MATHEMATICS OF COMPUTATION
VOLUME 62, NUMBER 206
APRIL 1994, PAGES 497–530

WHY NONCONSERVATIVE SCHEMES CONVERGE TO WRONG SOLUTIONS: ERROR ANALYSIS

THOMAS Y. HOU AND PHILIPPE G. LE FLOCH

ABSTRACT. This paper attempts to give a qualitative and quantitative description of the numerical error introduced by using finite difference schemes in nonconservative form for scalar conservation laws. We show that these schemes converge strongly in L^1_{loc} norm to the solution of an inhomogeneous conservation law containing a Borel measure source term. Moreover, we analyze the properties of this Borel measure, and derive a sharp estimate for the L^1 error between the limit function given by the scheme and the correct solution. In general, the measure source term is of the order of the entropy dissipation measure associated with the scheme. In certain cases, the error can be small for short times, which makes it difficult to detect numerically. But generically, such an error will grow in time, and this would lead to a large error for large-time calculations. Finally, we show that a local correction of any high-order accurate scheme in nonconservative form is sufficient to ensure its convergence to the correct solution.

1. INTRODUCTION

The purpose of this paper is to analyze the error introduced by using nonconservative finite difference schemes for the approximation of conservation laws. Although it has been well known that a conservative scheme should be used in approximating hyperbolic conservation laws, it is still very tempting to use a nonconservative scheme in certain contexts because it may give a seemingly simpler or more convenient formulation (see, for instance, Zwas and Roseman [31], Moretti [26] and Karni [18]). We show in this paper that nonconservative schemes in general do not converge to the correct solution and derive the equation which the nonconservative schemes approximate. It has the form of an inhomogeneous conservation law containing a Borel measure source term. Further, we analyze the properties of the measure source term, and estimate the L^1 error between the limit function given by the scheme and the correct solution. We show that the measure source term in general does not vanish, and is of the order of the entropy dissipation measure (see §§3 and 4).

Received by the editor January 7, 1992 and, in revised form, July 15, 1992.

1991 *Mathematics Subject Classification.* Primary 35L65, 65N06.

Key words and phrases. Hyperbolic conservation law, entropy discontinuous solution, nonconservative scheme, numerical error.

The first author was partially supported by an NSF grant DMS-9003202, AFOSR grant AFOSR-90-0090 and by the Sloan Foundation research fellowship.

The second author was partially supported by NSF grants DMS-88-06731, DMS 92-09326, and DMS 93-96260.

Our analysis also indicates that, in certain cases, the error due to nonconservation can be small for short times, which makes it difficult to detect numerically. But this is deceptive because generically such an error will grow in time. This would lead to a large error for large-time calculations, such as steady-state calculations.

Another motivation for this study is to understand how to approximate nonlinear hyperbolic systems in nonconservative form by finite difference schemes. In this case, globally conservative schemes are not appropriate and some kind of nonconservative schemes must be used. Theoretically, this is a more difficult question, even at the continuous level; see Dal Maso, Le Floch, and Murat [6] and Le Floch and Liu [24] for more discussions. On the other hand, this question has been addressed computationally, see for instance Colombeau and Leroux [1], Trangenstein [28], and Trangenstein and Colella [29]. It would be very interesting to analyze the convergence of these schemes. But as a start, we focus on an easier problem in the present paper: nonconservative schemes for conservative equations. We believe that a qualitative understanding of nonconservative schemes, even for conservative equations, may shed some light on the understanding of nonconservative schemes for nonconservative equations.

Consider a scalar conservation law in one space dimension, i.e., an equation of the form

$$\partial_t u + \partial_x f(u) = 0, \qquad u(t, x) \in R,\ t > 0,\ x \in R, \tag{1.1}$$

with initial value $u(0, x) = u_0(x)$. The so-called flux function $f: R \to R$ is a given smooth function, and the initial data u_0 belongs to the space $\mathrm{BV}(R)$ of all functions of bounded variation. As is well known, nonlinear hyperbolic equations like (1.1) in general do not admit smooth solutions globally defined in time. Weak solutions defined in the sense of distributions must be considered. An entropy criterion is also needed to ensure uniqueness in the class of weak solutions. See Lax [21] for background on hyperbolic equations, Volpert [30] and Kružkov [19] for an existence and uniqueness result of the entropy weak solution to (1.1). Many numerical techniques have been developed for approximating (1.1). A class of most widely used schemes is called conservative schemes. It is well known from the Lax-Wendroff Theorem [22] that a conservative difference scheme—if it converges—converges to a weak solution of (1.1).

The focus of the paper is to understand the error introduced by using nonconservative schemes for scalar conservation laws. Specifically, we consider a general (possibly high-order accurate) $(2k + 1)$-point finite difference scheme written in an incremental form (and so in general in a nonconservative form). Under suitable positivity conditions and a CFL (Courant-Friedrichs-Lewy) stability condition on the incremental coefficients, the scheme is shown to be TVD (Total Variation Diminishing) by using Harten's lemma [14]. We easily obtain the L^1 strong convergence of the scheme from standard compactness arguments; cf., for instance, [4]. The limit function given by the scheme is a function of bounded variation, denoted below by v, that can be very different from the exact (entropy weak) solution u of (1.1). We prove the following three facts concerning this limit v.

(1) *The function v is a solution to a conservation law containing a measure source term.*

We prove in §2 that v satisfies the following inhomogeneous conservation law:

$$\partial_t v + \partial_x f(v) = \mu, \tag{1.2}$$

where μ is a Borel measure defined on $R_+ \times R$. We show that the measure μ vanishes in the regions where v is a smooth function and the scheme converges strongly. The support of μ is expected to be concentrated on the curves (in (t, x)-plane) of discontinuity of the function v.

This measure, in general, is not identically zero unless the scheme under consideration admits a conservative form. For instance, we check numerically (in §6) that μ is not identically zero even if the scheme coincides with a conservative scheme up to second-order terms. In particular, a nonconservative scheme does not necessarily converge to the correct solution of (1.1), even if it contains the same numerical viscosity as the one of a conservative scheme.

(2) *The measure source term can be estimated.*

We next assume that the scheme under consideration is in some sense close to a (conservative and entropy-satisfying) E-scheme, e.g., the modified Lax-Friedrichs scheme or the Godunov scheme. That restricts our attention to first-order accurate schemes, only. For these schemes, we can prove discrete entropy inequalities, which yield the following entropy inequalities for the function v:

$$\partial_t \eta(v) + \partial_x q(v) \leq \mu_\eta. \tag{1.3}$$

In (1.3), (η, q) is any entropy-entropy flux pair for equation (1.1), while μ_η is a Borel measure depending on the entropy η and satisfying the bound

$$|\mu_\eta| \leq \left(\sup_u |\eta'| \right) |\mu| \tag{1.4}$$

as Borel measures. Here the supremum is taken over all u in the interval $[\inf u_0, \sup u_0]$, and $|\mu_\eta|$ and $|\mu|$ denote the measure of total variation of μ_η and μ, respectively.

We are able to derive an error estimate between v and the exact entropy weak solution u of (1.1). Here we assume that the nonconservative scheme under consideration coincides with a conservative one up to pth-order terms. Combining (1.3)-(1.4) with a result of DiPerna and Majda [8] leads to an error estimate. We prove that for every time $t \geq 0$,

$$\int_R |v(t, x) - u(t, x)| \, dx \leq Ct(\text{ampl}(u_0))^{p-1} \text{TV}(u_0), \tag{1.5}$$

where C is a positive constant (depending on the scheme only), $\text{ampl}(u_0)$ denotes the amplitude of the initial data and $\text{TV}(u_0)$ its total variation. Note that the error bound (1.5) grows linearly in time; however we also derive in this paper an estimate similar to (1.5) which is uniform in time.

In §4, we also provide a lower bound of this measure source term for a nonconservative version of the upwind scheme and the modified Lax-Friedrichs scheme. We show that for the upwind scheme, the measure source term coincides with the entropy dissipation measure (see Lemma 4.2). Thus the measure

source term cannot vanish identically if the solution contains discontinuities. Our estimates show that the measure source term is of the same order as the entropy dissipation measure.

Inequality (1.5) also shows that the error due to nonconservation could be small if the initial data is close to a constant. In this case, the error is difficult to detect. This is illustrated by numerical examples in §6. However, the smallness of the error is deceptive since the error will typically grow with time. The error due to conservation may accumulate to a large amount for large-time calculations, such as steady-state calculations for aerodynamics equations.

(3) *A local correction of a nonconservative scheme is sufficient to ensure its convergence to the correct solution.*

This part considers again high-order accurate difference schemes. We propose in §5 to modify any nonconservative scheme by doing a local correction only in the neighborhood of the discontinuities of the solution. If the gradient of the numerical solution exceeds some given bound (depending on the mesh size) in some region, we switch from the nonconservative scheme to a conservative scheme. We emphasize that this correction is performed only *locally*. After correction, the resulting scheme is still in nonconservative form. We refer to Harten and Zwas [16], Leroux and Quesseveur [25], and Harabetian and Pego [13] for similar treatments in the case of conservative schemes. We do not know if our correction is useful from a practical standpoint. We prove in §5 that a nonconservative scheme after correction converges to the correct (entropy weak) solution of (1.1). In other words, $\mu = 0$ in equations (1.2) and $\mu_\eta = 0$ in (1.3), so that $v \equiv u$.

We mention a pioneering work on this type of question by DiPerna in [7]. He proved (in the context of *systems* of conservation laws) the convergence of a scheme in nonconservative form defined by hybridization of the Lax-Friedrichs scheme and the random choice scheme introduced by Glimm [10].

Our results show that nonconservative finite difference schemes in general do not converge to the correct solution, even if the scheme contains the same numerical viscosity as that of a consistent conservative scheme. This implies that it is not enough to correct at the formal level the numerical viscosity of a nonconservative scheme, as Karni suggested in [18]. This point will be further elaborated in §6. We point out that for nonconservative hyperbolic systems, it has been proved in [24] that the Glimm scheme converges to the correct solution.

Related work can also be found in the papers of Goodman and Lax [12] and Hou and Lax [17] on the convergence of dispersive schemes. In that situation, the weak limit of the oscillatory solutions given by dispersive schemes does not satisfy equation (1.1).

An outline of the paper is as follows. In §2, we derive the limit equation with measure source term associated with a scheme in nonconservative form. Section 3 gives an estimate of the error due to nonconservation. Several examples are given in §4, which show that the error source term does not vanish as the mesh size tends to zero, and is of the order of the entropy dissipation measure. Section 5 presents a method of correction of any nonconservative scheme. Finally, we present in §6 numerical experiments, which confirm the analytical results of this paper.

2. Limit equation associated with a scheme in nonconservative form

This section introduces the nonconservative schemes under consideration, and states their elementary properties of stability and convergence. We derive in Theorem 2.1 the equation satisfied by the solution of a nonconservative scheme in the limit. This equation contains a Borel measure source term. In Theorem 2.2, we analyze this measure in some detail, and show that the measure is supported along the curves of discontinuity of the solution.

Let h and τ be the space and time increments of the discretization, and set

$$x_i = ih, \quad x_{i+1/2} = (i+1/2)h, \quad t_n = n\tau \quad \text{for } i \in Z,\ n \in N.$$

The ratio $\lambda = \tau/h$ will always be kept constant. We consider approximate solutions to problem (1.1) that have the form

$$u^h(t, x) = u_i^n, \qquad x \in [x_{i-1/2}, x_{i+1/2}),\ t \in [t_n, t_{n+1}). \tag{2.1a}$$

We compute $\{u_i^0\}_{i\in Z}$ by L^2 projection of the initial data u_0, as follows:

$$u_i^0 = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} u_0(y)dy, \qquad i \in Z. \tag{2.1b}$$

We then determine $\{u_i^{n+1}\}_{i\in Z}$ from $\{u_i^n\}_{i\in Z}$ by the following $(2k+1)$-point finite difference scheme in incremental form:

$$u_i^{n+1} = u_i^n - C_{i-1/2}^n(u_i^n - u_{i-1}^n) + D_{i+1/2}^n(u_{i+1}^n - u_i^n), \qquad i \in Z, \tag{2.1c}$$

where

$$C_{i-1/2}^n = C(u_{i-k}^n, \dots, u_{i+k-1}^n; \lambda), \qquad D_{i+1/2}^n = D(u_{i-k+1}^n, \dots, u_{i+k}^n; \lambda). \tag{2.1d}$$

In (2.1d), the incremental coefficients C and D are Lipschitz continuous functions defined from $R^{2k} \times R_+$ into R. We also assume in what follows that the coefficients C and D are consistent with the flux function f of equation (1.1), in the sense that

$$C(\alpha_0, \dots, \alpha_0; \lambda) - D(\alpha_0, \dots, \alpha_0; \lambda) = \lambda f'(\alpha_0) \quad \text{for } \alpha_0 \in R. \tag{2.2}$$

We recall classical conditions on C and D which guarantee the uniform stability in L^∞ and BV norms for the scheme (2.1a)-(2.1d). We refer to Harten [14] for a proof. We emphasize that the proof of [14] does not require the scheme to be conservative. Throughout this paper, we denote the total variation in x of a function w by $\mathrm{TV}(w)$.

Lemma 2.1. *Suppose that the functions C and D satisfy the positivity property*

$$C(\alpha; \lambda) \geq 0, \qquad D(\alpha; \lambda) \geq 0 \quad \textit{for all } \alpha \in R^{2k}, \tag{2.3a}$$

and the CFL stability condition

$$C(\alpha; \lambda) \leq 1/2, \qquad D(\alpha; \lambda) \leq 1/2 \quad \textit{for all } \alpha \in R^{2k}. \tag{2.3b}$$

Then the scheme (2.1a)-(2.1d) *satisfies the local maximum principle*

$$\min(u_{i-k}^n, \dots, u_{i+k}^n) \leq u_i^{n+1} \leq \max(u_{i-k}^n, \dots, u_{i+k}^n), \qquad n \in N,\ i \in Z, \tag{2.4a}$$

the TVD (Total Variation Diminishing) property

$$\text{(2.4b)} \qquad \mathrm{TV}(u^h(t_{n+1})) = \sum_{i\in Z} |u_i^{n+1} - u_{i-1}^{n+1}| \le \mathrm{TV}(u^h(t_n)) = \sum_{i\in Z} |u_i^n - u_{i-1}^n|$$

and the L^1 Lipschitz continuity property

$$\|u^h(t_m) - u^h(t_n)\|_{L^1(R)} = \sum_{i\in Z} |u_i^m - u_i^n| h \le \frac{1}{\lambda} |t_m - t_n| \mathrm{TV}(u_0),$$

for any $n, m \in N$.

By means of Lemma 2.1 and Helly's compactness theorem, it is a classical matter to verify that (a subsequence of) $\{u^h\}$ defined by (2.1a) converges in L^1_{loc} to a function $v: R_+ \times R \to R$ which satisfies

$$\text{(2.5a)} \qquad \inf u_0 \le v(t, x) \le \sup u_0 \quad \text{for almost every } (t, x) \in R_+ \times R,$$

$$\text{(2.5b)} \qquad \mathrm{TV}(v(t')) \le \mathrm{TV}(v(t)) \quad \text{for } t' \ge t \ge 0,$$

and

$$\text{(2.5c)} \qquad \|v(t') - v(t)\|_{L^1(R)} \le \frac{1}{\lambda} |t' - t| \mathrm{TV}(u_0) \quad \text{for } t, t' \ge 0.$$

However, the function $v = \lim u^h$ need not be a solution of the conservation law (1.1). This is in strong contrast to the case of conservative schemes: namely for conservative schemes the property of L^1_{loc} (or almost everywhere) convergence is sufficient for the passage to the limit in the scheme (2.1c), and the function $v = \lim u^h$ must be a weak solution to (1.1) (see Lax and Wendroff [22]). This is not necessarily the case for schemes in nonconservative form.

Our aim now is to derive the equation satisfied by v.

Theorem 2.1. *Suppose the scheme* (2.1a)-(2.1d) *satisfies the properties* (2.2)-(2.3b). *Then the function $v = \lim u^h$ is a weak solution (in the sense of distributions) of the following conservation law with a source term:*

$$\text{(2.6)} \qquad \partial_t v + \partial_x f(v) = \mu,$$

where μ is a locally bounded real-valued Borel measure defined on $R_+ \times R$ and characterized as follows.

Let $\tilde{C}$ and $\tilde{D}$ be the incremental coefficients of any conservative scheme satisfying condition (2.2). *Then the measure μ in* (2.6) *can be characterized as the weak-star limit of the following sequence of functions, which is uniformly bounded in $L^1_{\mathrm{loc}}(R_+ \times R)$:*

$$\text{(2.7a)} \qquad w^h(t, x) = w_i^n, \qquad x \in [x_{i-1/2}, x_{i+1/2}), \quad t \in [t_n, t_{n+1}),$$

and

$$\text{(2.7b)} \qquad w_i^n = \frac{1}{h}(\tilde{C}_{i-1/2}^n - C_{i-1/2}^n)(u_i^n - u_{i-1}^n) - \frac{1}{h}(\tilde{D}_{i+1/2}^n - D_{i+1/2}^n)(u_{i+1}^n - u_i^n),$$

for $i \in Z$, $n \in N$.

Proof. With $\tilde{C}$ and $\tilde{D}$ being the incremental coefficients of any $(2l+1)$-point conservative scheme consistent with (1.1), we can rewrite the nonconservative scheme (2.1a)-(2.1d) in the form ($i \in Z$, $n \in N$)

$$u_i^{n+1} = u_i^n - \tilde{C}_{i-1/2}^n (u_i^n - u_{i-1}^n) + \tilde{D}_{i+1/2}^n (u_{i+1}^n - u_i^n) + h w_i^n,$$

where w_i^n is defined in (2.7b). Since $\tilde{C}$ and $\tilde{D}$ correspond to a conservative scheme, there exists by definition a numerical flux function $g: R^{2l} \times R_+ \to R$ such that

$$\tilde{C}_{i-1/2}^n(u_i^n - u_{i-1}^n) + \tilde{D}_{i+1/2}^n(u_{i+1}^n - u_i^n) = \lambda(g_{i+1/2}^n - g_{i-1/2}^n),$$

where $g_{i+1/2}^n = g(u_{i-l+1}^n, \ldots, u_{i+l}^n; \lambda)$. We thus obtain

$$\frac{1}{\tau}(u_i^{n+1} - u_i^n) + \frac{1}{h}(g_{i+1/2}^n - g_{i-1/2}^n) = w_i^n. \tag{2.8}$$

By the Lax-Wendroff theorem, the left-hand side of (2.8) has a limit in the sense of distributions:

$$\partial_t v + \partial_x f(v).$$

Concerning the right-hand side of (2.8), we have for any $T > 0$

$$\begin{aligned}
&\sum_{n\tau \le T} \sum_{i \in Z} |w_i^n| h\tau \\
&\quad \le \sum_{n\tau \le T} \sum_{i \in Z} (|\tilde{C}_{i-1/2}^n - C_{i-1/2}^n||u_i^n - u_{i-1}^n| + |\tilde{D}_{i+1/2}^n - D_{i+1/2}^n||u_{i+1}^n - u_i^n|)\tau \\
&\quad \le (\sup|\tilde{C}| + \sup|C| + \sup|\tilde{D}| + \sup|D|) \sum_{n\tau \le T} \mathrm{TV}(u^h(t_n))\tau,
\end{aligned}$$

where the suprema are taken over all u in the interval $[\inf u_0, \sup u_0]$ (cf. (2.4a)). In view of property (2.4b), we obtain the bound

$$\sum_{n\tau \le T} \sum_{i \in Z} |w_i^n| h\tau \le O(1) T \mathrm{TV}(u_0). \tag{2.9}$$

Estimate (2.9) shows that the sequence $\{w^h\}$ defined by (2.7a)-(2.7b) is uniformly bounded in L^1_{loc}. Therefore, it converges in the weak-star sense of bounded measures to a Borel measure denoted by μ. As a consequence, (2.8) implies (2.6). The proof is completed. □

Remark 2.1. It is worth noting that the definition of μ is independent of the conservative scheme considered. Let $\tilde{C}'$ and $\tilde{D}'$ be the incremental coefficients of (another) conservative scheme consistent with (1.1). Set

$$w'^h(t, x) = w'^n_i \quad \text{for } x \in [x_{i-1/2}, x_{i+1/2}),\ t \in [t_n, t_{n+1})$$

with

$$w'^n_i = \frac{1}{h}(\tilde{C}'^n_{i-1/2} - C_{i-1/2}^n)(u_i^n - u_{i-1}^n) - \frac{1}{h}(\tilde{D}'^n_{i+1/2} - D_{i+1/2}^n)(u_{i+1}^n - u_i^n).$$

The sequences $\{w^h\}$ and $\{w'^h\}$ converge in the sense of distributions to the same Borel measure. Indeed, one has

$$\begin{aligned}
w'^n_i - w_i^n = &\left\{\frac{1}{h}\tilde{C}'^n_{i-1/2}(u_i^n - u_{i-1}^n) - \frac{1}{h}\tilde{D}'^n_{i+1/2}(u_{i+1}^n - u_i^n)\right\} \\
&- \left\{\frac{1}{h}\tilde{C}_{i-1/2}^n(u_i^n - u_{i-1}^n) - \frac{1}{h}\tilde{D}_{i+1/2}^n(u_{i+1}^n - u_i^n)\right\},
\end{aligned}$$

where *each* expression in parentheses admits by assumption a conservative form, and hence converges in the sense of distributions to the same term $\partial_x f(v)$. This proves that the sequence $\{w'^h - w^h\}$ tends to zero weakly. □

Next, we analyze the support in the (t, x)-plane of the measure μ introduced in Theorem 2.1. We conjecture that this measure is concentrated on the curves of discontinuity of the BV function v. We do not prove here this result in its whole generality. But we show that μ must vanish identically in regions where u^h converges uniformly with respect to x. Moreover, we show that, in the regions of monotonicity and continuity of v, strong convergence is a consequence of the TVD property of the scheme. We also check that the measure μ is absolutely continuous with respect to the entropy dissipation measure associated with the scheme.

Theorem 2.2. 1) *Define the entropy dissipation measure* β *as the weak-star limit of the sequence*

$$b^h(t, x) = (u_i^n - u_{i-1}^n)^2/h, \qquad x \in [x_{i-1/2}, x_{i+1/2}), \quad t \in [t_n, t_{n+1}).$$

Then the measure μ *introduced in Theorem* 2.1 *is absolutely continuous with respect to the entropy dissipation measure* β.

2) *Let* Ω *be an open subset of the* (t, x)*-plane. Suppose that*

$$u^h \to u \quad \text{in } L^1((T_1, T_2), L^\infty(a, b)) \tag{2.10}$$

for all compact subsets $[T_1, T_2] \times [a, b] \subset \Omega$. *Then the measure* μ *found in Theorem* 2.1 *vanishes identically in the set* Ω, *i.e.,* $\mu(B) = 0$ *for each compact set* $B \subset \Omega$.

3) *In particular, the above assumption* (2.10) *is satisfied for any set* Ω *in which, for almost every time* t, $v(t)$ *is continuous with respect to* x *and each function* $u^h(t)$ *is nondecreasing* (*or nonincreasing*) *with respect to* x.

Proof. We assume first that (2.10) holds, and fix a subset $[T_1, T_2] \times [a, b]$ of Ω. Let θ be a test function with support in $(T_1, T_2) \times (a, b)$ and set $\theta_i^n = \theta(x_i, t_n)$. We are going to prove that $\int \theta \, d\mu = 0$. We have, by definition (2.7b),

$$\begin{aligned}
&\sum_{a \le x_i \le b} \sum_{T_1 \le t_n \le T_2} w_i^n \theta_i^n h\tau \\
&= \sum_{a \le x_i \le b} \sum_{T_1 \le t_n \le T_2} \{\theta_i^n (\tilde{C}_{i-1/2}^n - C_{i-1/2}^n) - \theta_{i-1}^n (\tilde{D}_{i-1/2}^n - D_{i-1/2}^n)\}(u_i^n - u_{i-1}^n)\tau \\
&= \sum_{a \le x_i \le b} \sum_{T_1 \le t_n \le T_2} \theta_i^n \{\tilde{C}_{i-1/2}^n - \tilde{D}_{i-1/2}^n - C_{i-1/2}^n + D_{i-1/2}^n\}(u_i^n - u_{i-1}^n)\tau \\
&\quad + \sum_{a \le x_i \le b} \sum_{T_1 \le t_n \le T_2} (\theta_i^n - \theta_{i-1}^n)(\tilde{D}_{i-1/2}^n - D_{i-1/2}^n)(u_i^n - u_{i-1}^n)\tau.
\end{aligned}$$

In view of (2.2) and the Lipschitz continuity of the incremental coefficients, we

deduce that

$$\left|\sum_{a\le x_i\le b}\sum_{T_1\le t_n\le T_2}\theta_i^n w_i^n h\tau\right|$$

$$\le O(1)\sum_{\substack{a\le x_i\le b\\ T_1\le t_n\le T_2}}|\theta_i^n||u_i^n-u_{i-1}^n|\sum_{j=-k}^{k-1}|u_{i+j}^n-u_i^n|\tau+O(1)\,h\sum_{\substack{a\le x_i\le b\\ T_1\le t_n\le T_2}}|u_i^n-u_{i-1}^n|\tau$$

$$\le O(1)\sum_{\substack{a\le x_i\le b\\ T_1\le t_n\le T_2}}|\theta_i^n||u_{i+1}^n-u_i^n|^2\tau+O(1)\,h\sum_{\substack{a\le x_i\le b\\ T_1\le t_n\le T_2}}|u_i^n-u_{i-1}^n|\tau$$

$$\le O(1)\sum_{T_1\le t_n\le T_2}\sup_{a\le x_i\le b}|u_{i+1}^n-u_i^n|\tau\mathrm{TV}(u_0)+O(1)h\mathrm{TV}(u_0),$$

where we have used (2.4b) for the last inequality. Because of the assumption (2.10), we obtain

$$\lim_{h\to 0}\int_a^b\int_{T_1}^{T_2}\theta w^h\,dxdt=\lim_{h\to 0}\left\{\sum_{a\le x_i\le b}\sum_{T_1\le t_n\le T_2}\theta_i^n w_i^n h\tau\right\}=0,$$

so the restriction $w^h|_{(a,b)\times(T_1,T_2)}$ tends to zero weakly, and

$$\mu=\lim w^h=0\quad\text{in } (a,b)\times(T_1,T_2).$$

The second assertion of the theorem follows. The statement 1) of the theorem is a consequence of the above inequalities. To see this, we note that the above derivation implies

$$\sum_{i\in Z}\sum_{n\in Z}\theta_i^n w_i^n h\tau\le O(1)\sum_{i\in Z}\sum_{n\in Z}|\theta_i^n||u_{i+1}-u_i|^2\tau$$

for any test function θ. Passing to the limit, we get

$$\left|\int\theta\,d\mu\right|=|\langle\mu,\theta\rangle|\le O(1)\langle\beta,|\theta|\rangle=O(1)\int|\theta|\,d\beta,$$

which means that the measure μ is absolutely continuous with respect to the entropy dissipation measure β.

We now prove the third assertion of the theorem. We suppose that in a set Ω and for almost every time t, the function $v(t)$ is continuous in x and that each function $u^h(t)$ is monotone in x. We fix a compact set $[T_1,T_2]\times[a,b]\subset\Omega$. In view of the monotonicity of the function $u^h(t_0,\cdot)$, the continuity of $v(t_0,\cdot)$, and the convergence almost everywhere of $u^h(t_0,\cdot)$, Dini's Theorem shows that, for all times t_0 in (T_1,T_2) except those in a set of measure zero,

$$u^h(t_0,\cdot)\to v(t_0,\cdot)\quad\text{uniformly in } [a,b]. \tag{2.11}$$

Since (2.11) holds for almost every time t in (T_1,T_2), the Lebesgue convergence theorem implies (2.10). The proof is complete. □

Theorem 2.2 can be applied in particular to the case that the initial data u_0 admits a finite number of points in which monotonicity changes. Since the scheme (2.1a)-(2.1d) is assumed to be TVD, an initially monotone region in the

initial data generates a monotone approximate solution u^h. In these regions, assumption (2.10) of Theorem 2.2 holds. Dealing with the points in which monotonicity changes would require further analysis, cf. Le Floch and Liu [23]. We summarize the results of Theorems 2.1 and 2.2 as follows:

Corollary 2.1. *The nonconservative scheme* (2.1a)-(2.1d) *converges in the* L^1_{loc} *norm strongly to a function of bounded variation* v *that is a solution to a conservation law with a measure source term* μ:

$$\partial_t v + \partial_x f(v) = \mu. \tag{2.12}$$

If u_0 *has a finite number of points in which monotonicity changes, and* v *is piecewise continuous, then, at least for small times, the measure* μ *is concentrated on the union of the curves of discontinuity of* v *and the curves of points in which monotonicity of* v *changes.*

The assumption of piecewise continuity of v is realistic, since it is known—at least for (1.1)—that solutions of conservation laws are generally piecewise smooth (Dafermos [5]). In our case, proving this property is difficult because we do not know explicitly the equation satisfied by v; so the proof by Dafermos does not generalize. Also, the method of proof introduced by Glimm and Lax [11] was for the Glimm scheme, and it is not clear how it could be extended to schemes with numerical viscosity, such as the difference schemes considered here. See, however, [23].

Studying the convergence of nonconservative schemes requires information on the local convergence of the scheme (in the spirit of Glimm and Lax). But such a result of local convergence is not known for difference schemes, even in the conservative case (Kuznetsov's error estimate, e.g., [20] and [27], is an L^1 estimate; here we need L^∞_{loc} convergence!). Furthermore, note that we only obtain that a subsequence of $\{u^h\}$ is convergent. This is due to the fact that we do not have a uniqueness theorem for equations with measure source term, like (2.12).

3. Estimates for the error due to nonconservation

In this section, we prove in Theorem 3.3 a L^1 estimate for the difference $v-u$ between the limit function $v = \lim u^h$ given by the scheme (2.1a)-(2.1d) and the entropy weak solution u of (1.1). To this end, we derive for the function v entropy inequalities containing error terms similar to (2.6); cf. Theorem 3.1. Next, in Theorem 3.2, we follow arguments due to DiPerna and Majda [8] to analyze the resulting entropy inequalities with measure source terms.

To begin with, we estimate the entropy dissipation of the scheme (2.1a)-(2.1d) by comparing the scheme with the (modified) conservative Lax-Friedrichs scheme, or more generally with any E-scheme. We recall that the modified Lax-Friedrichs scheme is characterized by its incremental coefficients:

$$\tilde{C}_{\text{LF}}(\alpha_1, \alpha_2; \lambda) = \frac{\lambda}{2}\frac{f(\alpha_2) - f(\alpha_1)}{\alpha_2 - \alpha_1} + \frac{1}{4}, \tag{3.1a}$$

$$\tilde{D}_{\text{LF}}(\alpha_1, \alpha_2; \lambda) = -\frac{\lambda}{2}\frac{f(\alpha_2) - f(\alpha_1)}{\alpha_2 - \alpha_1} + \frac{1}{4}, \qquad (\alpha_1, \alpha_2) \in R^2. \tag{3.1b}$$

An E-scheme is characterized by the following incremental coefficients:

$$\tilde{C}_E(\alpha_1, \alpha_2; \lambda) = \frac{\lambda}{2} \frac{f(\alpha_2) - g_E(\alpha_1, \alpha_2)}{\alpha_2 - \alpha_1}, \tag{3.1aa}$$

$$\tilde{D}_E(\alpha_1, \alpha_2; \lambda) = \frac{\lambda}{2} \frac{f(\alpha_1) - g_E(\alpha_1, \alpha_2)}{\alpha_2 - \alpha_1}, \tag{3.1bb}$$

where $g_E: R^2 \to R$ must satisfy

$$(\alpha_2 - \alpha_1)(g_E(\alpha_1, \alpha_2; \lambda) - f(w)) \leq 0$$

for all $(\alpha_1, \alpha_2) \in R^2$ and for every w between α_2 and α_1. We consider the scheme in incremental form (2.1a)-(2.1d) under the assumptions (2.2)-(2.3b). Moreover, we assume that

$$|C(\alpha; \lambda) - \tilde{C}_E(\alpha_0, \alpha_1; \lambda)| \leq A_1(\lambda) \sum_{j=-k+1}^{k-1} |\alpha_{j+1} - \alpha_j|^p, \tag{3.2a}$$

$$|D(\alpha; \lambda) - \tilde{D}_E(\alpha_0), \alpha_1; \lambda)| \leq A_2(\lambda) \sum_{j=-k+1}^{k-1} |\alpha_{j+1} - \alpha_j|^p, \tag{3.2b}$$

for all $\alpha = (\alpha_{-k+1}, \ldots, \alpha_k) \in R^{2k}$, $A_1 = A_1(\lambda)$ and $A_2 = A_2(\lambda)$ being two positive constants and $p \geq 1$ an integer. Assumption (3.2a)-(3.2b) expresses that the scheme (2.1a)-(2.1d) is "close enough" to an E-scheme (cf. examples in §4).

We now define a (nonnegative) locally bounded Borel measure $\hat{\mu}$ by

$$\hat{\mu} = \text{weak}^* \lim \hat{w}^h, \tag{3.3aa}$$

$$\hat{w}^h(t, x) = \hat{w}_i^n, \qquad x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}), \quad t \in [t_n, t_{n+1}), \tag{3.3bb}$$

and

$$\begin{aligned} \hat{w}_i^n = {} & \frac{A_1(\lambda)}{h} \sum_{j=-k+1}^{k-1} |u_{i+j}^n - u_{i+j-1}^n|^p |u_i^n - u_{i-1}^n| \\ & + \frac{A_2(\lambda)}{h} \sum_{j=-k+1}^{k-1} |u_{i+j+1}^n - u_{i+j}^n|^p |u_{i+1}^n - u_i^n|. \end{aligned} \tag{3.3cc}$$

(Eventually, one needs to extract a subsequence of $\hat{w}^h$.)

We recall that a Lipschitz continuous function $(\eta, q): R \to R^2$ is a convex entropy-entropy flux pair (or entropy pair, for short) for equation (1.1) if one has

$$\eta \text{ strictly convex, and } q' = \eta' f'.$$

Theorem 3.1. *Consider the scheme* (2.1a)-(2.1d) *under the assumptions* (2.2)-(2.3b) *and* (3.2a)-(3.2b), *together with its limit* $v = \lim u^h$. *Then, for each entropy pair* (η, q), *one has*

$$\partial_t \eta(v) + \partial_x q(v) \leq (\sup |\eta'|) \hat{\mu} \tag{3.4}$$

in the sense of distributions, where $\hat{\mu}$ *is defined by* (3.3aa)-(3.3cc) *and the supremum is taken over the interval* $[\inf u_0, \sup u_0]$.

Remark 3.1. In view of hypothesis (3.2a)-(3.2b) and the definition (3.3aa)-(3.3cc) of $\hat{\mu}$, the measure μ defined in (2.7a)-(2.7b) satisfies $-\hat{\mu} \le \mu \le \hat{\mu}$. Note that $\hat{\mu}$ is by definition nonnegative, while the measure μ has no specific sign. □

To prove Theorem 3.1, we derive discrete entropy inequalities for the scheme (2.1a)-(2.1d). Given an entropy pair (η, q), we denote by $\tilde{Q}_E$ a numerical entropy flux corresponding to the E-scheme (3.1aa)-(3.1bb).

Lemma 3.1. *Under the assumptions of Theorem* 3.1, *one has*

$$\eta(u_i^{n+1}) - \eta(u_i^n) - \lambda(\tilde{Q}_E(u_i^n, u_{i+1}^n) - \tilde{Q}_E(u_{i-1}^n, u_i^n)) \le \tau \hat{w}_i^n \sup|\eta'|. \tag{3.4a}$$

Lemma 3.2. *Under the assumptions of Theorem* 3.1, *and if the flux is convex, one has*

$$\sum_{i\in Z} \eta(u_i^{n+1}) - \sum_{i\in Z} \eta(u_i^n) + \frac{\lambda}{96}(\inf f'')(\inf \eta'') \sum_{i\in Z} |u_{i+1}^n - u_i^n|^3 \le \sum_{i\in Z} \tau \hat{w}_i^n \sup|\eta'|. \tag{3.4b}$$

If, moreover, the E-scheme of reference is chosen to be the Lax-Friedrichs scheme, and if the CFL condition

$$\lambda \sup|f'| \le \frac{1}{2} - \theta \quad \text{for some } \theta \in (0, 1/2)$$

is satisfied, then, without assuming convexity of f, *one has*

$$\begin{cases} \eta(u_i^{n+1}) - \eta(u_i^n) - \lambda(\tilde{Q}(u_i^n, u_{i+1}^n) - \tilde{Q}(u_{i-1}^n, u_i^n)) \\ \quad + \frac{\theta}{32}(\inf \eta'')\,(|u_i^n - u_{i-1}^n|^2 + |u_{i+1}^n - u_i^n|^2) \le \tau \hat{w}_i^n \sup|\eta'|. \end{cases} \tag{3.4c}$$

Proof of Lemma 3.1. We decompose the scheme (2.1a)-(2.1d) in the form

$$u_i^{n+1} = \tilde{u}_i^{n+1} + (u_i^{n+1} - \tilde{u}_i^{n+1}), \qquad i \in Z,\ n \in N,$$

where $\{\tilde{u}_i^{n+1}\}_{i\in Z}$ is obtained from $\{u_i^n\}_{i\in Z}$ by the E-scheme (cf. (3.1aa)-(3.1bb)). It is known that the latter satisfies discrete entropy inequalities:

$$\eta(\tilde{u}_i^{n+1}) - \eta(u_i^n) - \lambda(\tilde{Q}_E(u_i^n, u_{i+1}^n) - \tilde{Q}_E(u_{i-1}^n, u_i^n)) \le 0, \qquad i \in Z,\ n \in N. \tag{3.5}$$

The left-hand sides of inequalities (3.4a) and (3.5) only differ by the factor $\eta(u_i^{n+1}) - \eta(\tilde{u}_i^{n+1})$, which we now estimate. Since η is a convex function, and using (2.1c), we have

$$\begin{aligned} \eta(u_i^{n+1}) - \eta(\tilde{u}_i^{n+1}) &\le \eta'(u_i^{n+1})(u_i^{n+1} - \tilde{u}_i^{n+1}) \\ &= \eta'(u_i^{n+1})\{-(C_{i-1/2}^n - \tilde{C}_{i-1/2}^n)(u_i^n - u_{i-1}^n) + (D_{i+1/2}^n - \tilde{D}_{i+1/2}^n)(u_{i+1}^n - u_i^n)\}. \end{aligned}$$

Then we use (3.2a)-(3.2b) and (3.3cc) to deduce from the above inequalities that

$$\eta(u_i^{n+1}) - \eta(\tilde{u}_i^{n+1}) \le \tau \hat{w}_i^n \sup|\eta'|. \tag{3.5a}$$

Combining inequalities (3.5) and (3.5a) gives (3.4a). □

Proof of Lemma 3.2. We follow the argument in the proof of Lemma 3.1, but now use the stronger entropy inequalities proved by Coquel and Le Floch [2]. If $\{\tilde{u}_i^{n+1}\}$ is obtained from $\{u_i^n\}$ by an E-scheme, then [2] gives

$$\sum_{i\in Z}\eta(\tilde{u}_i^{n+1}) - \sum_{i\in Z}\eta(u_i^n) + \frac{\lambda}{96}(\inf f'')(\inf \eta'')\sum_{i\in Z}|u_{i+1}^n - u_i^n|^3 \le 0.$$

Combining this inequality with (3.5a) in the proof of Lemma 3.1 gives (3.4b).

In the special case of the Lax-Friedrichs scheme, it was proved in [2] that

$$\eta(\tilde{u}_i^{n+1}) - \eta(u_i^n) + \frac{\theta}{32}(\inf \eta'')(|u_i^n - u_{i-1}^n|^2 + |u_{i+1}^n - u_i^n|^2) \le 0. \tag{3.5b}$$

Thus, in that case, one obtains (3.4c). □

From the discrete entropy inequalities (3.4a) given by Lemma 3.1 we deduce immediately the result of Theorem 3.1: the function $v = \lim u^h$ satisfies entropy inequalities containing measure source terms. We omit the proof of Theorem 3.1. We are going to deduce from the entropy inequalities (3.4), satisfied by v, the desired L^1 error estimate. To this end, we recall a result of DiPerna and Majda [8] (therein, conservation laws with measure source term were useful to analyze the method of nonlinear geometric optics).

Theorem 3.2. *Let u be an entropy weak solution of* (1.1) *in* $L^\infty(R_+, \mathrm{BV}(R))$. *Let w be any function in* $L^\infty(R_+, \mathrm{BV}(R))$ *satisfying the entropy inequalities*

$$\partial_t \eta(w) + \partial_x q(w) \le (\sup|\eta'|)\bar{\mu} \tag{3.6}$$

for all convex entropy pairs (η, q), where $\bar{\mu}$ is an arbitrary nonnegative Borel measure on $R_+ \times R$ independent of the entropy η. Then, for any time $t \ge 0$, one has

$$\int_R |w(t,x) - u(t,x)|\,dx \le \int_R |w(0,x) - u(0,x)|\,dx + \bar{\mu}(R \times [0,t)). \tag{3.7}$$

The proof of Theorem 3.2 is easy from the arguments of [8]. We now choose $w = v$ and $\bar{\mu} = \hat{\mu}$ in Theorem 3.2. The main assumption (3.6) of Theorem 3.2 is precisely the conclusion (3.4) given by Theorem 3.1. We thus have the inequality (3.7) with $w = v$ (here, $v(0,x) = u(0,x)$ in view of (2.1b)). It only remains to estimate the mass $\hat{\mu}([0,t) \times R)$ of the measure $\hat{\mu}$ defined by (3.3aa)-(3.3cc) in terms of the size of the initial data u_0. This is done in the following theorem.

Theorem 3.3. *Consider the scheme* (2.1a)-(2.1d) *and its limit $v = \lim u^h$ under the assumptions* (2.2)-(2.3b). *Suppose that the scheme is close to an E-scheme, in the sense* (3.2a)-(3.2b). *Then the following error estimate holds, for every time $t \ge 0$:*

$$\int_R |v(t,x) - u(t,x)|\,dx \le 2k(A_1 + A_2)t\mathrm{TV}(u_0)(\mathrm{oscil}(u_0))^p, \tag{3.8}$$

where u is the entropy solution of (1.1) *with initial data u_0 and the oscillation of the initial data u_0 is defined by*

$$\mathrm{oscil}(u_0) = \sup u_0 - \inf u_0.$$

If, moreover, the flux function f is convex and the constant

$$K = \frac{\lambda(\inf f'')}{48} - \|u_0\|_{L^\infty}(\operatorname{oscil}(u_0))^{p-2}4k(A_1 + A_2) \tag{3.9}$$

is positive, we have the ***uniform in time*** *estimate*:

$$\int_R |v(t, x) - u(t, x)|\, dx \le \frac{2k\lambda}{K}(A_1 + A_2)\|u_0\|_{L^2}^2 (\operatorname{oscil}(u_0))^{p-2}. \tag{3.10}$$

In the special case of the Lax-Friedrichs scheme (even if f is not convex), if the CFL condition

$$\lambda \sup |f'| \le \frac{1}{2} - \theta \quad \text{for some } \theta \in (1, 1/2),$$

is satisfied and if the constant L defined by

$$L = \frac{\theta}{4} - \|u_0\|_{L^\infty}\operatorname{oscil}(u_0)^{p-1}4k(A_1 + A_2) \tag{3.11}$$

is positive, then we have

$$\int_R |v(t, x) - u(t, x)|\, dx \le \frac{2k}{L}(A_1 + A_2)\|u_0\|_{L^2}^2 (\operatorname{oscil}(u_0))^{p-1}. \tag{3.12}$$

Proof. By Theorem 3.1 and Theorem 3.2 it is clear that

$$\int_R |v(t, x) - u(t, x)|\, dx \le \hat{\mu}([0, t) \times R). \tag{3.13}$$

In view of (3.3aa)-(3.3cc), we have

$$\hat{\mu}([0, t) \times R) \le (A_1 + A_2) \lim_{h\to 0} \sum_{\substack{i\in Z \\ n\in N,\, t_n \le t}} \sum_{j=-k+1}^{k} |u_{i+j}^n - u_{i+j-1}^n|^p |u_i^n - u_{i-1}^n|\tau. \tag{3.14}$$

Using the maximum principle and the TVD property (2.4a), (2.4b), one finds

$$\begin{aligned}
\hat{\mu}([0, t) \times R) &\le (A_1 + A_2)2k(\operatorname{oscil}(u_0))^p \lim_{h\to 0} \sum_{i\in Z,\, n\in N} |u_i^n - u_{i-1}^n|\tau \\
&\le (A_1 + A_2)2k(\operatorname{oscil}(u_0))^p \lim_{h\to 0} \sum_{t_n \le t} \operatorname{TV}(u_0)\tau \\
&\le (A_1 + A_2)2kt(\operatorname{oscil}(u_0))^p \operatorname{TV}(u_0),
\end{aligned}$$

which proves estimate (3.8) in view of (3.13)-(3.14).

To get the estimate (3.10), we start from the discrete entropy inequality (3.4a),

$$\begin{aligned}
&\sum_{i\in Z} \frac{1}{2}(u_i^{n+1})^2 - \sum_{i\in Z} \frac{1}{2}(u_i^n)^2 + \frac{\lambda}{96}(\inf f'') \sum_{i\in Z} |u_{i+1}^n - u_i^n|^3 \\
&\qquad \le \sup |\eta'| \sum_{i\in Z} (A_1 + A_2) \sum_{j=-k+1}^{k} |u_{i+j}^n - u_{i+j-1}^n|^p |u_i^n - u_{i-1}^n|,
\end{aligned}$$

where for definiteness we have chosen $\eta(u) = u^2/2$. After summation with respect to n, we have

$$\sum_{\substack{i\in Z\\ n\in Z,\, t_n\le t}} \frac{1}{2}(u_i^{n+1})^2 - \sum_{i\in Z}\frac{1}{2}(u_i^0)^2 + \frac{\lambda}{96}(\inf f'')\sum_{\substack{n\in N\\ i\in Z}}|u_{i+1}^n - u_i^n|^3$$
$$\le \|u^0\|_{L^\infty}(A_1 + A_2)(\operatorname{oscil}(u_0))^{p-2}2k\sum_{\substack{i\in Z\\ n\in N}}|u_i^n - u_{i-1}^n|^3.$$

Since the first term in the left-hand side of the above inequality is nonnegative, we obtain the uniform estimate

$$K\sum_{\substack{n\in N\\ i\in Z}}|u_{i+1}^n - u_i^n|^3 h \le \sum_{i\in Z}(u_i^0)^2 h \le \|u_0\|_{L^2(R)}^2, \tag{3.15}$$

where the constant K is defined by (3.9) and is assumed to be positive.

Finally, from (3.14) and (3.15), we deduce

$$\hat{\mu}([0,t)\times R) \le (A_1 + A_2)(\operatorname{oscil}(u_0))^{p-2}2k\lim_{h\to 0}\left\{\sum_{\substack{i\in Z\\ n\in N\\ t_n\le t}}|u_{i+1}^n - u_i^n|^3\tau\right\}$$
$$= (A_1 + A_2)(\operatorname{oscil}(u_0))^{p-2}2k\lambda\lim_{h\to 0}\left\{\sum_{\substack{i\in Z\\ n\in N}}|u_{i+1}^n - u_i^n|^3 h\right\}$$
$$\le (A_1 + A_2)(\operatorname{oscil}(u_0))^{p-2}2k\lambda\frac{1}{K}\|u_0\|_{L^2(R)}^2,$$

which completes the proof of (3.10). The proof of (3.12) can be obtained in a similar fashion. □

Remark 3.2. (1) If u_0, u and v are of order $O(\varepsilon)$, then estimates (3.8) or (3.12) give a bound of order $O(\varepsilon^{p+1})$, while (3.10) is of order $O(\varepsilon^p)$.

(2) When K given by (3.9) is positive, the entropy dissipation of the E-scheme is larger than the error to conservation made by using the nonconservative scheme. This explains why a uniform in time estimate can be derived in that case.

(3) It would be quite interesting to obtain an error estimate in terms of the *amplitude of the discontinuities* only, instead of the amplitude of the initial data. Such a result would be more realistic since the nonconservative scheme is expected to converge if the solution has no discontinuity. Unfortunately, as in §2, to prove such a result, we would need to be able to work with L^∞_{loc} instead of L^1.

(4) The assumption of convexity made in this section can probably be relaxed. See Coquel and Le Floch [3], where discrete entropy inequalities, similar to the ones used here, are derived in the case of a general (i.e., not necessarily convex) flux function.

4. Examples of nonconservative schemes

We focus here on two examples of nonconservative schemes, that we call nonconservative upwinding scheme, and nonconservative Lax-Friedrichs scheme,

respectively. We check, in particular, that these schemes satisfy the main assumptions (3.2a) and (3.2b) needed in §3 to get the error estimate.

We assume for simplicity that

$$f' > 0 \quad \text{and} \quad f'' > 0. \tag{4.1}$$

In particular, we may consider the Burgers equation with positive data:

$$f(u) = \frac{u^2}{2} \quad \text{with } u > 0. \tag{4.2}$$

We introduce the nonconservative upwinding scheme

$$u_i^{n+1} = u_i^n - \lambda\, a(u_{i-1}^n,\, u_i^n)\,(u_i^n - u_{i-1}^n), \qquad n \in N,\ i \in Z, \tag{4.3}$$

where the "numerical speed" $a\colon R^2 \to R$ is a given Lipschitz function. The choice

$$a_1(\alpha,\, \beta) = \frac{f(\beta) - f(\alpha)}{\beta - \alpha}, \qquad (\alpha,\, \beta) \in R^2, \tag{4.4}$$

corresponds to the (usual) conservative upwinding scheme. A function a which differs from (4.4) yields a scheme which does not admit a conservative form. In general, we only assume that a is consistent with the flux f, i.e.,

$$a(\alpha,\, \alpha) = f'(\alpha), \qquad \alpha \in R, \tag{4.5}$$

so that

$$a(\alpha,\, \beta) = a_1(\alpha,\, \beta) + O(|\alpha - \beta|). \tag{4.6}$$

We also want to consider the case of functions, say a_2, which coincide with a_1 up to the first-order terms, i.e.,

$$a_2(\alpha,\, \beta) = a_1(\alpha,\, \beta) + O(|\alpha - \beta|^2). \tag{4.7}$$

For definiteness, we construct a class of such functions a_2 as follows. Let $(U,\, F)\colon R \to R^2$ be a convex and increasing entropy-entropy flux pair for equation (1.1), i.e.,

$$U' > 0, \quad U'' > 0 \quad \text{and} \quad F' = f'U'. \tag{4.8}$$

Again, in the case of Burgers's equation (4.2), we can choose $(U,\, F)$ as follows:

$$U(u) = \frac{u^2}{2} \quad \text{and} \quad F(u) = \frac{u^3}{3} \quad \text{with } u > 0. \tag{4.9}$$

Then we set

$$a_2(\alpha,\, \beta) = \frac{F(\beta) - F(\alpha)}{U(\beta) - U(\alpha)}, \qquad (\alpha,\, \beta) \in R^2. \tag{4.10}$$

It is not hard to verify that property (4.7) holds in this case because of (4.8).

Another way to compare the speed a_2 given by (4.10) with the conservative choice (4.4) is provided by the "equivalent equation", which is formally derived from a scheme by performing Taylor expansion.

The equivalent equation for the scheme (4.3) with a satisfying only (4.5) is found to be

$$\partial_t w + \partial_x f(w) = \frac{h}{2}\left\{-f''(w) + 2\frac{\partial a}{\partial \alpha}(w,\, w)\right\}(w_x)^2 + \frac{h}{2}\partial_x\left((1 - \lambda f'(w))f'(w)w_x\right).$$

For the function $a = a_2$, it can be easily verified by using (4.8) that

$$\frac{\partial a}{\partial \alpha}(w, w) = f''(w)/2.$$

Thus, we obtain in the latter case,

$$\partial_t w + \partial_x f(w) = \frac{h}{2}\partial_x \left((1 - \lambda f'(w))\, f'(w)\partial_x w\right). \tag{4.11}$$

Equation (4.11) is independent of the specific entropy pair (U, F) and in particular, coincides with the equivalent equation associated with the conservative upwinding scheme. Roughly speaking, the scheme (4.3) with $a = a_2$ contains the *same* numerical viscosity terms as those of the scheme (4.3) with $a = a_1$. Naively, one may think this property is sufficient to ensure that the nonconservative scheme converges to the correct solution. This is because the viscosity terms are expected to "control" the formation and the evolution of the discontinuities in the solution. We will see in §6 that this need not be true, although there may be exceptions in some very special cases.

We next want to estimate the measure μ associated with the scheme (4.3). By Theorem 2.1, μ is the weak* limit of the sequence $w^h = \{w_i^n\}$ given by

$$w_i^n = \{a(u_{i-1}^n, u_i^n) - a_1(u_{i-1}^n, u_i^n)\}\frac{(u_i^n - u_{i-1}^n)}{h}. \tag{4.12}$$

In view of (4.5), condition (4.6) holds, and thus (4.12) gives

$$|w_i^n| \le O(1)\frac{1}{h}|u_i^n - u_{i-1}^n|^2.$$

This gives an immediate proof of the following lemma.

Lemma 4.1. *Assume that the function a in scheme* (4.3) *satisfies* (4.5). *Then the measure μ is absolutely continuous with respect to the measure of entropy dissipation of the scheme.*

Lemma 4.2. *The measure μ associated with the scheme* (4.3) *under the assumption* (4.5) *does not vanish in general.*

Proof. It is sufficient to give one example, say

$$a(\alpha, \beta) = k(\beta - \alpha) + \frac{f(\alpha) - f(\beta)}{\alpha - \beta}$$

with $k > 0$. Then one has

$$w_i^n = \frac{k}{h}(u_i^n - u_{i-1}^n)^2,$$

thus

$$\sum_{i,n} |w_i^n| h\tau = k \sum_{i,n} |u_i^n - u_{i-1}^n|^2 \tau.$$

The last term is the entropy dissipation of the scheme. In that case, the measure μ and the entropy dissipation measure *coincide*. But the entropy dissipation measure cannot vanish identically if the solution contains discontinuities. □

If $a = a_2$, then using (4.10) and (4.12) gives

$$w_i^n = \left\{\frac{F(U_i^n) - F(U_{i-1}^n)}{U(u_i^n) - U(u_{i-1}^n)} - \frac{f(u_i^n) - f(u_{i-1}^n)}{u_i^n - u_i^n}\right\}\frac{u_i^n - u_{i-1}^n}{h}. \tag{4.13}$$

By means of (4.8), it is tedious but not difficult to verify the following lemma.

Lemma 4.3. *The sequence* $\{w_i^n\}$ *defined by* (4.13) *satisfies*

$$w_i^n = \int_0^1 \int_0^1 G''((1-\sigma)U_{i-1/2}^n(s) + \sigma U(u_{i-1/2}^n(s))) \cdot \big(U(u_{i-1/2}^n(s)) - U_{i-1/2}^n(s)\big)\, d\sigma ds \frac{(u_i^n - u_{i-1}^n)}{h} \tag{4.14}$$

with $U_{i-1/2}^n(s) = (1-s)U(u_{i-1}^n) + sU(u_i^n)$ *and* $u_{i-1/2}^n(s) = (1-s)u_{i-1}^n + su_i^n$, *and* $G\colon R \to R$ *is defined by* $G(\alpha) = F(U^{-1}(\alpha))$.

Formula (4.14) can also be written as

$$w_i^n = -\frac{1}{h}(u_i^n - u_{i-1}^n)^3 \int_{[0,1]^4} G''(g_1(\sigma, s))U''(g_2(\tau, r, s))s(1-s)\tau\, ds d\tau dr d\sigma,$$

where

$$g_1(\sigma, s) = (1-\sigma)U_{i-1/2}^n(s) + \sigma U(u_{i-1/2}^n(s))$$

and

$$g_2(\tau, r, s) = u_i^n + \tau(1 - r + sr)(u_i^n - u_{i-1}^n).$$

The explicit formula (4.14) for the error term w_i^n is especially interesting because it implies the following result:

Lemma 4.4. *Consider the scheme* (4.3) *and* (4.10) *under assumption* (4.8). *Suppose that* $\{u_i^n\}$ *is nonincreasing with respect to* i *for all* n; *then the error term* w_i^n *defined in* (4.13) *satisfies the following bounds*:

$$M_1 \frac{(u_i^n - u_{i-1}^n)^3}{h} \le w_i^n \le M_2 \frac{(u_i^n - u_{i-1}^n)^3}{h} \le 0, \tag{4.15}$$

where M_1 *and* M_2 *are two positive constants depending only on the scheme and the* L^∞ *norm of the initial data.*

Proof. It is not hard to verify that the function G satisfies

$$G''(\alpha) = f''(U^{-1}(\alpha))/U'(U^{-1}(\alpha)) \quad \text{for all } \alpha \in R.$$

This implies that G is a strictly convex function because f is strictly convex and U' is positive. By (4.8), U is also strictly convex, thus the result follows from (4.14). □

By Lemma 4.4, proving that the scheme (4.3) and (4.10) does not converge to the correct solution is equivalent to showing that the cubic term

$$\sum_{i,n} (u_i^n - u_{i-1}^n)^3 \tau \tag{4.16}$$

does not vanish as $h \to 0$. Here the sum is over a compact set of the (t, x)-plane. It seems difficult to establish this fact for the scheme since we do not know any analytical structure of the numerical solution a priori.

Recall that the difference scheme under consideration contains a viscous term of order $O(h)$. Thus, it is natural to study the properties of the following equation

$$\partial_t u^\varepsilon + \partial_x f(u^\varepsilon) = \varepsilon u_{xx}^\varepsilon$$

with smooth data. By a classical argument, one can show that

$$\varepsilon \int_0^T \int_R (\partial_x u^\varepsilon(t, x))^2 \, dxdt \le O(1)\|u(0, \cdot)\|_{L^2},$$

and that the left-hand side of the above inequality cannot vanish if $\lim u^\varepsilon$ contains a discontinuity. Now, let us consider the continuous analogue of (4.16), i.e.,

$$\varepsilon^2 \int_0^T \int_R (\partial_x u^\varepsilon(t, x))^3 \, dxdt. \tag{4.17}$$

We would like to show that the term in (4.17) does not vanish in general. For simplicity, we assume that u^ε is a viscous travelling wave solution, i.e.,

$$u^\varepsilon(t, x) = \bar{u}\left(\frac{x - Vt}{\varepsilon}\right)$$

for some velocity V and for $\bar{u}$ such that

$$-V\bar{u}' + f(\bar{u})' = \bar{u}'',$$

and

$$\lim_{\xi \to -\infty} \bar{u}(\xi) = u_L, \qquad \lim_{\xi \to +\infty} \bar{u}(\xi) = u_R.$$

Here $u_L > u_R$ and $-V(u_R - u_L) + f(u_R) - f(u_L) = 0$. For such u^ε, we have

$$\varepsilon^2 \int_0^T \int_R (\partial_x u^\varepsilon(t, x))^3 \, dxdt = \varepsilon^2 \int_0^T \int_R \left(\frac{1}{\varepsilon}\bar{u}'\right)^3 d\xi dt = T \int_R (\bar{u}')^3 \, d\xi,$$

which does not depend on ε and is in general *not* equal to zero. Thus the term in (4.17) need not vanish as $\varepsilon \to 0$ for viscous solutions of conservation laws. By analogy, we expect that this observation also applies to the discrete scheme. This would imply that the error term (4.16) given by the nonconservative scheme (4.3) and (4.10) does not vanish in general as $h \to 0$.

The above observation leads us to conjecture that a nonconservative scheme, which contains the same viscosity terms as those of a conservative scheme, need not converge to the correct solution of the conservative equation in general. Numerical evidence for this conjecture will be given in §6. This is in contradiction with Karni's conclusion in [18]. We also present numerical evidence in §6 which shows that Karni's correction scheme is not sufficient to correct the error to conservation.

Proposition 4.1. *Assume that the speed a in the nonconservative scheme* (4.3) *satisfies the condition* (4.7); *then the error estimates given in Theorem* 3.3 *apply to this scheme.*

This is easy to prove since the assumption (3.2a)-(3.2b) of §3 (with $p = 2$) is satisfied for this scheme. One example which satisfies the assumption of Proposition 4.1 is given by the case when $a = a_2$ in (4.10).

Finally, we turn to another example of nonconservative schemes. This is a Lax-Friedrichs scheme in nonconservative form and is defined by its incremental coefficients

$$\begin{cases} C(\alpha, \beta) = \dfrac{\lambda}{2}\dfrac{F(\beta) - F(\alpha)}{U(\beta) - U(\alpha)} + \dfrac{1}{4}, \\ D(\alpha, \beta) = -\dfrac{\lambda}{2}\dfrac{F(\beta) - F(\alpha)}{U(\beta) - U(\alpha)} + \dfrac{1}{4}, \end{cases} \qquad (\alpha, \beta) \in R^2, \tag{4.18}$$

where (U, F) is an entropy-entropy flux satisfying conditions (4.8). Then, the same conclusions as before (for (4.3), (4.10)) can be obtained easily for this scheme. Moreover, the equivalent equation of this scheme is

$$\partial_t w + \partial_x f(w) = \frac{h}{2\lambda}\partial_x \left(\left(\frac{1}{2} - \lambda^2 f'(w)^2 \right) \partial_x w \right), \tag{4.19}$$

which indeed is independent of (U, F) and coincides with the equivalent equation of the original Lax-Friedrichs scheme. For the scheme (4.18), one can perform the same error analysis as before.

5. Correction for a scheme in nonconservative form

This section shows that a slight correction of any nonconservative scheme like (2.1a)-(2.1d) ensures its convergence to the entropy weak solution of problem (1.1). The method of correction used here follows ideas from DiPerna [7] as well as Harten and Zwas [16] and Leroux and Quesseveur [25] (also, see a recent paper by Harabetian and Pego [13]). We define a hybrid scheme from (2.1a)-(2.1d) by switching to any given conservative scheme in the neighborhood of discontinuities of the solution. We use here any (conservative and first-order accurate) E-scheme in the region of discontinuities. In fact, we can also use even the Glimm scheme (which is not conservative!) as in [7]. This section is only theoretical and we do not know if the corrected scheme is useful for computational purposes.

We consider approximate solutions $\{u^h\}$ defined by (2.1a)-(2.1d) with (2.1c) replaced by the following $(n \in N, i \in Z)$:

$$u_i^{n+1} = \begin{cases} u_i^n - C_{i-1/2}^n (u_i^n - u_{i-1}^n) + D_{i+1/2}^n (u_{i+1}^n - u_i^n), \\ \qquad\qquad \text{if } |u_i^n - u_{i-1}^n| + |u_{i+1}^n - u_i^n| \le b h^a, \\ u_i^n - \tilde{C}_{i-1/2}^n (u_i^n - u_{i-1}^n) + \tilde{D}_{i+1/2}^n (u_{i+1}^n - u_i^n) \quad \text{otherwise.} \end{cases} \tag{5.1}$$

In (5.1), C and D are the incremental coefficients of any (possibly high-order accurate) consistent scheme in nonconservative form; $\tilde{C}$ and $\tilde{D}$ are taken to be the coefficients of any E-scheme. The constants $b > 0$ and $a \in (0, 1)$ are fixed and control the switching between the two schemes under consideration. Heuristically, in the regions where the solution is smooth, one has $|u_i^n - u_{i-1}^n| = O(h) \le O(h^a)$, so that (2.1c) is always used there. The E-scheme is used only in the neighborhoods of the points of discontinuity of the solution.

The following theorem proves that the slight correction of (2.1c) given by (5.1) is sufficient to ensure convergence of the nonconservative scheme to the correct solution.

Theorem 5.1. *Let* $\{u^h\}$ *be the solution constructed by the hybrid scheme in nonconservative form* (5.1). *Suppose that* (2.2)-(2.3b) *and the CFL condition* $\lambda \sup |f'| \le \frac{1}{2}$ *are satisfied. Moreover, we assume that the constant* a *in* (5.1) *satisfies*

$$a \in (0, 1). \tag{5.2}$$

Then $\{u^h\}$ *converges in* L^1_{loc} *norm to the entropy weak solution* u *of problem* (1.1).

Proof. It is not difficult to see that the results of §§2 and 3 still apply to (5.1). In view of Theorem 2.1 and Theorem 3.1, the main difficulty is to prove that the error to conservation (μ in (2.6) and $\hat{\mu}$ in (3.4)) vanish. Let

$$\Omega_T = \{(i, n) \in Z \times N : \ |u_i^n - u_{i-1}^n| + |u_{i+1}^n - u_i^n| \le b\, h^a \text{ and } n\tau \le T\}$$

and consider the term

$$E^h = \sum_{(i,n)\in\Omega_T} \left\{ |C_{i-1/2}^n - \tilde{C}_{i-1/2}^n|\, |u_i^n - u_{i-1}^n| + |D_{i+1/2}^n - \tilde{D}_{i+1/2}^n|\, |u_{i+1}^n - u_i^n| \right\} \tau,$$

which clearly bounds the error to conservation. We must verify that E^h tends to zero with h. It is then easy to deduce from this result that μ and $\hat{\mu}$ in (2.12) and (3.4) vanish identically and to conclude that $\lim u^h = v = u$. We omit these details.

It remains to prove $\lim E^h = 0$. One has, from (3.2) and the definition of Ω_T,

$$E^h \le O(1)h^a \sum_{n\tau\le T} \mathrm{TV}(u^h(t_n))\, \tau = O(1)h^a \tau(\mathrm{TV}(u_0)) \to 0,$$

because of (5.2), which completes the proof. □

In practice, it would be useful to know how to evaluate the numerical values for the constants a and b in (5.1) (no clear strategy seems to exist for that). This makes the use of (5.1) difficult for practical computations. A smoother switch than (5.1) could be considered. We also refer to Harten and Zwas [16] and Harabetian and Pego [13] for related matters and to Engquist and Sjogreen [9] for another procedure of correcting the error to conservation in a finite difference scheme.

6. Numerical experiments

This section is intended to illustrate our results of §§2 to 5, as well as to provide numerical evidence for the location of the support of the measure μ.

We consider five numerical examples in this section. The first example is the upwinding scheme in nonconservative form for the Burgers equation (cf. (4.2)-(4.3) and (4.9)-(4.10)). We choose the following Riemann shock initial data

$$u_0(x) = \begin{cases} 1.5 & \text{if } x \le 0, \\ 0.5 & \text{if } x > 0. \end{cases} \tag{6.1}$$

In Figure 1a (next page), we plot the solution at $t = 1$ with space increments $h = 0.005$, 0.0025, 0.00125, and 0.000625, respectively. In all these calculations, the time increment is $\tau = 0.2h$. For this initial data, the correct shock speed is equal to 1. It is quite clear from Figure 1a that the numerical solution converges strongly to an incorrect shock position, although the error is not so large.

The numerical shock position can be computed by evaluating the source term on the right side of (2.6). To this end, we just need to sum the right-hand side of (4.13) near the curve of discontinuity in space and time. Our calculations show that as h decreases to zero, the numerical shock position converges to $x = 1.00426$. So the relative error of the shock position is about 0.43%.

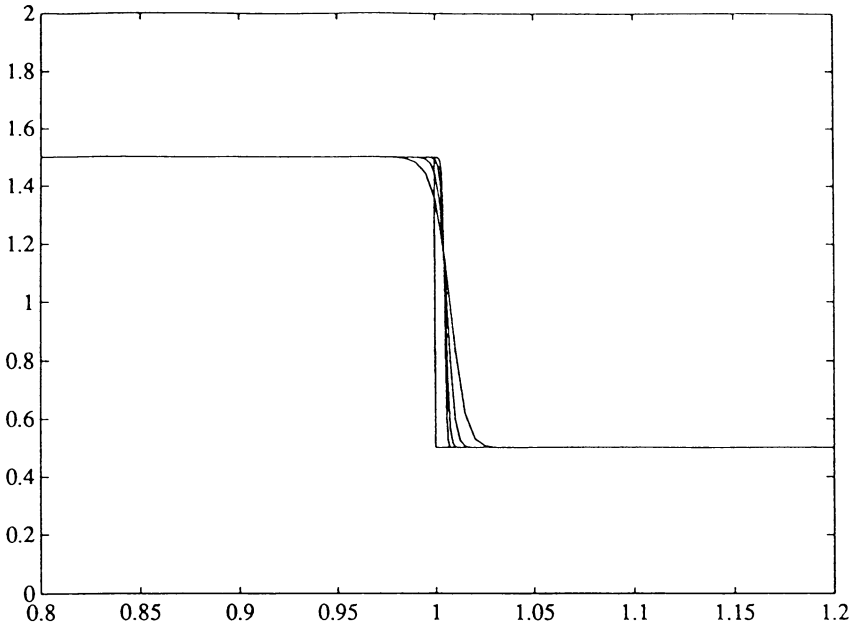


FIGURE 1a. Upwind scheme in nonconservative form, $t = 1$, $dt/dx = 0.2$

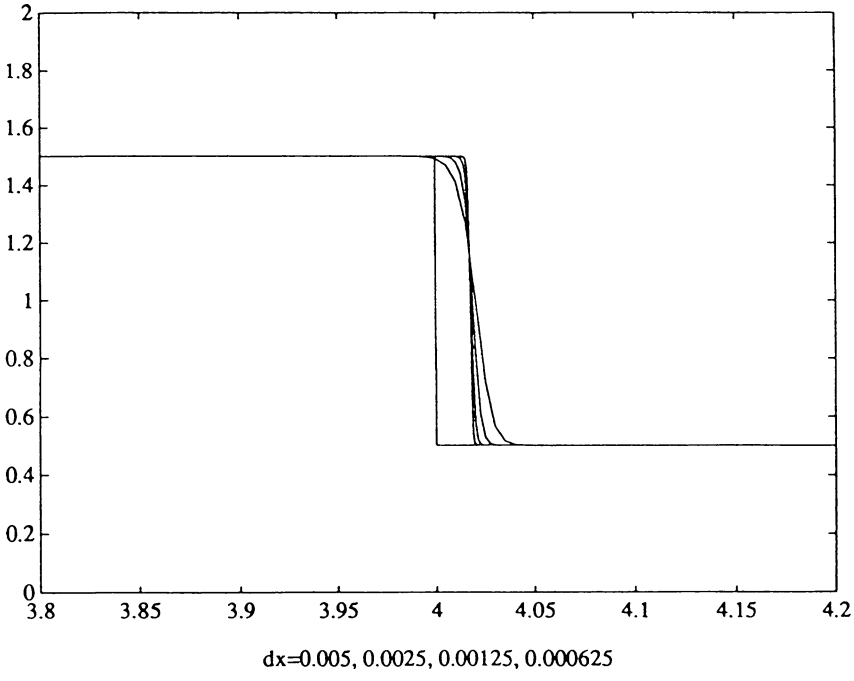


FIGURE 1b. Upwind scheme in nonconservative form, $t = 4$, $dt/dx = 0.2$

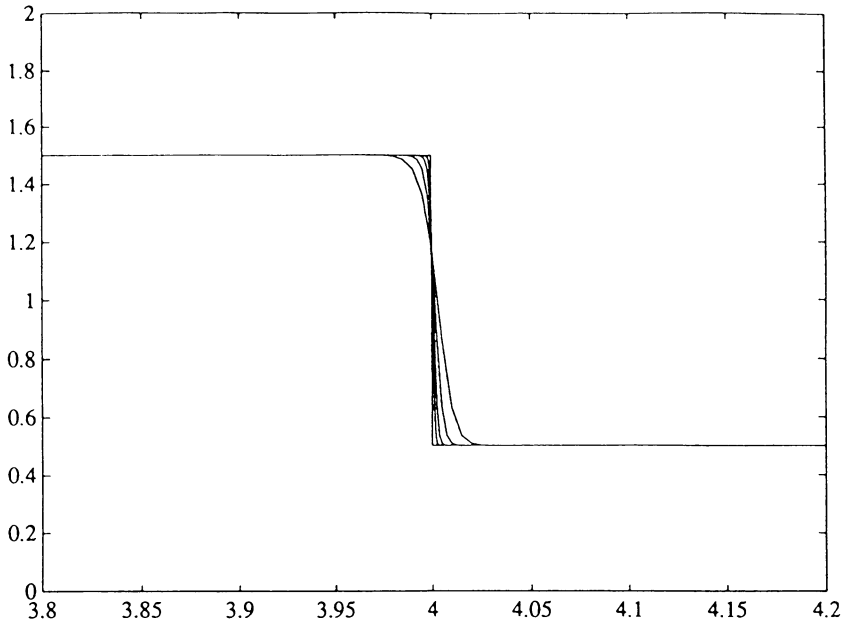


FIGURE 1c. Upwind scheme in conservative form, $t = 4$, $dt/dx = 0.2$

In Figure 1b, we plot the same calculations at time $t = 4$. As we can see, the error grows bigger in time. It is almost linear in time. If we measure the error in the shock position, we find that it is about 1.7% as $h \to 0$. In order to illustrate that this error in the numerical shock position is not due to truncation error, we plot in Figure 1c the same calculations at $t = 4$ using the upwinding scheme in conservative form. In this case, it is clear that the numerical solution converges to the correct shock position as the mesh sizes tend to zero. And the error in the nonconservative scheme is much larger than the truncation error made by the conservative scheme. This confirms the conjecture of §4, as well as the results of §§2 and 3.

Next we would like to demonstrate that the error due to nonconservation cannot be corrected by using a high-order scheme. To this end, we compare two nonconservative approximations of the convection term uu_x with first-order and second-order accuracy, respectively. The first-order method is an upwinding scheme in nonconservative form, i.e., (4.3) with the choice of numerical speed given by $a(\alpha, \beta) = f'(\beta)$. The second-order scheme is a nonconservative second-order ENO (Essentially Nonoscillatory) scheme, cf. [15], which for Burgers's equation reads

$$\frac{(u_i^{n+1} - u_i^n)}{\tau} = -u_i^n D_- \left(u_i^n + \frac{h}{2} \operatorname{minmod}[D_- u_i^n, D_+ u_i^n] \right) \quad \text{if } u_i^n > 0,$$

$$\frac{(u_i^{n+1} - u_i^n)}{\tau} = -u_i^n D_+ \left(u_i^n + \frac{h}{2} \operatorname{minmod}[D_- u_i^n, D_+ u_i^n] \right) \quad \text{if } u_i^n \le 0,$$

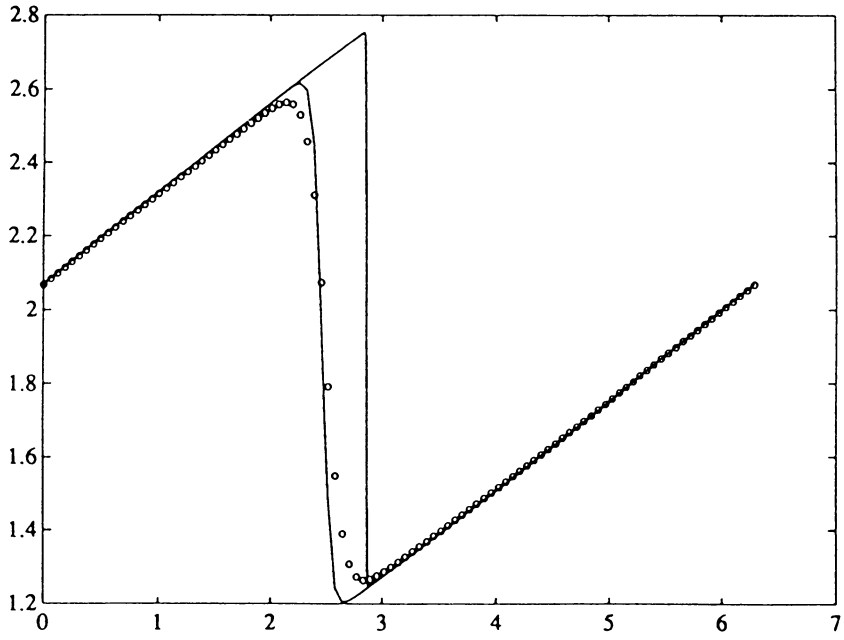


FIGURE 2a. Nonconserv. 2nd-order ENO vs. nonconserv. upwinding, $t = 3$

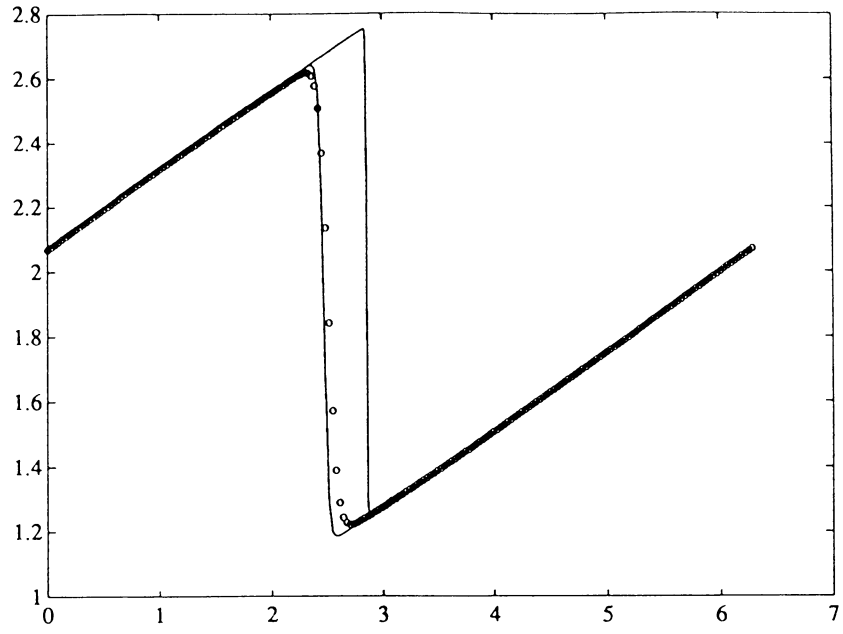
where D_- and D_+ are the standard backward and forward divided difference operators, respectively. The minmod function is defined by

$$\text{minmod}(a, b) = \begin{cases} 0 & \text{if} \quad ab < 0, \\ \text{sgn}(a)\min[|a|, |b|] & \text{if} \quad ab \geq 0. \end{cases}$$

In Figures 2a and 2b, we compare the numerical solution of the nonconservative upwinding scheme with that of the nonconservative second-order ENO scheme, using $m = 100$ and $m = 200$, respectively. The initial data is given by

$$u_0(x) = 2 + \sin x. \tag{6.2}$$

The numerical solutions are plotted at time $t = 3$. The correct solution is also plotted using the conservative second-order van Leer scheme with $m = 2000$ as a reference for the "true" solution. In these pictures, the line with circles corresponds to the first-order nonconservative upwinding scheme, while the solid line next to the line with circles corresponds to the nonconservative second-order ENO scheme. The solid line on the right side corresponds to the solution of the van Leer scheme. As we can see, the second-order nonconservative scheme only sharpens the numerical shock profile. But it still converges to the same wrong shock position as the first-order nonconservative scheme.



m=200, compared with van Leer scheme using m=2000

FIGURE 2b. Nonconserv. 2nd-order ENO vs. nonconserv. upwinding, $t = 3$

There are some proposals for correcting the error due to nonconservative form of the scheme in a formal way. Among others is a correction scheme proposed by Karni [18]. Karni's idea is to correct the numerical viscous form in a nonconservative scheme so that the corrected scheme has the same formal viscous form as that of a consistent and conservative scheme. For example, let us consider the Lax-Friedrichs scheme for the scalar conservation law

$$u_t + f(u)_x = 0. \tag{6.3}$$

It has the form

$$u_i^{n+1} = \frac{1}{2}(u_{i-1}^n + u_{i+1}^n) - \frac{\lambda}{2}(f_{i+1}^n - f_{i-1}^n), \tag{6.4}$$

where $f_i^n = f(u_i^n)$ and $\lambda = \tau/h$. The numerical viscous form (or equivalent equation) for this scheme reads

$$u_t + f(u)_x = \frac{\tau}{2}(u_{xx}/\lambda^2 - u_{tt}). \tag{6.5}$$

Now let $w = w(u)$ be a new dependent variable and let $T = dw/du$. The equation for w is given formally by

$$w_t + A(w)w_x = 0, \tag{6.6}$$

where $A(w) = f'(u(w))$. Karni realized that if one uses a "Lax-Friedrichs type" scheme

$$w_i^{n+1} = \frac{1}{2}(w_{i-1}^n + w_{i+1}^n) - \frac{\lambda}{4}(A_{i+1}^n + A_{i-1}^n)(w_{i+1}^n - w_{i-1}^n), \tag{6.7}$$

to approximate (6.6), it would not give seemingly a consistent weak solution for u. This is because the viscous form for (6.7),

$$w_t + A(w)_x = \frac{\tau}{2}(w_{xx}/\lambda^2 - w_{tt}), \tag{6.8}$$

is not consistent with the viscous form (6.5) for (6.3). To correct this inconsistency, Karni proposed to apply a "Lax-Friedrichs type" scheme to the following modified equation:

$$w_t + A(w)w_x = \frac{\tau}{2}D \tag{6.9}$$

with

$$D = \frac{dw}{du}(u_{xx}/\lambda^2 - u_{tt}) - (w_{xx}/\lambda^2 - w_{tt}). \tag{6.10}$$

The motivation for using (6.9)-(6.10) is that the viscous form of the Lax-Friedrichs scheme for this modified equation is *formally* consistent with that of (6.4).

We remark that the correction on the formal level for the viscous form is not enough to ensure that the corrected scheme would converge to the correct weak solution (cf. §3). In the following, we will test Karni's idea numerically for Burgers's equation with $w = u^2/2$. Again, we use the shock Riemann initial data (6.1). In Figure 3, we plot a sequence of numerical solutions obtained by applying the "Lax-Friedrichs type " scheme to equations (6.9)-(6.10). The correction term D is discretized by a centered difference approximation as suggested by Karni. The three curves on the left correspond to the numerical solutions of Karni's scheme using $h = 0.005, 0.0025$, and 0.00125, respectively. The correct shock is at $x = 1$. As we can see, Karni's scheme does not converge to the correct weak solution. The curve with circles on the right corresponds to the approximation without any correction, i.e., D is set to be zero in (6.9). Obviously, it gives the wrong solution. It is interesting to note that the corrected scheme produces an error of the same order as that of the uncorrected scheme. What is even more interesting is that if we replace D in (6.9) by $D/2$, it seems to give the correct solution; see the curve in the middle marked with '+' signs. We do not have a good explanation for this behavior.

Next we consider the modified Lax-Friedrichs scheme for Burgers's equation. It has the form

$$u_i^{n+1} = \frac{1}{4}(u_{i-1}^n + 2u_i^n + u_{i+1}^n) - \frac{\lambda}{2}(f(u_{i+1}^n) - f(u_{i-1}^n)). \tag{6.11}$$

There are two distinct features for this version of the Lax-Friedrichs scheme.

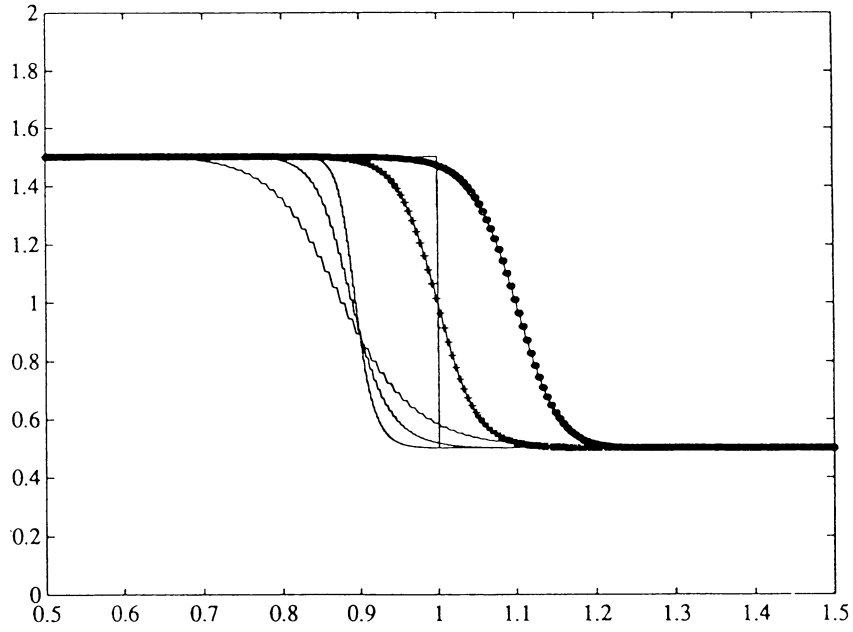


FIGURE 3. Lax-Friedrichs scheme in nonconservative form, $t = 1$

First it satisfies a stronger version of entropy inequality as stated in the proof of Lemma 3.2. It is an E-scheme. Secondly, its viscous form reads

$$u_t + f(u)_x = \frac{\tau}{2}(u_{xx}/(2\lambda^2) - u_{tt}). \tag{6.12}$$

Thus, according to Theorem 3.3, we expect that the error due to nonconservation is bounded uniformly in time. But what we found numerically in this case is even better than what is stated in the theorem.

We now apply scheme (4.18) to Burgers's equation with the quadratic entropy $U(u) = u^2$. For the shock Riemann initial data (6.1), we found that the error in the shock position is only about 0.06% for time $t = 1$ and about 0.2% at time $t = 4$ as the mesh sizes tend to zero. In Figure 4a (next page), we plot the solutions at time $t = 1$ using $h = 0.005, 0.0025, 0.00125$, and 0.000625, respectively. The errors seem to be too small to tell from the "eye norm". But the smallness of the error for short times is deceptive because the error grows almost linearly in time. By the time $t = 16$, the error is of the order 0.8%.

We also computed with the modified Lax-Friedrichs scheme (6.11), using the variable $w = u^2/2$ as in (6.6) with Karni's correction. Recall there seems to be a mysterious factor of 2 missing in the Lax-Friedrichs type scheme with Karni's correction. Now, with the modified Lax-Friedrichs scheme, we do have a factor of 2 in the viscous form (6.10). This seems to make a big difference. In Figure 4b, we approximate the Burger's equation by the modified Lax-Friedrichs scheme using the variable $w = u^2/2$ with Karni's correction (6.9)-(6.10). The $O(1)$ error in the shock position seems to be substantially reduced in this case.

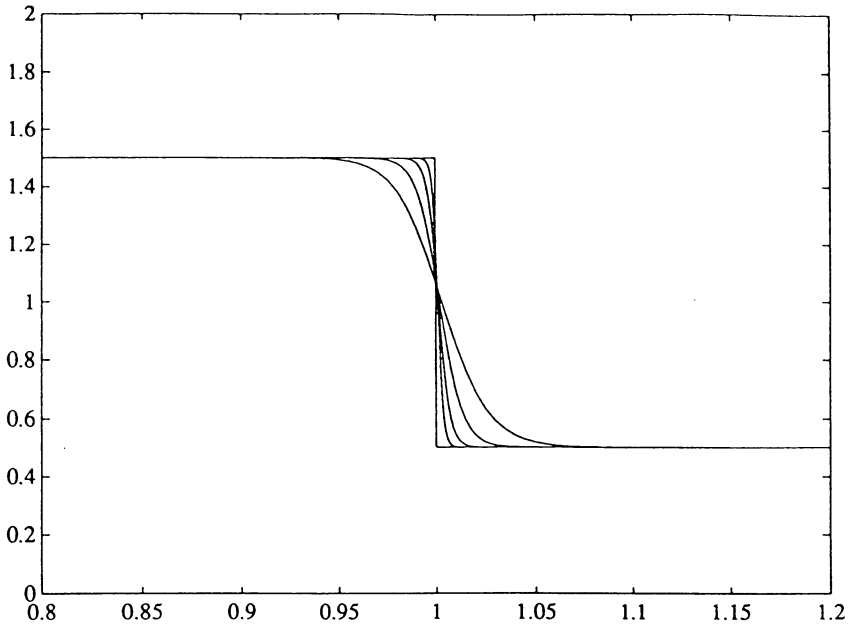


FIGURE 4a. Modified Lax-Friedrichs scheme in nonconservative form, $t = 1$

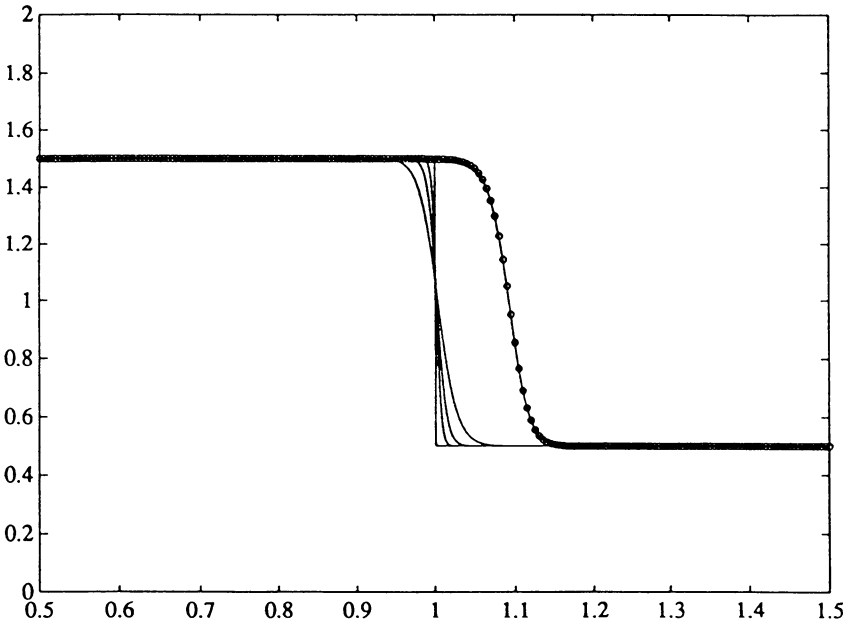
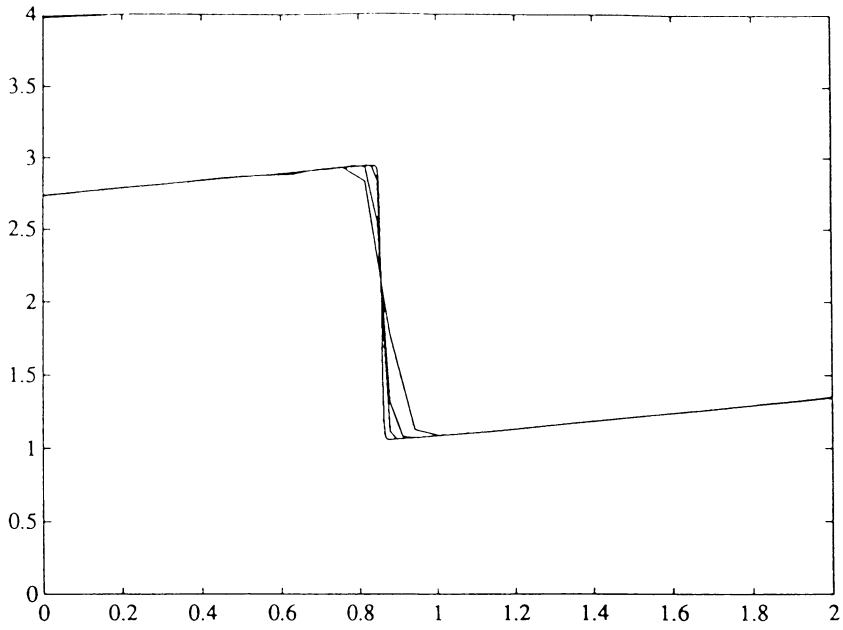


FIGURE 4b. Modified Lax-Friedrichs scheme in nonconservative form, $t = 1$

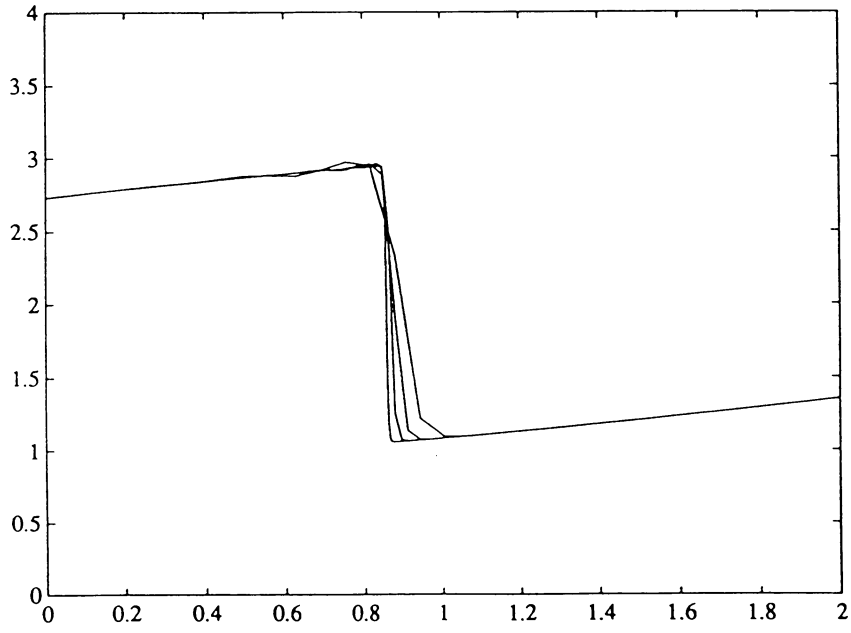


$a = 0.8$, $b = 0.5$, compared with van Leer scheme using $m = 2000$

FIGURE 5a. Hybrid Lax-Wendroff scheme with van Leer scheme, $t = 2$, $m = 100, 200, 400$

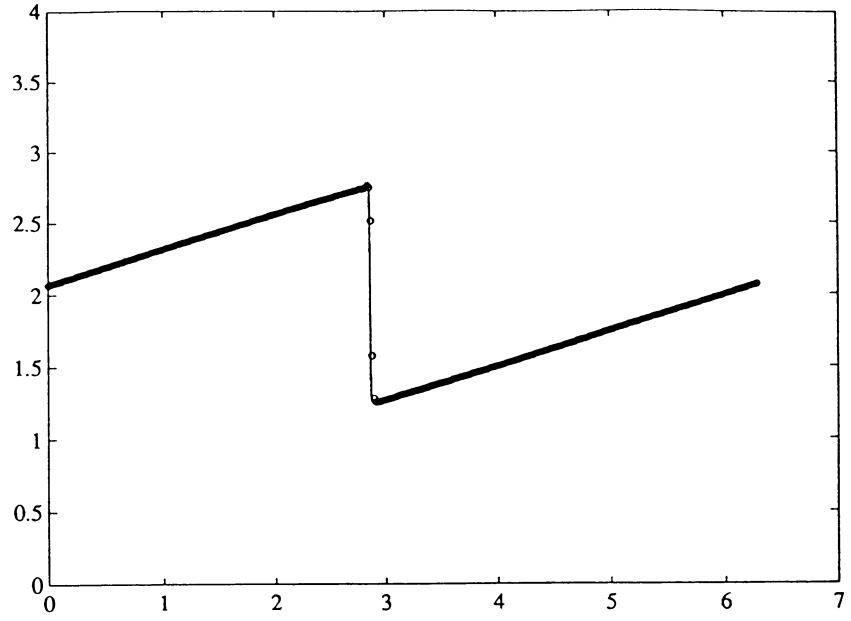
Again, the smallness of the error is only true for short times. The error will grow with time. (In a revised version of [18], Karni agreed with our conclusions here.)

Lastly, we test our correction scheme proposed in §5. We choose two approaches in forming the hybrid scheme. The first choice is to use the Lax-Wendroff scheme in the smooth regions and van Leer's scheme in the region containing a discontinuity. The switching strategy is given by (5.1) in §5. We remark that a similar strategy has been used by Harten and Zwas [16] and by Harabetian and Pego [13]. Note also that Lax-Wendroff *does not* satisfy the TVD assumption of Theorem 5.1. We could use a TVD scheme here as well, but this would not change the conclusions below. In Figures 5a and 5b, we plot the solution at time $t = 2$ with initial data given by (6.2). In Figure 5a, we choose the switching parameters a and b in (5.1) to be $a = 0.8$ and $b = 0.5$. We plot the solutions using $m = 100, 200$, and 400, respectively, and compare them with the solution obtained by using the second-order conservative van Leer scheme with $m = 2000$. We can see clearly that, as the mesh sizes tend to zero, the numerical solution converges to the correct weak solution. In Figure 5b (next page), we perform the same calculations but with a different choice of switching parameters $a = 0.6$, $b = 1$. The decreasing of a corresponds to fewer points in the inner region containing the shock discontinuity. We could still see that the numerical solution converges to the physical solution as we increase the numerical resolution, although the coarse-grid calculation is a bit rough. In Figure 5c (next page), we compute the solution at time $t = 3$, now with a slightly bigger value for a, $a = 0.8$ and the same value for b, $b = 1$, but less grid points, $m = 400$. The numerical



$a = 0.6$, $b = 1$, compared with van Leer scheme using $m = 2000$

FIGURE 5b. Hybrid Lax-Wendroff scheme with van Leer scheme, $t = 2$, $m = 100, 200, 400$



$a = 0.8$, $b = 1$, compared with van Leer scheme using $m = 2000$

FIGURE 5c. Hybrid Lax-Wendroff scheme with van Leer scheme, $t = 3$, $m = 400$

solution converges beautifully to the physical solution. This confirms our theory in §5. It also demonstrates that the choices of the switching parameters a and b are not very sensitive in the calculations.

Our second choice of hybridization is between a Leap-Frog scheme and the Glimm scheme [10]. Naturally, we use the Glimm scheme near the region of shock discontinuity. In the calculations, we choose the initial data as $u_0(x) = \sin x$. The solution is computed at time $t = 0.4$. In Figure 6a, we plot the numerical solution using 400 grid points. The inner region has the width $\delta = 0.1$ in this case. There are about 40 grid points in the inner region. The shock position is clearly very well approximated. In Figure 6b, we reduce the width of the inner region to $\delta = 0.025$. There are about 10 grid points in the inner region. The shock position is still well captured. Moreover, because the outer region becomes larger, the smooth part of the solution is better approximated than in the case of Figure 6a.

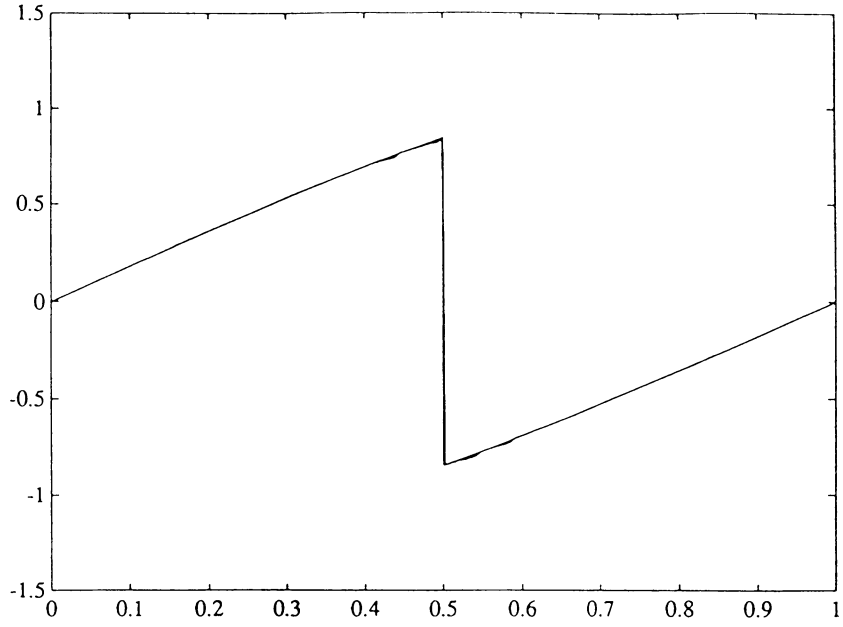


FIGURE 6a. Hybrid centered difference scheme with Glimm scheme, $t = 0.4$

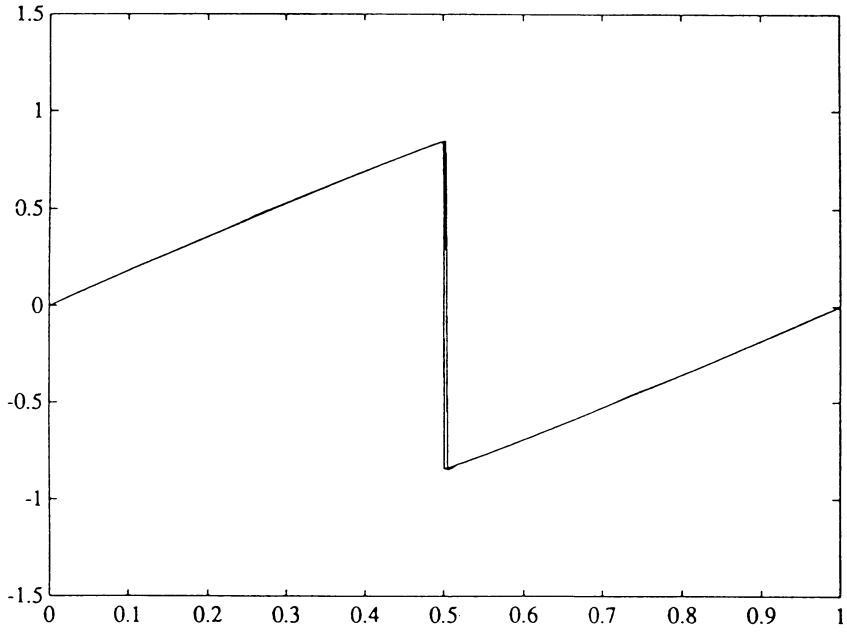


FIGURE 6b. Hybrid centered difference scheme with Glimm scheme, $t = 0.4$

ACKNOWLEDGMENTS

The authors want to thank Peter D. Lax and Robert Kohn for their support of, and interest in, this work. The work of the first author has been supported in part by an NSF grant DMS-90-03202, an Air Force Grant AFOSR-90-0090, and a Sloan research fellowship. This work was done while the second author was a Courant Instructor at the Courant Institute of Mathematical Sciences, and on leave from a research position at the Ecole Polytechnique, and supported in part by an NSF grant DMS-88-06731, and the Centre National de la Recherche Scientifique.

BIBLIOGRAPHY

1. J. F. Colombeau and A. Y. Leroux, *Numerical methods for hyperbolic systems in nonconservative form using products of distributions*, Advances in Computer Methods for P.D.E., vol. 6 (R. Vichnevetsky and R. S. Stepleman, eds.), Inst. Math. Appl. Conf. Series, Oxford Univ. Press, New York, 1987, pp. 28–37.
2. F. Coquel and Ph. Le Floch, *Convergence of finite difference schemes for conservation laws in several space dimensions: the corrected antidiffusive flux approach*, Math. Comp. **57** (1991), 169–210.
3. ———, *On the finite volume method for multidimensional conservation laws*, preprint, December 1991, Courant Institute, New York University (unpublished).
4. M. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp. **34** (1980), 1–21.

5. C. M. Dafermos, *Characteristics in hyperbolic conservation laws. A study of the structure and the asymptotic behavior of solutions* (R. J. Knops, ed.), Heriot-Watt University, Nonlinear Analysis and Mechanics: Heriot-Watt Symposium Volume I, 1981, pp. 1–58.
6. G. Dal Maso, Ph. Le Floch, and F. Murat, *Definition and weak stability of nonconservative products*, Preprint CMAP, Ecole Polytechnique, Palaiseau (France).
7. R. DiPerna, *Finite difference scheme for conservation laws*, Comm. Pure Appl. Math. **25** (1982), 379–450.
8. R. J. DiPerna and A. Majda, *The validity of nonlinear geometric optics for weak solutions of conservation laws*, Comm. Math. Phys. **98** (1985), 313–347.
9. B. Engquist and B. Sjogreen, *Numerical approximation of hyperbolic conservation laws with stiff terms*, UCLA Computational and Applied Mathematics Report 89-07, 1989.
10. J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of conservation laws*, Comm. Pure Appl. Math. **18** (1965), 695–715.
11. J. Glimm and P. D. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc. No. 101, Amer. Math. Soc., Providence, RI, 1970.
12. J. Goodman and P. D. Lax, *On dispersive difference schemes*, Comm. Pure Appl. Math. **41** (1988), 591–613.
13. E. Harabetian and R. Pego, *Efficient hybrid shock capturing scheme*, IMA Preprint Series No. 743, Minneapolis, MN, 1990.
14. A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys. **49** (1983), 357–393.
15. A. Harten and S. Osher, *Uniformly high order accurate non-oscillatory schemes* I, SIAM J. Numer. Anal. **24** (1987), 279–309.
16. A. Harten and G. Zwas, *Self-adjusting fluid schemes for shock computations*, J. Comput. Phys. **9** (1972), 568–583.
17. T. Y. Hou and P. D. Lax, *Dispersive approximations in fluid dynamics*, Comm. Pure Appl. Math. **44** (1991), 1–40.
18. S. Karni, *Viscous shock profiles and primitive formulations*, SIAM J. Numer. Anal. **29** (1992), 1592–1609.
19. S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb. **10** (1970), 217–243.
20. N. N. Kuznetsov, *Accuracy of some approximate method for computing the weak solutions of a first order quasilinear equation*, USSR Comput. Math. and Math. Phys. **16** (1976), 105–119.
21. P. D. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, CBMS Monographs, vol. 11, SIAM, Philadelphia, PA, 1973.
22. P. D. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math. **13** (1960), 217–237.
23. Ph. Le Floch and J.-G. Liu, *Entropy and monotonicity (EMO) consistent schemes for conservation laws*, preprint, 1993.
24. Ph. Le Floch and T.-P. Liu, *Existence theory for nonlinear hyperbolic systems in nonconservative form*, Forum Math. **5** (1993), 261–280.
25. A. Y. Leroux and P. Quesseveur, *Convergence of an antidiffusive Lagrange-Euler scheme for quasilinear equations*, SIAM J. Numer. Anal. **21** (1984), 985–994.
26. G. Moretti, *The λ-scheme*, Comput. & Fluids **7** (1979), 191–205.
27. E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal. **28** (1991), 891–906.
28. J. A. Trangenstein, *A second-order algorithm for the dynamic response of soils*, Impact Comput. Sci. Engrg. **2** (1990), 1–39.
29. J. A. Trangenstein and P. Colella, *A higher-order Godunov method for modeling finite deformations in elastic-plastic solids*, Comm. Pure Appl. Math. **44** (1991), 41–100.

30. A. I. Volpert, *The space BV and quasilinear equations*, Math. USSR Sb. **2** (1967), 225–267.
31. G. Zwas and J. Roseman, *The effect of nonlinear transformations on the computation of weak solutions*, J. Comput. Phys. **12** (1973), 179–186.

California Institute of Technology, Applied Mathematics, 217-50, Pasadena, CA 91125
E-mail address: hou@ama.caltech.edu.

Department of Mathematics, University of Southern California, Los Angeles, CA 90089
E-mail address: lefloch@math.usc.edu.