x \leqq 1 ;$$

$$p = 1000 \quad \text{for} \quad 0 \leqq x \leqq 0.1; \ 0.01 \text{ for } 0.1 < x \leqq 0.9; \ 100 \text{ for } 0.9 < x < 1 .$$

We take $a_1 = 32.56, a_2 = 169, a_3 = 411.8397$ with 400 zones and CFL = 0.25. The second-order relaxing scheme with van Leer slope limiter is applied with $\varepsilon = 10^{-8}$. We display the results at $t = 0.01$ in Figure 3(a), $t = 0.03$ in Figure 3(b), and $t = 0.038$ in Figure 3(c).

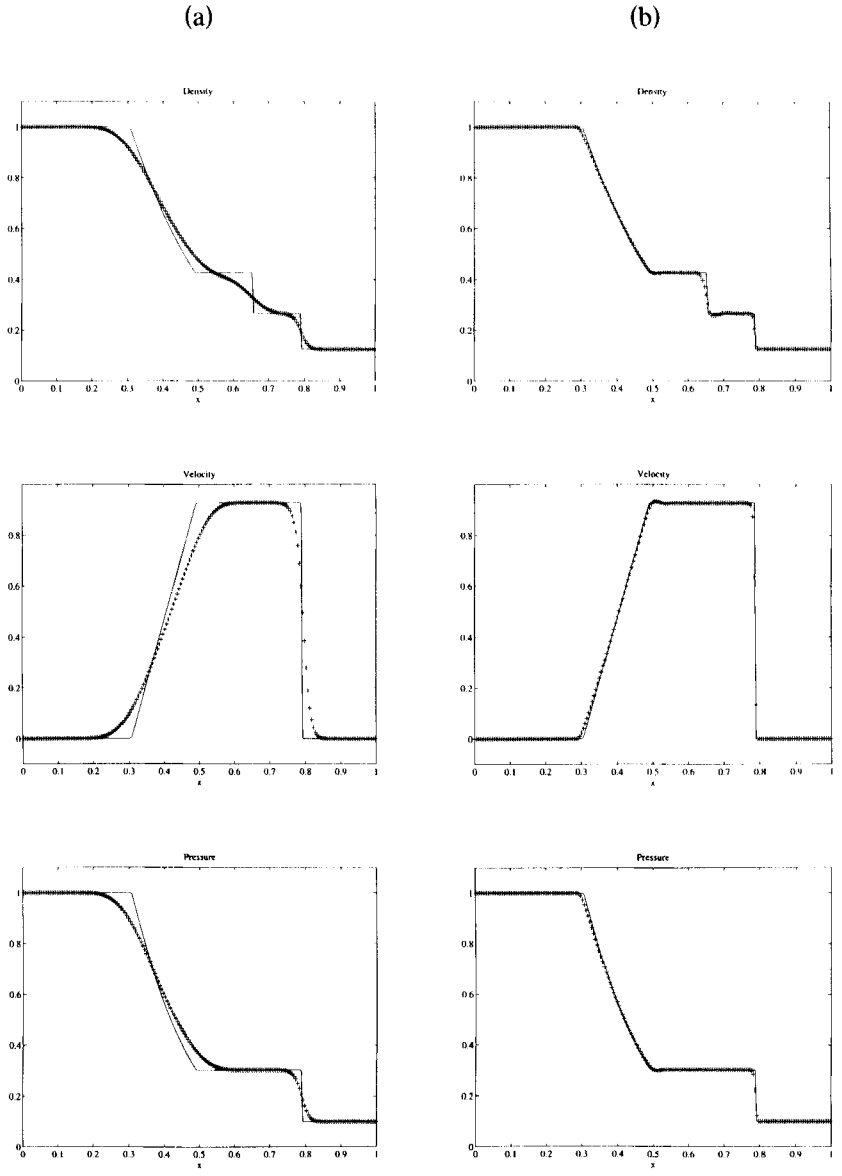



Figure 1. Sod shock tube at $t = 0.1644$ on a grid of 200 zones with CFL = 0.75. (a) The first-order relaxing scheme on the relaxation system with distinct eigenvalues, $\varepsilon = 10^{-8}$. (b) The second-order relaxing scheme with van Leer's slope limiter on the relaxation system with distinct eigenvalues, $\varepsilon = 10^{-8}$.

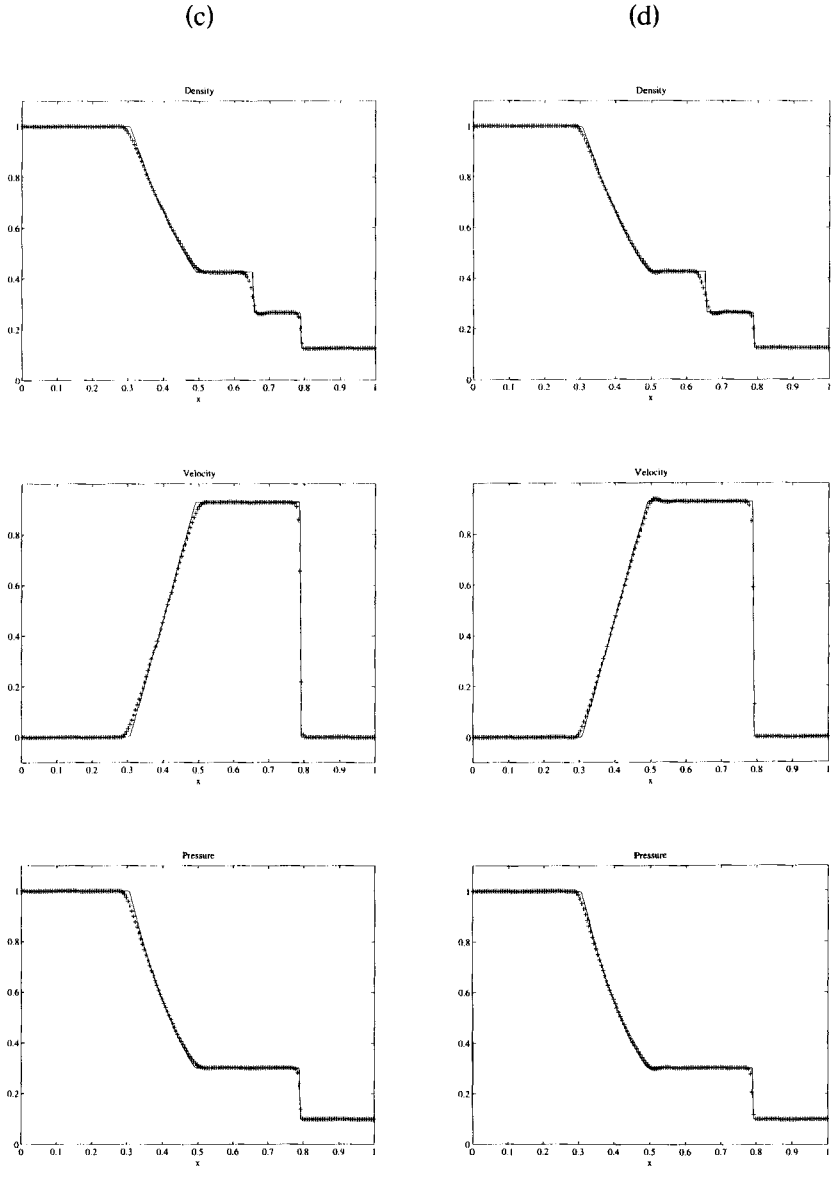



Figure 1 (cont'd). (c) Same test as in Figure 1(b) except for $\varepsilon = 10^{-4}$. (d) Same test as in Figure 1(b) except on a relaxing system with $a_1 = a_2 = a_3$.

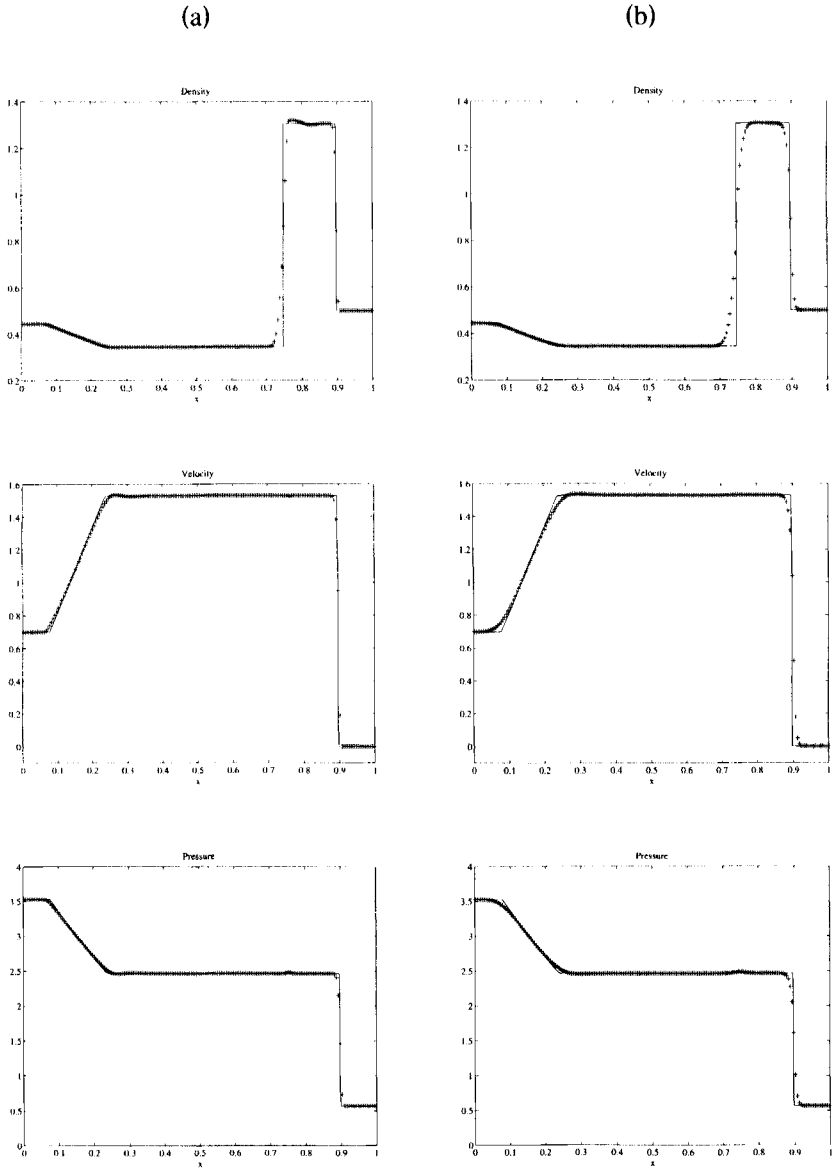



Figure 2. Lax shock tube at $t = 0.16$ by the second-order relaxing schemes. $\varepsilon = 10^{-8}$, 200 zones, CFL = 0.75. (a) van Leer's slope limiter. (b) the minmod slope limiter.

(a)

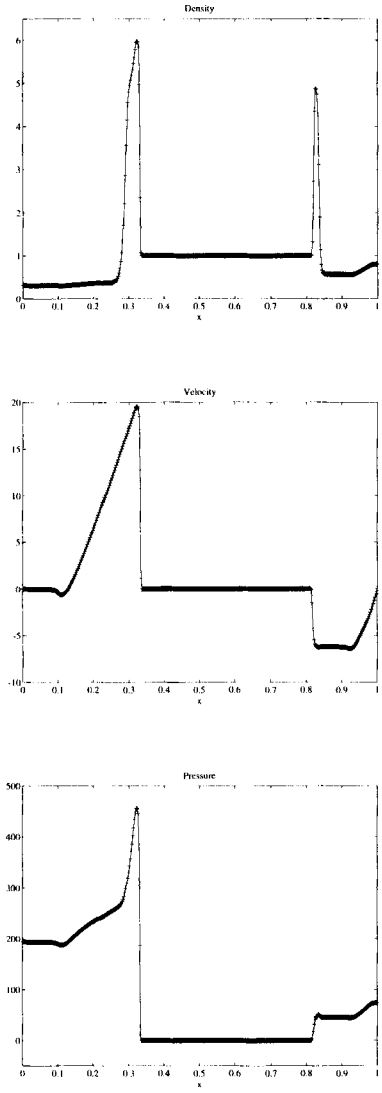



Figure 3. Interacting blast wave by the second-order relaxing scheme using van Leer's slope limiter with $\varepsilon = 10^{-8}$, 400 zones and CFL = 0.25. Both solid line and the '+' denote the same numerical solution. (a) $t = 0.01$.

(b)

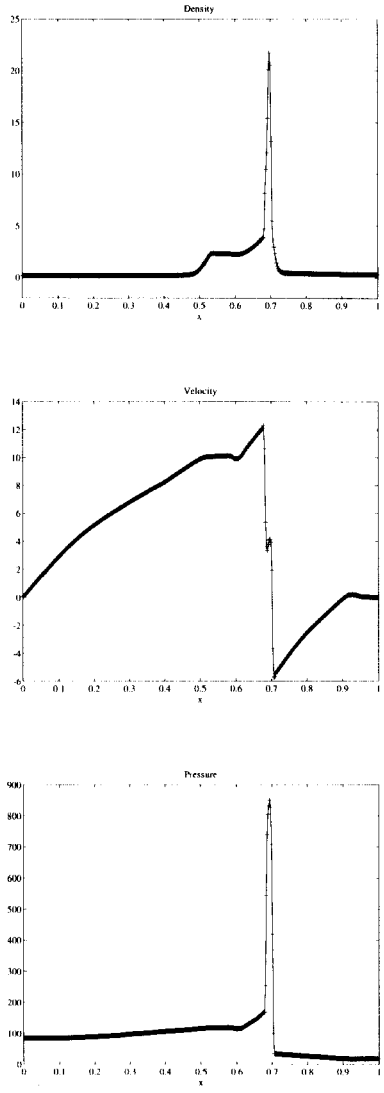



Figure 3 (cont'd). (b) $t = 0.03$.

(c)

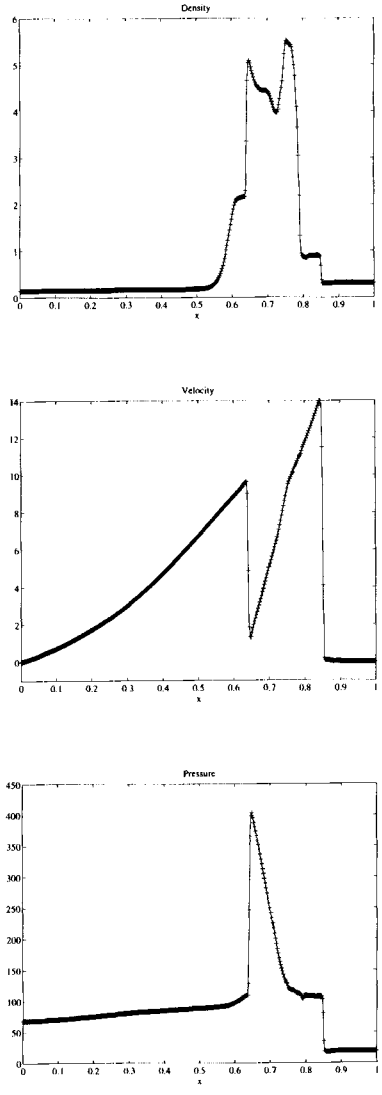



Figure 3 (cont'd). (c) $t = 0.038$.

**2-D Tests**

We want to obtain approximate solutions to the 2-D Euler equations for gas dynamics:

$$
\begin{aligned}
&\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}m + \frac{\partial}{\partial x}n = 0\,,\\
&\frac{\partial}{\partial t}m + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0\,,\\
&\frac{\partial}{\partial t}n + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0\,,\\
&\frac{\partial}{\partial t}E + \frac{\partial}{\partial x}((E+p)u) + \frac{\partial}{\partial y}((E+p)v) = 0\,.
\end{aligned}
\tag{5.7}
$$

Here $\rho$, $(u, v)^t$, $(m, n)^t = (\rho u, \rho v)^t$, $p$ and $E$ are respectively density, velocity, momentum, pressure and energy. The equation of state for a polytropic gas is given by

$$
p = (\gamma - 1)(E - \tfrac{1}{2}\rho(u^2 + v^2))\,. \tag{5.8}
$$

In air we take $\gamma = 1.4$. Let

$$
\mathbf{u} = \begin{pmatrix} \rho \\ m \\ n \\ E \end{pmatrix}\,,\quad F(\mathbf{u}) = \begin{pmatrix} m \\ \rho u^2 + p \\ \rho uv \\ (E+p)u \end{pmatrix}\,,\quad G(\mathbf{u}) = \begin{pmatrix} n \\ \rho uv \\ \rho v^2 + p \\ (E+p)v \end{pmatrix}\,. \tag{5.9}
$$

Then one can rewrite (5.7) as

$$
\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) + \frac{\partial}{\partial y}G(\mathbf{u}) = 0\,. \tag{5.10}
$$

The relaxation system for (5.10) is

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{v} + \frac{\partial}{\partial y}\mathbf{w} = 0\,,\\
&\frac{\partial}{\partial t}\mathbf{v} + A\frac{\partial}{\partial x}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{v} - F(\mathbf{u}))\,,\\
&\frac{\partial}{\partial t}\mathbf{w} + B\frac{\partial}{\partial y}\mathbf{u} = -\frac{1}{\varepsilon}(\mathbf{w} - G(\mathbf{u}))\,.
\end{aligned}
\tag{5.11}
$$

Here $A = \operatorname{diag}\{a_1, a_2, a_3, a_4\}$ and $B = \operatorname{diag}\{b_1, b_2, b_3, b_4\}$ with $a_i > 0, b_i > 0$ for $1 \leqq i \leqq 4$. For this problem we always use the second-order relaxing scheme with van Leer's slope limiter (3.14) and take $\varepsilon = 10^{-8}$. The CFL number is defined as

$$
\text{CFL number} = \max\left(\max_i \sqrt{a_i}\,\frac{k}{\Delta x}, \max_i \sqrt{b_i}\,\frac{k}{\Delta y}\right)\,. \tag{5.12}
$$

Here $\Delta x$ and $\Delta y$ are the mesh sizes in the $x$ and $y$ directions respectively.

(a) *Regular Shock Reflection.* The computational domain is a rectangle of length 4 and height 1 divided into $60 \times 20$ rectangular grids with $\Delta x = \frac{1}{15}$, $\Delta y = \frac{1}{20}$. Dirichlet conditions are imposed on the left and upper boundaries as

$$(\rho, u, v, p)\big|_{(0,y,t)} = (1,\ 2.9,\ 0,\ 1/1.4)\ ,$$

$$(\rho, u, v, p)\big|_{(x,1,t)} = (1.69997,\ 2.61934,\ -0.50633,\ 1.52819)\ .$$

The bottom boundary is a reflecting wall and the supersonic outflow condition is applied along the right boundary. Initially, the solution in the entire domain is set to be that at the left boundary [4]. By taking $a_1 = a_2 = a_3 = a_4 = 3.74118, b_1 = b_2 = b_3 = b_4 = 1.62810$, we iterate for 1000 time steps using $k = 0.005$, at which time the solution reaches a steady state.

In Figure 4(a) we show a 30 equi-distributed contour plot of the pressure. In Figure 4(b) we plot the profile of the solution at $y = 0.525$.

(b) *Double Mach Reflection.* A planar shock is incident on an oblique surface, with the surface at a 30° angle to the direction of propagation of the shock. Our test problem involves a Mach 10 shock, ahead of it is the undisturbed air with a density of 1.4 and a pressure of 1. We use the boundary conditions described in [29]. In this problem we take $a_1 = b_1 = 98, a_2 = b_2 = 162, a_3 = b_3 = 242, a_4 = b_4 = 338$. The density at $t = 0.2$ is plotted in Figure 5(a) with $\Delta x = \Delta y = \frac{1}{60}$, $k = 0.0004$ and in Figure 5(b) with $\Delta x = \Delta y = \frac{1}{120}$, $k = 0.0002$. In each plot 30 equi-distributed contours are shown. There are some oscillations behind the shocks, but this is standard for any second-order method in 2-D, since no method higher than first order can be TVD [10].

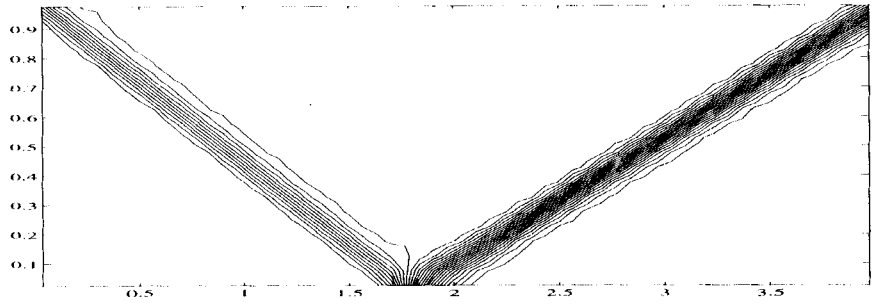
*Remarks.*

(i) The choices of $A$ and $B$ in all these numerical examples are based on a rough estimate of the characteristics of the original Euler equations such that the characteristics of the relaxation system interlace with those of the Euler equations, or we simply take the largest eigenvalue of the Euler equations. Other choices can be made as long as the numerical stability is maintained, and usually larger $A$ and $B$ give more numerical viscosity.

(ii) All the numerical results presented here were obtained using uniform Cartesian grids with the standard discretizations discussed in this paper. No other techniques, such as artificial compression, subcell resolution, adaptive mesh and flattening were applied to sharpen the results. Although these techniques can be combined with the relaxing schemes and are under consideration, the purpose of this paper is to provide the framework for a general approach.

(iii) All the results with $\varepsilon = 10^{-8}$ can be almost reproduced with about equal quality by using the relaxed schemes and numerical convergence of the relaxing schemes to the relaxed schemes as $\varepsilon \to 0$ is observed for the Euler equations. Thus for strictly hyperbolic systems one can just use the relaxed schemes, which are much easier to implement with more efficiency and much less memory. We present the results for the relaxing schemes just to demonstrate their

(a)




(b)

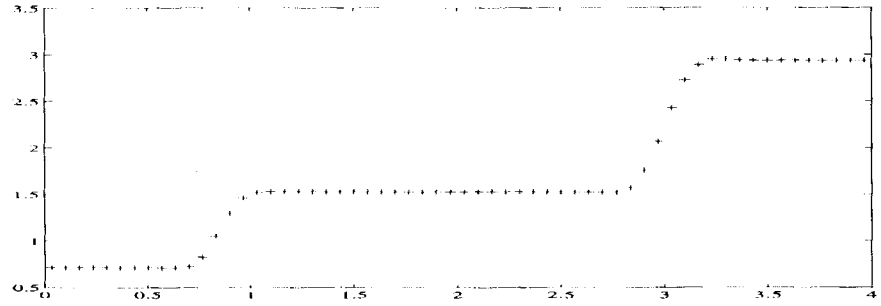
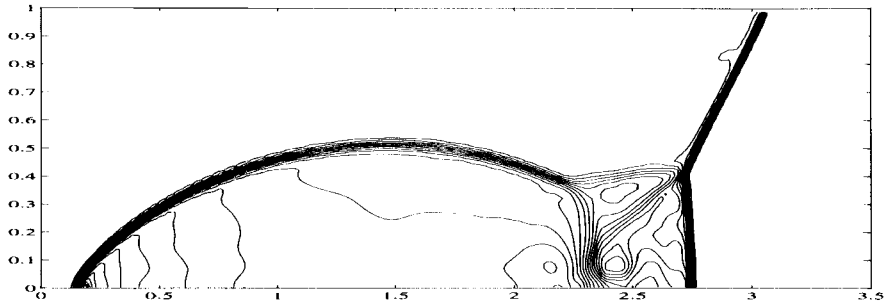



Figure 4. Steady state solution of the regular shock reflection by the second-order relaxing scheme with van Leer's slope limiter using $\Delta x = \frac{1}{15}$, $\Delta y = \frac{1}{20}$. $\varepsilon = 10^{-8}$. (a) Pressure contour. Here 30 equi-distributed contours are shown. (b) Pressure profile along the line $y = 0.525$.

(a)




(b)

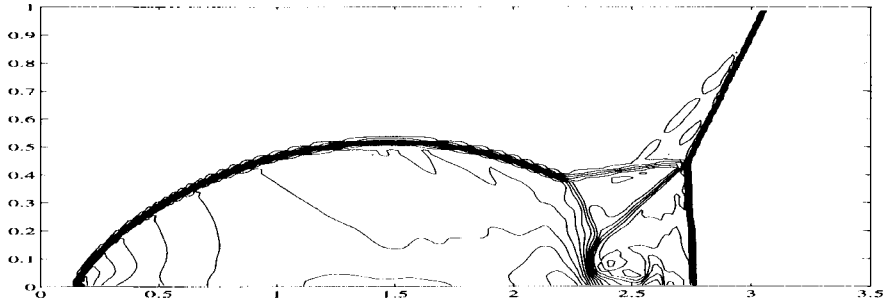



Figure 5. The double mach reflection at $t = 0.2$ by the second-order relaxing scheme with van Leer's slope limiter. $\varepsilon = 10^{-8}$. The mesh is rectangular with: (a) $\Delta x = \Delta y = \frac{1}{60}$; (b) $\Delta x = \Delta y = \frac{1}{120}$. In each plot 30 equi-distributed contours are shown.

correct zero relaxation limit, and the relaxing schemes are also of great interest for their potential applications to hyperbolic systems with stiff source terms. For degenerated hyperbolic systems or mixed hyperbolic-elliptic systems it is not clear whether the zero relaxation limit is valid.

## 6. Conclusions

A class of numerical schemes based on the local relaxation approximation for systems of conservation laws in several variables have been introduced and analyzed. The main feature of this class of schemes is its simplicity and generality. It uses neither Riemann solver spatially nor nonlinear systems of algebraic equations solvers temporally, yet can achieve high order accuracy and picks up the right weak solutions. The second-order relaxed schemes for the 1-D scalar equation are shown to be TVD. Both numerical experiments and theoretical analysis indicate that the relaxation schemes proposed in this paper have a great deal of advantages.

Here we concentrate on developing the basic method. Other ideas such as local adaptive mesh refinement and shock tracking techniques could be incorporated into the basic framework here to give more delicate results. These are currently under investigation. Further theoretical studies, such as cell entropy conditions for higher order relaxing and relaxed schemes, are also under consideration.

We would like to emphasize that this approach is not restricted to hyperbolic problems. Indeed, the basic formulation of the relaxation system in this paper imposes no condition on the original conservation laws. We are pursuing the application of this approach to degenerate hyperbolic equations and equations that may change type, in which the understanding of basic wave patterns is extremely hard to achieve analytically and numerically. The prospects in all these directions are very encouraging, and we hope to present more challenging results in due course.

## Appendix: Dissipation Conditions

In this appendix we shall provide the proofs of Propositions 2.1 and 2.2 in Section 2 which ensure the dissipation structures of the reduced systems (2.4) and (2.18) respectively. We start with the 1-D systems.

Proof of Proposition 2.1: Using the entropy variable $\mathbf{w} = q'(\mathbf{u})$, one can rewrite system (2.4) as

$$\begin{aligned} &\frac{\partial}{\partial t}\mathbf{w} + \eta''(\mathbf{u})F'(\mathbf{u})\eta''(\mathbf{u})^{-1}\frac{\partial}{\partial x}\mathbf{w} \\ &\quad = \varepsilon\,\eta''(\mathbf{w})\frac{\partial}{\partial x}\left(\eta''(\mathbf{u})^{-1}[aI - \eta''(\mathbf{u})F'(\mathbf{u})^2\eta''(\mathbf{u})^{-1}]\frac{\partial}{\partial x}\mathbf{w}\right), \end{aligned} \tag{A.1}$$

where assumption (2.9) has been used. Therefore, it suffices to show that

$$aI - (\eta''(\mathbf{u})F'(\mathbf{u}))(F'(\mathbf{u})\eta''(\mathbf{u})^{-1}) \geqq 0\,. \tag{A.2}$$

Since both $\eta''F'$ and $F'(\eta'')^{-1}$ are symmetric matrices (hereafter we will omit the dependence on $\mathbf{u}$), there exist orthonormal bases in $\mathbb{R}^n$, $\{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\}$ and $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ such that

$$F'(\eta'')^{-1}\mathbf{r}_i = \lambda_i \mathbf{r}_i\,, \quad \eta''F'\mathbf{l}_i = \mu_i \mathbf{l}_i\,, \quad i = 1, 2, \ldots, n\,, \tag{A.3}$$

where $\lambda_i$ and $\mu_i$ are the generalized eigenvalues of $F'$ defined in (2.7), and $\mathbf{l}_i$ and $\mathbf{r}_i$ are the column eigen-vectors in $\mathbb{R}^n$. Now for any vector $\mathbf{y} \in \mathbb{R}^n$, one has

$$\begin{aligned} &\mathbf{y} = \sum_{i=1}^{n} \alpha_i \mathbf{r}_i\,, \qquad \mathbf{y}^t = \sum_{j=1}^{n} \beta_j \mathbf{l}_j^t\,, \quad \text{and} \\ &\|\mathbf{y}\|^2 = \sum_{i=1}^{n} \alpha_i^2 = \sum_{j=1}^{n} \beta_j^2 = \|\mathbf{y}^t\|^2 \end{aligned} \tag{A.4}$$

where $\mathbf{y}^t$ is the transpose of $\mathbf{y}$. It follows from (A.3), (A.4) and (2.7) that

$$\begin{aligned} &\mathbf{y}^t(aI - (\eta''F')(F'(\eta'')^{-1})\mathbf{y} \\ &= \sum_{i,j=1}^{n} \beta_i \alpha_j \mathbf{l}_i^t (aI - (\eta''F')(F'(\eta'')^{-1}))\mathbf{r}_j \\ &= \sum_{i,j=1}^{n} \beta_i \alpha_j (a - \mu_i \lambda_j)\, \mathbf{l}_i^t \mathbf{r}_j = a\|\mathbf{y}\|^2 - \sum_{i,j=1}^{n} \mu_i \lambda_j \beta_i \alpha_j\, \mathbf{l}_i^t \mathbf{r}_j \\ &= a\|\mathbf{y}\|^2 - \left(\sum_{i=1}^{n} \beta_i \mu_i \mathbf{l}_i^t\right) \cdot \left(\sum_{j=1}^{n} \alpha_j \lambda_j \mathbf{r}_j\right) \\ &\geqq a\|\mathbf{y}\|^2 - \|\sum_{i=1}^{n} \beta_i \mu_i \mathbf{l}_i^t\| \, \|\sum_{j=1}^{n} \alpha_j \lambda_j \mathbf{r}_j\| \\ &= a\|\mathbf{y}\|^2 - \left(\sum_{i=1}^{n} \mu_i^2 \beta_i^2\right)^{1/2} \left(\sum_{i=1}^{n} \lambda_j^2 \alpha_j^2\right)^{1/2} \geqq (a - \lambda\mu)\|\mathbf{y}\|^2\,. \end{aligned}$$

Thus one can conclude (A.2). This completes the proof of Proposition 2.1.

Similarly one can prove Proposition 2.2. To this end, one first needs an elementary lemma.

LEMMA A.1. *Let $B \equiv diag\{b_1, \ldots, b_n\}$ be a positive diagonal matrix, and $\mathbf{g} = (g_1, \ldots, g_n)^t$ be any column vector in $\mathbb{R}^n$. Then the matrix*

$$B - \mathbf{g} \cdot \mathbf{g}^t \tag{A.5}$$

*is nonnegative definite if and only if*

$$\frac{g_1^2}{b_1} + \frac{g_2^2}{b_2} + \cdots + \frac{g_n^2}{b_n} \leqq 1\,. \tag{A.6}$$

Proof: Observe that $B - \mathbf{g}\cdot\mathbf{g}^t \geqq 0$ if and only if

$$\sqrt{B^{-1}}(B - \mathbf{g}\cdot\mathbf{g}^t)\sqrt{B^{-1}} \equiv I - (\sqrt{B^{-1}}\mathbf{g})\cdot(\sqrt{B^{-1}}\mathbf{g})^t \geqq 0\,. \tag{A.7}$$

First, suppose (A.7) holds. In particular,

$$\left(\sqrt{B^{-1}}\,\mathbf{g}\right)^t \left(I - (\sqrt{B^{-1}}\mathbf{g})\cdot(\sqrt{B^{-1}}\mathbf{g})^{-1}\right)\left(\sqrt{B^{-1}}\,\mathbf{g}\right) \geqq 0\,,$$

i.e.

$$\left(\sum_{i=1}^n \frac{g_i^2}{b_i}\right)\left(1 - \sum_{i=1}^n \frac{g_i^2}{b_i}\right) \geqq 0\,.$$

This implies (A.6). Conversely, one needs to show the matrix $M = I + (\sqrt{B^{-1}}\mathbf{g})\cdot(\sqrt{B^{-1}}\mathbf{g})^t$ is nonnegative definite. If $\mathbf{e}_1 = \sqrt{B^{-1}}\mathbf{g}$ is a zero vector, there is nothing to prove. Otherwise, one may find vectors $\mathbf{e}_2,\ldots,\mathbf{e}_n$ such that $\{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n\}$ forms an orthogonal basis for $\mathbb{R}^n$. Then direct computations show that

$$\begin{cases} \mathbf{e}_1^t M\mathbf{e}_1 = |\mathbf{e}_1|^2 - |\mathbf{e}_1|^4 = \left(\sum_{i=1}^n \frac{g_i^2}{b_i}\right)\left(1 - \sum_{i=1}^n \frac{g_i^2}{b_i}\right) \geqq 0\,, \\ \mathbf{e}_i^t M\mathbf{e}_i = |\mathbf{e}_i|^2 - (\mathbf{e}_i^t\cdot\mathbf{e}_1)(\mathbf{e}_1^t\mathbf{e}_i) = |\mathbf{e}_i|^2\,, \qquad i \neq 1\,, \\ \mathbf{e}_i^t M\mathbf{e}_j = 0\,, \qquad i \neq j \end{cases} \tag{A.8}$$

where (A.6) has been used. Now for any given vector $\mathbf{y} = \sum_{i=1}^n y_i\mathbf{e}_i \in \mathbb{R}^n$, one has from (A.8) that

$$\mathbf{y}^t M\mathbf{y} = \sum_{i=1}^n y_i^2\mathbf{e}_i^t M\mathbf{e}_i \geqq 0\,.$$

Consequently, $B - \mathbf{g}\cdot\mathbf{g}^t$ is nonnegative definite.

Proof of Proposition 2.2: To show (2.23), one can choose $2m$ sets of orthonormal bases for $\mathbb{R}^n$, $\{\mathbf{l}_1^{(i)},\ldots,\mathbf{l}_n^{(i)}\}$ and $\{\mathbf{r}_1^{(i)},\ldots,\mathbf{r}_n^{(i)}\}$ $(i = 1,\ldots,m)$, such that

$$F_i'(\eta'')^{-1}\mathbf{r}_k^{(i)} = \lambda_k^{(i)}\mathbf{r}_k^{(i)}\,, \quad \eta'' F_i'\mathbf{l}_k^{(i)} = \mu_k^{(i)}\mathbf{l}_k^{(i)}\,, \quad k = 1,\ldots,n,\ i = 1,\ldots,m\,. \tag{A.9}$$

It follows that

$$(\mathbf{l}_k^{(i)})^t(\eta'' F_i')(F_j'(\eta'')^{-1})\mathbf{r}_l^{(j)} = \mu_k^{(i)}\lambda_l^{(j)}(\mathbf{l}_k^{(i)})^t(\mathbf{r}_l^{(i)})\,. \tag{A.10}$$

Now for any $Y \in \mathbb{R}^{mn}$, one can write $Y = (\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(m)})$ such that $\mathbf{y}^{(k)} \in \mathbb{R}^n$ and $|Y|^2 = \sum_{i=1}^n |\mathbf{y}^{(i)}|^2$. Since $\mathbf{y}^{(i)} \in \mathbb{R}^n$, one has the following decompositions

$$\begin{aligned} &\mathbf{y}^{(i)} = \sum_{k=1}^n \alpha_k^{(i)}\mathbf{r}_k^{(i)}\,, \qquad (\mathbf{y}^{(i)})^t = \sum_{\mathbf{l}=1}^n \beta_l^{(i)}(\mathbf{l}_l^{(j)})^t\,, \quad \text{and} \\ &|\mathbf{y}^{(i)}|^2 = \sum_{k=1}^n (\alpha_k^{(i)})^2 = \sum_{\mathbf{l}=1}^n (\beta_l^{(i)})^2\,. \end{aligned} \tag{A.11}$$

Using (A.10) and (A.11), one can compute that

$$
\begin{aligned}
&Y^t(a_i I\delta_{ij} - (\eta'' F_i')(F_j'(\eta'')^{-1}))Y \\
&= \sum_{i,j=1}^{n} (\mathbf{y}^{(i)})^t (a_i I\delta_{ij} - (\eta'' F_i')(F_j'(\eta'')^{-1}))\mathbf{y}^{(j)} \\
&= \sum_{i,j=1}^{m} (\delta_{ij} a_i (\mathbf{y}^{(i)})^t \mathbf{y}^{(j)} - \left( (\mathbf{y}^{(i)})^t \eta'' F_i' \right) \cdot \left( F_j (\eta'')^{-1} \mathbf{y}^{(j)} \right) \\
&= \sum_{i=1}^{m} a_i |\mathbf{y}^{(i)}|^2 - \sum_{i,j=1}^{m} \left( \sum_{l=1}^{n} \mu_l^{(i)} \beta_l^{(i)} (\mathbf{l}_l^{(i)})^t \right) \left( \sum_{k=1}^{n} \lambda_k^{(j)} \alpha_k^{(j)} \mathbf{r}_k^{(j)} \right) \\
&\geqq \sum_{i=1}^{m} a_i |(\mathbf{y}^{(i)})^t| \, |\mathbf{y}^{(i)}| - \sum_{i,j=1}^{n} | \sum_{l=1}^{n} \mu_l^{(i)} \beta_l^{(i)} (\mathbf{l}_l^{(i)})^t | \, | \sum_{k=1}^{n} \lambda_k^{(j)} \alpha_k^{(j)} \mathbf{r}_k^{(j)} | \\
&\geqq \sum_{i=1}^{m} a_i |(\mathbf{y}^{(i)})^t| \, |\mathbf{y}^{(i)}| - \sum_{i,j=1}^{n} \lambda^{(i)} \lambda^{(j)} |(\mathbf{y}^{(i)})^t| \, |\mathbf{y}^{(j)}| \\
&= \left( |\mathbf{y}^{(1)}|, \ldots, |\mathbf{y}^{(m)}| \right) \\
&\quad \times \cdot \left( \operatorname{diag}\{a_1, \ldots, a_m\} - \begin{pmatrix} \lambda^{(1)} \\ \vdots \\ \lambda^{(m)} \end{pmatrix} (\lambda^{(1)}, \ldots, \lambda^{(m)}) \right) \begin{pmatrix} |\mathbf{y}^{(1)}| \\ \vdots \\ |\mathbf{y}^{(m)}| \end{pmatrix} .
\end{aligned}
\tag{A.12}
$$

Then by the lemma, the last term on the right-hand side above is nonnegative if and only if (2.25) holds. This completes the proof of Proposition 2.2.

**Acknowledgments.** The authors are grateful for the encouragement and support of Peter Lax, David Levermore, and George Papanicolaou. We also thank Cathleen Morawetz and Philip Roe for their careful reading of the manuscript and their suggestive comments, and Jian-Guo Liu for many helpful discussions.

The research of the first author was supported by AFOSR Grant F49620-92-J0098 while he was a visiting member at Courant Institute. The research of the second author was supported in part by a Sloan Foundation Fellowship, National Science Foundation Grant DMS-90-02286, and Department of Energy Grant DE-FG02-88ER-25053.

## Bibliography

[1] Caflisch, R., Jin, S., and Russo, G., *Uniformly accurate schemes for hyperbolic systems with relaxation,* SIAM J. Numer. Anal., submitted.

[2] Chapman, S., and Cowling, T. G., *The Mathematical Theory of Nonuniform Gases,* 3rd Edition, Cambridge Univ. Press, 1970.

[3] Chen, G. Q., Levermore, C. D., and Liu, T. P., *Hyperbolic conservation laws with stiff relaxation terms and entropy,* Comm. Pure Appl. Math., 47, 1994, pp. 787–830.

[4] Colella, P., *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys. 87, 1990, pp. 171–200.
[5] Crandall, M. G., and Majda, A., *The method of fractional steps for conservation laws*, Math. Comput. 34, 1980, pp. 285–314.
[6] Deshpande, S. M., *A second order accurate, kinetic-theory based, method for inviscid compressible flows*, NASA Langley Tech. paper No. 2613, 1986.
[7] Doolen, G. D., (ed.), *Lattice Gas Methods for PDE's: Theory, Applications, and Hardware*, Physica D 47(1–2), 1991.
[8] Friedrichs, K. O., and Lax, P. D., *Systems of conservation laws with a convex extensions*, Proc. Nat. Acad. Sci. USA 68, 1971, pp. 1686–1688.
[9] Giga, Y., and Miyakawa, T., *A kinetic construction of global solutions of first order quasilinear equations*, Duke Math. J. 50, 1983, pp. 505–515.
[10] Goodman, J. B., and LeVeque, R. J., *On the accuracy of stable schemes for 2D scalar conservation laws*, Math. Comp. 45, 1985, pp. 15–21.
[11] Harten, A., Lax, P. D., and van Leer, B., *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev. 25, 1983, pp. 35–61.
[12] Jin, S., *Implicit numerical schemes for hyperbolic conservation laws with stiff relaxation terms*, J. Comput. Phys., submitted.
[13] Jin, S., and Levermore, C. D., *Numerical schemes for hyperbolic systems with stiff relaxation terms*, J. Comput. Phys., submitted.
[14] Lax, P. D., *Weak solutions of nonlinear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math. 7, 1954, pp. 159–193.
[15] LeVeque, R. J., *Numerical Methods for Conservation Laws*, rev. ed., Birkhäuser, Basel, 1992.
[16] Liu, T. P., *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys. 108, 1987, pp. 153–175.
[17] Majda, A., and Pego, R., *Stable viscosity matrices for systems of conservation laws*, J. Differential Equations 56, 1985, pp. 229–269.
[18] Morawetz, C. S., *On a weak solution for a transonic flow problem*, Comm. Pure Appl. Math. 38, 1985, pp. 797–817.
[19] Nessyahu, H., and Tadmor, E., *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys. 87, 1990, pp. 408–463.
[20] Perthame, B., *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal. 29, 1992, pp. 1–29.
[21] Perthame, B., and Tadmor, E., *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys. 136, 1991, pp. 501–517.
[22] Pullin, D. I., *Direct simulation methods for compressible inviscid ideal-gas-flow*, J. Comput. Phys. 34, 1980, pp. 231–244.
[23] Reitz, R. D., *One-dimensional compressible gas dynamics calculations using the Boltzmann equation*, J. Comput. Phys. 42, 1981, pp. 108–123.
[24] Shu, C. W., and Osher, S., *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys. 77, 1988, pp. 439–471.
[25] Sod, G., *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws*, J. Comput. Phys. 27, 1978, pp. 1–31.
[26] Sweby, P. R., *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal. 21, 1984, pp. 995–1011.
[27] van Leer, B., *Towards the ultimate conservative difference schemes V. A second-order sequal to Godunov's method*, J. Comput. Phys. 32, 1979, pp. 101–136.
[28] Whitham, G. B., *Linear and Nonlinear Waves*, Wiley, New York, 1974.
[29] Woodward, P., and Colella, P., *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys. 54, 1984, pp. 115–173.

Received May 1993.