Second-Order Conservative Schemes and the Entropy Condition

By Maria E. Schonbek

Abstract. We consider numerical approximations to solutions of systems of hyperbolic conservation laws of the form $\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$, $u \in \mathbb{R}^n$ and $f: \mathbb{R}^n \to \mathbb{R}^n$ smooth. We show that conservative three-point second-order accurate methods cannot satisfy a local entropy inequality.

1. Introduction. We consider numerical approximations to solutions of systems of hyperbolic conservation laws of the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0,$$ (1.1)

where $u \in \mathbb{R}^n$ and $f: \mathbb{R}^n \to \mathbb{R}^n$ is smooth. It is well-known that, even with smooth initial data, globally defined classical solutions do not exist, due to the development of shock waves. In the enlarged class of discontinuous solutions, uniqueness is lost in general, and one is faced with the problem of choosing the unique physically relevant solution. The traditional criterion for admissibility for solutions is based on entropy dissipation and requires that

$$\frac{\partial}{\partial t} \eta(u) + \frac{\partial}{\partial x} q(u) \leq 0,$$ (1.2)

where $\eta, q$ is a generalized convex entropy pair in the sense of Lax [2]. In order to establish convergence of numerical methods of classical finite-difference schemes it is necessary to satisfy the entropy condition above. One obvious attack is to derive a discrete entropy inequality of the form

$$D_t \eta + D_x q \leq 0.$$ (1.3)

This can be achieved for a class of first-order methods and provides a convenient way of verifying the entropy condition (1.2). It was shown by Lax [2] that if the system of conservation laws (1.1) admits a strictly convex entropy, then all weak solutions of (1.1) which are limits of the Lax-Friedrichs scheme will satisfy the discrete entropy inequality (1.3), provided the Courant-Friedrichs-Lewy constant is small enough. It is natural to ask if this result can be extended to second-order accurate methods. In this paper we provide a negative answer to this question.
Specifically, for conservative three-point second-order accurate methods we show that an inequality of the form
\[ \frac{\eta(u) - \eta(z)}{\Delta t} + \frac{Q(w, z) - Q(z, v)}{\Delta x} \leq 0 \]
cannot hold for any convex entropy \( \eta \). Here \( u = u(x, t + \Delta t) \), \( v = u(x - \Delta x, t) \), \( w = u(x + \Delta x, t) \), \( z = u(x, t) \) and the numerical flux \( Q \) satisfies the standard consistency condition
\[ Q(u, u) = q(u). \]

Thus the analysis of the entropy production for second-order accurate methods is not a purely local problem. In [4], Majda and Osher construct a simple modification of the Lax-Wendroff scheme which retains the features of conservation form, three-point scheme, and second-order accuracy. The scheme is obtained by adding a nonlinear viscosity term to the classical Lax-Wendroff operator. They show that if the solutions to this scheme converge boundedly a.e. to a function \( u \), then there is a weak solution of (1.1) which satisfies the entropy condition (1.2) with \( \eta(u) = \frac{1}{2} u^2 \) and \( q(u) = \int_0^u s f'(s) \, ds \). Finally, we would like to mention that basic \( L^2 \) stability estimates for conservative second-order accurate schemes have been established by Engquist and Osher [1] and by Majda and Osher [4]. These estimates provide a starting point on the attack of the convergence problem using the theory of compensated compactness [5], [6], [7].

2. The Entropy Inequality for Second-Order Schemes. We consider numerical approximations to weak solutions of strictly hyperbolic systems of conservation laws of the form
\[ u_t + f(u)_x = 0. \]
Here \( u \in \mathbb{R}^n \) and \( f: \mathbb{R}^n \to \mathbb{R}^n \) is a smooth nonlinear mapping. The condition of strict hyperbolicity requires the Jacobian \( Df = \partial f / \partial u \) to have \( n \) linearly independent eigenvectors. The conservation laws under consideration admit a strictly convex entropy function \( \eta(u) \). We recall that a pair of functions \( \eta(u) \), \( q(u) \) is called a generalized entropy, entropy flux pair for (2.1) if
\[ \eta(u)_t + q(u)_x = 0 \]
holds identically for any smooth vector field \( u(x, t) \) which satisfies (2.1). We note that (2.2) is equivalent to the compatibility condition
\[ \nabla \eta(u) \nabla f(u) = \nabla q(u). \]
This condition follows carrying out the differentiation in (2.3) and multiplying Eq. (2.1) by \( \nabla \eta \), i.e.,
\[ \nabla \eta u_t + \nabla \eta \nabla f u_x = 0, \quad \nabla \eta u_t + \nabla q u_x = 0. \]

As is well-known, the initial value problem (2.1) does not have, in general, global classical solutions and one has to look for weak solutions (bounded measurable functions which satisfy (2.1) in the sense of distributions). In this setting, uniqueness is lost. In the class of genuine nonlinear hyperbolic conservation laws which admit a strictly convex entropy, the physical relevant solution is selected by the following criterion.
Entropy admissibility criterion: Let \( \eta(u) \) be a strictly convex entropy for (2.1) with entropy flux \( q(u) \). A weak solution \( u(x, t) \) of (2.1) is called admissible if it satisfies
\[
(2.4) \quad \eta(u)_t + q(u)_x \leq 0
\]
in the sense of distributions.

We will show that a discrete version of (2.4) for conservative three-point second-order accurate schemes cannot be achieved for any strictly convex entropy \( \eta \). Specifically, we consider explicit three-point conservation schemes [3]:
\[
\frac{\Delta u}{\Delta t} = -\frac{\Delta g}{\Delta x}.
\]
Here \( \Delta u \) is the forward time difference
\[
\Delta u = u(x, t + \Delta t) - u(x, t),
\]
g is a vector-valued function of two vector arguments which reduces to \( f \) when its arguments are equal
\[
(2.6) \quad g(u, v) = f(u)
\]
and \( \Delta g \) is the symmetric space difference
\[
\Delta g = g(u(x + \Delta x), u(x, t)) - g(u(x, t), u(x - \Delta x, t)).
\]

We recall that condition (2.6) ensures the consistency of the scheme (2.5) with the system (2.1) in the following sense. Let \( v(x, t) \) be a solution of the difference scheme (2.5) (define for the sake of convenience \( v(x, t) = v(x, t') \), when \( t \) is a noninteger multiple of \( \Delta t \), where \( t' = \Delta t[t/\Delta t] \). If \( v(x, t) \) converges boundedly a.e. to some function \( u(x, t) \) as \( \Delta x \) and \( \Delta t \) tend to zero then \( u(x, t) \) is a weak solution of (2.1) [3].

We recall that in [2], Lax shows under the appropriate restrictions for the Courant-Friedrichs-Lewy number, that the Lax-Friedrichs scheme for systems satisfies the entropy inequality (2.4). The method used is to show that the discrete entropy inequality
\[
D_t \eta + D_x q \leq 0
\]
holds. Here \( D_t \) and \( D_x \) are the difference operators corresponding to the Lax-Friedrichs scheme, i.e.,
\[
D_t = \frac{1}{\Delta t} \left( T(\Delta t) - \frac{S(\Delta x) + S(-\Delta x)}{2} \right), \quad D_x = \frac{S(\Delta x) - S(-\Delta x)}{2\Delta x},
\]
where \( T(h) \) is translation in \( t \) by the amount \( h \) and \( S(k) \) is translation in \( x \) by the amount \( k \). We will show that the method used for the L-F scheme will not extend to three-point conservative second-order accurate schemes. Specifically, we show that a discrete entropy inequality of the form
\[
(2.7) \quad \frac{\eta(u) - \eta(z)}{\Delta t} + \frac{Q(w, z) - Q(z, v)}{\Delta x} \leq 0
\]
cannot hold for any strictly convex entropy \( \eta \). Here, the numerical flux is supposed to satisfy the standard consistency condition
\[
Q(u, u) = q(u)
\]
and \( u = u(x, t + \Delta t), v = u(x - \Delta x, t), w = u(x + \Delta x, t), z = u(x, t) \).
System (2.1) was assumed hyperbolic, that is, \( \nabla f \) has \( n \) linearly independent eigenvectors \( r_j(u) \) satisfying
\[
(2.8) \quad \nabla f(u)r_j(u) = \lambda_j(u)r_j(u),
\]
where \( \lambda_j(u) \) are the eigenvalues of \( \nabla f \).

**Theorem 2.1.** Let \( \eta, q: \mathbb{R}^n \to \mathbb{R} \) be an entropy, entropy flux pair corresponding to (2.1) with \( \eta \in C^2 \). Let \( g: (\mathbb{R}^n)^2 \to \mathbb{R} \) satisfy the consistency property \( g(u, u) = f(u) \). Define \( \Gamma_{\eta, q}(\mathbb{R}^n)^3 \times \mathbb{R} \to \mathbb{R} \) by
\[
(2.9) \quad \Gamma_{\eta, q}(z, v, w, \lambda) = \eta(u) - \eta(z) + \lambda [Q(w, z) - Q(v, z)],
\]
where \( Q: (\mathbb{R}^n)^2 \to \mathbb{R} \) reduces on the diagonal to \( q \), i.e.,
\[
Q(u, u) = q(u),
\]
and \( u \) is defined by
\[
u = z - \lambda [g(w, z) - g(z, v)].
\]
Then for all strictly convex \( \eta(u) \) the following inequality does not hold
\[
\Gamma_{\eta, q}(z, v, w, \lambda) \leq 0.
\]

**Proof.** For notational convenience we let
\[
\Gamma_{\eta, q}(z, v, w, \lambda) = \Gamma(z, v, w).
\]
We note that with the choice of \( v = u(x - \Delta x, t), w = u(x + \Delta x, t), z = u(x, t) \) and \( \lambda = \Delta t/\Delta x \), the theorem states that the discrete entropy inequality (2.7) does not hold. We will show that for any fixed \( z \) there exist points \( v \) and \( w \) near \( z \) such that
\[
\Gamma(v, w) = \Gamma(z, v, w) > 0.
\]
We choose \( v \) and \( w \) so that they are connected by a \( j \)-simple wave through the point \( z \). Such points always exist since \( j \)-waves are integral curves of ordinary differential equations. For completeness, we recall the definition of a \( j \)-simple wave.

**Definition.** A \( j \)-simple wave curve through \( u_0 \) is a solution \( u(e, u_0) \) of
\[
\frac{d}{de} u(e, u_0) = r_j(u(e, u_0)), \quad u(0, u_0) = u_0,
\]
where \( r_j \) is the \( j \)th right eigenvector of \( \nabla f \), i.e., it satisfies (2.8).

Since \( v \) and \( w \) are connected through \( z \) by a \( j \)-simple wave we can express them as
\[
v = v(e) = v(0) + \dot{v}(0)e + O(e^2), \quad w = w(\delta) = w(0) + \dot{w}(0)\delta + O(\delta^2),
\]
where \( v(0) = w(0) = z \) and \( \dot{v}(0) = \dot{w}(0) = r_j \). Hence \( \Gamma(v, w) \) can be considered as a function of the parameters \( e \) and \( \delta \), i.e.,
\[
\Gamma(e, \delta) = \Gamma(v(e), w(\delta)).
\]
To prove the theorem it suffices to show that \( \Gamma(e, \delta) \) does not attain a maximum at \((0, 0)\). Expanding \( \Gamma(e, \delta) \) in a Taylor series around the origin yields
\[
(2.10) \quad \Gamma(e, \delta) = \Gamma(0, 0) + \Gamma_\epsilon(0, 0)e + \Gamma_\delta(0, 0)\delta + O(e^2, \delta^2)
\]
\[
= \Gamma_\epsilon(0, 0)e + \Gamma_\delta(0, 0)\delta + O(e^2, \delta^2),
\]
SECOND-ORDER CONSERVATIVE SCHEMES

where the last equality follows since \( \Gamma(0, 0) = \Gamma(z, z) = 0 \). If either \( \Gamma\varepsilon(0, 0) \) or \( \Gamma\delta(0, 0) \) is nonzero, then (2.10) will imply that \( \Gamma(\varepsilon, \delta) \) changes of sign for \( \varepsilon \) and \( \delta \) near zero and \( \Gamma(0, 0) = 0 \) cannot be a maximum. Hence, we impose conditions which ensure that \( \Gamma\varepsilon(0, 0) = \Gamma\delta(0, 0) = 0 \). More precisely, we show that a necessary and sufficient condition for \( \Gamma\varepsilon(0, 0) = \Gamma\delta(0, 0) = 0 \) is that \( \nabla \Gamma(0, 0) = 0 \). The sufficiency is obvious since

\[
\Gamma\varepsilon(0, 0) = \nabla_n \Gamma(0, 0) \rho_j \quad \text{and} \quad \Gamma\delta(0, 0) = \nabla_w \Gamma(0, 0) \rho_j,
\]

where \( \nabla_n \Gamma = (\Gamma_{v1}, \ldots, \Gamma_{vn}), \) \( \nabla_w \Gamma = (\Gamma_{w1}, \ldots, \Gamma_{wn}) \). We note that we are free to choose \( v, w \) and \( z \) on any of the \( n \) \( j \)-waves going through \( z \). Hence, if \( \Gamma\varepsilon \) and \( \Gamma\delta \) are to be zero at the origin, then \( \nabla_n \Gamma(0, 0) \rho_j = 0 \) and \( \nabla_w \Gamma(0, 0) \rho_j = 0 \) for \( j = 1, \ldots, n \). Recalling that the \( \rho_j \) are \( n \) linearly independent vectors, it follows that \( \nabla \Gamma(0, 0) = 0 \). Hence, we assume that \( \Gamma_{v1}(0, 0) = \Gamma_{w1}(0, 0) = 0, \) \( i = 1, \ldots, n \), or equivalently by (2.10),

\[
\Gamma_{v1}(0, 0) = \nabla_n u_{v1} - \lambda Q_{v1} = \lambda \left[ \nabla_n g_{u_{v1}} - Q_{v1} \right] = 0
\]

and

\[
\Gamma_{w1}(0, 0) = \nabla_n u_{w1} + \lambda Q_{w1} = -\lambda \left[ \nabla_n g_{u_{w1}} - Q_{w1} \right] = 0,
\]

where all the terms on the right-hand side are evaluated at \( z \). Therefore, if \( \Gamma\varepsilon \) and \( \Gamma\delta \) vanish at the origin, the following conditions hold on the diagonal

(2.11) \( \nabla_n(z)g_{v1}(z, z) = Q_{v1}(z, z) \),

(2.12) \( \nabla_n(z)g_{w1}(z, z) = Q_{w1}(z, z) \).

Under these conditions, we show that

(2.13) \( \Gamma_{v1}(0, 0) + \Gamma_{w1}(0, 0) > 0 \)

and hence, \( \Gamma(0, 0) \) is not a local maximum. A straightforward computation yields

\[
\Gamma_{v1} = \nabla_v^2 \Gamma \varepsilon^2 + \nabla_v \Gamma \varepsilon \rho, \quad \Gamma_{w1} = \nabla_w^2 \Gamma \delta^2 + \nabla_w \Gamma \delta \rho.
\]

Since \( \varepsilon(0) = \delta(0) = \rho_j \) and \( \Gamma(0, 0) = \Gamma(0, 0) = 0 \), we have

(2.14) \( \Gamma_{v1}(0, 0) + \Gamma_{w1}(0, 0) = \left[ \nabla_v^2 \Gamma(0, 0) + \nabla_w^2 \Gamma(0, 0) \right] \rho_j^2 = \Lambda \rho_j^2 \).

Hence, (2.13) is equivalent to showing that \( \Lambda \) is positive definite. We note that the entries of \( \Lambda \) are of the form \( \Gamma_{v1v1}(0, 0) + \Gamma_{w1w1}(0, 0) \). Taking the pertinent derivatives in (2.9), it follows that

\[
\Gamma_{v1}(w, v) = \nabla_n(u)u_{v1} - Q_{v1}(z, v) \lambda = \lambda \left[ \nabla_n(u)g_{v1}(z, v) - Q_{v1}(z, v) \right],
\]

\[
\Gamma_{w1}(w, z) = \nabla_n(u)u_{w1} + Q_{w1}(w, z) \lambda = -\lambda \left[ \nabla_n(u)g_{w1}(w, z) - Q_{w1}(w, z) \right];
\]

hence

(2.15) \( \Gamma_{v1v1}(0, 0) = \lambda^2 \nabla^2 g_{v1v1} + \lambda \left[ \nabla_n g_{v1v1} - Q_{v1v1} \right], \)

(2.16) \( \Gamma_{w1w1}(0, 0) = \lambda^2 \nabla^2 g_{w1w1} - \lambda \left[ \nabla_n g_{w1w1} - Q_{w1w1} \right], \)

where all the terms on the right-hand side are evaluated at \( z \). Differentiating Eq. (2.11) with respect to \( z_i \) and Eq. (2.12) with respect to \( z_j \) yields

\[
\frac{\partial}{\partial z_j}(\nabla n)g_{v1} + \nabla g_{v1v1} + \nabla g_{v1w1} = Q_{v1w1} + Q_{v1v1},
\]

\[
\frac{\partial}{\partial z_j}(\nabla n)g_{w1} + \nabla g_{v1v1} + \nabla g_{w1w1} = Q_{w1w1} + Q_{w1v1}.
\]
Multiplying the last equations by $\lambda$ and subtracting gives

\begin{equation}
\lambda\left[Q_{w,w_j} - Q_{v,v_j}\right] - \lambda\left[\nabla\lambda g_{w,w_j} - g_{v,v_j}\right] = \lambda\left[\frac{\partial}{\partial z_i} (\nabla \eta) g_{w_j} - \frac{\partial}{\partial z_j} (\nabla \eta) g_{v_i}\right],
\end{equation}

where all the terms are evaluated at $z$. Combining (2.15) with (2.16) yields

\begin{equation}
\Gamma_{v,v_j}(0,0) + \Gamma_{w,w_j}(0,0) = \lambda^2 \nabla^2 \eta (g_{v_i} g_{v_j} + g_{w_i} g_{w_j}) + \lambda \left[\frac{\partial}{\partial z_i} (\nabla \eta) g_{w_j} - \frac{\partial}{\partial z_j} (\nabla \eta) g_{v_i}\right] = I + II.
\end{equation}

It is shown in [3] that a conservative scheme which approximates (2.1) is second-order accurate if and only if the leading term of the numerical flux is of the form

\begin{equation}
G(w, z) = \frac{f(z) + f(w)}{2} + (\lambda/2) A(w - z),
\end{equation}

where the quantity $A$ is taken to be $(\nabla f(w)^2 + \nabla f(z)^2)/2$ for the sake of symmetry. Any other choice would make a difference that is quadratic in $w - z$. Hence, for conservative second-order schemes, the numerical flux can be expressed as

\begin{equation}
g(w, z) = G(w, z) + O(|w - z|^2).
\end{equation}

It follows that

\begin{align*}
g_{v_i}(z, v) &= \frac{1}{2} f_{u_i}(v) + \frac{1}{2} \lambda A_{v_i}(z - v) - \frac{1}{2} \lambda A e_i + O(|z - v|), \\
g_{w_i}(w, z) &= \frac{1}{2} f_{u_i}(w) + \frac{1}{2} \lambda A_{w_i}(w - z) - \frac{1}{2} \lambda A e_i + O(|z - w|),
\end{align*}

where $e_i^t$ is the transpose of the $i$th vector in the canonical basis of $R^n$, e.g., $e_1^t = (0,1,0,\ldots,0)^t$. From here on, all functions will be evaluated at $z$; thus, to simplify the notation the argument $z$ will be omitted. From the last equalities it follows that for $i = 1,\ldots,n$,

\begin{align*}
g_{v_i} &= \frac{1}{2} f_{u_i} - \frac{1}{2} \lambda (\nabla f)^2 e_i = \frac{1}{2} (B - \lambda B^2) e_i, \\
g_{w_i} &= \frac{1}{2} f_{u_i} + \frac{1}{2} \lambda (\nabla f)^2 e_i = \frac{1}{2} (B + \lambda B^2) e_i,
\end{align*}

where $B = \nabla f(z)$. The equalities on the right-hand side follow since $f_{u_i} = B e_i$, $i = 1,\ldots,n$.

The term $I$ in (2.17) can be rewritten as

\begin{align*}
I &= \frac{1}{4} \lambda e_i \left[(B' - \lambda (B^2)' \nabla^2 \eta (B - \lambda B^2) \\
&\quad + (B' + \lambda (B^2)' \nabla^2 \eta (B + \lambda B^2)\right] e_i' \\
&= \frac{1}{2} \lambda e_i [B' \nabla^2 \eta B + \lambda^2 (B^2)' \nabla^2 \eta B^2] e_i'.
\end{align*}

Thus $I$ is the $(i, j)$th entry of the matrix

\begin{equation}
M_1 = \frac{1}{2} \lambda^2 \left[B' \nabla^2 \eta B + \lambda^2 (B^2)' \nabla^2 \eta B^2\right].
\end{equation}
To rewrite II in (2.17), we note that
\[
\frac{\partial}{\partial z_j} \nabla \eta = e_j \nabla^2 \eta, \quad j = 1, \ldots, n.
\]
Therefore,
\[
(2.19) \quad II = \frac{1}{2} \lambda \left( e_j \nabla^2 \eta (B + \lambda B^2) e_j - e_j \nabla^2 \eta (B - \lambda B^2) e_j \right)
\]
\[= \frac{1}{2} \lambda e_j \left( \nabla^2 \eta (B + \lambda B^2) - (B^T - \lambda (B^2)^T) \nabla^2 \eta \right) e_j.
\]
Hence II is the \((i, j)\)th entry of the matrix
\[
(2.20) \quad M_2 = \frac{1}{2} \lambda \left( \nabla^2 \eta B - B^T \nabla^2 \eta \right) + \lambda \left( \nabla^2 \eta B^2 + (B^2)^T \nabla^2 \eta \right).
\]
Combining (2.17), (2.18) and (2.20) yields
\[
(2.21) \quad \left[ \nabla \phi \Gamma(0,0) + \nabla \phi' \Gamma(0,0) \right] r_j^2 = [M_1 + M_2] r_j^2,
\]
where \(M_1\) and \(M_2\) are defined by (2.18) and (2.20). We recall that \(r_j\) is the right \(j\)th eigenvector of \(\nabla f\), i.e.,
\[
\nabla f r_j = Br_j = \lambda_j r_j,
\]
and hence \(r_j^T\) is the \(j\)th left eigenvector of \(B^T\), i.e.,
\[
r_j^T B = (B r_j)^T = r_j^T \lambda_j.
\]
Hence,
\[
M_1 r_j^2 = \alpha \nabla^2 \eta r_j^2,
\]
where \(\alpha = \frac{1}{2} \lambda (\lambda_j^2 + \lambda_j^2)\). Noting that the symmetry of \(\nabla^2 \eta\) implies that
\[
(\nabla^2 \eta B - B^T \nabla^2 \eta') r_j^2 = \lambda_j (\nabla^2 \eta - \nabla^2 \eta') r_j^2 = 0,
\]
it follows that
\[
M_2 r_j^2 = \beta (\nabla^2 \eta + \nabla^2 \eta') r_j^2,
\]
where \(\beta = \frac{1}{2} \lambda^2 \lambda_j\). Plugging these expressions of \(M_1\) and \(M_2\) into (2.20) yields
\[
(2.22) \quad \left[ \nabla \phi \Gamma(0,0) + \nabla \phi' \Gamma(0,0) \right] r_j^2 = \left[ (\alpha + \beta) \nabla^2 \eta + \beta \nabla^2 \eta' \right] r_j^2.
\]
The strict convexity of the entropy \(\eta\) implies that the matrices \(\nabla^2 \eta\) and \(\nabla^2 \eta'\) are positive definite. Since \(\alpha\) and \(\beta\) are positive, it follows from (2.14) and (2.22) that
\[
\Gamma_{ee}(0,0) + \Gamma_{\theta \theta}(0,0) = \left[ (\alpha + \beta) \nabla^2 \eta + \beta \nabla^2 \eta' \right] r_j^2 > 0.
\]
Thus, either \(\Gamma_{ee}(0,0)\) is positive or \(\Gamma_{\theta \theta}(0,0)\) is positive, ruling out any maxima for \(\Gamma\) at the origin.

Department of Mathematics
Duke University
Durham, North Carolina 27706


