Discrete entropy inequalities via an optimization process

Nina Aguillon, Emmanuel Audusse, Vivien Desveaux, Julien Salomon

ESAIM: M2AN 58 (2024) 363–391
https://doi.org/10.1051/m2an/2023098

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

DISCRETE ENTROPY INEQUALITIES *VIA* AN OPTIMIZATION PROCESS

Nina Aguillon[1,*], Emmanuel Audusse[2], Vivien Desveaux[3] and Julien Salomon[4]

Abstract. The solutions of hyperbolic systems may contain discontinuities. These weak solutions verify not only the original PDEs, but also an entropy inequality that acts as a selection criterion determining whether a discontinuity is physical or not. Obtaining a discrete version of these entropy inequalities when approximating the solutions numerically is crucial to avoid convergence to unphysical solutions or even unstability. However such a task is difficult in general, if not impossible for schemes of order 2 or more. In this paper, we introduce an optimization framework that enables us to quantify a posteriori the decrease or increase of entropy of a given scheme, locally in space and time. We use it to obtain maps of numerical diffusion and to prove that some schemes do not have a discrete entropy inequality. A special attention is devoted to the widely used second order MUSCL scheme for which almost no theoretical results are known.

Mathematics Subject Classification. 35L03, 65M08, 76M12, 35L40, 76-10.

Received May 19, 2023. Accepted November 30, 2023.

1. Introduction

Many physical phenomena can be described by means of a hyperbolic system, also called a system of conservation laws. Some famous hyperbolic systems include the Lighthill–Whitham–Richards or Aw-Rascle models for traffic flow, the shallow water equations of Barr de Saint-Venant, the Euler equation for fluid dynamics and the inviscid magnetohydrodynamics equation.

This class of partial differential equations (PDEs) does not contain any regularization terms such as diffusion or dispersion. Their solutions typically develop discontinuities in finite time. These discontinuities are observed in traffic jams, during floods caused by dam breaks, at hydraulic jumps, or in aeronautics. The PDE should be understood in the weak sense to allow such discontinuous solutions. Doing so makes it possible to construct infinitely many discontinuous solutions for the same initial data. An additional criterion should consequently be included to select only the physical weak solution. It generally takes the form of an entropy (or energy) inequality and is related to the second law of thermodynamics.

Keywords and phrases. Numerical diffusion, finite volume methods, discrete entropy inequality, MUSCL scheme.

[1] Sorbonne-Université and INRIA Paris, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), 75005 Paris, France.

[2] Université Sorbonne Paris Nord and INRIA Paris, CNRS, Laboratoire Analyse, Géométrie et Applications (LAGA), 99 av. J.-B. Clément, 93430 Villetaneuse, France.

[3] Université de Picardie Jules Verne, CNRS, LAMFA, 33 rue Saint-Leu, 80039 Amiens Cedex 1, France.

[4] INRIA Paris, ANGE Project-Team, 75589 Paris Cedex 12, France and Laboratoire Jacques-Louis Lions, Sorbonne Université, CNRS, 75005 Paris, France.

*Corresponding author: nina.aguillon@sorbonne-universite.fr

Discretizing the PDE to obtain a numerical approximation of the solution can be done in several ways. In this paper we focus on finite volume schemes which are well adapted to the low regularity of the solution and built around the idea of conservation laws. In the design of such schemes, it seems important that the entropy also decreases at the numerical level. This condition ensures that the scheme will not converge towards a nonphysical solution of the PDE. The loss of entropy in each cell during one time step is called the numerical diffusion. Discrete entropy inequalities have mainly been obtained for first order schemes in space and time. Realistic codes use high-order discretizations and splitting techniques and usually incorporate ideas and knowledge from the well-understood first-order framework. For this reason, they probably verify a discrete entropy inequality in most cases. However, no explicit formulas are known in practice.

This paper proposes to quantify the numerical diffusion with an a posteriori minimization technique where the scheme is used as a black box. Our primary goal is to obtain maps of numerical diffusion which quantify in space and time the loss of entropy coming from the choice of discretization. As a practical motivation and a long term objective, let us mention numerical oceanic circulation models where the numerical diffusion is related to artificial changes of salinity, density and temperature between two adjacent distinct water masses, and is usually called spurious mixing. This phenomenon is identified as a major issue in numerical cores for climate application [13]. We refer the reader to [5, 18] for the quantification of spurious mixing in simplified configurations and [17] in a realistic setting.

This paper provides a mathematical insight on the quantification of numerical diffusion in realistic codes but is still far away from oceanic applications. Maps of numerical diffusion are obtained by minimizing a functional which takes into account the consistency of the numerical entropy fluxes and the fact that the entropy should decrease at each time step. This minimization couples all the cells of the mesh, but we also propose a local and cheap quantification that gives qualitatively good results. A different perspective allows us to construct the worst initial data in terms of entropy by minimizing a different but related functional. We apply this procedure to prove that no discrete entropy inequality exists for most of the versions of the widely used MUSCL approach with a 2 steps Runge-Kutta time discretization. This was suspected in [2]. A notable exception is the limitation in the entropy variables with a HLL first order scheme [1].

We now present the mathematical framework on hyperbolic PDEs and their discretization with finite volume schemes, with an emphasis on discrete entropy inequality which is at the core of this work. Our main results and the outline of the paper are described hereafter.

Fundamentals on discrete entropy inequalities

Consider a hyperbolic system in 1 dimension (1D) in space

$$\partial_t u(x,t) + \partial_x f(u(x,t)) = 0, \quad x \in \mathbf{R}, t \in \mathbf{R}^+, \tag{1}$$

where the vectorial unknown u belongs to some convex domain $\Omega \subset \mathbf{R}^d$. The flux $f : \Omega \to \mathbf{R}^d$ is $\mathcal{C}^1$-regular and its Jacobian matrix Df is diagonalizable with real eigenvalues. We are only interested in weak solutions of (1) that additionally satisfy the entropy inequality (or energy inequality)

$$\partial_t \eta(u) + \partial_x G(u) \leq 0, \tag{2}$$

where the entropy $\eta : \Omega \to \mathbf{R}$ is strictly convex. The entropy flux G is linked to the entropy η through the relation on their Jacobian matrices $D\eta Df = DG$. Such hyperbolic systems arise in particular in the modeling of nonviscous flows. In this paper, we consider the scalar ($d = 1$) Burgers equations related to the Lighthill–Whitham–Richards model for traffic flows and the Euler equations of inviscid gas dynamics for which $d = 3$.

We now turn to the numerical discretization of (1) with a finite volume technique. For the sake of simplicity a space interval $[a, b]$ with periodic boundary conditions is considered throughout the paper. A finite number of cells M is fixed with size $\Delta x = \dfrac{b-a}{M}$ and we note $x_{j-1/2} = a + (j-1)\Delta x$ for $j \in [\![1, M]\!]$. The points $a = x_{1/2}$ and $b = x_{M+1/2}$ are identified and all the space subscript are considered modulo M. For the sake of simplicity, every family $(X)_{j\in[\![1,M]\!]}$ of quantities indexed by the set of cells will be denoted by $(X_j)_j$.

We also consider a discretization in time $0 = t^0 < t^1 < \dots t^n < \dots$ A Courant–Friedrichs–Lewy (CFL) condition is imposed at each time step. It reads, for some CFL $\alpha \in (0,1)$ depending on the scheme,

$$\left(t^{n+1} - t^n\right) \max_{1 \le j \le M} \rho\left(Df\left(u_j^n\right)\right) = \alpha \Delta x, \tag{3}$$

where $\rho\left(DF\left(u_j^n\right)\right)$ is the spectral radius of the Jacobian matrix $DF\left(u_j^n\right)$. The time step varies with the time iteration but for the sake of simplicity, we will denote it independently of n as $t^{n+1} - t^n = \Delta t$.

A finite volume scheme writes

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(f_{j+1/2}^n - f_{j-1/2}^n\right), \tag{4}$$

where the vectors u_j^n and $f_{j+1/2}^n$ correspond to the numerical approximations of the mean of the exact solution u and flux $f(u)$

$$u_j^n \approx \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u\left(x, t^n\right) dx, \;\; f_{j+1/2}^n \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f\left(u\left(x_{j+1/2}, s\right)\right) ds.$$

Definition 1. A consistent finite volume scheme with a stencil of $s_L \in \mathbf{N}$ cells to the left and $s_R \in \mathbf{N}$ cells to the right is a formula expressing the numerical flux $f_{j+1/2}^n$ in terms of its $s_L + s_R$ neighbor cells $f_{j+1/2}^n = \mathcal{F}\left(u_{j-s_L+1}^n, \dots, u_{j+s_R}^n\right)$ such that, for all $u \in \Omega$, $\mathcal{F}\left(u_{-s_L+1}, \dots, u_{s_R}\right)$ tends to $f(u)$ as $\left(u_{-s_L+1}, \dots, u_{s_R}\right)$ tends to $(u, \dots, u)$.

Equations (1) and (2) are understood in a weak sense to allow discontinuous solutions. Inequality (2) does not hold for every discontinuity but selects only entropy satisfying shocks. The fact that the scheme (4) is stable and computes the entropy solution, with only entropy satisfying shocks, is strongly related to the existence of a numerical counterpart of (2) at the discrete level, called a discrete entropy inequality

$$\forall j \in [\![1, M]\!], \quad \eta\left(u_j^{n+1}\right) \le \eta\left(u_j^n\right) - \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^n - G_{j-1/2}^n\right). \tag{5}$$

In the spirit of Definition 1, we introduce the notion of entropy satifying scheme.

Definition 2. Consider a consistent finite volume scheme of $s_L \in \mathbf{N}$ cells to the left and $s_R \in \mathbf{N}$ to the right. It is a consistent entropy satisfying scheme if there exists a numerical entropy flux function $\mathcal{G}$ such that the following two conditions are satisfied.

- Inequality (5) holds for all data $\left(u_k^n\right)_k$ with the choice $G_{j+1/2}^n = \mathcal{G}\left(u_{j-s_L+1}^n, \dots, u_{j+s_R}^n\right)$. The remainder

$$D_j^n = \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^n - G_{j-1/2}^n\right) \tag{6}$$

  is called numerical diffusion and is nonpositive in all cells.
- The numerical entropy flux function $\mathcal{G}$ is consistent: for all $u \in \Omega$, $\mathcal{G}\left(u_{-s_L+1}, \dots, u_{s_R}\right)$ tends to $G(u)$ as $\left(u_{-s_L+1}, \dots, u_{s_R}\right)$ tends to $(u, \dots, u)$.

There exist several first order schemes (4) with a stencil $s_L = s_R = 1$ for which an explicit formula for $\mathcal{G}$ yielding to nonpositive diffusion D_j^n is known:

- for scalar equations $d = 1$, monotone schemes, see [14];
- for hyperbolic systems $d \ge 2$ the Godunov and HLL schemes are entropy satisfying; see [14]. In the specific case of gas dynamics, the HLLC scheme and some relaxation or kinetic schemes are also entropy satisfying see [3, 20] and references therein.

For schemes of order larger than 1 the specific form of (5) seems out of reach for hyperbolic systems and the question is still largely open. Some works present results in that direction, mainly for second order schemes.

- In [4,9] and [1], inequality (5) is slightly modified. The schemes are either difficult to implement or there is no guarantee that they capture entropy solution.
- The local discrete entropy inequality (5) is replaced by the weaker global condition $\sum_j \eta\left(u_j^{n+1}\right) \leq \sum_j \eta\left(u_j^n\right)$ in [10] for the multilayer shallow water equations, in [20] and [16] for gas dynamics and in [6] for a conservation law with nonconvex flux.
- A different approach consists in using a second order scheme and to go back to first order if (5) does not hold. The MOOD technique (see [7,11]) was initially developed to ensure the positiveness of some quantities like the density and the pressure, as well as some discrete maximum principle. This method was later extended in [2] to ensure the discrete entropy inequalities (5) hold. However it is limited to the gas dynamics (23).

On the other hand many schemes of order 2 or more are employed in the applications for their realistic results. They are typically designed to be of high order when the solution is smooth and include corrections to ensure stability, such as limiters or explicit numerical diffusion. Positivness and lack of spurious oscillations are also often taken into account. However, the existence of discrete entropy inequality often remains an open question, as is the case for the Piecewise Parabolic Method (PPM) [8] or the (Weighted) Essentially Non Oscillatory method (ENO/WENO) [15,21,26].

In this work, we are concerned with the *a posteriori* determination of the numerical entropy fluxes $G_{j+1/2}^n$ and numerical diffusions D_j^n of Definition 2. We do not attempt to find an analytical choice of $\mathcal{G}$. Instead, we follow a different approach and construct an adequate functional $\mathcal{J}$ that is minimized to obtain

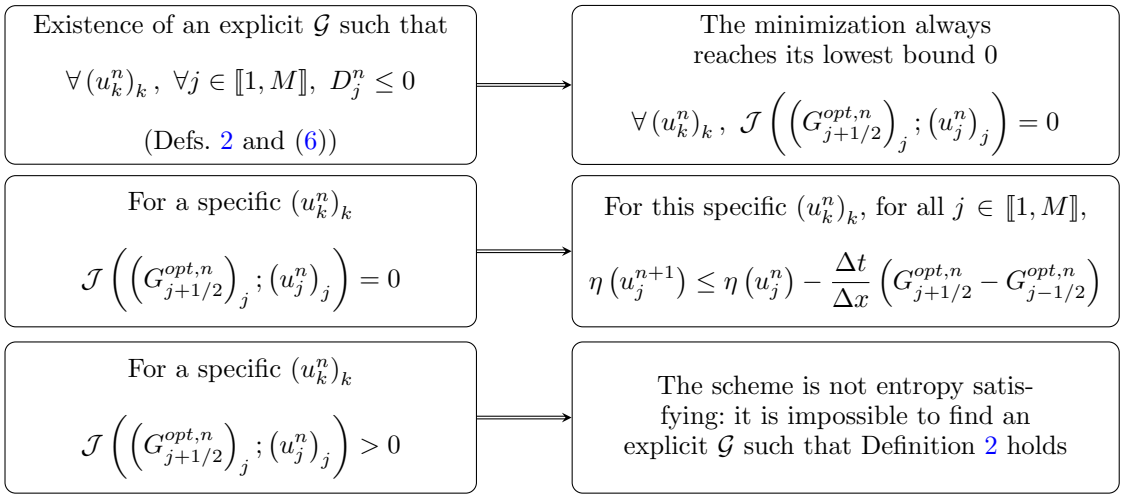
$$\left(G_{j+1/2}^{opt,n}\right)_j = \underset{\gamma=\left(\gamma_{j+1/2}^n\right)_j \in \mathbf{R}^M}{\arg\min} \left\{\mathcal{J}\left(\gamma; \left(u_j^n\right)_j\right)\right\} \tag{7}$$

and the corresponding a posteriori numerical diffusion

$$D_j^{opt,n} = \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^{opt,n} - G_{j-1/2}^{opt,n}\right). \tag{8}$$

The first objective is to be able to quantify the numerical diffusion in settings where finding an explicit formula for $\mathcal{G}$ is out of reach. The functional is built so that the two constraints of Definition 2, namely the nonpositivity of the diffusion and the consistency of the fluxes, are satisfied "as much as possible" and can be applied to any explicit finite volume scheme satisfying Definition 1.

This a posteriori approach offers a novel method to construct numerical entropy fluxes. There are strong links between Definition 2 of entropy satisfying scheme and the fact that the minimum of the functional $\mathcal{J}$ is 0. Our main results are summarized in the diagram below.



The paper is organized as follows. The construction of the functional $\mathcal{J}$ and the minimization procedure are introduced in Section 2. In Section 3, we prove several theoretical results about the minimizers of the functional $\mathcal{J}$, including the two first implications of the previous diagram. The first one ensures that the minimization procedure is efficient on the class of entropy satisfying schemes where an expression for the numerical entropy fluxes $G_{j+1/2}$ in (5) can be found in the literature. The second one is more practical, since it is applied on a specific initial data for a generic scheme and allows to visualize the distribution of numerical diffusion by making the best choice for $G_{j+1/2}$. We also prove a Lax-Wendroff theorem adapted to our minimization procedure in the sense that if a numerical scheme converges and we can find numerical entropy fluxes such that $\mathcal{J}\left(\left(G^{opt,n}_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$, then the limit is an entropy solution.

The minimization procedure is illustrated on various testcases in Section 4 both for the Burgers equation and for the Euler equations of gas dynamics. The third implication of the previous diagram is nothing but the contraposition of the first one. We exploit it in Section 5 to build initial data that cannot satisfy any discrete entropy inequality. As an illustration, we prove that most variants of the MUSCL scheme with a second-order Runge-Kutta time discretization for the gas dynamics are not entropy satisfying.

2. Minimization procedure

The minimization is carried out at some fixed time iteration n which is referred to as "the initial data". In other words $\left(u^n_j\right)_j$ is fixed by the user and u^{n+1}_j is then obtained with Scheme (4). The superscript n plays no particular role and one can fix, *e.g.*, $n=0$.

In this section, the numerical entropy flux is computed by a minimization procedure (7). The function $\mathcal{J}:\mathbf{R}^M\to\mathbf{R}$ is defined by the sum of two contributions $\mathcal{J}=\mathcal{J}^D+\mathcal{J}^C$. The first part $\mathcal{J}^D$ gathers the contributions of positive numerical diffusion

$$\mathcal{J}^D\left(\gamma;(u^n_j)_j\right)=\sum_{j=1}^{M}\max\left(0,\eta\left(u^{n+1}_j\right)-\eta\left(u^n_j\right)+\frac{\Delta t}{\Delta x}\left(\gamma^n_{j+1/2}-\gamma^n_{j-1/2}\right)\right)^2.$$

Indeed the quantity $\eta\left(u^{n+1}_j\right)-\eta\left(u^n_j\right)+\dfrac{\Delta t}{\Delta x}\left(\gamma^n_{j+1/2}-\gamma^n_{j-1/2}\right)$ corresponds to the numerical diffusion (6) in the j-th cell if the numerical entropy fluxes are given by $G_{j+1/2}=\gamma_{j+1/2}$. A negative value is consistent with the Definition 2 of an entropy satisfying scheme; thus only positive contributions are kept in the function $\mathcal{J}^D$. Overall, we have $J^D\left(\gamma;(u^n_j)_j\right)=0$ if and only if Inequality (5) holds with the choice $G=\gamma$.

It remains to take into account the consistency property of the numerical entropy fluxes of Definition 2, which is the role of the second part of the functional. We make use of *a priori* consistency bounds on the numerical entropy fluxes $m^n_{j+1/2}\le G^n_{j+1/2}\le M^n_{j+1/2}$ depending on s_L cells on the left and s_R on the right. These bounds are defined in Lemma 1 and satisfy the following consistency property

$$u^n_{j-s_L+1}=\cdots=u^n_{j+s_R}=u\implies m^n_{j+1/2}=M^n_{j+1/2}=G(u).$$

This motivates the choice

$$\mathcal{J}^C\left(\gamma;\left(u^n_j\right)_j\right)=\left(\frac{\Delta t}{\Delta x}\right)^2\sum_{j=0}^{M}\left(\max\left(0,\gamma^n_{j+1/2}-M^n_{j+1/2}\right)^2+\max\left(0,m^n_{j+1/2}-\gamma^n_{j+1/2}\right)^2\right),$$

which vanishes if and only if $\forall j\in\{1/2,\dots,M+1/2\},\quad m^n_{j+1/2}\le\gamma^n_{j+1/2}\le M^n_{j+1/2}$. In the particular case where $u^n_{j-s_L+1}=\cdots=u^n_{j+s_R}=u$, the j-th term of the sum in $\mathcal{J}^C\left(\gamma;\left(u^n_j\right)_j\right)$ vanishes if and only if $\gamma^n_{j+1/2}=G(u)$, meaning that γ is consistent. The function $\mathcal{J}$ is nonnegative, $\mathcal{C}^1$-regular and convex with respect to γ, but the set where it vanishes has no reason either to exist or to be reduced to a single point.

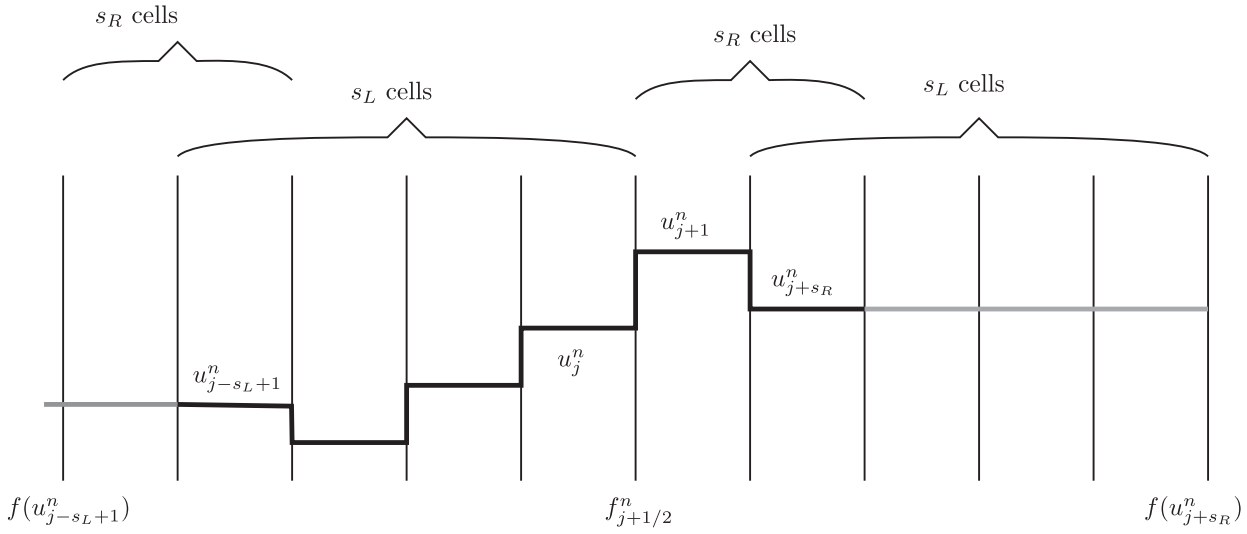


FIGURE 1. Modification of the initial data away from the stencil. Consistency can be used to compute the fluxes at interfaces $j+3/2-s_L-s_R$ and $j-1/2+s_L+s_R$ and the value of $f^n_{j+1/2}$ remains unchanged.

We now define the bounds $m^n_{j+1/2}$ and $M^n_{j+1/2}$. To do so, we zoom on the interface $j+1/2$ and construct a modified initial data $\left(\tilde{u}_k^{n,j+1/2}\right)_k$ as illustrated on Figure 1 in such a way that the consistency property can be used as close as possible to the interface $j+1/2$ without changing the value of the numerical flux $f_{j+1/2}$. The easiest way to proceed consists in extending the initial data by constant values outside of the stencil.

Then the finite volume scheme (4) applied to this new initial data gives numerical fluxes $\left(\tilde{f}_{k+1/2}^{n,j+1/2}\right)_k$. They are used to compute the update $\left(\hat{u}_k^{n,j+1/2}\right)_k$. The benefit brought by this modified data is that the evolution of the total entropy on the left and the right of interface $j+1/2$ can be expressed only in terms of $G^n_{j+1/2}$, $\left(\tilde{u}_k^{n,j+1/2}\right)_k$ and $\left(\hat{u}_k^{n,j+1/2}\right)_k$. It yields bounds on the individual numerical entropy flux $G^n_{j+1/2}$ that can be constructed from the initial data $(u^n_k)_k$ and depends on the chosen numerical scheme.

Lemma 1. *Consider a consistent entropy satisfying scheme in the sense of Definition 2. Then, for all* $j \in [\![1, M]\!]$, *we have* $m^n_{j+1/2} \leq G^n_{j+1/2} \leq M^n_{j+1/2}$, *with s*

$$\begin{cases} M^n_{j+1/2} = G\left(u^n_{j-s_L+1}\right) + \dfrac{\Delta x}{\Delta t} \displaystyle\sum_{k=j-s_L-s_R+2}^{j} \eta\left(\tilde{u}_k^{n,j+1/2}\right) - \eta\left(\hat{u}_k^{n,j+1/2}\right), \\ m^n_{j+1/2} = G\left(u^n_{j+s_R}\right) + \dfrac{\Delta x}{\Delta t} \displaystyle\sum_{k=j+1}^{j+s_L+s_R-1} \eta\left(\hat{u}_k^{n,j+1/2}\right) - \eta\left(\tilde{u}_k^{n,j+1/2}\right), \end{cases} \tag{9}$$

where

$$\tilde{u}_k^{n,j+1/2} = u^n_{\min(\max(k,j-s_L+1),j+s_R)} \tag{10}$$

and for all $k \in [\![1, M]\!]$,

$$\hat{u}_k^{n,j+1/2} = \tilde{u}_k^{n,j+1/2} - \frac{\Delta t}{\Delta x}\Big(\mathcal{F}\left(\tilde{u}_{k-s_L+1}^{n,j+1/2}, \dots, \tilde{u}_k^{n,j+1/2}, \tilde{u}_{k+1}^{n,j+1/2}, \dots, \tilde{u}_{k+s_R}^{n,j+1/2}\right) - \mathcal{F}\left(\tilde{u}_{k-s_L}^{n,j+1/2}, \dots, \tilde{u}_{k-1}^{n,j+1/2}, \tilde{u}_k^{n,j+1/2}, \dots, \tilde{u}_{k+s_R-1}^{n,j+1/2}\right)\Big).$$

Proof. The flux at the interface $j+1/2$ is independent of the values in cells $j-s_L$ and smaller, and in cells $j+s_R+1$ and larger. Thus, we modify the initial data by extending it by $u_{j-s_L+1}^n$ on its left and by $u_{j+s_R}^n$ on its right

$$\tilde{u}_k^{n,j+1/2} = \begin{cases} u_{j-s_L+1}^n & \text{if } k \le j - s_L, \\ u_k^n & \text{if } j - s_L + 1 \le k \le j + s_R, \\ u_{j+s_R}^n & \text{if } j + s_R + 1 \le k, \end{cases}$$

which corresponds to (10). We update this modified initial data with the chosen finite volume scheme $\mathcal{F}$ and obtain $\hat{u}_k^{n,j+1/2}$.

The interest of the modified initial data $\left(\tilde{u}_k^{n,j+1/2}\right)_k$ is that the fluxes at interfaces $j - s_L - s_R + 3/2$ (and before) and $j + s_R + s_L - 1/2$ (and after) are given by consistency:

$$\mathcal{F}\left(\tilde{u}_{j-2s_L-s_R+2}^{n,j+1/2}, \dots, \tilde{u}_{j-s_L-s_R+1}^{n,j+1/2}, \tilde{u}_{j-s_L-s_R+2}^{n,j+1/2}, \dots, \tilde{u}_{j-s_L+1}^{n,j+1/2}\right) = f\left(u_{j-s_L+1}^n\right),$$

$$\mathcal{F}\left(\tilde{u}_{j+s_R}^{n,j+1/2}, \dots, \tilde{u}_{j+s_L+s_R-1}^{n,j+1/2}, \tilde{u}_{j+s_L+s_R}^{n,j+1/2}, \dots, \tilde{u}_{j+s_L+2s_R-1}^{n,j+1/2}\right) = f\left(u_{j+s_R}^n\right).$$

Now, as the scheme is entropy satisfying

$$\forall k,\ \eta\left(\hat{u}_k^{n,j+1/2}\right) \le \eta\left(\tilde{u}_k^{n,j+1/2}\right) - \frac{\Delta t}{\Delta x}\left(\tilde{G}_{k+1/2}^{n,j+1/2} - \tilde{G}_{k-1/2}^{n,j+1/2}\right),$$

where $\tilde{G}_{k+1/2}^{n,j+1/2} = \mathcal{G}\left(\tilde{u}_{k-s_L+1}^{n,j+1/2}, \dots, \tilde{u}_k^{n,j+1/2}, \tilde{u}_{k+1}^{n,j+1/2}, \dots, \tilde{u}_{k+s_R}^{n,j+1/2}\right)$ for some function $\mathcal{G}$, see Definition 2. These inequalities are not of much use because we do not know an explicit formula for $\mathcal{G}$. However, we deduced from the construction of $\left(\tilde{u}_k^{n,j+1/2}\right)_k$ several consequences. The consistency yields at interfaces $j - s_L - s_R + 3/2$ and $j + s_L + s_R - 1/2$

$$\tilde{G}_{j-s_L-s_R+3/2}^{n,j+1/2} = G\left(u_{j-s_L+1}^n\right) \qquad \text{and} \qquad \tilde{G}_{j+s_L+s_R-1/2}^{n,j+1/2} = G\left(u_{j+s_R}^n\right).$$

On the other hand the flux at interface $j+1/2$ remains unchanged

$$\tilde{G}_{j+1/2}^{n,j+1/2} = \mathcal{G}\left(\tilde{u}_{j-s_L+1}^{n,j+1/2}, \dots, \tilde{u}_{j+s_R}^{n,j+1/2}\right) = \mathcal{G}\left(u_{j-s_L+1}^n, \dots, u_{j+s_R}^n\right) = G_{j+1/2}^n.$$

We eliminate the other numerical entropy fluxes by summation:

$$\sum_{k=j+1}^{j+s_L+s_R-1} \eta\left(\hat{u}_k^{n,j+1/2}\right) \le \left(\sum_{k=j+1}^{j+s_L+s_R-1} \eta\left(\tilde{u}_k^{n,j+1/2}\right)\right) - \frac{\Delta t}{\Delta x}\left(G\left(u_{j+s_R}^n\right) - G_{j+1/2}^n\right),$$

$$\sum_{k=j-s_L-s_R+2}^{j} \eta\left(\hat{u}_k^{n,j+1/2}\right) \le \sum_{k=j-s_L-s_R+2}^{j} \eta\left(\tilde{u}_k^{n,j+1/2}\right) - \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^n - G\left(u_{j-s_L+1}^n\right)\right).$$

We can now bound $G_{j+1/2}^n$ from above and below and conclude. □

Remark 1. In the case of two points scheme $s_L = s_R = 1$, Lemma 1 correspond to the notion of "interface entropy inequality" in Definition 2.7 of [3]. In this specific case it implies the desired discrete entropy inequality (5) at the cost of a time step Δt twice smaller Proposition 2.9 of [3].

3. Main results

In this section, we state some properties of the minimizers of $\mathcal{J}$ and revisit some standard concepts of finite volume scheme theory from our a posteriori point of view.

3.1. Entropy dissipation and zero minimization

Proposition 1. *Consider a finite volume scheme (4) that admits a discrete entropy inequality (5) for some numerical entropy fluxes* $\left(G^n_{j+1/2} = \mathcal{G}\left(u^n_{j-s_L+1}, \dots, u^n_{j+s_R}\right)\right)_j$. *Then for all initial data* $\left(u^n_j\right)_j$ *we have* $\mathcal{J}\left(\left(G^n_{j+1/2}\right)_j; \left(u^n_j\right)_j\right) = 0$.

Proof. This property follows from the definition of the functional. If (5) holds, then

$$\mathcal{J}^D\left(\left(G^n_{j+1/2}\right)_j; \left(u^n_j\right)_j\right) = 0.$$

It remains to prove that the second part also vanishes, which is the case if and only if $m^n_{j+1/2} \leq G^n_{j+1/2} \leq M^n_{j+1/2}$. This holds by construction on the bounds $m^n_{j+1/2}$ and $M^n_{j+1/2}$. The scheme is also entropy diminishing on the modified initial data $\left(\tilde{u}^{n,j+1/2}_k\right)_k$. The lower (resp. upper) bound follows from the diminution of total entropy in the $s_L + s_R - 1$ cells on the left (resp. right cells) during the time step. Details are given in the proof of Lemma 1 in the previous section. □

Remark 2. The contraposition of Proposition 1 states that if there exists an initial data $(u^n_j)_j$ such that for all $(G^n_{j+1/2})_j$, $\mathcal{J}\left(\left(G^n_{j+1/2}\right)_j; \left(u^n_j\right)_j\right) > 0$ then the scheme is not entropy satisfying. We further exploit this fact in Section 5 to build another minimization procedure which constructs the worst initial data in terms of entropy and use it to prove that some schemes are non entropy satisfying.

Proposition 2. *Fix the initial data* $\left(u^n_j\right)_j$. *Suppose that there exists* $\left(G^{opt,n}_{j+1/2}\right)_j$ *such that the functional vanishes* $\mathcal{J}\left(\left(G^{opt,n}_{j+1/2}\right)_j; \left(u^n_j\right)_j\right) = 0$. *Then the corresponding scheme verifies*

$$\forall j \in [\![1, M]\!], \qquad \eta\left(u^{n+1}_j\right) \leq \eta\left(u^n_j\right) - \frac{\Delta t}{\Delta x}\left(G^{opt,n}_{j+1/2} - G^{opt,n}_{j-1/2}\right).$$

The scheme is also consistent in the sense that if $u^n_{j-s_L+1} = \dots = u^n_j = \dots = u^n_{j+s_R}$, *then* $G^{opt,n}_{j+1/2} = G\left(u^n_j\right)$.

Let us insist that this result only shows that there exist numerical entropy fluxes yielding to nonpositive numerical diffusion for the particular choice of initial data $\left(u^n_j\right)_j$, which does not mean that the scheme is always entropy satisfying.

Proof. Suppose that $\mathcal{J}\left(\left(G^{opt,n}_{j+1/2}\right)_j; \left(u^n_j\right)_j\right) = 0$. The first contribution $\mathcal{J}^D$ is zero thus

$$\forall j \in [\![1, M]\!], \quad \eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right) + \frac{\Delta t}{\Delta x}\left(G^{opt,n}_{j+1/2} - G^{opt,n}_{j-1/2}\right) \leq 0,$$

which is exactly (5). The term $\mathcal{J}^C$ is also zero, thus $m^n_{j+1/2} \leq G^{opt,n}_{j+1/2} \leq M^n_{j+1/2}$. If $\left(u^n_{j-s_L+1}, \dots, u^n_{j+s_R}\right) = \left(u^n_j, \dots, u^n_j\right)$, then $\left(\tilde{u}^{n,j+1/2}_k\right)_k$ is constant equal to u^n_j, as well as $\left(\hat{u}^{n,j+1/2}_k\right)_k$, so that $m^n_{j+1/2} = M^n_{j+1/2} = G\left(u^n_j\right)$ and $G^{opt,n}_{j+1/2} = G\left(u^n_j\right)$. □

3.2. Discrepancy between global minimizers

When the set where the functional $\mathcal{J}$ vanishes is nonempty, it most likely contains several solutions. We first quantify how close they are from each other.

Proposition 3. *Suppose that* $\mathcal{J}\left(\left(G^n_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$ *and* $\mathcal{J}\left(\left(\bar{G}^n_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$, *and that the numerical flux* $\mathcal{F}$ *used in* (4) *is consistent and Lipschitz regular. Then for all* j,

$$|G^n_{j+1/2}-\bar{G}^n_{j+1/2}|=\sum_{k\in\{j-s_L+1,\dots,j+s_R\}}O\left(|u^n_k-u^n_j|^2\right).$$

When the scheme is known to satisfy a given discrete entropy inequality and when the solution is smooth, this result shows that the difference between the numerical entropy flux found in the literature and the one returned by the minimization procedure is of order Δx^2. Note that the consistency property alone gives an order Δx.

Proof. Suppose that the minimization procedure has two different global minimizers $\mathcal{J}\left(\left(G^n_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$ and $\mathcal{J}\left(\left(\bar{G}^n_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$. Then $m^n_{j+1/2}\le G^n_{j+1/2}\le M^n_{j+1/2}$ and $m^n_{j+1/2}\le \bar{G}^n_{j+1/2}\le M^n_{j+1/2}$, thus $|\bar{G}^n_{j+1/2}-G^n_{j+1/2}|\le M^n_{j+1/2}-m^n_{j+1/2}$, where

$$\begin{aligned}M^n_{j+1/2}-m^n_{j+1/2}&=G\left(u^n_{j-s_L+1}\right)-G\left(u^n_{j+s_R}\right)\\&\quad+\frac{\Delta x}{\Delta t}\sum_{k=j-s_L-s_R+2}^{j+s_L+s_R-1}\eta\left(\tilde{u}^{n,j+1/2}_k\right)-\eta\left(\hat{u}^{n,j+1/2}_k\right).\end{aligned}\tag{11}$$

By convexity of the entropy η,

$$\begin{aligned}\eta\left(\hat{u}^{n,j+1/2}_k\right)&=\eta\left(\tilde{u}^{n,j+1/2}_k-\frac{\Delta t}{\Delta x}\left(\tilde{f}^{n,j+1/2}_{k+1/2}-\tilde{f}^{n,j+1/2}_{k-1/2}\right)\right),\\&\ge\eta\left(\tilde{u}^{n,j+1/2}_k\right)-\frac{\Delta t}{\Delta x}D\eta\left(\tilde{u}^{n,j+1/2}_k\right)\left(\tilde{f}^{n,j+1/2}_{k+1/2}-\tilde{f}^{n,j+1/2}_{k-1/2}\right).\end{aligned}$$

Combining this inequality with (11), we get

$$M^n_{j+1/2}-m^n_{j+1/2}\le G\left(u^n_{j-s_L+1}\right)-G\left(u^n_{j+s_R}\right)+\sum_{k=j-s_L-s_R+2}^{j+s_L+s_R-1}D\eta\left(\tilde{u}^{n,j+1/2}_k\right)\left(\tilde{f}^{n,j+1/2}_{k+1/2}-\tilde{f}^{n,j+1/2}_{k-1/2}\right),$$

which rewrites as

$$\begin{aligned}M^n_{j+1/2}-m^n_{j+1/2}&\le G\left(u^n_{j-s_L+1}\right)-G\left(u^n_{j+s_R}\right)+D\eta\left(u^n_j\right)\left(f\left(u^n_{j+s_R}\right)-f\left(u^n_{j-s_L+1}\right)\right)\\&\quad+\sum_{k=j-s_L-s_R+2}^{j+s_L+s_R-1}\left(D\eta\left(\tilde{u}^{n,j+1/2}_k\right)-D\eta\left(u^n_j\right)\right)\left(\tilde{f}^{n,j+1/2}_{k+1/2}-\tilde{f}^{n,j+1/2}_{k-1/2}\right).\end{aligned}$$

If the numerical flux is Lipschitz regular, the last sum is $\displaystyle\sum_{k=j-s_L+1}^{j+s_R}O\left(|u^n_k-u^n_j|^2\right)$. A first order expansion at point u^n_j of the other contributions is

$$\left(DG\left(u^n_j\right)-D\eta\left(u^n_j\right)Df\left(u^n_j\right)\right)\left(u^n_{j-s_L+1}-u^n_{j+s_R}\right)+O\left(|u^n_{j-s_L+1}-u^n_j|^2\right)+O\left(|u^n_{j+s_R}-u^n_j|^2\right).$$

The result then follows from $D\eta Df=DG$. □

3.3. Lax-Wendroff theorem

One of the main theoretical results about numerical schemes for systems of conservation laws is the Lax-Wendroff theorem, which states that if a numerical scheme converges in a certain sense, then the limit is a weak solution. In addition, if the scheme satisfies relevant discrete entropy inequalities, then the limit is an entropy solution.

The latter statement usually requires the numerical entropy flux to be a continuous and consistent function $\mathcal{G}$ of the neighboring approximations. This is not the case in this work, since the numerical entropy fluxes $G^n_{j+1/2}$ are obtained through a minimization procedure and we do not use a entropy flux function $\mathcal{G}$. However, it is possible to adapt the Lax-Wendroff theorem to our framework. This extension somehow means it is possible to replace the classical Definition 2 of entropy satisfying scheme by the requirement "for all initial data $\left(u^n_j\right)_j$, there exists $G^{opt,n}$ such that$\mathcal{J}\left(G^{opt,n},\left(u^n_j\right)_j\right)=0$". The latter condition differs on the treatment of the consistency property but leads to an identical Lax–Wendroff theorem. Note that proving such a result has no reason to be easier than finding an explicit function $\mathcal{G}$ and is possibly more difficult as $\mathcal{G}$ has s_L+s_R+1 variables while $\mathcal{J}$ is defined on $\mathbf{R}^M$.

Since the time discretization will be important in this section, the objects related to the optimization problem (7) at time t^n, *i.e.*, with initial data $\left(u^n_j\right)_j$, will be noted with a superscript n. The following Theorem is given on a bounded space domain $[a,b]$ with periodic boundary conditions since it is the framework of this paper. It can easily be generalized on the unbounded space domain $\mathbf{R}$.

For a discretization $\Delta=(\Delta x,\Delta t)$, we introduce the piecewise constant function u_Δ defined almost everywhere (a.e). in $[a,b]\times[0,+\infty)$ by

$$u_\Delta(x,t)=u^n_j,\quad x_{j-1/2}<x<x_{j+1/2},\quad t^n\le t<t^{n+1}.$$

Theorem 1. *Assume the scheme (4) is consistent in the sense of Definition 1, with $\mathcal{F}$ continuous. Consider a sequence $\Delta_k=(\Delta x_k,\Delta t_k)$ which converges to $(0,0)$ with the ratio $\lambda=\frac{\Delta t_k}{\Delta x_k}$ being constant. Assume that*

- *there exists a compact $K\subset\Omega$ such that $u_{\Delta_k}(x,t)\in K$ for a.e. $(x,t)\in[a,b]\times[0,+\infty)$ and for all $k\in\mathbf{N}$;*
- *the sequence u_{Δ_k} converges in $L^1_{loc}([a,b]\times(0,+\infty))$ to a function u.*

Then u is a weak solution of (1).

Furthermore, if for a given entropy pair (η,G), for all discretization $\Delta_k=(\Delta x_k,\Delta t_k)$ and for all $n\in\mathbf{N}$, there exists a family $\left(G^{opt,n}_{j+1/2}\right)_j$ such that $\mathcal{J}\left(\left(G^{opt,n}_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$, then the solution u satisfies the entropy inequality (2) in the sense of distributions on $[a,b]\times(0,+\infty)$.

Proof. For the sake of simplicity, the subscript k in Δ_k will be omitted all along the proof. The proof of the convergence of u_Δ toward a weak solution of (1) is exactly the same as in the original Lax-Wendroff theorem. The reader is referred for instance to [14, 19] for a complete proof.

Concerning the entropy inequality, for a fixed discretization Δ, let us consider a compactly supported test function $\varphi\in C^1_0([a,b]\times(0,+\infty))$, with $\varphi\ge0$. Let us notice that the function φ is assumed to vanish when $t=0$. This choice is made for the sake of simplicity of the proof. However, the extension to the case where φ does not vanish at $t=0$ is immediate.

Since $\mathcal{J}^D\left(\left(G^{opt,n}_{j+1/2}\right)_j;\left(u^n_j\right)_j\right)=0$, we have for all j and n

$$\eta\left(u^{n+1}_j\right)-\eta\left(u^n_j\right)+\lambda\left(G^{opt,n}_{j+1/2}-G^{opt,n}_{j-1/2}\right)\le0.$$

Multiplying this inequality by $\Delta x\varphi(x_j,t^n)$, summing over j and n and performing a summation by parts, we obtain

$$\Delta x\sum_{j,n}\eta\left(u^{n+1}_j\right)\left(\varphi\left(x_j,t^{n+1}\right)-\varphi(x_j,t^n)\right)+\Delta t\sum_{j,n}G^{opt,n}_{j+1/2}\left(\varphi(x_{j+1},t^n)-\varphi(x_j,t^n)\right)\ge0.$$

Introducing the piecewise constant functions

$$\begin{aligned}\varphi_\Delta(x,t) &= \varphi(x_j,t^n), \quad x_{j-1/2} < x < x_{j+1/2}, \quad t^n \le t < t^{n+1},\\ G_\Delta^{opt}(x,t) &= G_{j+1/2}^{opt,n}, \quad x_j < x < x_{j+1}, \quad t^n \le t < t^{n+1},\end{aligned}$$

where $x_j = \frac{x_{j-1/2}+x_{j+1/2}}{2}$, the last inequality writes

$$\int_{[a,b]\times[\Delta t,+\infty)} \eta(u_\Delta(x,t)) \frac{\varphi_\Delta(x,t)-\varphi_\Delta(x,t-\Delta t)}{\Delta t}\mathrm{d}x\mathrm{d}t + \int_{[a,b]\times\mathbf{R}^+} G_\Delta^{opt}(x,t) \frac{\varphi_\Delta(x+\Delta x/2)-\varphi_\Delta(x-\Delta x/2)}{\Delta x}\mathrm{d}x\mathrm{d}t \ge 0. \quad (12)$$

As in the classical Lax-Wendroff theorem, the first integral converges as $\Delta \to (0,0)$ to

$$\int_{[a,b]\times\mathbf{R}^+} \eta(u(x,t))\,\partial_t\varphi(x,t)\,\mathrm{d}x\mathrm{d}t.$$

The difference with the classical Lax-Wendroff theorem lies in the second integral of (12). First, let us introduce the following piecewise constant functions:

$$\begin{aligned}M_\Delta(x,t) &= M_{j+1/2}^n, \quad x_j < x < x_{j+1}, \quad t^n \le t < t^{n+1},\\ m_\Delta(x,t) &= m_{j+1/2}^n, \quad x_j < x < x_j, \quad t^n \le t < t^{n+1},\\ \tilde{u}_\Delta^{j+1/2}(x,t) &= \tilde{u}_k^{j+1/2}, \quad x_{k-1/2} < x < x_{k+1/2}, \quad t^n \le t < t^{n+1},\\ \tilde{f}_\Delta^{j+1/2}(x,t) &= \tilde{f}_{k+1/2}^{j+1/2}, \quad x_k < x < x_{k+1}, \quad t^n \le t < t^{n+1}, \quad (13)\\ \hat{u}_\Delta^{j+1/2}(x,t) &= \hat{u}_k^{j+1/2}, \quad x_{k-1/2} < x < x_{k+1/2}, \quad t^n \le t < t^{n+1}. \quad (14)\end{aligned}$$

Since φ is smooth, the term $\frac{\varphi_\Delta(x+\Delta x/2)-\varphi_\Delta(x-\Delta x/2)}{\Delta x}$ uniformly converges to $\partial_x\varphi(x,t)$. Moreover, according to (9) and $\tilde{u}_k^{n,j+1/2}$ and since the ratio λ is constant, $M_{j+1/2}^n$ and $m_{j+1/2}^n$ are continuous functions of a finite number of u_k^n, that all lie in the compact K. Therefore the functions M_Δ and m_Δ are uniformly bounded. Since $\mathcal{J}^C\left(\left(G_{j+1/2}^{opt,n}\right)_j;\left(u_j^n\right)_j\right)=0$, it follows that the inequalities

$$m_\Delta^n(x,t) \le G_\Delta^{opt}(x,t) \le M_\Delta^n(x,t) \quad (15)$$

hold for a.e. $(x,t)\in[a,b]\times[0,+\infty)$. As a consequence, the function G_Δ^{opt} is also uniformly bounded and therefore

$$\int_{[a,b]\times\mathbf{R}^+} G_\Delta^{opt}(x,t)\left(\frac{\varphi_\Delta\left(x+\frac{\Delta x}{2}\right)-\varphi_\Delta\left(x-\frac{\Delta x}{2}\right)}{\Delta x}-\partial_x\varphi(x,t)\right)\mathrm{d}x\mathrm{d}t \to 0. \quad (16)$$

Next we have

$$\begin{aligned}\tilde{u}_\Delta^{j+1/2}(x,t) &= u_\Delta\left(x+\min\left(\max\left(k,j-s_L+1\right),j+s_R\right)\Delta x - k\Delta x,t\right),\\ \tilde{f}_\Delta^{j+1/2}(x,t) &= \mathcal{F}\left(\tilde{u}_\Delta^{j+1/2}\left(x-(s_L-1/2)\Delta x,t\right),\cdots,\tilde{u}_\Delta^{j+1/2}\left(x+(s_R-1/2)\Delta x,t\right)\right),\\ \hat{u}_\Delta^{j+1/2}(x,t) &= \tilde{u}_\Delta^{j+1/2}(x,t)-\lambda\left(\tilde{f}_\Delta^{j+1/2}(x+\Delta x/2,t)-\tilde{f}_\Delta^{j+1/2}(x-\Delta x/2,t)\right). \quad (17)\end{aligned}$$

It follows from (9) that

$$M_\Delta(x,t) = G\left(u_\Delta\left(x-(s_L-1/2)\,\Delta x,t\right)\right) + \lambda \sum_{k=j-s_L-s_R+2}^{j} \left(\eta\left(\tilde{u}_\Delta^{j+1/2}\left(x-(j+1/2-k)\,\Delta x,t\right)\right) - \eta\left(\hat{u}_\Delta^{j+1/2}\left(x-(j+1/2-k)\,\Delta x,t\right)\right)\right). \tag{18}$$

Since u_Δ converges in $L^1_{loc}\left([a,b]\times(0,+\infty)\right)$ to u, then up to a subsequence, $\tilde{u}_\Delta^{j+1/2}$ converges to u a.e. Thanks to the continuity and the consistency of $\mathcal{F}$, we get from (13) that $\tilde{f}_\Delta^{j+1/2}$ converges a.e. to $f(u)$. Then according to (14) and (17), we deduce that $\hat{u}_\Delta^{j+1/2}$ also converges a.e. to u. Finally, it follows from (18) that M_Δ converges a.e. to $G(u)$. A similar process shows that up to a subsequence, m_Δ converges a.e. to $G(u)$. We deduce from (15) that up to a subsequence, G_Δ^{opt} converges a.e. to $G(u)$. The dominated convergence theorem then ensures that

$$\int_{[a,b]\times\mathbf{R}^+} G_\Delta^{opt}(x,t)\,\partial_x\varphi(x,t)\,\mathrm{d}x\mathrm{d}t \to \int_{[a,b]\times\mathbf{R}^+} G(u)\,\partial_x\varphi(x,t)\,\mathrm{d}x\mathrm{d}t. \tag{19}$$

Summing (16) and (19), we obtain

$$\int_{[a,b]\times\mathbf{R}^+} G_\Delta^{opt}(x,t)\,\frac{\varphi_\Delta(x+\Delta x/2)-\varphi_\Delta(x-\Delta x/2)}{\Delta x}\mathrm{d}x\mathrm{d}t \to \int_{[a,b]\times\mathbf{R}^+} G(u)\,\partial_x\varphi(x,t)\,\mathrm{d}x\mathrm{d}t.$$

Hence the limit of (12) writes

$$\int_{[a,b]\times\mathbf{R}^+} \eta(u(x,t))\,\partial_t\varphi(x,t)\,dxdx + \int_{[a,b]\times\mathbf{R}^+} G(u)\,\partial_x\varphi(x,t)\,\mathrm{d}x\mathrm{d}t \geq 0,$$

which concludes the proof. □

4. A posteriori quantification of the numerical diffusion: numerical results

We now apply the minimization of the functional $\mathcal{J}$ to obtain maps of numerical diffusion in different settings[1]. The numerical entropy fluxes $G_{j\pm1/2}^{opt,n}$ are the results of the minimization (7) and we define the corresponding a posteriori numerical diffusion according to (6) by

$$D_j^{opt,n} = \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^{opt,n} - G_{j-1/2}^{opt,n}\right). \tag{20}$$

In practice the minimization is performed using the blackbox fminunc of MATLAB [23], with the initial guess $\gamma_{j+1/2} = \frac{1}{2}\left(m_{j+1/2}^n + M_{j+1/2}^n\right)$. The values for $D_j^{opt,n}$ are often small, in particular on fine meshes or when the solution is smooth. However, these values remain far away from the machine precision estimated at 10^{-16}.

4.1. Continuous solution of the Burgers equation

In this first test case, we consider the scalar ($d=1$) Burgers equation

$$\partial_t u + \partial_x\left(\frac{u^2}{2}\right) = 0 \quad \text{and} \quad \partial_t\left(u^2\right) + \partial_x\left(\frac{2u^3}{3}\right) \leq 0. \tag{21}$$

[1] In all the following numerical tests, we use a personal laptop, equipped with a 2.6 GHz Intel Core i7 6-core processor of 16 GB and 2667 MHz DDR4 memory.

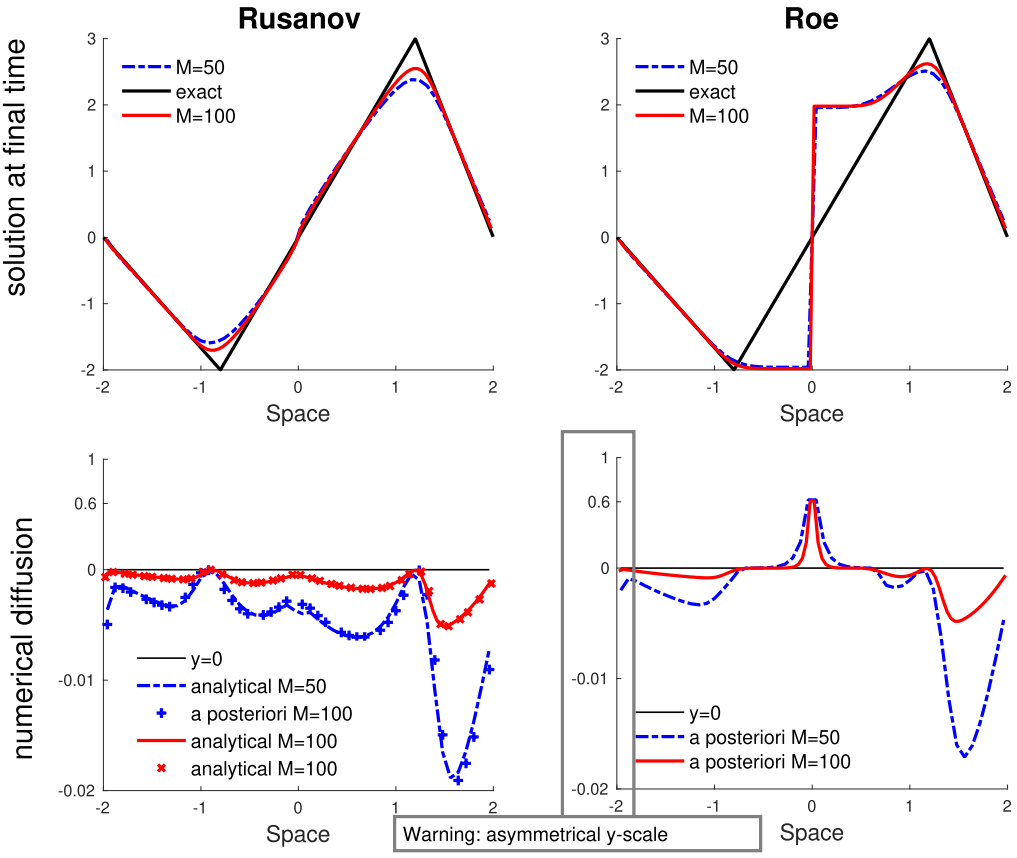


FIGURE 2. Quantification of the numerical diffusion on Testcase (22) for the entropy satisfying Rusanov scheme and the entropy violating Roe scheme.

The choice of $\eta(u) = u^2$ is arbitrary; for the scalar Burgers equation any convex entropy η could be considered. The initial data is

$$u^0(x) = \begin{cases} -2 - x & \text{if } -2 < x \leq 0, \\ 3 - \frac{3}{2}x & \text{if } 0 < x \leq 2, \end{cases} \tag{22}$$

and we use periodic boundary conditions. The exact solution is piecewise linear, and for $t < \frac{2}{3}$ is given by

$$u(x,t) = \begin{cases} \dfrac{-(x+2)}{1-t} & \text{if } -2 \leq x \leq -2t, \\ \dfrac{x}{t} & \text{if } -2t \leq x \leq 3t, \\ \dfrac{-3(x-2)}{2-3t} & \text{if } 3t \leq x \leq 2. \end{cases}$$

The space interval $[-2, 2]$ is discretized with 50 or 100 cells, and the functional $\mathcal{J}$ is minimized at the last iteration at $T = 0.4$. The parameter α in the Courant–Friedrichs–Levy condition (3) is set to $\alpha = 0.5$.

We compared the entropy satisfying Rusanov scheme and the entropy violating Roe scheme with a forward Euler march in time, see the appendix for the definitions of the corresponding fluxes $\mathcal{F}$. The stencil is $s_L = s_R = 1$. The minimization procedure detects the two previous behaviors: $D_j^{opt,n}$ remains nonpositive everywhere for

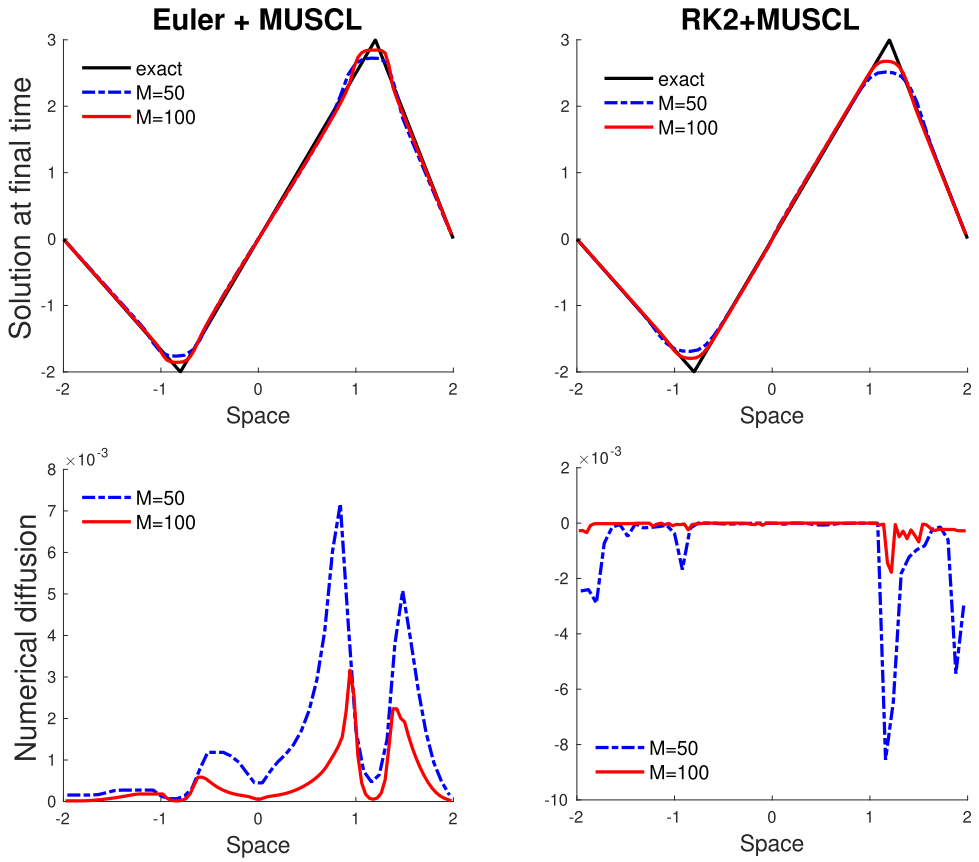


FIGURE 3. Influence of the time discretization on the numerical diffusion for the MUSCL scheme Testcase (22).

the Rusanov scheme, while it has large strictly positive values localized around the stationary entropy creating shock at the sonic point for the Roe schemes.

The two plots of Figure 3 show the results for the MUSCL second order flux in space based on a Rusanov flux and a minmod limiter. We illustrate here the well-known fact that increasing the order in space without increasing the order in time is detrimental in terms of entropy, whereas increasing both orders is efficient. The reader will find more specific details related to the MUSCL scheme in the following section. Together with a forward Euler march in time, (5) is grossly false in the sense that the total entropy increases with time: $\sum_j \eta\left(u_j^{n+1}\right) > \sum_j \eta\left(u_j^n\right)$. Combined with a second order Runge-Kutta time stepping, the a posteriori procedure finds a discrete entropy inequality (5).

To conclude this section, let us illustrate Proposition 3. The Rusanov scheme with an Euler method is entropy satisfying. With the choice of $\mathcal{G}$ given in the appendix (A.2) the discrete entropy inequality (5) holds. We denote by D_j^n the associated nonpositive numerical diffusion (6). We compare the total discrepancy $\Delta x \sum \left|D_j^{opt,n} - D_j^n\right|$ between this quantity and the one obtained through minimization (20) and obtained as expected that is decreases one order faster than the total diffusion $\Delta x \sum_j \left|D_j^n\right| = \Delta x \sum_j \left|\eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right)\right|$ as the mesh becomes finer and finer. Note that this quantity does not depend on the numerical entropy fluxes $G_{j+1/2}^n$.

4.2. Gas dynamics

We now turn to the Euler equations. The unknown is $u = (\rho, \rho v, E)$, where ρ is the density of the fluid, v is its velocity and E its total energy. This system reads

$$\begin{cases} \partial_t \rho + \partial_x (\rho v) = 0 \\ \partial_t (\rho v) + \partial_x \left(\rho v^2 + p\right) = 0 \\ \partial_t E + \partial_x (v (E + p)) = 0. \end{cases} \tag{23}$$

The pressure force p is related to ρ and E through an ideal gas equation of state

$$p = (\gamma - 1) \left(E - \frac{\rho v^2}{2} \right) \quad \text{where} \quad \gamma \in (1, 3] .$$

Both the density and the pressure should remain nonnegative, thus

$$\Omega = \left\{ (\rho, \rho v, E) \in \mathbf{R}^3 : \ \rho \geq 0 \ \text{ and } \ E \geq \frac{\rho v^2}{2} \right\}.$$

There exists an infinite number of entropy inequalities for these equations, see [1]. We consider the classical inequality on the specific entropy $s = \dfrac{p}{\rho^\gamma}$

$$\partial_t (-\rho \ln (s)) + \partial_t (-v \rho \ln (s)) \leq 0. \tag{24}$$

The minimization method for quantifying the numerical diffusion does not depend on the number of unknowns d, since the quantification of the numerical diffusion only concerns the scalar equation on the entropy evolution (24). We focus on a widely chosen scheme of order two in space and time, namely the Van Leer version of the MUSCL scheme [25].

In this scheme the piecewise constant in space approximation $\left(u_j^n\right)_j$ is replaced by a reconstructed piecewise affine data. For scalar equations $d = 1$, the reconstruction procedure heavily relies on the fact that the exact solutions of (1) verify a maximum principle property and are total variation diminishing (TVD). A family of functions called limiters is considered to determine the slopes in each cell [14] and allows to keep these features at the discrete level. Both properties are lost for hyperbolic systems $d \geq 2$. Limiters are also used but several choices are possible. We investigate the effects of some of them in this section.

It remains to decide how the evolution in time of the reconstructed initial data is performed. A first possibility is to compute the exact flux of (1) by solving generalized Riemann problems. This is only feasible for some particular hyperbolic systems and very costly from a computational point of view. This path is followed in [4] for scalar equations and in [9] for systems. A more convenient approach is to combine the reconstruction in space with a second order method in time (typically a second order Runge-Kutta method (RK2)) to obtain a precise enough approximation of the fluxes. Each individual substep is based on a first order solver.

In any case the discrete entropy inequalities found in the literature for the MUSCL approach differ from (5). The quantity $\eta\left(u_j^n\right)$ is replaced with a linear approximation in equation (1.9) of [4] for scalar equations or with a nonlinear entropy diminishing projection operator in Theorem 2.9 of [9] for systems. In [1], $\eta\left(u_j^n\right)$ is replaced by a convex combination of three terms that not only depends on u_j^n but also on u_{j-1}^n and u_{j+1}^n, see equations (2.7) and (2.10) of [1]. None of these modified entropy inequality is sufficient to prove a Lax–Wendroff theorem. Several numerical simulations [2] even indicate that the MUSCL+RK2 scheme may converge to incorrect solutions.

We first reproduce the Sod tube testcase of Toro [24]

$$\begin{cases} \rho^0 (x) = \mathbf{1}_{x<0} + 0.125 \times \mathbf{1}_{x \geq 0}, \\ u^0 (x) = 0.75 \times \mathbf{1}_{x<0}, \\ p^0 (x) = \mathbf{1}_{x<0} + 0.1 \times \mathbf{1}_{x \geq 0}, \end{cases} \tag{25}$$

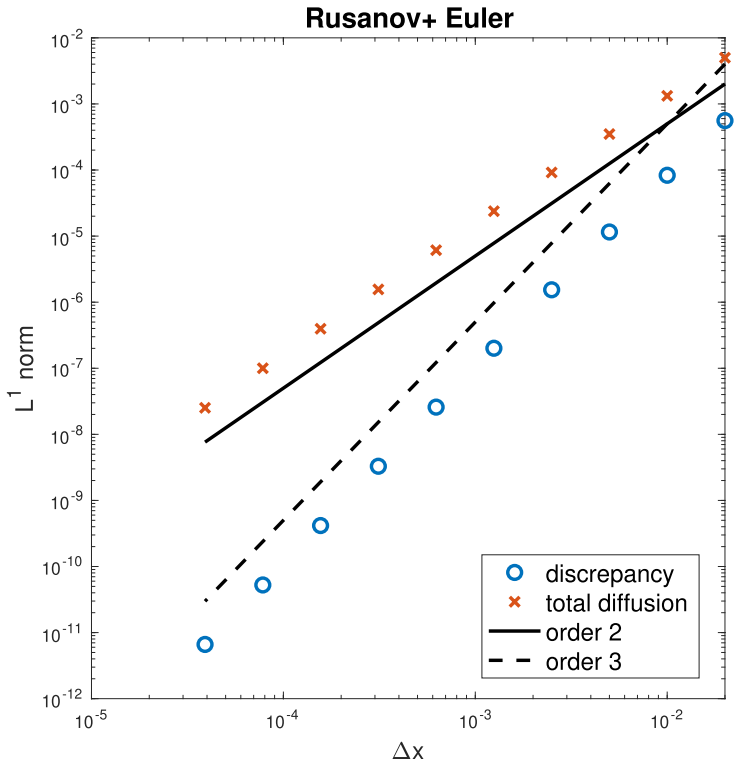


FIGURE 4. Total numerical diffusion and total discrepancy between two numerical diffusions on the Testcase (22).

on the time interval $[0, 0.2]$ and the space interval $[-1, 1]$. The CFL number is $\alpha = 1/6$ and the number of cells is $M = 400$. The discontinuity at $x = 0$ creates a 1-rarefaction wave, a 2-contact discontinuity and a 3-shock. We stick to periodic boundary conditions, so there is another discontinuity at $x = -1$. It creates a 1-shock, a 2-contact discontinuity and a 3-rarefaction wave.

On Figure 5 we compare the first order HLLC scheme Section 10.4.2 of [24] and the Roe scheme without entropy fix Section 11.2 of [24] with a forward Euler time stepping. Then we consider the second order MUSCL scheme with a RK2 time stepping. The slopes are limited on the primitive variable (ρ, u, p) and the underlying first order scheme is the HLLC scheme. We compare the results obtained with a minmod limiter and a superbee limiter.

The a posteriori quantification of the numerical diffusion gives once again positive values near the stationary nonphysical shock created by the Roe scheme. It also detects the overcompressive behavior of the superbee limiter, with a spike of positive numerical diffusion located on the oscillations in density around the central contact discontinuity. The superbee limiter is often too strong and may prevent the scheme from converging, see [2]. For second order schemes, the numerical diffusion is located around one or two spikes around each discontinuity. This depends on the initialization of the optimization algorithm, see Figure 8 below.

Then we consider a testcase where the solution does not contain a shock, but only a contact discontinuity. The velocity and the pressure are initially constant, equal to 0.1 and 1. The initial density is

$$\rho^0(x) = 1 + 0.2x + 0.05\sin(6\pi x) + 0.4 \times \mathbf{1}_{x<0}. \tag{26}$$

The final time is 2 seconds, the CFL number is $\alpha = 0.75$. On Figure (6), we compare the HLL scheme and the HLLC scheme and find unsurprisingly that the latter is much less diffusive. As usual, we observe this fact

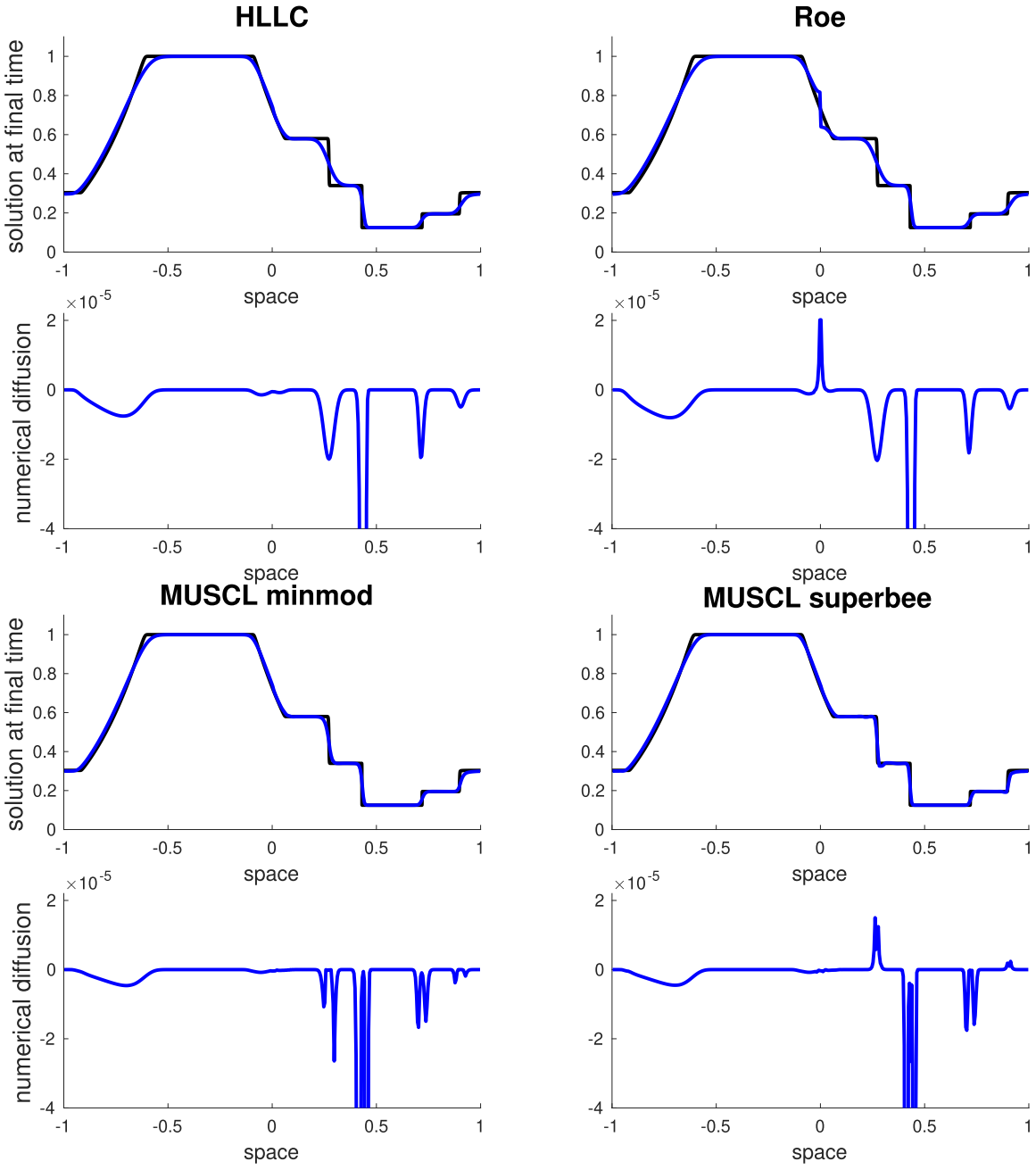


FIGURE 5. Densities (lines 1 and 3) and a posteriori numerical diffusion (lines 2 and 4) for several numerical schemes on Testcase (25).

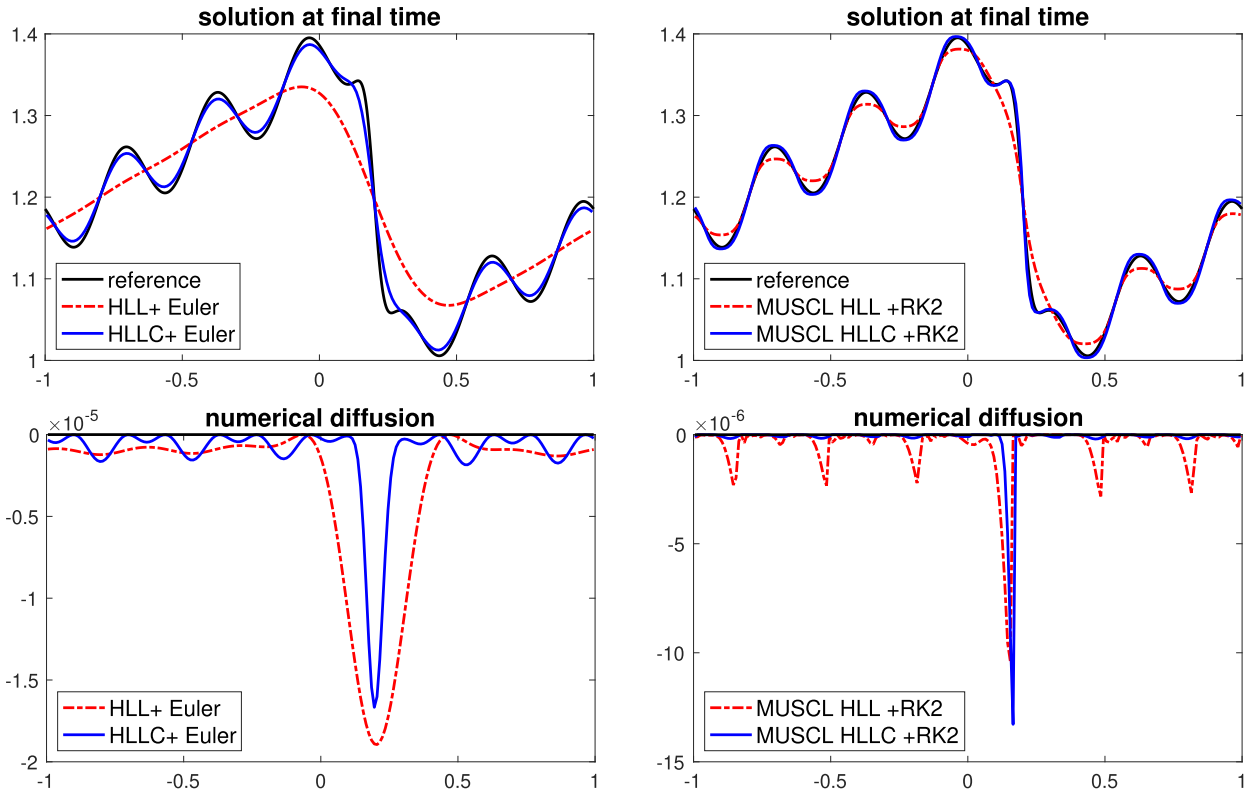


FIGURE 6. Densities (first line) and numerical diffusion (second line) for Testcase (26). The discontinuity is a slowly moving contact.

through the higher level of detail captured by the HLLC scheme at the final time. The quantification of numerical diffusion confirms it quantitatively, with lower and more localized $D^{opt,n}$ values. The same holds for their second order extensions with a MUSCL procedure, using either the HLL or the HLLC flux as the underlying first order scheme. The slope limitation is on the conservative variables $(\rho, \rho u, E)$ and we use a minmod limiter.

4.3. A naive *a priori* quantification of the numerical diffusion

In this section we turn to the practical consideration of the computational cost of the minimization presented above. The functional $\mathcal{J}^D$ couples all the cells of the mesh because each numerical entropy flux $\gamma_{j+1/2}$ appears in two consecutive terms of the sum. The cost grows quadratically with the meshsize, and becomes larger than the minute for meshes with more than 4000 cells, see Figure 7, left.

This observation disqualifies the use of this procedure in realistic codes were the quantification of D_j^n would be useful. In ocean global circulation models (OGCMs) for example, the meshes contain billion of cells and the numerical diffusion under scrutiny here is strongly linked with the question of spurious mixing identified as a key issue in such codes [13]. This explains why several works deal with its quantification. Some of them ([5, 18] and others) are directly related to the mathematical theory of discrete entropy inequality in the sense that specific choices of numerical entropy fluxes are imposed in (6) to defined the numerical diffusion D_j^n. Here we propose a complementary approach: while these works focus on $G_{j+1/2}^n$ at the risk of making a suboptimal choice and quantifying D_j^n incorrectly, we present a more robust choice of D_j^n at the cost of forgetting the numerical entropy fluxes $G_{j+1/2}$. As a consequence the theoretical results of the previous section are almost completely lost; however the quantitative results are very satisfactory and the computational cost is drastically

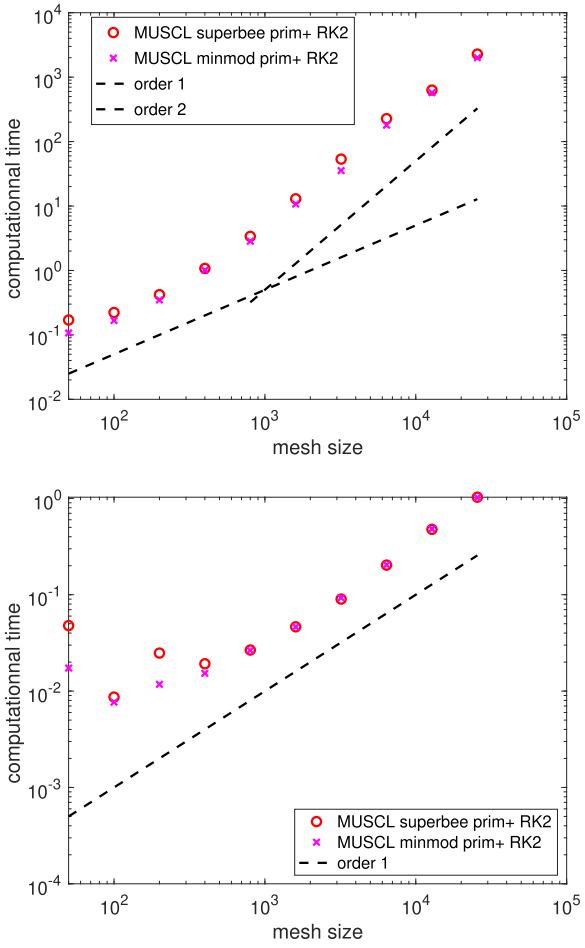


FIGURE 7. Computational cost of the minimization procedure (left) and of the *a priori* guess (28) at the first iteration of Testcase (26).

diminished, see Figure 7, right. We believe this fastest quantification could be used for some applications in the future, including the quantification of spurious mixing in OGCM or as a metric for mesh refinement.

In the derivation of the consistency part of the functional $\mathcal{J}^C$, we proposed bounds on the numerical entropy flux $G^n_{j+1/2} \in [m^n_{j+1/2}, M^n_{j+1/2}]$. Under the hypothesis that the bounds are correctly ordered, it follows that $\underline{D}^n_j \leq D^{opt,n}_j \leq \overline{D}^n_j$, with

$$\underline{D}^n_j = \eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right) + \lambda\left(m^n_{j+1/2} - M^n_{j-1/2}\right) \tag{27}$$

and

$$\overline{D}^n_j = \eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right) + \lambda\left(M^n_{j+1/2} - m^n_{j-1/2}\right).$$

An important point is that $\underline{D}^n_j$ and $\overline{D}^n_j$ are computationally affordable, see right of Figure 7. Indeed, $m^n_{j+1/2}$ and $M^n_{j+1/2}$ are computed with the scheme (4) on a small initial data centered around the interface $j+1/2$ with $s_L + s_R$ cells on its left and on its right, see Lemma 1.

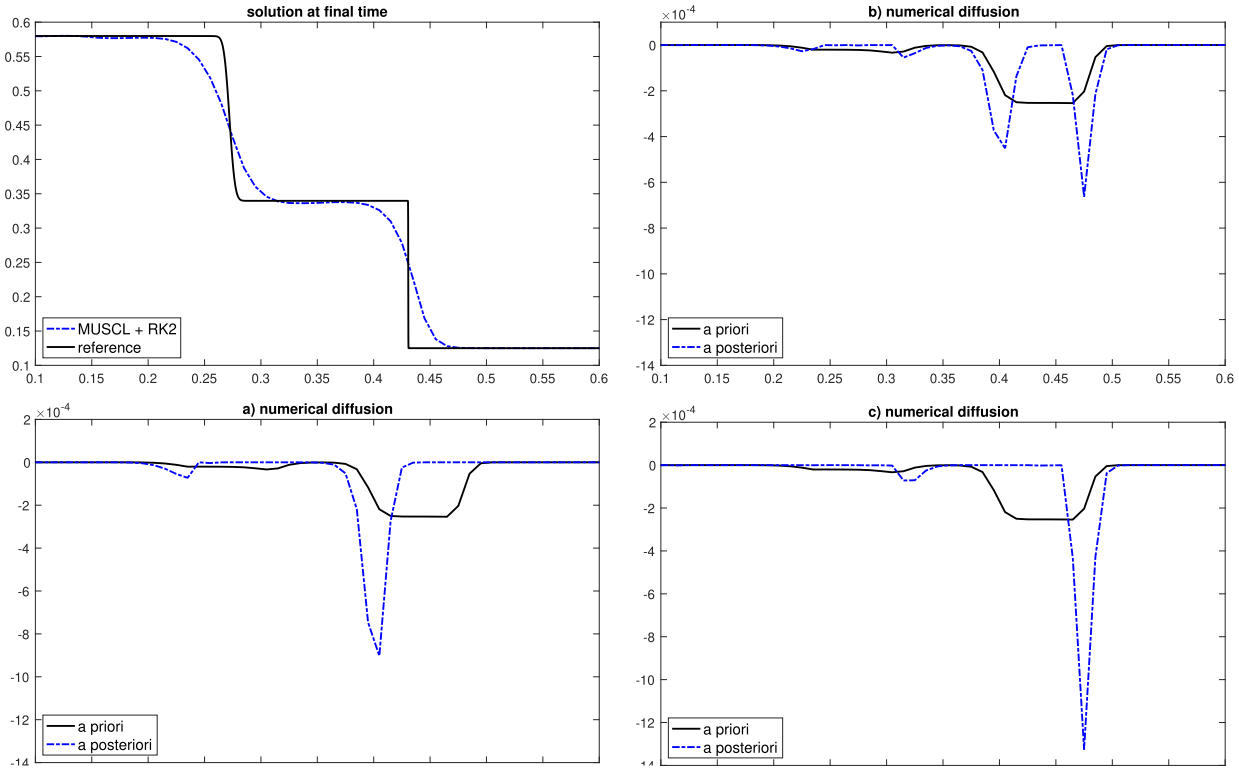


FIGURE 8. Comparison of $D_j^{opt,n}$ and $D_j^{a\,priori,n}$ on Testcase (25). For the computation of $D_j^{opt,n}$ the optimization is initialized with (a) $\gamma_{j+1/2}^n = m_{j+1/2}^n$ or (b) $\gamma_{j+1/2}^n = \frac{1}{2}\left(m_{j+1/2}^n + M_{j+1/2}^n\right)$ or (c) $\gamma_{j+1/2}^n = M_{j+1/2}^n$.

We observe that the lower bound (27) is particularly interesting for two reasons. First, if the left hand side is positive it indicates that (5) cannot hold on this particular cell. We detail and capitalize on that idea in Section 5. Second, even though it is much lower than $D_j^{opt,n}$, the variations of $\underline{D}_j^n$ are similar.

As a final stage, we perform a naive renormalization by a constant coefficient α and define $D_j^{a\,priori,n} = \alpha \underline{D}_j^n$ in such a way that the total amount of numerical diffusion is correct:

$$\sum_j D_j^{a\,priori,n} = \alpha \sum_j \underline{D}_j^n = \sum_j \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right).$$

It yields to the *a priori* quantification of the numerical diffusion

$$D_j^{a\,priori,n} = \frac{\sum_k \eta\left(u_k^{n+1}\right) - \eta\left(u_k^n\right)}{\sum_k \underline{D}_k^n} \underline{D}_j^n. \tag{28}$$

On Figure 8 we compare $D_j^{a\,priori,n}$ and $D_j^{opt,n}$. The a posteriori quantification $D_j^{opt,n}$ depends on the initialization of the minimization procedure and produces narrow spikes of numerical diffusion. The *a priori* quantification $D_j^{a\,priori,n}$ gives smoother results, with a diffusion spread out the whole length of the discrete

discontinuity. The computation of this quantity only requires the bounds $m^n_{j+1/2}$ and $M^n_{j+1/2}$ and does not require any optimization. We insist on the fact $D_j^{a\,priori,n}$ is pertinent if one is only interested in the numerical diffusion and not in the numerical entropy fluxes, which are not provided by this approach.

Remark 3. One could think that the upper bound $\bar{D}^n_j$ is also interesting. In particular if this quantity was nonpositive in all cells it would be a strong indication that the scheme is entropy satisfying. However, we observed that $\bar{D}^n_j$ is in most cases strictly positive, even for first order entropy satisfying scheme. It makes this quantity far less interesting both qualitatively and theoretically.

5. An entropy stress test for numerical schemes

In this section we go back to the classical question of the existence of discrete entropy inequality (5). We explore further the implication of the *a priori* consistency bounds $m^n_{j+1/2} \leq G^n_{j+1/2} \leq M^n_{j+1/2}$. It leads to a novel criterion and an algorithm which allows us to detect that entropy inequality (5) is not verified.

5.1. Worst initial data in terms of entropy

Proposition 4. *Consider a numerical flux $\mathcal{F}$ with a stencil of s_L points on the left and s_R on the right. If there exists an initial data*

$$\left(\dots, u^0_{-s_L}, u^0_{-s_L}, u^0_{-s_L+1}, \dots, u^0_0, \dots, u^0_{s_R-1}, u^0_{s_R}, u^0_{s_R}, \dots\right) \tag{29}$$

such that

$$\min\left(\eta\left(u^0_0\right) - \frac{\Delta t}{\Delta x}\left(m^0_{1/2} - M^0_{-1/2}\right) - \eta\left(u^1_0\right), M^0_{1/2} - m^0_{1/2}\right) < 0, \tag{30}$$

where $M^0_{1/2}$, $m^0_{1/2}$ and $M^0_{-1/2}$ are given by (9) with $n = 0$, then the finite volume scheme does not have a discrete entropy inequality.

Proof. By contraposition, let us assume that the scheme has a discrete entropy inequality. Therefore, for all initial data $\left(u^n_j\right)_j$, there exists consistent entropy numerical fluxes $\left(G_{j+1/2}\right)_j$ such that

$$\eta(u^1_j) \leq \eta(u^0_j) - \frac{\Delta t}{\Delta x}\left(G_{j+1/2} - G_{j-1/2}\right).$$

According to Lemma 1, we have $m^0_{j+1/2} \leq G_{j+1/2} \leq M^0_{j+1/2}$. It follows especially $m^0_{1/2} \leq M^0_{-1/2}$ and

$$\eta(u^1_0) \leq \eta(u^0_0) - \frac{\Delta t}{\Delta x}\left(m^0_{1/2} - M^0_{-1/2}\right).$$

Therefore the opposite of (30) is satisfied. This is true in particular for the initial data of the form (29). □

As a consequence, to determine if a scheme is entropy satisfying or not, we introduce the new functional

$$\mathcal{E}(u^0_{-s_L}, \dots, u^0_{s_R}) = \min\left(\eta\left(u^0_0\right) - \frac{\Delta t}{\Delta x}\left(m^0_{1/2} - M^0_{-1/2}\right) - \eta\left(u^1_0\right), m^0_{1/2} - M^0_{1/2}\right).$$

For a hyperbolic system of d equations, it has $d(s_L + s_R + 1)$ unknowns in Ω. Following Proposition 4, if the minimization of $\mathcal{E}$ returns a strictly negative result then the scheme is not entropy satisfying. In other words this new optimization constructs the worst initial data in terms of entropy.

Remark 4. The first part of (30) is similar to the definition (27) of $\underline{D}^0_0$. We already knew that if this quantity is positive the scheme is not entropy satisfying. This result goes further in that direction by restricting the size of the data to $s_L + s_R + 1$ values and by shifting the optimization procedure on the initial data.

TABLE 1. Summary of the obtained results for different variants of MUSCL scheme.

	1st order scheme	reconstructed variables	entropy satisfying	most negative value of $\mathcal{E}$
(a)	HLL	conservative $(\rho, \rho u, E)$	no	-2.228
(b)	HLL	primitive (ρ, u, p)	no	-4.352×10^{-5}
(c)	HLL	entropy (ρ, u, s)	yes?	≥ 0
(d)	HLLC	conservative $(\rho, \rho u, E)$	no	-5.257×10^{-1}
(e)	HLLC	primitive (ρ, u, p)	no	-3.518×10^{-4}
(f)	HLLC	entropy (ρ, u, s)	no	-2.73×10^{-4}

5.2. Exploration of the RK2+MUSCL scheme

We apply this procedure to the gas dynamic (23) approximated by the common MUSCL scheme in space and a two step Runge-Kutta march in time. Some details are given in Section 4.2 and in the appendix. The six versions have been tested; counterexamples are constructed for five of them. The results are summarized in Table 1.

More precisely, we consider the limitations in primitive variables (ρ, v, p), entropy variables (ρ, v, s) and conservative variables (ρ, q, E) described in [1] where the positiveness of the pressure and density are guaranteed. The underlying first order scheme is either the Rusanov scheme, the HLL or the HLLC scheme. We implemented two versions of the latter: one based on a pressure estimate and the other one on the estimation of extremal wave speeds. The CFL number is fixed to $\alpha = 0.1$. Details and references are given in the appendix. For the overall scheme we have $s_L = s_R = 4$ and the minimization of Proposition 4 has a total of 18 parameters with the positiveness of density and pressure as constraints.

We focus on counterexamples with small total variation. The unknowns are constrained to

$$\forall j \in \{-4, \cdots, 4\} \quad \left(\rho_j^0, u_j^0, p_j^0\right) \in (\rho_0, u_0, p_0) + [-0.1, 0.1]^3, \tag{31}$$

where (ρ_0, u_0, p_0) is selected randomly in a larger domain. The main reason for considering almost constant initial data is that if

$$u^0_{-s_L} = u^0_{-s_L+1} = \ldots u^0_0 = \cdots = u^0_{s_R-1} = u^0_{s_R},$$

then $M^0_{\pm 1/2} = m^0_{\pm 1/2}$, $\eta\left(u_0^1\right) = \eta\left(u_0^0\right)$ and the three components of Proposition 4 vanish. Thus counterexamples may be found in the vicinity of constant states if the scheme is not entropy satisfying. This region somehow corresponds to regular solutions.

We chose randomly 30 000 initializations in the region

$$(\rho_0, u_0, p_0) \in [0.11, 5] \times [-0.2, 10] \times [0.11, 10] \tag{32}$$

and another 30 000 initializations in the smaller region

$$(\rho_0, u_0, p_0) \in [0.8, 1.2] \times [-0.2, 1] \times [0.8, 1.2]. \tag{33}$$

On local extremum the MUSCL scheme is identical to the chosen first order scheme. Thus it seems more likely to find counterexamples on monotonic data. Half of the initialization are modified so that each variable is increasing or decreasing. The rearrangement is made either in the primitive, entropy or conservative variables, in equal proportion.

The minimization procedure constructs numerous initial data that violate (5) in the first time step. They are gathered on Figure 9 when (ρ_0, u_0, p_0) belongs to the large region (32) and on Figure 10 when it belongs to the smaller region (33). Each point on these Figures represents the mean value

$$\left(\frac{1}{9}\sum_{k=-4}^{4} \rho_j^0, \frac{1}{9}\sum_{k=-4}^{4} v_j^0, \frac{1}{9}\sum_{k=-4}^{4} p_j^0\right)$$

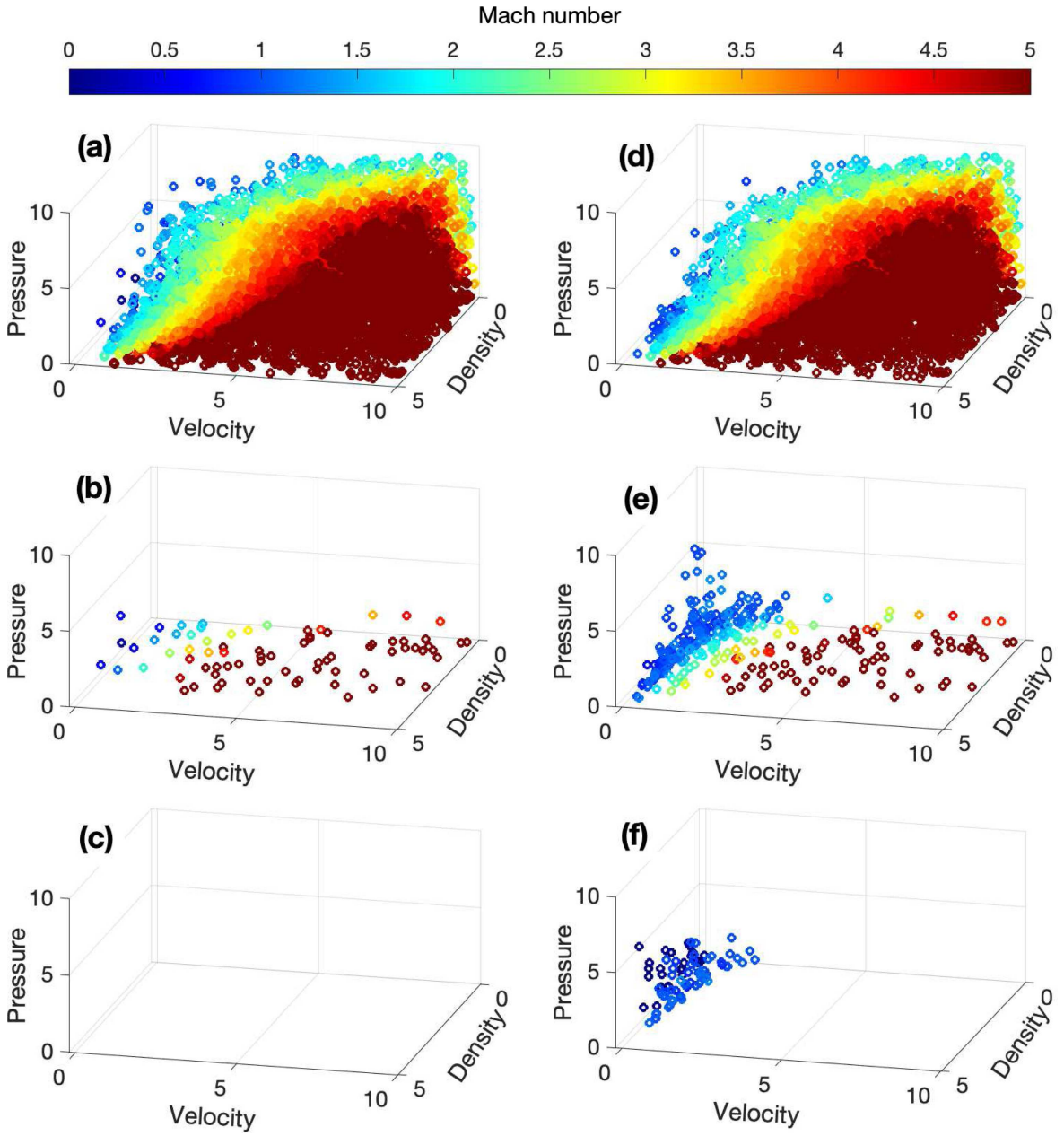


FIGURE 9. Counterexamples on the region $[0.11, 5] \times [-0.2, 10] \times [0.11, 10]$. The limited variables are (ρ, q, E) in the first line, (ρ, u, p) in the second line, (ρ, u, s) in the third line. The first order underlying scheme is HLL on the left and HLLC on the right.

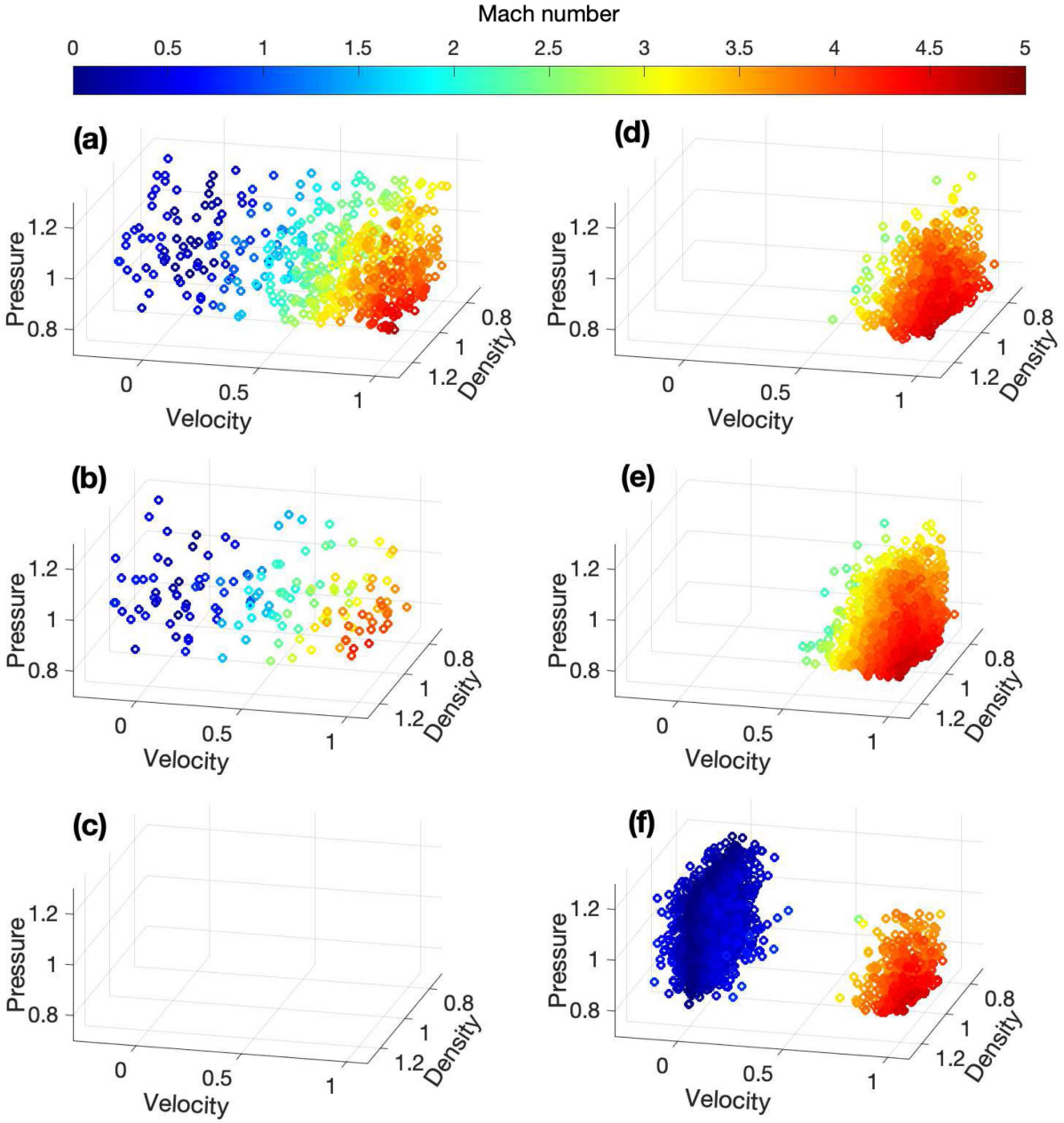


FIGURE 10. Zoom on a the region $[0.8, 1.2] \times [-0.2, 1] \times [0.8, 1.2]$. The limited variables are (ρ, q, E) in the first line, (ρ, u, p) in the second line, (ρ, u, s) in the third line. The first order underlying scheme is HLL on the left and HLLC on the right.

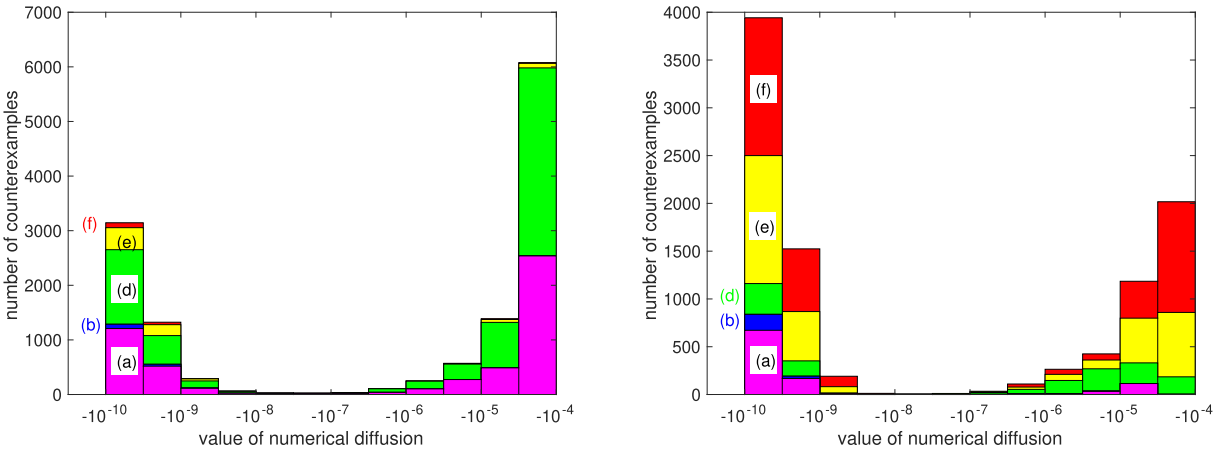


FIGURE 11. Histogram of the values of $\mathcal{E}(u^0_{-4}, \dots, u^0_4)$ for the counterexamples found in the large region (32) (left) and in the small region (33) (right). A small number of counterexamples for scheme (b) and (f) (left) or (a) and (b) (right) are found in the region $\mathcal{E} \approx -10^{-4}$. The histogram are stopped artificially at -10^{-4}.

of a counterexample. The results for the two versions of the HLLC schemes are very similar and we include only the version based on the wave-speed estimates. Similarly the Rusanov and HLL schemes behave similarly and we only include the HLL scheme. We used the fmincon of MATLAB [23] to minimize the functional $\mathcal{E}$ under the constraints (31). The histograms of the values corresponding to (30) for those counterexamples are given on Figure 11. We only kept counterexamples where this value is greater than -10^{-10} which is several order of magnitude larger than the machine precision estimated at 2.2×10^{-16}.

The influence of the choice of the set variables which are limited is striking on Figure 9, with more counterexamples for the conservative variables (ρ, q, E) than for the primitive variables (ρ, v, p). There are even fewer counterexamples for the entropy variables (ρ, v, s). This hierarchy is less clear on the zoom of Figure 10. The distribution on counterexamples around $|v| = 0$ depends on the numerical choices.

The counterexample isolated on Figure 12 has a very small total variation and a null velocity. We checked that it remains a counterexample for smaller and smaller timestep (or equivalently for smaller and smaller CFL number). This may indicate that the limit $\Delta t \to 0$ in (5) does not hold for the MUSCL+RK2 scheme based on a HLL Riemann solver with a limitation on the conservative variables.

On the other hand we see many counterexamples with large Mach number $|v|/\sqrt{\frac{\gamma p}{\rho}}$ on Figure 9. The chosen first order scheme is not to blame for the lack of discrete entropy inequality. Indeed in that case the exact flux at interface $j + 1/2$ is either $f\left(u^0_j\right)$ or $f\left(u^0_{j+1}\right)$ depending on the sign of the velocity. Most numerical schemes reproduce that.

Eventually Functional (30) remains nonnegative for all the random initializations when the limitation is in the entropy variable (ρ, v, s) and the first order scheme is HLL or Rusanov. To explore further if this scheme is entropy satisfying, we relaxed the constraint on the total variation and searched for counterexamples in the much larger domain

$$\forall j \in \{-4, \cdots, 4\} \quad \left(\rho^0_j, u^0_j, p^0_j\right) \in [0.001, 10] \times [-50, 50] \times [0.001, 20].$$

This search was unsuccessful with 57 000 random initializations. Interestingly it holds for the overlimited version of the MUSCL scheme of [1] but also for the simpler original scheme. It supports the following conjecture.

Conjecture 1. *The RK2+MUSCL scheme for the gas dynamic equations* (23) *with*

ρ	0.9985	0.9991	0.9998	0.9998	1.0005	1.0005	1.0009	1.0018
v	0	0	0	0	0	0	0	0
p	1.2519	1.2531	1.2549	1.2566	1.2584	1.2601	1.2618	1.2635

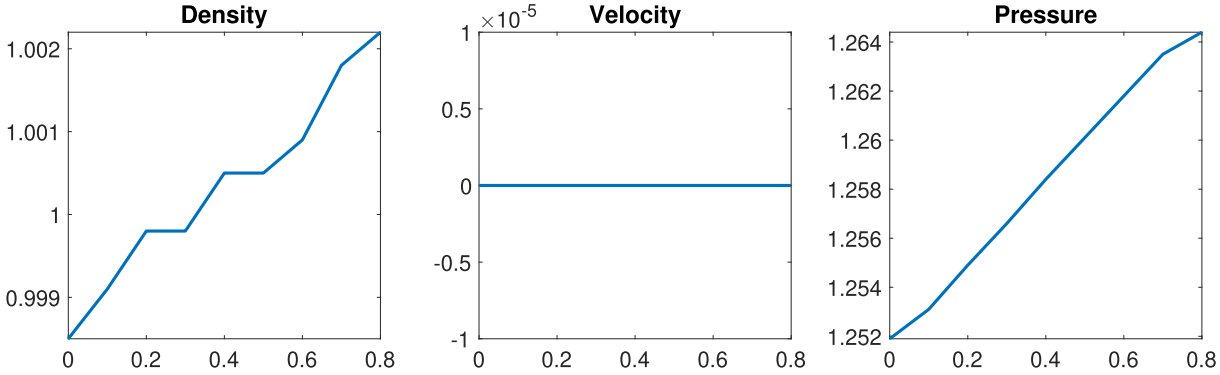


FIGURE 12. MUSCL+RK2 scheme with this initial data is non entropy satisfying, with a HLL first order scheme, a minmod limiter and a reconstruction in (ρ, u, p). In this case $\mathcal{E}(u^0_{-4}, \ldots, u^0_4) \approx -4.10^{-9}$.

(1) a MUSCL scheme in space based on a HLL first order scheme with a minmod limiter on the entropy variables (ρ, u, s)
(2) the RK2 march in time described in the Appendix

verifies a discrete entropy inequality for a CFL number $\alpha = 0.1$.

The same type of numerical experiment has been performed for the scalar Burgers equation (21). Three entropy satisfying two points schemes (Rusanov, Osher and Godunov) and three non entropy satisfying schemes (Roe, Lax-Wendroff and Mac Cormack) have been tested. We performed 30 000 optimization of the functional $\mathcal{E}$ and found no negative values for the entropy satisfying schemes, and counterexamples for the non entropy satisfying schemes, which supports the following conjecture.

Conjecture 2. *Consider the RK2+MUSCL scheme for the Burgers equation (21) with a minmod limiter on u and the RK2 march in time described in the appendix. For a Courant number up to 1, this scheme is entropy satisfying when the chosen underlying first order scheme is entropy satisfying, and non entropy satisfying when it is not.*

APPENDIX A. DESCRIPTION OF THE NUMERICAL SCHEMES

The Rusanov scheme (first order)

This scheme is one of the simplest approximate Riemann Solver, see [14]. The numerical flux is

$$\mathcal{F}(u_L, u_R) = \frac{f(u_L) + f(u_R)}{2} - \frac{A(u_L, u_R)}{2}(u_R - u_L) \tag{A.1}$$

and the numerical entropy flux is

$$\mathcal{G}(u_L, u_R) = \frac{G(u_L) + G(u_R)}{2} - \frac{A(u_L, u_R)}{2}(\eta(u_R) - \eta(u_L)). \tag{A.2}$$

The scalar quantity $A(u_L, u_R)$ should be large enough. For scalar equation, we set $A(u_L, u_R) = \max(|f'(u_L)|, |f'(u_R)|)$ and for the equations of gas dynamic (23) we take

$$A(u_L, u_R) = \max\left(|v_L| + \sqrt{\frac{\gamma p_L}{\rho_L}}, |v_R| + \sqrt{\frac{\gamma p_R}{\rho_R}}\right).$$

The HLL scheme (first order)

This scheme is also a simple two waves approximate Riemann solver. We present the scheme for gas dynamic. Following Section 10.3 of [24] we introduce the sound speeds $a_L = \sqrt{\gamma p_L/\rho_L}$ and $a_R = \sqrt{\gamma p_R/\rho_R}$ and the Roe averages $\bar{a} = \dfrac{\sqrt{\rho_L}a_L + \sqrt{\rho_R}a_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}$ and $\bar{v} = \dfrac{\sqrt{\rho_L}v_L + \sqrt{\rho_R}v_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}$. The wave speed are estimates as $S_L = \min(v_L - a_L, v_R - a_R, \bar{v} - \bar{a})$ and $S_R = \min(v_L + a_L, v_R + a_R, \bar{v} + \bar{a})$ and the flux is given by equation (10.21) of [24]. With $-S_L = S_R = A(u_L, u_R)$ the HLL scheme is the Rusanov scheme.

The HLLC scheme for gas dynamics (first order)

This scheme is a commonly used approximate Riemann solver for (23). This solver is exact on some important particular solutions of (23), the stationary contact discontinuities. We implemented two versions of the schemes and obtained similar results. The first one corresponds to Paragraph 10.4.2 of [24]. The extremal wave speed are estimated with (10.49) of [24] and the pressure and velocities are constant in the "star region". The second one is based on a pressure estimate and corresponds to Paragraph 10.6, variant 1 of [24].

The Roe scheme (first order)

In the case of scalar conservation laws, the numerical scheme is

$$\mathcal{F}(u_L, u_R) = f(u_L)\,\mathbf{1}_{\sigma\geq 0} + f(u_R)\,\mathbf{1}_{\sigma<0}, \qquad \sigma = \frac{f(u_R) - f(u_L)}{u_R - u_L}.$$

We refer the reader to [22] and Chapter 11 of [24] for the presentation of the Roe scheme for the Euler equations (23). This scheme does not have a discrete entropy inequality since it preserves entropy violating shock ($f(u_L) = f(u_R)$, $G(u_R) > G(u_L)$). In can also produce negative pressure [12], in which case the simulation fails entirely.

The MUSCL scheme (first order)

The MUSCL procedure is a commonly used procedure to obtain a second ordre scheme in space, and can be easily combined with a second order time scheme such as a 2 step Runge-Kutta method. It is a 4-points flux with $s_L = s_R = 2$.

For scalar conservation laws, the procedure is the following. At the beginning of each time step, the constant value u_j^n in cell $\mathcal{C}_j = [x_j - \frac{\Delta x}{2}, x_j - \frac{\Delta x}{2}]$ is replaced by the affine function

$$x \mapsto u_j^n + \sigma_j^n (x - x_j),$$

where $\sigma_j \in \mathbf{R}$ is a slope, determined in such a way that no new extrema are created and the scheme remains total variation diminishing. A common choice is

$$\begin{aligned}\sigma_j^n &= \operatorname{minmod}\left(u_j^n - u_{j-1}^n, u_{j+1}^n - u_j^n\right),\\ &= \max\left(0, \min\left(u_j^n - u_{j-1}^n, u_{j+1}^n - u_j^n\right)\right) + \min\left(0, \max\left(u_j^n - u_{j-1}^n, u_{j+1}^n - u_j^n\right)\right).\end{aligned}$$

Other limiters are possible, see Section 13.7.3 of [24]. The MUSCL flux is given by

$$F_{j+1/2} = \mathcal{F}\left(u_{j,+}^n, u_{j+1,-}^n\right), \quad u_{j,\pm}^n = u_j^n \pm \frac{\Delta x}{2}\sigma_j^n, \tag{A.3}$$

where $\mathcal{F}$ is a first order two points scheme.

In the case of hyperbolic system with $p > 1$, this strategy is mimicked componentwise. For the Euler equation (23) we followed the strategy of [1]. The reconstruction with a minmod limiter can be applied on the three conservative variables $(\rho, \rho v, E)$. It is also common to reconstruct the primitive variables (ρ, v, p) or in the entropy variables $\left(\rho, v, s = \dfrac{p}{\rho^\gamma}\right)$ and to deduce the values $(\rho v)^n_{j,\pm} = \rho^n_{j,\pm} v^n_{j,\pm}$ with

$$E^n_{j,\pm} = \frac{p^n_{j,\pm}}{\gamma - 1} + \frac{\rho^n_{j,\pm} \left(v^n_{j,\pm}\right)^2}{2} \quad \text{or} \quad E^n_{j,\pm} = \frac{(\rho^n_{j,\pm})^\gamma s^n_{j,\pm}}{\gamma - 1} + \frac{\rho^n_{j,\pm} \left(v^n_{j,\pm}\right)^2}{2}.$$

A variation of the discrete entropy inequality (5) is obtained in [1] but is not sufficient to obtain a Lax-Wendroff theorem.

A second order Runge-Kutta method in time

To achieve second order in time one can use a two step Runge-Kutta method in time. Consider a numerical flux $\mathcal{F}$ with a stencil of s_L points to the left and s_R points to the right. Two substeps are computed

$$\bar{u}^n_j = u^n_j - \frac{\Delta t}{\Delta x}\left(\mathcal{F}\left(u^n_{j-s_L+1}, \ldots, u^n_{j+s_R}\right) - \mathcal{F}\left(u^n_{j-s_L}, \ldots, u^n_{j+s_R-1}\right)\right),$$

$$\bar{\bar{u}}^n_j = \bar{u}^n_j - \frac{\Delta t}{\Delta x}\left(\mathcal{F}\left(\bar{u}^n_{j-s_L+1}, \ldots, \bar{u}^n_{j+s_R}\right) - \mathcal{F}\left(\bar{u}^n_{j-s_L}, \ldots, \bar{u}^n_{j+s_R-1}\right)\right).$$

The final update is $u^{n+1}_j = \dfrac{u^n_j + \bar{\bar{u}}^n_j}{2}$.

The whole procedure rewrites in the compact form (4) with

$$\bar{\mathcal{F}}\left(u^n_{j-2s_L+1}, \ldots, u^n_{j+2s_R}\right) = \frac{\mathcal{F}\left(u^n_{j-s_L+1}, \ldots, u^n_{j+s_R}\right) + \mathcal{F}\left(\bar{u}^n_{j-s_L+1}, \ldots, \bar{u}^n_{j+s_R}\right)}{2}.$$

If the numerical flux $\mathcal{F}$ is entropy satisfying with a numerical entropy flux $\mathcal{G}$, the same holds for $\bar{\mathcal{F}}$ with the numerical entropy flux

$$\bar{\mathcal{G}}\left(u^n_{j-2s_L+1}, \ldots, u^n_{j+2s_R}\right) = \frac{\mathcal{G}\left(u^n_{j-s_L+1}, \ldots, u^n_{j+s_R}\right) + \mathcal{G}\left(\bar{u}^n_{j-s_L+1}, \ldots, \bar{u}^n_{j+s_R}\right)}{2}.$$

For scalar conservation laws $d = 1$, the maximum of the wave speeds decreases with time, thus the CFL restriction for the computation of $\bar{u}_j$ is more restrictive than the one for $\bar{\bar{u}}$. This is not true when $p \geq 2$, and we adopt the time evolution of equation (2.12) of [1], with the slight modification that Δt_2 cannot exceed Δt_1.

References

[1] C. Berthon, Stability of the MUSCL schemes for the Euler equations. *Commun. Math. Sci.* **3** (2005) 133–157.

[2] C. Berthon and V. Desveaux, An entropy preserving MOOD scheme for the Euler equations. *Int. J. Finite.* **11** (2014) 39.

[3] F. Bouchut, Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-balanced Schemes for Sources. Frontiers in Mathematics. Birkhäuser Verlag, Basel (2004).

[4] F. Bouchut, C. Bourdarias and B. Perthame, A MUSCL method satisfying all the numerical entropy inequalities. *Math. Comput.* **65** (1996) 1439–1461.

[5] H. Burchard and H. Rennau, Comparative quantification of physically and numerically induced mixing in ocean models. *Ocean Model.* **20** (2008) 293–311.

[6] C. Chalons and P.G. LeFloch, A fully discrete scheme for diffusive-dispersive conservation laws. *Numer. Math.* **89** (2001) 493–509.

[7] S. Clain, S. Diot and R. Loubere, A high-order finite volume method for systems of conservation laws-multi-dimensional optimal order detection (mood). *J. Comput. Phys.* **230** (2011) 4028–4050.

[8] P. Colella and P.R. Woodward, The piecewise parabolic method (ppm) for gas-dynamical simulations. *J. Comput. Phys.* **54** (1984) 174–201.
[9] F. Coquel and P.G. LeFloch, An entropy satisfying MUSCL scheme for systems of conservation laws. *Numer. Math.* **74** (1996) 1–33.
[10] F. Couderc, A. Duran and J.P. Vila, An explicit asymptotic preserving low froude scheme for the multilayer shallow water model with density stratification. *J. Comput. Phys.* **343** (2017) 235–270.
[11] S. Diot, S. Clain and R. Loubere, Improved detection criteria for the multi-dimensional optimal order detection (mood) on unstructured meshes with very high-order polynomials. *Comput. Fluids* **64** (2012) 43–63.
[12] B. Einfeldt, C. Munz, P.L. Roe and B. Sjogreen, On godunov-type methods near low-densities. *J. Comput. Phys.* **92** (1991) 273–295.
[13] B. Fox-Kemper, A. Adcroft, C.W. Böning, E.P. Chassignet, E. Curchitser, G. Danabasoglu, C. Eden, M.H. England, R. Gerdes, R.J. Greatbatch and S.M. Griffies, Challenges and prospects in ocean circulation models. *Front. Mar. Sci.* **6** (2019) 65.
[14] E. Godlewski and P.-A. Raviartm, Numerical approximation of hyperbolic systems of conservation laws, 2nd edition. In Vol. 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York (2021).
[15] A. Harten and S. Osher, Uniformly high-order accurate nonoscillatory schemes .1. *SIAM J. Numer. Anal.* **24** (1987) 279–309.
[16] A. Hiltebrand and S. Mishra, Entropy stable shock capturing space-time discontinuous Galerkin schemes for systems of conservation laws. *Numer. Math.* **126** (2014) 103–151.
[17] R.M. Holmes, J.D. Zika, S.M. Griffies, A.M.C.C. Hogg, A.E. Kiss and M.H. England, The geography of numerical mixing in a suite of global ocean models. *J. Adv. Model. Earth Syst.* **13** (2021) e2020MS002333.
[18] K. Klingbeil, M. Mohammadi-Aragh, U. Graewe and H. Burchard, Quantification of spurious dissipation and mixing - discrete variance decay in a finite-volume framework. *Ocean Model.* **81** (2014) 49–64.
[19] P. Lax and B. Wendroff, Systems of conservation laws. *Commun. Pure Appl. Math.* **13** (1960) 217–237.
[20] L. Martaud, M. Badsi, C. Berthon, A. Duran and K. Saleh, Global entropy stability for class of unlimited high-order schemes for hyperbolic systems of conservation laws. Preprint: arXiv:hal-03206727 (2021).
[21] B. Perthame and C.W. Shu, On positivity preserving finite volume schemes for Euler equations. *Numer. Math.* **73** (1996) 119–130.
[22] P.L. Roe, Approximate riemann solvers, parameter vectors, and difference-schemes. *J. Comput. Phys.* **43** (1981) 357–372.
[23] The Mathworks, Inc., Natick, Massachusetts. MATLAB version 9.11.0.1769968 (R2021b) (2021).
[24] E.F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, 3rd edition. A practical introduction, Springer-Verlag, Berlin (2009).
[25] B. Van Leer, Towards the ultimate conservative difference scheme. 5. 2nd-order sequel to Godunovs method. *J. Comput. Phys.* **32** (1979) 101–136.
[26] X. Zhang and C.-W. Shu, Positivity-preserving high order finite difference weno schemes for compressible Euler equations. *J. Comput. Phys.* **231** (2012) 2245–2258.