HOW TO FIND A DISCRETE ENTROPY INEQUALITY WHEN YOU DON'T KNOW IF IT EXISTS

Nina Aguillon*[1], Emmanuel Audusse†[2], Vivien Desveaux‡[3], Julien Salomon§[1]

[1]Inria, Project-Team ANGE, Sorbonne Université and Université de Paris, CNRS, Laboratoire Jacques-Louis Lions (LJLL), 75589 Paris Cedex 12.
[2]Université Sorbonne Paris Nord, CNRS, Laboratoire Analyse, Géométrie et Applications (LAGA),99 av. J.-B. Clément, 93430 Villetaneuse,
[3]Université de Picardie Jules Verne, CNRS, LAMFA, 33 rue Saint-Leu 80039 Amiens Cedex 1.

December 2, 2022

Abstract. The solutions of hyperbolic systems 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. In this paper, we introduce an optimization framework that enable to quantify a posteriori entropy. We use it to quantify numerical diffusion and detect non-entropic schemes.

Introduction

Many physical phenomenon can be described with a hyperbolic system, also called system of conservation laws. Some famous hyperbolic systems are the Lighthill–Whitham–Richards or Aw-Rascle models for traffic flow model, the shallow water equations of Barré de Saint-Venant, the Euler equation for fluids dynamics and the inviscid magneto hydrodynamics equation.

This class of partial differential equation (PDE) does not contain any regularization term 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 understand in the weak sense to allow such discontinuous solutions. Doing so, it becomes possible to construct infinitely many discontinuous solutions for the same initial data. An additional criterion should consequently be imposed 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 which states that the entropy of the solution decreases with time.

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

*nina.aguillon@sorbonne-universite.fr
†audusse@math.univ-paris13.fr
‡vivien.desveaux@u-picardie.fr
§julien.salomon@inria.fr

entropy also decreases at the numerical level. This condition ensures that the scheme will not converge towards an 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 been mainly obtained for first order schemes in space and time. Realistic codes use higher order discretizations and splitting techniques and usually incorporates ideas and knowledge from the first order theory. For this reason they probably verify a discrete entropy inequality in most cases. However, no explicit formula 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 motivation is to obtain maps of numerical diffusion which quantify in space and time the loss of energy coming from the choice of discretization. In numerical oceanic circulation models, the numerical diffusion is linked to the undesirable changes of salinity, density and temperature between two adjacent distinct water masses, and is usually called spurious mixing. It is identified as a major issue in numerical cores for climate application [13]. Ideally this spurious mixing would be one order of magnitude below the physical mixing. The large time and space scales considered in this field makes it impossible to handle with the current computational cost constraints. Another approach would be to somehow parameterize the numerical diffusion in order to re-inject the correct amount of energy into the system. It would require the precise quantification of the spurious mixing which is still an open issue in ocean global circulation model. 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 complex 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 from one time step to the other. This minimization couples every cells of the mesh, but we also propose a local and cheap quantification that gives qualitatively good results. A different perspective on the same functional allows us to construct the worst initial data in terms of entropy. We show that for gas dynamics no discrete entropy inequality exists for the widely used MUSCL approach with a 2 steps Runge-Kutta time discretization, as suspected in [2]. A notable exception is the limitation in the entropic variables with a HLL first order scheme which passes this entropy stresstest.

The minimization procedure is presented in Section 2 after a presentation of the mathematical framework of discrete entropy inequalities for finite volume schemes in Section 1. The links between the existence of a discrete entropy inequality and the minimization are presented in Section 3 and numerical results for first and second order schemes are presented in Section 4. In Section 5, we present another minimization procedure which constructs initial data for which no discrete entropy inequality holds at the first iteration in time.

1. Fundamentals on discrete entropy inequalities

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

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

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 verify 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.

$$\begin{cases} \partial_t u + \partial_x \left(\dfrac{u^2}{2} \right) = 0 \\ \partial_t \left(u^2 \right) + \partial_x \left(\dfrac{2u^3}{3} \right) \leq 0 \end{cases} . \tag{3}$$

We also consider the Euler equations of inviscid gas dynamics for which $d = 3$ and $u = (\rho, \rho v, E)$, where ρ is the density of the fluid, v is its velocity and E its total energy. It writes

$$\begin{cases} \partial_t \rho + \partial_x (\rho v) = 0 \\ \partial_t (\rho v) + \partial_x \left(\rho v^2 + p \right) = 0 \\ \partial_t E + \partial_x (v (E + p)) = 0 \end{cases} . \tag{4}$$

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

$$p = (\gamma - 1) \left(E - \frac{\rho v^2}{2} \right)$$

where $\gamma \in (1, 3]$. 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]. In the examples considered in this paper, we focus on

$$\partial_t (-\rho \ln(s)) + \partial_t (-v\rho \ln(s)) \leq 0 \tag{5}$$

where the specific entropy s is defined as $s = \dfrac{p}{\rho^\gamma}$.

We now turn to the numerical discretization of (1) with a finite volume technique. The space interval $[a, b]$ is discretized into M intervals of the same size $\frac{b-a}{M}$ for simplicity, and we denote $x_{j-1/2} = a + (j-1)\Delta x$, $j \in \{1, \dots, M\}$ the end points of the cells:

$$a = x_{1/2} < x_{3/2} < \cdots < x_{M-1/2} < x_{M+1/2} = b.$$

We also consider a discretization in time

$$0 = t^0 < t^1 < \dots t^n < \dots$$

A Courant–Friedrichs–Lewy condition is imposed at each time step. It reads, for some CFL $\alpha \in (0, 1)$ depending on the scheme,

$$\left(t^{n+1} - t^n \right) \max_{1 \leq j \leq M} \rho \left(DF \left(U_j^n \right) \right) \leq \alpha \Delta x$$

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 we will denote it independently of n as $t^{n+1}-t^n=\Delta t$ for the sake of simplicity.

A finite volume scheme writes

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

In Equation (6), the vectorial quantity u_j^n is a numerical approximation of the mean of the exact solution u on the j-th cell at the n-th time step:

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

while $f_{j+1/2}^n$ is an approximation of the mean flux passing through the interface $x_{j+1/2}$ on the time interval $\left(t^n,t^{n+1}\right)$:

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

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 the choice of a formula expressing the numerical flux $f_{j+1/2}^n$ in terms of its s_L+s_R neighbor cells

$$f_{j+1/2}^n=\mathcal{F}\left(u_{j-s_L+1}^n,\ldots,u_{j+s_R}^n\right)$$

such that, for all $u\in\Omega$,

$$\mathcal{F}\left(u_{-s_L+1},\ldots,u_{s_R}\right)\to f\left(u\right)\quad\text{ as }\quad\left(u_{-s_L+1},\ldots,u_{s_R}\right)\to\left(u,\ldots,u\right).$$

Equations (1) and (2) are understood in a weak sense to allow discontinuous solutions. Notably the entropy inequality (2) does not hold for every discontinuity but selects only entropy satisfying shocks. The fact that the scheme (6) is stable and computes the physical 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

$$\eta\left(u_j^{n+1}\right)\leq\eta\left(u_j^n\right)-\frac{\Delta t}{\Delta x}\left(G_{j+1/2}^n-G_{j-1/2}^n\right). \tag{7}$$

In this expression, u_j^{n+1} and u_j^n are known and linked by the choice of scheme (6). The function η is the continuous entropy function of (2). The difficulty is to find the numerical entropy fluxes $G_{j+1/2}^n$ and $G_{j-1/2}^n$ such that the inequality holds. 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 $s_R\in\mathbf{N}$. It $\mathcal{F}$ is an consistent entropy satisfying scheme if there exists a numerical entropy flux function $\mathcal{G}$ such that the following two points are verified.

- Inequality (7) holds for all j with $G_{j+1/2}^n=\mathcal{G}\left(u_{j-s_L+1}^n,\ldots,u_{j+s_R}^n\right)$. In other words, the numerical diffusion on cell j at time t^n

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

  is nonpositive everywhere.
- The numerical entropy flux function $\mathcal{G}$ is consistent: for all $u\in\Omega$,

$$\mathcal{G}\left(u_{-s_L+1},\ldots,u_{s_R}\right)\to G\left(u\right)\quad\text{ as }\quad\left(u_{-s_L+1},\ldots,u_{s_R}\right)\to\left(u,\ldots,u\right).$$

There exists several schemes (6) for which an explicit formula for $\mathcal{G}$ yielding to nonpositive diffusion D_j^n is known. They mainly fall into the two following categories.

- for scalar equations $d = 1$, monotone schemes of order 1, see [14];
- for hyperbolic systems $d \geq 2$ 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. All those schemes are first order and $s_L = s_R = 1$.

For schemes of ordrer larger than 1 the specific form of (7) 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 (7) is slightly modified. The schemes are either difficult to implement or there is no guarantee that they capture physical solution.
- The local discrete entropy inequality (7) can be replaced by the decay of the total entropy

$$\sum_j \eta\left(u_j^{n+1}\right) \leq \sum_j \eta\left(u_j^n\right) \tag{9}$$

  This is achieved 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 (7) 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 (7) hold. However it is limited to the gas dynamics (4).

On the other hand many schemes of order 2 or more are employed in trusted codes for their good results despite the lack of discrete entropy inequality. Typically they are designed to be of high order when the solution is smooth. They integrate ideas to ensure stability, such as limiters or explicit numerical diffusion. Positivity and lack of spurious oscillations are also often taken into account. In other words, a great deal of work is done in the direction of a discrete entropy inequality even though there is no guarantee that there is indeed one. We can mention for instance the Piecewise Parabolic Method (PPM) [8] or the (Weighted) Essentially Non Oscillatory method (ENO/WENO); see [15, 21, 25].

In this work we are concerned with the *a posteriori* quantification of the numerical diffusion D_j^n and numerical entropy fluxes $G_{j+1/2}^n$. These quantities are related by the definition of the numerical diffusion (8). We fix a PDE of the form described in (1) endowed with an entropy inequality (2) and the choice of a consistent numerical flux $\mathcal{F}$ used in the update (6). We attempt to find numerical entropy fluxes $\left(G_{j+1/2}^n\right)_j$ such that (7) holds by a minimization procedure. In the case where an explicit formula for $\mathcal{G}$ (and thus for $G_{j+1/2}^n$) is known, the minimization will reach its lower bound 0 and return consistent numerical entropy flux and nonpositive

numerical diffusions. In general the minimization ensures that these two properties are satisfied as much as possible.

2. Minimization procedure

We consider a regular space discretization in 1D with M cells as described in Section 1. For the sake of simplicity, we consider periodic boundary conditions, i.e., $f^n_{1/2} = f^n_{M+1/2}$. For j outside of $\{1, \cdots, M\}$, u^n_j is set by $u^n_j = u^n_{mod(j,M)}$. We also impose periodic boundary condition on the numerical entropy fluxes, and one should read $\gamma^n_{1/2} = \gamma^n_{M+1/2}$ and $G^{opt,n}_{1/2} = G^{opt,n}_{M+1/2}$ when it appears in the sequel. We recall that the sequence $\left(u^n_j\right)_j$ is known and Scheme (6) is used to obtain $\left(u^{n+1}_j\right)_j$. In this section the time iteration n is fixed and referred to as "the initial data". The exponent n plays no particular role and one can fix $n = 0$. The question in that case is "is the scheme entropy satisfying at the first iteration?".

In this section, the numerical entropy flux is given by a minimization procedure

$$\left(G^{opt,n}_{j+1/2}\right)_{1\leq j\leq M} = \underset{\gamma=\left(\gamma^n_{3/2},\ldots,\gamma^n_{M+1/2}\right)\in\mathbf{R}^M}{\arg\min} \left\{\mathcal{G}\left(\gamma\right)\right\}. \tag{10}$$

The function $\mathcal{G} : \mathbf{R}^M \to \mathbf{R}$ is defined by the sum of two contributions

$$\mathcal{G}\left(\gamma\right) = \mathcal{G}^D\left(\gamma\right) + \mathcal{G}^C\left(\gamma\right). \tag{11}$$

The first part $\mathcal{G}^D$ gathers the undesirable contribution of positive numerical diffusions

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

Following (8), if the numerical entropy fluxes were $G^n_{j+1/2} = \gamma^n_{j+1/2}$ the associated numerical diffusion would be $\eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right) + \frac{\Delta t}{\Delta x}\left(\gamma^n_{j+1/2} - \gamma^n_{j-1/2}\right)$. In particular, $\mathcal{G}^D\left(\gamma\right)$ is null if and only if this choice of numerical entropy fluxes only gives nonpositive numerical diffusion D^n_j.

It remains to take into account the consistency property of the numerical entropy fluxes of Definition 2 and this is the role of the second part of the functional. The key point is that we are able to find some a priori bounds $M^n_{j+1/2}$ and $m^n_{j+1/2}$ on the flux $G^n_{j+1/2}$, which depend on the $s_L + s_R$ neighbor cells

$$M^n_{j+1/2} = \mathcal{M}^C\left(u^n_{j-s_L+1}, \ldots, u^n_{j+s_R}\right) \quad m^n_{j+1/2} = m^C\left(u^n_{j-s_L+1}, \ldots, u^n_{j+s_R}\right).$$

These bounds verify the following consistency property

$$\forall u \in \Omega, \quad \mathcal{M}^C\left(u, \ldots, u\right) = m^C\left(u, \ldots, u\right) = G\left(u\right).$$

This motivates the choice

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

which vanishes if and only if

$$\forall j \in \{1/2, \ldots, M + 1/2\}, \quad m^n_{j+1/2} \leq \gamma^n_{j+1/2} \leq M^n_{j+1/2}.$$

In the particular case where $u^n_{j-s_L+1} = \cdots = u^n_{j+s_R} = u$, the j-th term of $\mathcal{G}^C(\gamma)$ vanishes if and only if $\gamma^n_{j+1/2} = G(u)$, which is nothing but the consistency requirement for the numerical entropy flux.

The function $\mathcal{G}$ is nonnegative, $\mathcal{C}^1$-regular and convex with respect to γ, but the set where it vanishes has no reason neither to exist nor to be reduced to a single point. In the next sections we precise the construction of the consistency bounds $\mathcal{M}^C$ and m^C. We also explore the strong link between the existence of a discrete entropy inequality (7) and the fact that an optimum $G^{opt,n}$ such that $\mathcal{G}(G^{opt,n}) = 0$ can be found.

2.1. **Consistency bounds for 2-points fluxes.** We first consider the simplest case of a flux that depends only on its two neighboring cells ($s_L = s_R = 1$):

$$f^n_{j+1/2} = \mathcal{F}\left(u^n_j, u^n_{j+1}\right).$$

The formulas are shorter than in the general case of Section 2.2 but the fundamental ideas are the same.

Lemma 3. *Consider a two points flux $\mathcal{F}$, and suppose that the finite volume scheme* (6) *is entropy satisfying in the sense of Definition 2. Then for all j, $m^n_{j+1/2} \le G^n_{j+1/2} \le M^n_{j+1/2}$ with*

$$\begin{cases} M^n_{j+1/2} = G\left(u^n_j\right) + \dfrac{\Delta x}{\Delta t}\left(\eta\left(u^n_j\right) - \eta\left(\hat{u}^{j+1/2}_j\right)\right) \\ m^n_{j+1/2} = G\left(u^n_{j+1}\right) + \dfrac{\Delta x}{\Delta t}\left(\eta\left(\hat{u}^{j+1/2}_{j+1}\right) - \eta\left(u^n_{j+1}\right)\right) \end{cases} \tag{12}$$

where

$$\begin{cases} \hat{u}^{j+1/2}_j = u^n_j - \dfrac{\Delta t}{\Delta x}\left(f^n_{j+1/2} - f\left(u^n_j\right)\right) \\ \hat{u}^{j+1/2}_{j+1} = u^n_{j+1} - \dfrac{\Delta t}{\Delta x}\left(f\left(u^n_{j+1}\right) - f^n_{j+1/2}\right) \end{cases}$$

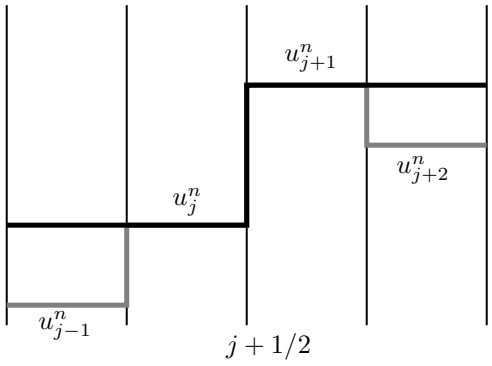

FIGURE 1. The fluxes $f^n_{j+1/2}$ and $G^n_{j+1/2}$ at interface $j+1/2$ are the same regardless of the values in cells $j-1$ and $j+2$. The use of the modified data (black) instead of the initial data (grey) allows to use the consistency at interfaces $j-1/2$ and $j+3/2$

Remark 4. Lemma 3 is a necessary condition for a scheme to be entropy satisfying. This is what F. Bouchut calls an "interface entropy inequality" in [3, Definition 2.7]. It implies the desired discrete entropy inequality (7) at the cost of a time step Δt twice smaller [3, Proposition 2.9].

Proof. The numerical flux $f^n_{j+1/2}$ and numerical entropy flux $G^n_{j+1/2}$ at interface $j+1/2$ are the same if we start with the initial data

$$u^n = \left(\dots, u^n_{j-1}, u^n_j, u^n_{j+1}, u^n_{j+2}, \dots\right)$$

or the modified initial data

$$\begin{aligned}\tilde{u}^{j+1/2} = \left(\tilde{u}_k^{j+1/2}\right)_k &= \left(\dots, u^n_j, u^n_j, u^n_{j+1}, u^n_{j+1}, \dots\right)\\ &= \begin{cases} u^n_j & \text{if } k \le j \\ u^n_{j+1} & \text{if } k > j. \end{cases}\end{aligned}$$

The evolution in time with this modified initial data is particularly simple. All the fluxes $\hat{f}^{j+1/2}_{k+1/2}$ are computed by consistency except at interface $j+1/2$ where it is the same that for the original initial data

$$\tilde{f}^{j+1/2}_{k+1/2} = \mathscr{F}\left(\tilde{u}_k^{j+1/2}, \tilde{u}_{k+1}^{j+1/2}\right) = \begin{cases} f\left(u^n_j\right) & \text{if } k < j \\ f^n_{j+1/2} & \text{if } k = j \\ f\left(u^n_{j+1}\right) & \text{if } k > j+1 \end{cases}.$$

It yields the update

$$\hat{u}_k^{j+1/2} = \begin{cases} u^n_j & \text{if } k < j \\ u^n_j - \dfrac{\Delta t}{\Delta x}\left(f^n_{j+1/2} - f\left(u^n_j\right)\right) & \text{if } k = j \\ u^n_{j+1} - \dfrac{\Delta t}{\Delta x}\left(f\left(u^n_{j+1}\right) - f^n_{j+1/2}\right) & \text{if } k = j+1 \\ u^n_{j+1} & \text{if } k > j+1 \end{cases}.$$

The consistency can be applied similarly at interfaces $j-1/2$ and $j+3/2$ for the entropy evolution. If the scheme is entropy satisfying, then the two following inequalities hold

$$\begin{cases} \eta\left(\hat{u}_j^{j+1/2}\right) \le \eta\left(u^n_j\right) - \dfrac{\Delta t}{\Delta x}\left(G^n_{j+1/2} - G\left(u^n_j\right)\right) \\ \eta\left(\hat{u}_{j+1}^{j+1/2}\right) \le \eta\left(u^n_{j+1}\right) - \dfrac{\Delta t}{\Delta x}\left(G\left(u^n_{j+1}\right) - G^n_{j+1/2}\right) \end{cases}$$

and the bounds on $G^n_{j+1/2}$ of Lemma 3 follow. □

2.2. **Extension to larger stencils.** We now generalize Lemma 3 or the notion of interface entropy inequality of [3] to arbitrary stencils. We follow exactly the same steps than for two points schemes. First, a modified initial data $\left(\tilde{u}_k^{j+1/2}\right)_k$ is constructed for each interface $j+1/2$ in such a way that the consistency property can be used for interfaces far enough from $j+1/2$. Then the finite volume scheme (6) applied to this new initial data gives numerical fluxes $\left(\tilde{f}^{j+1/2}_{k+1/2}\right)_k$. They are used to compute the update $\left(\hat{u}_k^{j+1/2}\right)_k$. Eventually two local budgets of entropy allows to bound $G^n_{j+1/2}$.

Lemma 5. *Consider a consistent entropy satisfying scheme in the sense of Definition 2. Then for all $j \in \{0, \dots, M\}$ we have $m^n_{j+1/2} \le G^n_{j+1/2} \le M^n_{j+1/2}$, with the bounds*

$$\begin{cases} M^n_{j+1/2} = G\left(u^n_{j-s_L+1}\right) + \dfrac{\Delta x}{\Delta t} \displaystyle\sum_{k=j-s_L-s_R+2}^{j} \eta\left(\tilde{u}^{j+1/2}_k\right) - \eta\left(\hat{u}^{j+1/2}_k\right) \\ m^n_{j+1/2} = G\left(u^n_{j+s_R}\right) + \dfrac{\Delta x}{\Delta t} \displaystyle\sum_{k=j+1}^{j+s_L+s_R-1} \eta\left(\hat{u}^{j+1/2}_k\right) - \eta\left(\tilde{u}^{j+1/2}_k\right) \end{cases} \tag{13}$$

where

$$\tilde{u}^{j+1/2}_k = u^n_{\min(\max(k, j-s_L+1), j+s_R)} \tag{14a}$$

$$\tilde{f}^{j+1/2}_{k+1/2} = \mathcal{F}\left(\tilde{u}^{j+1/2}_{k-s_L+1}, \dots, \tilde{u}^{j+1/2}_k, \tilde{u}^{j+1/2}_{k+1}, \dots, \tilde{u}^{j+1/2}_{k+s_R}\right) \tag{14b}$$

$$\hat{u}^{j+1/2}_k = \tilde{u}^{j+1/2}_k - \frac{\Delta t}{\Delta x}\left(\tilde{f}^{j+1/2}_{k+1/2} - \tilde{f}^{j+1/2}_{k-1/2}\right) \tag{14c}$$

Proof. We focus on the flux at the interface $j + 1/2$. This flux is indifferent to 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 prolongated it by $u^n_{j-s_L+1}$ on its left and by $u^n_{j+s_R}$ on its right

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

which writes (14a) in short. The numerical fluxes applied to this modified data are

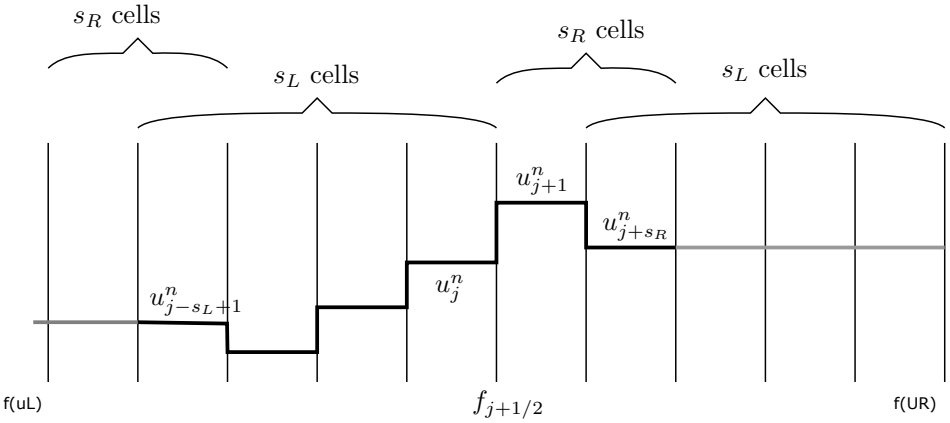

FIGURE 2. Modification of the initial data away from the stencil. Consistency can be used to compute the fluxes at interfaces $j + 3/2 - s_L - s_R$ and $j - 1/2 + s_L + s_R$.

given by (14b) and yields the update (14c).

The interest of the modified initial data $\left(\tilde{u}^{j+1/2}_k\right)_k$ 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:

$$\begin{aligned}\tilde{f}^{j+1/2}_{j-s_L-s_R+3/2} &= \mathcal{F}\left(\tilde{u}^{j+1/2}_{j-2s_L-s_R+2},\dots,\tilde{u}^{j+1/2}_{j-s_L+1}\right)\\ &= \mathcal{F}\left(u^n_{j-s_L+1},\dots,u^n_{j-s_L+1}\right) = f\left(u^n_{j-s_L+1}\right)\end{aligned}$$

and

$$\begin{aligned}\tilde{f}^{j+1/2}_{j+s_L+s_R-1/2} &= \mathcal{F}\left(\tilde{u}^{j+1/2}_{j+s_R},\dots,\tilde{u}^{j+1/2}_{j+s_L+2s_R-1}\right)\\ &= \mathcal{F}\left(u^n_{j+s_R},\dots,u^n_{j+s_R}\right) = f\left(u^n_{j+s_R}\right).\end{aligned}$$

It follows that $\hat{u}^{j+1/2}_k = \tilde{u}^{j+1/2}_k$ for all $k \le j-s_L-s_R+1$ and for all $k \ge j+s_L+s_R$.

Now, suppose that the scheme is entropy satisfying, i.e. that

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

where once again

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

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, the construction of $\left(\tilde{u}^{j+1/2}_k\right)_k$ gives us three information. The consistency yields

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

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

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

We eliminate the other numerical entropy fluxes by summation:

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

and

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

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

3. Main results

3.1. Entropy dissipation and zero minimization.

Proposition 6. *Consider a finite volume scheme* (6) *that admits a discrete entropy inequality* (7) *for some numerical entropy fluxes* $\left(G^n_{j+1/2}\right)_j$. *Then* $\mathcal{J}\left(\left(G^n_{j+1/2}\right)_j\right) = 0$ *for all initial data* $\left(u^n_j\right)_j$, *where the functional* $\mathcal{J}$ *is defined with* (6) *and* (13).

Proof. This property follows from the definition of the functional. If (7) holds, clearly $\mathcal{J}^D\left(\left(G^n_{j+1/2}\right)_j\right) = 0$. It remains to prove that the second part also vanishes, which is the case if and only if $m^n_{j+1/2} \le G^n_{j+1/2} \le M^n_{j+1/2}$. 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 $\left(\tilde{u}_k^{j+1/2}\right)_k$. The lower (resp. upper) bound comes from the diminution of total entropy during 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 5 in the previous section. □

Proposition 7. *Fix the initial data $\left(u_j^n\right)_j$. Suppose that there exists $\left(G_{j+1/2}^{opt,n}\right)_j$ such that $\mathcal{G}\left(\left(G_{j+1/2}^{opt,n}\right)_j\right)=0$. Then the scheme dissipates the entropy:*

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

It is also consistent in the sense that if $u_{j-s_L+1}^n = \cdots = u_j^n = \cdots = u_{j+s_R}^n$, then $G_{j+1/2}^{opt,n} = G\left(u_j^n\right)$.

This results only says that it is possible to find numerical entropy fluxes yielding to nonpositive numerical diffusion for the particular choice of initial data $\left(u_j^n\right)_j$. It does not mean that the scheme is always entropy satisfying.

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

$$\forall j \quad \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^{opt,n} - G_{j-1/2}^{opt,n}\right) \leq 0$$

which is exactly (7). The last contribution $\mathcal{G}^C$ is also zero, thus

$$m_{j+1/2}^n \leq G_{j+1/2}^{opt,n} \leq M_{j+1/2}^n.$$

If $\left(u_{j-s_L+1}^n, \ldots, u_{j+s_R}^n\right) = \left(u_j^n, \ldots, u_j^n\right)$, the modified initial data $\left(\tilde{u}_k^{j+1/2}\right)_k$ is constant equals to u_j^n and so is $\left(\hat{u}_k^{j+1/2}\right)_k$. Thus $m_{j+1/2}^n = M_{j+1/2}^n = G\left(u_j^n\right)$ and $G_{j+1/2}^{opt,n} = G\left(u_j^n\right)$. □

3.2. **Discrepancy between global minimizers.** The function $\mathcal{G}\left(\left(G_{j+1/2}^n\right)_j\right)$ vanishes if and only if (7) holds and for all j, $m_{j+1/2}^n \leq G_{j+1/2}^n \leq M_{j+1/2}^n$. When this set of inequalities is nonempty it most likely contains several solutions. We first quantify how close they are from each other.

Proposition 8. *Suppose that $\mathcal{G}\left(\left(G_{j+1/2}^n\right)_j\right)=0$ and $\mathcal{G}\left(\left(\bar{G}_{j+1/2}^n\right)_j\right)=0$, and that the numerical flux $\mathcal{F}$ used in (6) is consistent and Lipschitz regular. Then for all j,*

$$|G_{j+1/2}^n - \bar{G}_{j+1/2}^n| = \sum_{k\in\{j-s_L+1,\ldots,j+s_R\}} O\left(|u_k^n - u_j^n|^2\right)$$

When the scheme is known to have a discrete entropy inequality, this shows that when the solution is smooth, the difference between the numerical entropy flux found in the literature and the one returned by the minimization procedure is of order Δx^2.

Proof. Suppose that the minimization procedure has two different global minimizers $\mathcal{G}\left(G^n_{j+1/2}\right) = 0$ and $\mathcal{G}\left(\bar{G}^n_{j+1/2}\right) = 0$. Then $m^n_{j+1/2} \leq G^n_{j+1/2} \leq M^n_{j+1/2}$ and $m^n_{j+1/2} \leq \bar{G}^n_{j+1/2} \leq M^n_{j+1/2}$, thus

$$|\bar{G}^n_{j+1/2} - G^n_{j+1/2}| \leq M^n_{j+1/2} - m^n_{j+1/2}.$$

Let us extend the latter term

$$M^n_{j+1/2} - m^n_{j+1/2} = G\left(u^n_{j-s_L+1}\right) - G\left(u^n_{j+s_R}\right) + \frac{\Delta x}{\Delta t} \sum_{k=j-s_L-s_R+2}^{j+s_L+s_R} \eta\left(\tilde{u}^{j+1/2}_k\right) - \eta\left(\hat{u}^{j+1/2}_k\right)$$

By convexity of the entropy η,

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

We arrive at

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

If the numerical flux is Lipschitz regular, the last sum is

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

A second order expansion at point u^n_j of the first line is

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

It yields the result because $D\eta Df = DG$.

□

3.3. **Lax-Wendroff theorem.** One of the main theoretical results about numerical schemes for systems of conservation laws is the Lax-Wendroff theorem. This result ensures 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.

This last statement usually requires the numerical entropy flux to be a continuous and consistent function of the neighboring approximations. This is not the case in this work, since the numerical entropy fluxes are obtained through a minimization procedure. However, it is possible to adapt the Lax-Wendroff theorem to the framework of this paper.

Since the time discretization will be important in this section, the objects related to the optimization problem (10) at time t^n, i.e., with initial data $\left(u_j^n\right)_j$, will be noted with an exponent n.

For a discretization $\Delta = (\Delta x, \Delta t)$, we introduce the piecewise constant function u_Δ defined a.e. in $\mathbf{R} \times [0, +\infty)$ by

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

Theorem 1. *Assume the numerical flux $\mathcal{F}$ to be continuous and consistent. 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. $(x,t) \in \mathbf{R} \times [0,+\infty)$ and for all $k \in \mathbf{N}$;*
- *the sequence u_{Δ_k} converges in $L^1_{loc}(\mathbf{R} \times (0,+\infty))$ to a function u.*

Then u is a weak solution of (1).

Furthermore, if for a given entropy pair (η, G) and for all $n \in \mathbf{N}$, there exists a sequence $\left(G_{j+1/2}^{opt,n}\right)_j$ such that $\mathcal{J}\left(\left(G_{j+1/2}^{opt,n}\right)_j\right) = 0$, then the solution u satisfies the entropy inequality (2) in the sense of distributions on $\mathbf{R} \times (0,+\infty)$.

Proof. For the sake of simplicity, the subscript k in Δ_k will be ommited all along the proof. The first part of the theorem is exactly the same as in the original Lax-Wendroff theorem. The reader is referred for instance to [19, 14] for a complete proof.

Concerning the entropy inequality, let us consider a test function $\varphi \in C_c^1(\mathbf{R} \times (0,+\infty))$, with $\varphi \geq 0$. Since $\mathcal{J}^D\left(\left(G_{j+1/2}^{opt,n}\right)_j\right) = 0$, we have for all j and n

$$\eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \lambda\left(G_{j+1/2}^{opt,n} - G_{j-1/2}^{opt,n}\right) \leq 0.$$

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

$$\Delta x \sum_{j,n} \eta\left(u_j^{n+1}\right)\left(\varphi\left(x_j, t^{n+1}\right) - \varphi\left(x_j, t^n\right)\right) + \Delta t \sum_{j,n} G_{j+1/2}^{opt,n}\left(\varphi\left(x_{j+1}, t^n\right) - \varphi\left(x_j, t^n\right)\right) \geq 0.$$

Introducing the piecewise constant functions

$$\varphi_\Delta(x,t) = \varphi\left(x_j, t^n\right), \quad x_{j-1/2} < x < x_{j+1/2}, \quad t^n \leq t < t^{n+1},$$

$$G_\Delta^{opt,n}(x,t) = G_{j+1/2}^{opt,n}, \quad x_j < x < x_{j+1}, \quad t^n \leq t < t^{n+1},$$

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

$$\int_{\mathbf{R} \times [\Delta t, +\infty)} \eta\left(u_\Delta(x,t)\right) \frac{\varphi_\Delta(x,t) - \varphi_\Delta(x, t-\Delta t)}{\Delta t} dx dt + \int_{\mathbf{R} \times \mathbf{R}^+} G_\Delta^{opt}(x,t) \frac{\varphi_\Delta(x + \Delta x/2) - \varphi_\Delta(x - \Delta x/2)}{\Delta x} dx dt \geq 0. \tag{15}$$

As in the classical Lax-Wendroff theorem, the first integral converges to

$$\int_{\mathbf{R}\times\mathbf{R}^+} \eta\left(u\left(x,t\right)\right)\partial_t\varphi\left(x,t\right)dxdt.$$

The difference with the classical Lax-Wendroff theorem lies in the second integral of (15). Since φ is smooth, the term

$$\frac{\varphi_\Delta\left(x+\Delta x/2\right)-\varphi_\Delta\left(x-\Delta x/2\right)}{\Delta x}$$

uniformly converges to $\partial_x\varphi\left(x,t\right)$. Moreover, according to (13) and (**??**) and since the ratio λ is constant, $M^{C,n}_{j+1/2}$ and $m^{C,n}_{j+1/2}$ are continuous functions of a finite number of u^n_k, that all lie in the compact K. Therefore the $M^{C,n}_{j+1/2}$ and $m^{C,n}_{j+1/2}$ are uniformly bounded. Since $\mathcal{G}^C\left(\left(G^{opt,n}_{j+1/2}\right)_j\right)=0$, it follows that the inequalities

$$m^n_\Delta\left(x,t\right)\le G^{opt}_\Delta\left(x,t\right)\le M^n_\Delta\left(x,t\right) \tag{16}$$

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

$$\int_{\mathbf{R}\times\mathbf{R}^+} G^{opt}_\Delta\left(x,t\right)\left(\frac{\varphi_\Delta\left(x+\frac{\Delta x}{2}\right)-\varphi_\Delta\left(x-\frac{\Delta x}{2}\right)}{\Delta x}-\partial_x\varphi\left(x,t\right)\right)dxdt\to 0. \tag{17}$$

Next, we need to introduce the following piecewise constant functions:

$$M_\Delta\left(x,t\right)=M^{C,n}_{j+1/2},\quad x_j<x<x_{j+1},\quad t^n\le t<t^{n+1},$$

$$m_\Delta\left(x,t\right)=m^{C,n}_{j+1/2},\quad x_j<x<x_j,\quad t^n\le t<t^{n+1},$$

$$\tilde{u}^{j+1/2}_\Delta\left(x,t\right)=\tilde{u}^{j+1/2}_k,\quad x_{k-1/2}<x<x_{k+1/2},\quad t^n\le t<t^{n+1},$$

$$\tilde{f}^{j+1/2}_\Delta\left(x,t\right)=\tilde{f}^{j+1/2}_{k+1/2},\quad x_k<x<x_{k+1},\quad t^n\le t<t^{n+1}, \tag{18}$$

$$\hat{u}^{j+1/2}_\Delta\left(x,t\right)=\hat{u}^{j+1/2}_k,\quad x_{k-1/2}<x<x_{k+1/2},\quad t^n\le t<t^{n+1}. \tag{19}$$

According to (**??**), we have

$$\tilde{u}^{j+1/2}_\Delta\left(x,t\right)=u_\Delta\left(x+\min\left(\max\left(k,j-s_L+1\right),j+s_R\right)\Delta x-k\Delta x,t\right),$$

$$\tilde{f}^{j+1/2}_\Delta\left(x,t\right)=\mathcal{F}\left(\tilde{u}^{j+1/2}_\Delta\left(x-\left(s_L-1/2\right)\Delta x,t\right),\cdots,\tilde{u}^{j+1/2}_\Delta\left(x+\left(s_R-1/2\right)\Delta x,t\right)\right),$$

$$\hat{u}^{j+1/2}_\Delta\left(x,t\right)=\tilde{u}^{j+1/2}_\Delta\left(x,t\right)-\lambda\left(\tilde{f}^{j+1/2}_\Delta\left(x+\Delta x/2,t\right)-\tilde{f}^{j+1/2}_\Delta\left(x-\Delta x/2,t\right)\right).$$

It follows from (13) that

$$\begin{aligned}M_\Delta\left(x,t\right)=&\,G\big(u_\Delta\left(x-\left(s_L-1/2\right)\Delta x,t\right)\big)\\&+\lambda\left[\sum_{k=j-s_L-s_R+2}^{j}\eta\left(\tilde{u}^{j+1/2}_\Delta\left(x-\left(j+1/2-k\right)\Delta x,t\right)\right)\right.\\&\left.-\eta\left(\hat{u}^{j+1/2}_\Delta\left(x-\left(j+1/2-k\right)\Delta x,t\right)\right)\right].\end{aligned} \tag{20}$$

Since u_Δ converges in $L^1_{loc}\left(\mathbf{R}\times(0,+\infty)\right)$ to u, then up to a subsequence, $\tilde{u}^{j+1/2}_\Delta$ converges to u almost everywhere (a.e.). Thanks to the continuity and the consistency of $\mathcal{F}$, we get from (18) that $\tilde{f}^{j+1/2}_\Delta$ converges a.e. to $f\left(u\right)$. Then according

to (19), we deduce that $\hat{u}_\Delta^{j+1/2}$ also converges a.e. to u. Finally, it comes from (20) 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 (16) that up to a subsequence, G_Δ^{opt} converges a.e. to $G(u)$. The dominated convergence theorem then ensures that

$$\int_{\mathbf{R}\times\mathbf{R}^+} G_\Delta^{opt}(x,t)\,\partial_x\varphi(x,t)\,dxdt \to \int_{\mathbf{R}\times\mathbf{R}^+} G(u)\,\partial_x\varphi(x,t)\,dxdt. \tag{21}$$

Summing (17) and (21), we obtain

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

Hence the limit of (15) writes

$$\int_{\mathbf{R}\times\mathbf{R}^+} \eta(u(x,t))\,\partial_t\varphi(x,t)\,dxdx + \int_{\mathbf{R}\times\mathbf{R}^+} G(u)\,\partial_x\varphi(x,t)\,dxdt \geq 0,$$

which concludes the proof. □

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

4.1. **Continuous solution of the Burgers equation.** In this first test case, we consider the Burgers equation (3) with the initial data

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

and periodic boundary conditions. The inner discontinuity creates a rarefaction fan, its left extremity travels at speed -2 and its right extremity at speed 3. The exact solution is piecewise linear, and for $t < \frac{2}{3}$ is given by

$$u^0(x) = \begin{cases} \dfrac{-(x+2)}{1-t} & \text{if } -2 \leq x \leq -2t \\ \dfrac{x}{t} & \text{if } -2t \leq x \leq 3t \\ \dfrac{-3(x-2)}{2-3t} & \text{if } 3t \leq x \leq 2 \end{cases}.$$

The results for different numerical schemes are given on Figure **??**. The space interval $[-2,2]$ is discretized with 50 or 100 cells, and the functional (11) is minimized at the last iteration at $T = 0.4$. The time step is restricted with the CFL condition

$$\Delta t = \frac{0.5\Delta x}{\max_j |u_j^n|}$$

and we use periodic boundary conditions.

Four first order schemes are compared: the Rusanov, Godunov, Roe and MacCormack schemes 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 first two schemes are entropy satisfying: an explicit formula for $\mathcal{G}$ yielding nonpositive numerical diffusions (8) is known. The last two schemes are known to fail in sonic rarefactions (smooth part of the solution crossing $u = 0$) because they create nonphysical stationary discontinuities that violate the second equation of (3). We

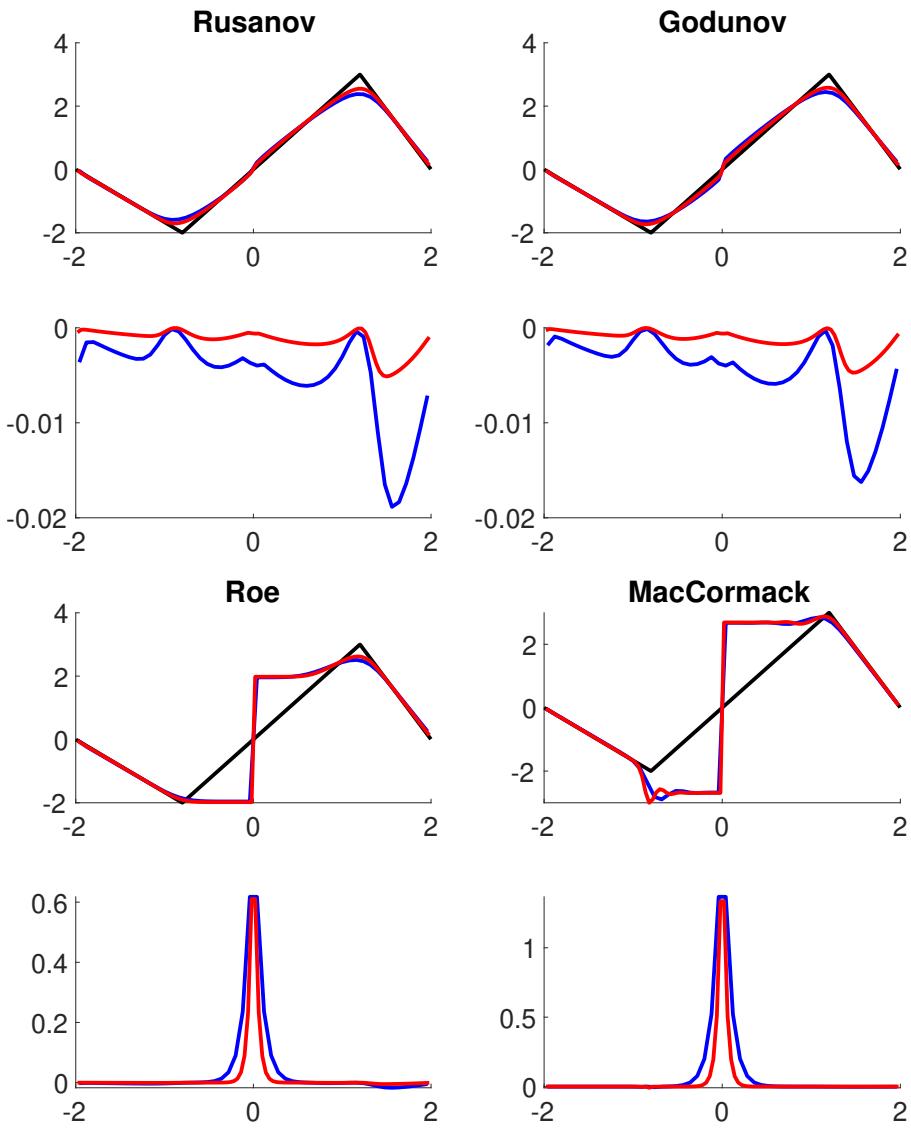

FIGURE 3. Entropy satisfying and entropy violating numerical schemes on Testcase (22). Lines 1 and 3: approximate solution at the last iteration in time. Lines 2 and 4: a posteriori quantification of the numerical diffusion $D^{opt,n}$. Results for $M = 50$ in blue, $M = 100$ in red, exact solution in black.

also tested the Osher and Lax-Wendroff fluxes. They give results very similar to the Rusanov and MacCormack schemes respectively and are not represented on Figure 3.

We define the a posteriori entropy fluxes as the results of the minimization (10), and define the a posteriori numerical diffusion according to (8) as

$$D_j^{opt,n} = \eta\left(u_j^{n+1}\right) - \eta\left(u_j^n\right) + \frac{\Delta t}{\Delta x}\left(G_{j+1/2}^{opt,n} - G_{j-1/2}^{opt,n}\right). \tag{23}$$

The minimization procedure detects the two previous behaviors: $D_j^{opt,n}$ remains nonpositive everywhere for the Rusanov and Godunov flux, while it has large strictly positive values localized around the stationary entropy creating shocks at the sonic point.

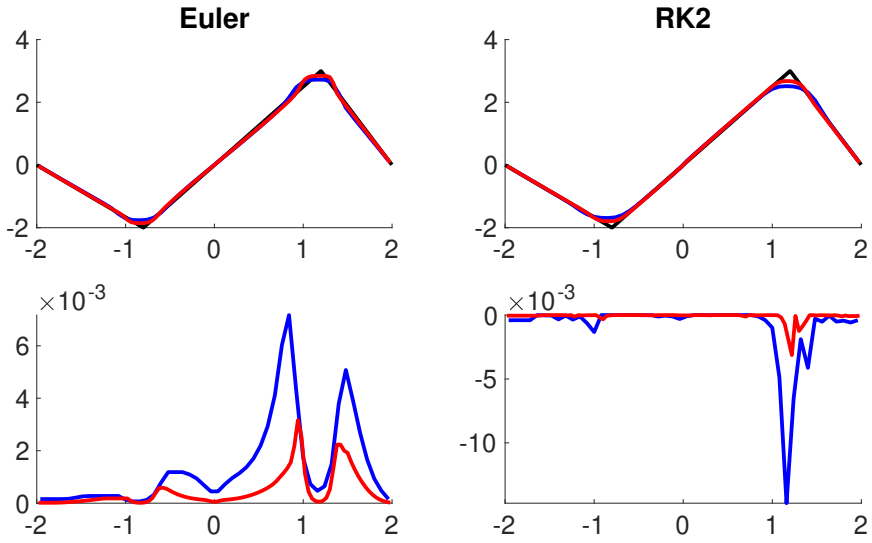

FIGURE 4. Influence of the time discretization for a second order scheme in space. First line: approximate solution at the last iteration in time. Second line: a posteriori quantification of the numerical diffusion $D^{opt,n}$. Results for $M = 50$ in blue, $M = 100$ in red, exact solution in black.

The two plots of Figure 4 show the results for MUSCL second order flux in space based a Rusanov flux and a minmod limiter (see the appendix for details). Together with a forward Euler march in time, (7) is grossly false in the sense that the total energy increases with time: $\sum_j \eta\left(u_j^{n+1}\right) > \sum_j \eta\left(u_j^n\right)$. Combined with a RK2 time stepping, the a posteriori procedure finds a discrete entropy inequality (7).

In the case of the Rusanov flux (32) combined with a forward Euler march in time, we indicate on Figure 5 the evolution of the total diffusion

$$\Delta x \left| \sum_j \left(u_j^{n+1}\right)^2 - \left(u_j^n\right)^2 \right|$$

and the total discrepancy $\Delta x \sum \left| D_j^{opt,n} - D_j^n \right|$. The numerical diffusion D_j^n is given by (8) with the standard numerical entropy fluxes (33), for which the discrete entropy inequality (7) holds. The figure illustrates Proposition 8: the exact solution is regular, the scheme is entropy satisfying, and the discrepancy decreases one

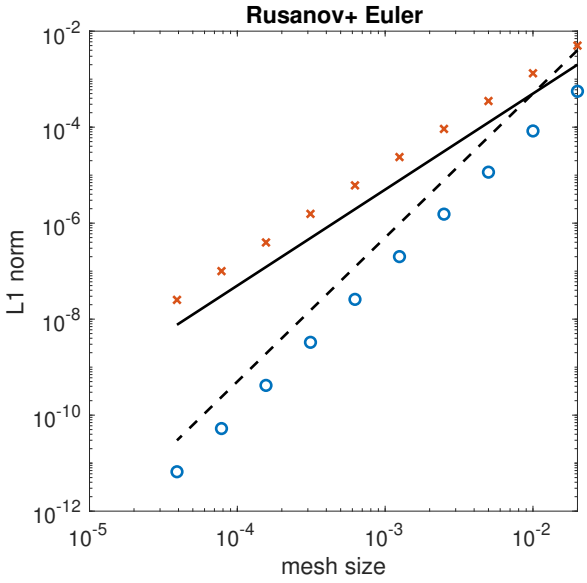

FIGURE 5. Total numerical diffusion (red crosses) and total discrepancy between the diffusion (blue circles) at the last iteration of the Testcase 22 with finer and finer grids.

order faster than the total diffusion. Here the orders are 3 and 2 because they are integrated on a grid of size Δx and on a time step of $\Delta t = O(\Delta x)$.

4.2. **Gas dynamics.** We now turn to the Euler equations (4). The method is indifferent to the number of unknowns d, since the a posteriori quantification of the numerical diffusion only concerns the scalar equation (2). We focus on a widely chosen scheme of order two in space and time, namely the Van Leer version of the MUSCL scheme [24].

In this scheme the piecewise constant in space approximation $\left(u_j^n\right)_j$ by a reconstructed piecewise affine data. In spirit the fluxes would be computed as the exact flux of (1) with this new detailed initial data. This is only feasible for some particular hyperbolic systems and incredibly costly; a second order approximation is sufficient and can be obtained with a two step Runge Kutta (RK2) time discretization. Bibliographic references and some details are given in the Appendix.

For scalar equations $d = 1$, this 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.

The discrete entropy inequalities found in the literature for the MUSCL approach differ from (7). The quantity $\eta\left(u_j^n\right)$ is replaced with a linear approximation in [4, Equation (1.9)] for scalar equations or with a nonlinear entropy diminishing projection operator in [9, Theorem 2.9] for systems. Those schemes rely on

generalized Riemann problems and are difficult to implement in practice. In the more convenient strategy described above, Berthon obtained some variation of (7) for the Euler equation, where $\eta\left(u_j^n\right)$ is replaced by a convex combination of three terms that depends not only of u_j^n but also on u_{j-1}^n and u_{j+1}^n, see [1, Equations (2.7) and (2.10)]. However, these modified entropy inequalities are not sufficient to prove a Lax–Wendroff like theorem. In [2], several numerical simulations indicates that there is most likely no such theorem, and that the MUSCL+RK2 scheme may converge to incorrect solutions.

We first reproduce the Sod tube testcase of Toro

$$\begin{cases} \rho^0(x) = \mathbf{1}_{x<0} + 0.125 \times \mathbf{1}_{x\geq 0} \\ u^0(x) = 0.75 \times \mathbf{1}_{x<0} \\ p^0(x) = \mathbf{1}_{x<0} + 0.1 \times \mathbf{1}_{x\geq 0} \end{cases} \tag{24}$$

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

On Figure 6 we compare the first order HLLC scheme [23, Section 10.4.2] and the Roe scheme without entropy fix [23, Section 11.2] 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 in two spikes around each discontinuity. This depends on the initialization, see Figure 9 below.

Then we consider a testcase where the solution does not contain a shock, but only a contact discontinuity. In this case the entropy should be conserved on the choosen time interval. The velocity and the pressure are initially constant, equal to 0.1 and 1. The initial density is

$$\rho^0(x) = 1 + 0.2x + 0.05\sin(6\pi x) + 0.4 \times \mathbf{1}_{x<0} \tag{25}$$

The final time is 2 seconds, the CFL number is 0.75. We compare the Rusanov scheme and the HLLC scheme, and find unsurprisingly that the latter is much less diffusive. The same holds for their second order extensions with a MUSCL procedure, using either the Rusanov 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.** The minimization procedure presented above finds numerical entropy fluxes and couples all the cells of the mesh. We used the blackbox fminunc of matlab, with the initial

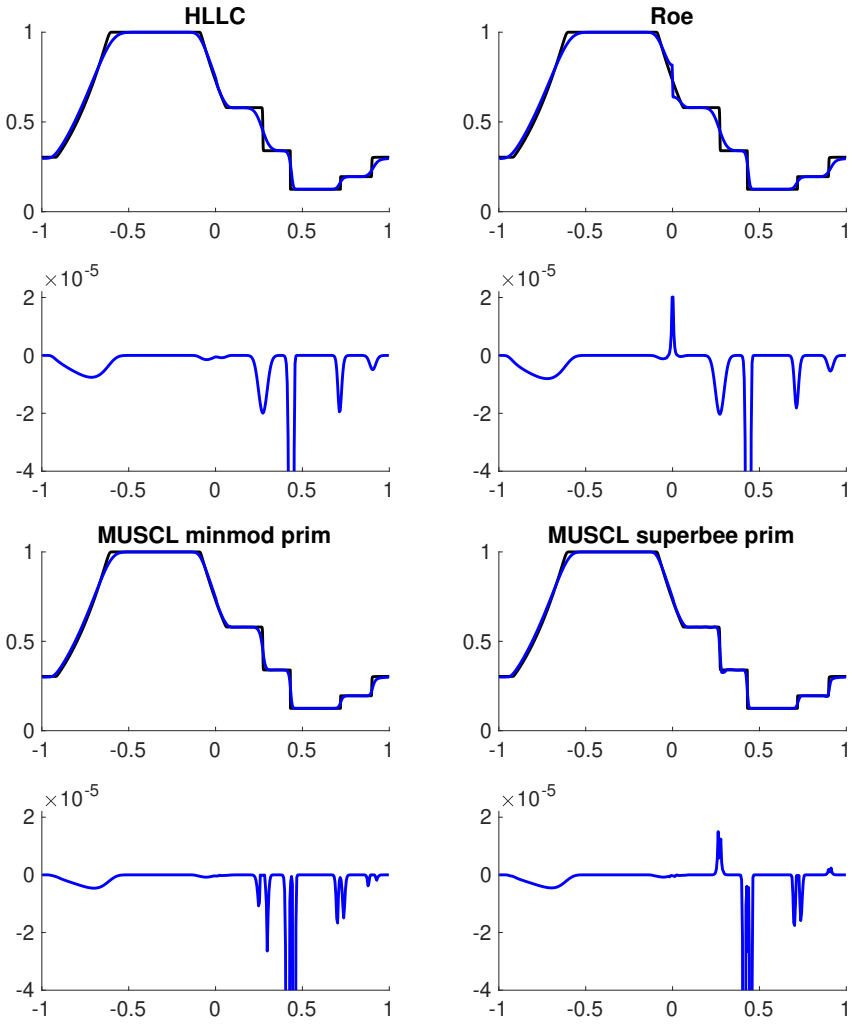

FIGURE 6. Densities (lines 1 and 3) and a posteriori numerical diffusion (lines 2 and 4) for several numerical schemes on Test-case (24).

guess $\gamma_{j+1/2} = \frac{1}{2}\left(m^n_{j+1/2} + M^n_{j+1/2}\right)$. The cost grows quadratically with the mesh-size, and becomes larger than the minute for meshes with more than 4 000 cells, see Figure 8, left. Here we present a much cheaper way to quantify the numerical diffusion. It does not inherit the good mathematical properties of the minimization procedure, but the numerical results are quite similar and the computational cost is drastically reduced.

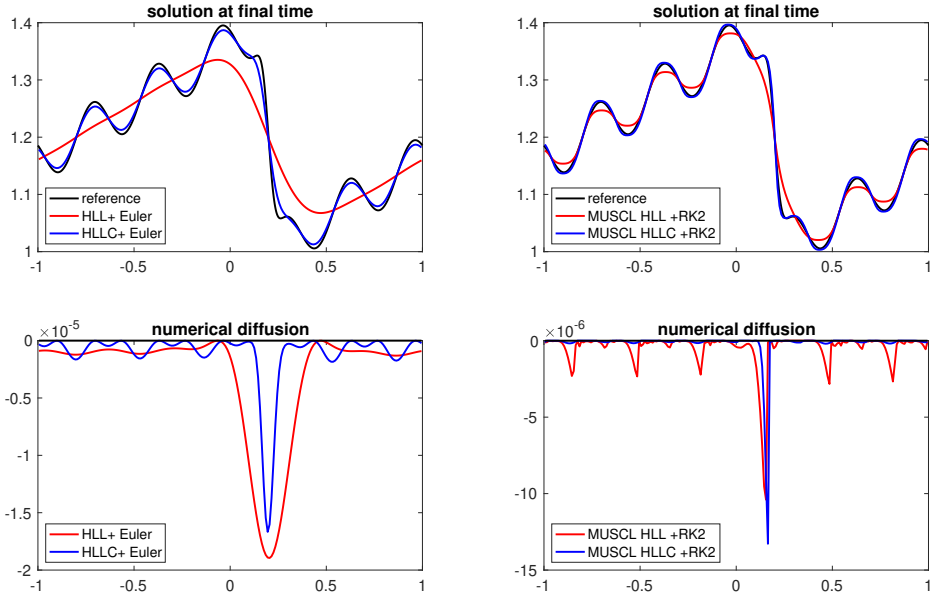

FIGURE 7. Densities (first line) and numerical diffusion (second line) for Testcase (25). The discontinuity is a slowly moving contact.

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 $\underline{D}^n_j \leq D^{opt,n}_j \leq \bar{D}^n_j$, with

$$\underline{D}^n_j = \eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right) + \lambda\left(m^n_{j+1/2} - M^n_{j-1/2}\right) \tag{26}$$

and

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

An important point is that $\underline{D}^n_j$ and $\bar{D}^n_j$ are computationally affordable. Indeed, $m^n_{j+1/2}$ and $M^n_{j+1/2}$ are computed with the scheme (6), 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 5.

Numerically, we observed than the lower bound (26) is particularly interesting for two reasons. First, if the left hand side is positive it indicates than (7) does not hold on this particular cell. We detail and capitalize on that idea in Section 5. Second, even though it is much lower than $D^{opt,n}_j$, the variations of $\underline{D}^n_j$ are similar.

As a final stage, we perform a naive renormalization of $\underline{D}^n_j$ by a constant coefficient α, in such a way that the total amount of numerical diffusion is correct:

$$\alpha \sum_j \underline{D}^n_j = \sum_j D^{opt,n}_j = \sum_j \eta\left(u^{n+1}_j\right) - \eta\left(u^n_j\right).$$

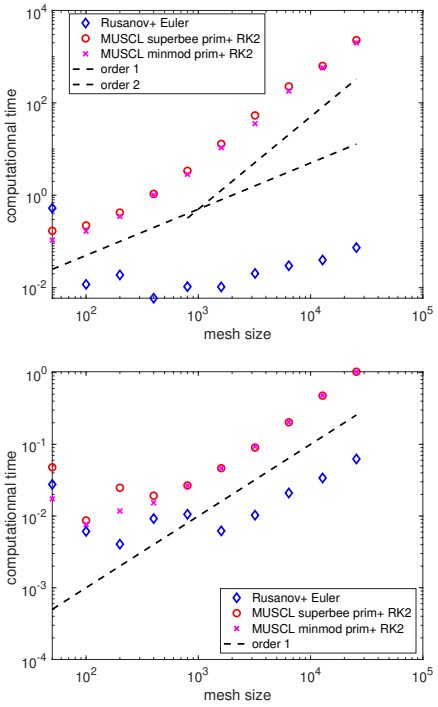

FIGURE 8. Computational cost of the minimization procedure (left) and of the a priori guess (right) at the first iteration of Test-case (25)

It yields to the a priori quantification of the numerical diffusion

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

On Figure 9 we compare $D_j^{priori,n}$ 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 $D_j^{priori,n}$ gives smoother results, with a diffusion spread out the whole length of the discrete discontinuity.

To conclude, the computation of $D_j^{priori,n}$ is pertinent if one is only interested in the numerical diffusion and not in the numerical entropy fluxes, which have been completely forgotten here. The results are visually satisfactory. Note that we are not able to extend Proposition 8.

5. An entropy stress test for numerical schemes

5.1. Worst initial data in terms of entropy.

Proposition 9. *Consider a numerical flux $\mathcal{F}$ with a stencil of s_L points on the left and and s_R on the right. Then the finite volume scheme does not have a discrete*

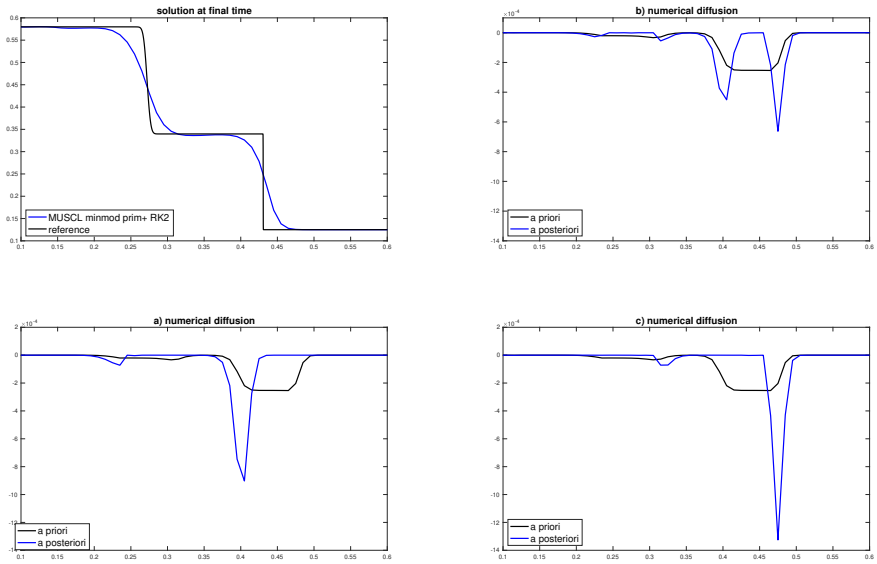

FIGURE 9. Comparison of $D_j^{opt,n}$ and $D_j^{priori,n}$ on Testcase (24). The optimization is initialized with (a) $\gamma_{j+1/2}^n = m_{j+1/2}^n$ or (b) $\gamma_{j+1/2}^n = \frac{1}{2}\left(m_{j+1/2}^n + M_{j+1/2}^n\right)$ or (c) $\gamma_{j+1/2}^n = M_{j+1/2}^n$

entropy inequality if and only if there exists an initial data

$$\left(u_{-s_L}^0, u_{-s_L+1}^0, \dots, u_0^0, \dots, u_{s_R-1}^0, u_{s_R}^0\right)$$

such that

(27)
$$\min\left(\eta\left(u_0^0\right) - \frac{\Delta t}{\Delta x}\left(m_{1/2}^0 - M_{-1/2}^0\right) - \eta\left(u_0^1\right), m_{1/2}^0 - M_{1/2}^0, m_{-1/2}^0 - M_{-1/2}^0\right) < 0$$

where $M_{1/2}^0$, $m_{1/2}^0$, $M_{-1/2}^0$ *and* $m_{-1/2}^0$ *are given by* (13) *with* $n = 0$.

Proof. If the scheme if entropy satisfying, for all choice of initial data

$$\left(u_{-s_L}^0, u_{-s_L+1}^0, \dots, u_0^0, \dots, u_{s_R-1}^0, u_{s_R}^0\right),$$

there exist two consistent entropy numerical fluxes $G_{1/2}^0$ and $G_{-1/2}^0$ such that

$$\eta\left(u_0^1\right) \leq \eta\left(u_0^0\right) - \frac{\Delta t}{\Delta x}\left(G_{1/2}^0 - G_{-1/2}^0\right). \tag{28}$$

Moreover, we know from Lemma 5 that $G_{\pm 1/2}^0 \in [m_{\pm 1/2}^0, M_{\pm 1/2}^0]$.

Thus there are two possibilities for a scheme to be non entropy satisfying. First, a set of bounds can be incorrectly ordered $M_{1/2}^0 < m_{1/2}^0$ or $M_{-1/2}^0 < m_{-1/2}^0$. Going back to the definition it simply means that there is a $s_L + s_R$ initial data

$$\left(u_{-s_L}^0, u_{-s_L+1}^0, \dots, u_0^0, \dots, u_{s_R-1}^0\right)$$

such that the total entropy increases in one iteration

$$\sum_{k=-s_L}^{s_R-1} \eta\left(u_k^1\right) > \sum_{k=-s_L}^{s_R-1} \eta\left(u_k^0\right) - \frac{\Delta t}{\Delta x}\left(G\left(u_{s_R-1}^0\right) - G\left(u_{s_L}^0\right)\right)$$

where the Neumann boundary conditions

$$u_k^0 = \begin{cases} u_{-s_L}^0 & \text{for } k < -s_L \\ u_{s_R-1}^0 & \text{for } k > s_R - 1 \end{cases}$$

are used to update the solution. The second possibility is that the bounds are correctly ordered but

$$\eta\left(u_0^1\right) > \eta\left(u_0^0\right) - \frac{\Delta t}{\Delta x}\left(m_{1/2}^0 - M_{-1/2}^0\right).$$

In that case for all choice $\left(G_{-1/2}^0, G_{1/2}^0\right) \in [m_{-1/2}^0, M_{-1/2}^0] \times [m_{1/2}^0, M_{1/2}^0]$,

$$\eta\left(u_0^1\right) > \eta\left(u_0^0\right) - \frac{\Delta t}{\Delta x}\left(G_{1/2}^0 - G_{-1/2}^0\right),$$

and (7) cannot hold.

□

5.2. **RK2+MUSCL.** The procedure of Section 5.1 is applied to the equation of gas dynamic (4) approximated by the common MUSCL scheme in space and a two step Runge-Kutta march in time. We consider the limitations in primitive variables (ρ, v, p), entropic variables (ρ, v, s) and conservative variables (ρ, q, E) of [1]. 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. Details and references are given in the appendix. For the overall scheme we have $s_L = s_R = 4$ and the minimization of Proposition 9 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 constraint to belong to

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

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

$$\left(u_{-s_L}^0 = u_{-s_L+1}^0 = \ldots, u_0^0 = \cdots = u_{s_R-1}^0 = u_{s_R}^0\right)$$

then $M_{\pm 1/2}^0 = m_{\pm 1/2}^0$ and $\eta\left(u_0^1\right) = \eta\left(u_0^0\right)$ and the three components of Proposition 9 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

$$(\rho_0, u_0, p_0) \in [0.11, 5] \times [-0.2, 10] \times [0.11, 10]. \tag{29}$$

and another 30 000 initializations in the smaller region

$$(\rho_0, u_0, p_0) \in [0.8, 1.2] \times [-0.2, 1] \times [0.8, 1.2]. \tag{30}$$

On local extremum the MUSCL scheme is identical to the chosen first order scheme. Thus is seems more likely to find counterexamples on monotonic data. Thus half of

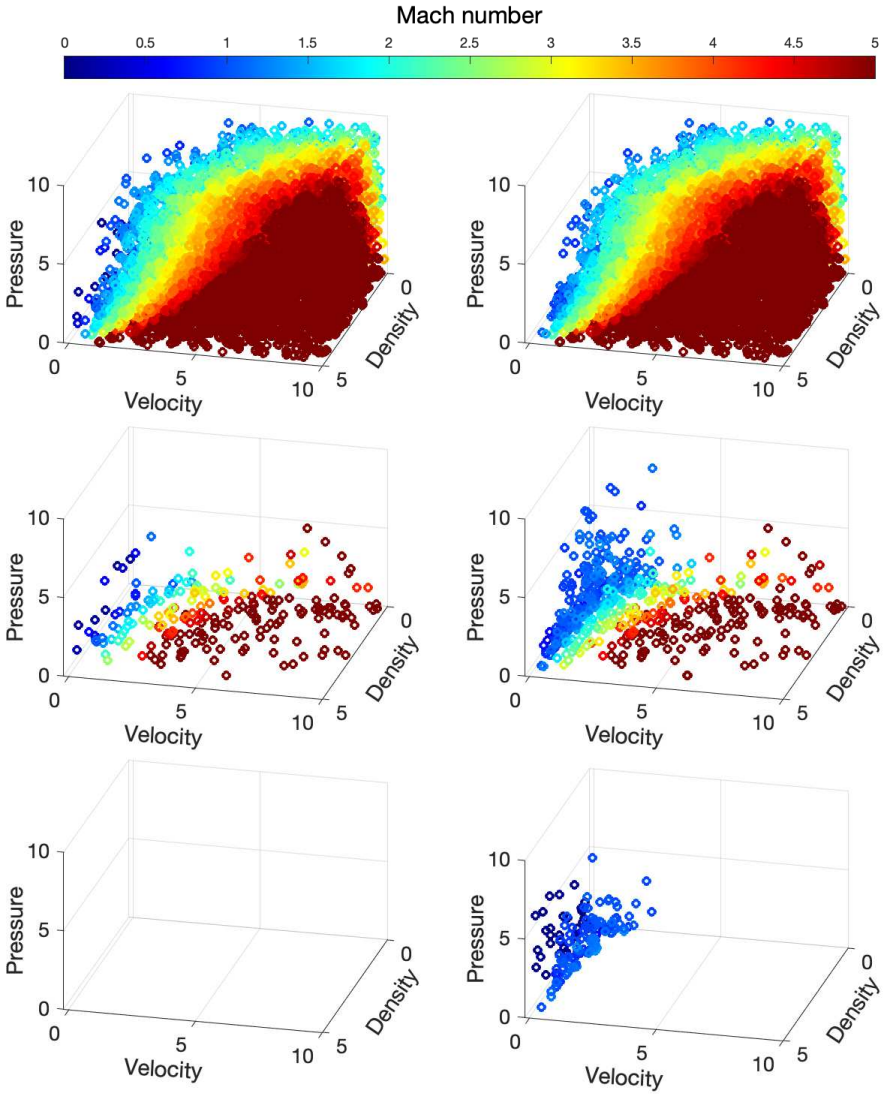

FIGURE 10. Counterexamples on the region $[0.11, 5] \times [-0.2, 10] \times [0.11, 10]$. The limited variables are (ρ, q, E) in the first line, (ρ, u, p) in the second line, (ρ, u, s) in the third line. The first order underlying scheme is HLL on the left and HLLC on the right.

the initialization are modified so that each variable is increasing or decreasing. The rearrangement is made either in the primitive, entropic or conservative variables, in equal proportion.

The minimization procedure constructs numerous initial data that violate (7) in the first time step. They are gathered on Figure 10 when (ρ_0, u_0, p_0) belongs to the large region (29) and on Figure 11 when it belongs to the smaller region (30). For

the latter the Mach number $|v|/\sqrt{\frac{\gamma p}{\rho}}$ does not exceed 1.04 while it can reach 60 in the large region.

Each point on these Figures represents the mean value

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

of a counterexample. The results for the two versions of the MUSCL 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.

The influence of the choice of the set variables which are limited is striking on Figure 10, with more counterexamples for the conservative variables (ρ, q, E) than for the primitive variables (ρ, v, p). There is even less counterexamples for the entropic variables (ρ, v, s). This hierarchy is less clear on the zoom of Figure 11. The distribution on counterexamples around $|v| = 0$ depends on the numerical choices: sometimes there is a gap around $v = 0$ and sometimes not.

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 (7) 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 on Figure 10. The chosen first order scheme is not to blame for the lack of discrete entropy inequality. Indeed in that case the exact flux is $f^n_{j+1/2}$ is either $f\left(u^n_j\right)$ or $f\left(u^n_{j+1}\right)$ depending on the sign of the velocity and most numerical schemes reproduce that.

Eventually Functional (27) remains nonnegative for all the random initializations when the limitation is in the entropic 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 search for counterexamples in the much larger domain

$$\forall j \in \{-4, \cdots, 4\} \ \left(\rho_j^0, u_j^0, p_j^0\right) \in [0.001, 10] \times [-50, 50] \times [0.001, 20].$$

This search was unsuccessful with 25 000 random initializations. Interestingly it holds for the overlimited version of the MUSCL scheme of [1] for which a modified version of (7) is proven, but also for the simpler original scheme. It supports the following conjecture.

Conjecture. *The RK2+MUSCL scheme for the gas dynamic equation* (4) *with*

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

verifies a discrete entropy inequality for a CFL number α *of* 0.1.

The same type of numerical experiment has been performed for the scalar Burgers equation 3. Three entropy satisfying (Rusanov, Osher and Godunov) and three non entropy satisfying (Roe, Lax-Wendroff and Mac Cormack) first order schemes have been tested.

Conjecture. *Consider the RK2+MUSCL scheme for the Burgers equation* (4) *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.*

References

[1] C. Berthon. Stability of the MUSCL Schemes for the Euler Equations. Communications in Mathematical Sciences, 3(2):133 – 157, 2005.

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

[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. Comp., 65(216):1439–1461, 1996.

[5] H. Burchard and H. Rennau. Comparative quantification of physically and numerically induced mixing in ocean models. Ocean Modelling, 20(3):293–311, 2008.

[6] C. Chalons and P. G. LeFloch. A fully discrete scheme for diffusive-dispersive conservation laws. Numer. Math., 89(3):493–509, 2001.

[7] S. Clain, S. Diot, and R. Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). J. Comput. Phys., 230(10):4028–4050, 2011.

[8] P. Colella and P. R. Woodward. The piecewise parabolic method (ppm) for gas-dynamical simulations. Journal of computational physics, 54(1):174–201, 1984.

[9] F. Coquel and P. G. LeFloch. An entropy satisfying MUSCL scheme for systems of conservation laws. Numer. Math., 74(1):1–33, 1996.

[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. Journal of Computational Physics, 343:235–270, 2017.

[11] S. Diot, S. Clain, and R. Loubère. Improved detection criteria for the multi-dimensional optimal order detection (MOOD) on unstructured meshes with very high-order polynomials. Comput. & Fluids, 64:43–63, 2012.

[12] B. Einfeldt, C.-D. Munz, P. L. Roe, and B. Sjögreen. On Godunov-type methods near low densities. J. Comput. Phys., 92(2):273–295, 1991.

[13] B. Fox-Kemper, A. Adcroft, C. W. Böning, E. P. Chassignet, et al. Challenges and prospects in ocean circulation models. Frontiers in Marine Science, 6, 2019.

[14] E. Godlewski and P.-A. Raviart. Numerical approximation of hyperbolic systems of conservation laws, volume 118 of Applied Mathematical Sciences. Springer-Verlag, New York, 2021. Second edition.

[15] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes. I. SIAM J. Numer. Anal., 24(2):279–309, 1987.

[16] A. Hiltebrand and S. Mishra. Entropy stable shock capturing space-time discontinuous Galerkin schemes for systems of conservation laws. Numer. Math., 126(1):103–151, 2014.

[17] R. M. Holmes, J. D. Zika, S. M. Griffies, A. M. Hogg, A. E. Kiss, and M. H. England. The geography of numerical mixing in a suite of global ocean models. Journal of Advances in Modeling Earth Systems, 13(7), 2021.

[18] K. Klingbeil, M. Mohammadi-Aragh, U. Gräwe, and H. Burchard. Quantification of spurious dissipation and mixing – discrete variance decay in a finite-volume framework. Ocean Modelling, 81:49–64, 2014.

[19] P. Lax and B. Wendroff. Systems of conservation laws. Communications on Pure and Applied Mathematics, 13(2):217–237, 1960.

[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, hal-03206727, 2021.

[21] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. Numer. Math., 73(1):119–130, 1996.
[22] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys., 43(2):357–372, 1981.
[23] E. F. Toro. Riemann solvers and numerical methods for fluid dynamics. Springer-Verlag, Berlin, third edition, 2009. A practical introduction.
[24] B. van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. J. Comput. Phys., 135(2):227–248, 1997.
[25] X. Zhang and C.-W. Shu. Positivity-preserving high order finite difference WENO schemes for compressible Euler equations. J. Comput. Phys., 231(5):2245–2258, 2012.

Description of the numerical schemes

In this section we describe the schemes used in the numerical simulations.

Time step restriction. The time step is fixed proportionally to the space: a Courant number $\alpha \in (0,1)$ is fixed and we take for the Burgers equation

$$\Delta t = t^{n+1} - t^n = \frac{\alpha \Delta x}{\max_j |u_j^n|}$$

and for gas dynamics

$$\Delta t = t^{n+1} - t^n = \frac{\alpha \Delta x}{\max_j \left(|v_j^n| + \sqrt{\gamma p_j^n / \rho_j^n} \right)}. \tag{31}$$

First order schemes.

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

$$\mathscr{F}(u_L, u_R) = \frac{f(u_L) + f(u_R)}{2} - \frac{A(u_L, u_R)}{2}(u_R - u_L) \tag{32}$$

and the numerical entropy flux is

$$\mathcal{G}(u_L, u_R) = \frac{G(u_L) + G(u_R)}{2} - \frac{A(u_L, u_R)}{2}(\eta(u_R) - \eta(u_L)). \tag{33}$$

The scalar quantity $A(u_L, u_R)$ should be large enough to stabilize the centered flux $\dfrac{f(u_L) + f(u_R)}{2}$. For scalar equation, we set

$$A(u_L, u_R) = \max(|f'(u_L)|, |f'(u_R)|).$$

For the equation of gas dynamic (4) (with the classical notation of [23, Chapter 4]) the optimal choice would be

$$\begin{aligned} A(u_L, u_R) = \max\Bigg(& \left| v_L - \sqrt{\frac{\gamma p_L}{\rho_L}} \right|, \left| v_* - \sqrt{\frac{\gamma p_*}{\rho_{L*}}} \right|, \\ & \left| v_* + \sqrt{\frac{\gamma p_*}{\rho_{R*}}} \right|, \left| v_R + \sqrt{\frac{\gamma p_R}{\rho_R}} \right| \Bigg) \end{aligned} \tag{34}$$

However, Formula (34) requires to solve a Riemann problem at each interface to compute the middle velocity and pressure (v_*, p_*), which is very costly. Here we simply use

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

The HLL scheme. This scheme is also a simple two waves approximate Riemann solver. We present the scheme for gas dynamic. Following [23, Section 10.3] we denote by a the sound speed

$$a_L = \sqrt{\gamma p_L/\rho_L} \quad \text{and} \quad a_R = \sqrt{\gamma p_R/\rho_R}$$

and the Roe averages

$$\bar{a} = \frac{\sqrt{\rho_L}a_L + \sqrt{\rho_R}a_R}{\sqrt{\rho_L} + \sqrt{\rho_R}} \quad \text{and} \quad \bar{v} = \frac{\sqrt{\rho_L}v_L + \sqrt{\rho_R}v_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}.$$

The wave speed are estimate as

$$S_L = \min(v_L - a_L, v_R - a_R, \bar{v} - \bar{a}) \quad \text{and} \quad S_R = \min(v_L + a_L, v_R + a_R, \bar{v} + \bar{a})$$

the flux is given by [23, Equation (10.21)].

The HLLC scheme for gas dynamics. This scheme is commonly used for numerical simulations of (4). This approximate Riemann solver contains two intermediate states and is by construction exact on isolated contact discontinuities. As a consequence it is much less diffusive than the Rusanov or HLL schemes?

It relies on an approximation of the exact solution of the solution with piecewise constant initial data centered in each interface of the mesh. The strength of this solver is that some important particular solution of (4), the stationary contact discontinuities, are exactly captured by this scheme. We implemented two versions of the schemes. The first one corresponds to [23, Paragraph 10.4.2]. The extremal wave speed are estimated with [23, (10.49)] and the pressure and velocities are constant in the "star region". The second one is based on a pressure estimate and corresponds to [23, Paragraph 10.6, variant 1].

The Roe scheme. In the case of scalar conservation laws the numerical scheme is

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

where $\sigma = \dfrac{f(u_R) - f(u_L)}{u_R - u_L}$. We refer the reader to [22] and [23, Chapter 11] for the presentation of the Roe scheme for the Euler equations (4).

This scheme does not have an 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] when applied to the equation of gas dynamics, in which case the simulation fails entirely. However, this scheme and its various correction behaves appropriately in some regimes, with a low numerical diffusion and a reasonable computational cost.

Lax Wendroff scheme and MacCormack scheme. For scalar conservation laws the Lax Wendroff scheme is given by

$$\mathcal{F}(u_L, u_R) = \frac{f(u_L) + f(u_R)}{2} - \frac{\lambda}{2} f'\left(\frac{u_L + u_R}{2}\right)\big(f(u_R) - f\big(u_L\big)\big)$$

The MacCormack scheme is given by

$$\mathcal{F}(u_L, u_R) = \frac{f(u_R) + f\left(u_L - \frac{\Delta t}{\Delta x}\big(f(u_R) - f(u_L)\big)\right)}{2}$$

These schemes do not have a discrete entropy inequality since they are exact on stationnary entropy violating shocks.

The second order MUSCL scheme. The MUSCL procedure is a commonly used procedure to obtain a second ordre scheme in space, and is easily combined with a second order time scheme such as a 2 step Runge-Kutta method.

For scalar conservation law, the procedure is the following. At the beginning of each time step, the constant value u_j^n in cell $C_j = [x_j - \frac{\Delta x}{2}, x_j - \frac{\Delta x}{2}]$ 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. For example, $\sigma_j^n = \operatorname{minmod}\left(u_j^n - u_{j-1}^n, u_{j+1}^n - u_j^n\right)$ where the minmod limiter is defined as

$$\operatorname{minmod}(a,b) = \max(0, \min(a,b)) + \min(0, \max(a,b)).$$

Other limiters are possible, see [23, Section 13.7.3]. The MUSCL flux is given by

$$F_{j+1/2} = \mathcal{F}\left(u_{j,+}^n, u_{j+1,-}^n\right), \quad u_{j,\pm}^n = u_j^n \pm \frac{\Delta x}{2}, \tag{35}$$

where $\mathcal{F}$ is a first order two points scheme. Note that $u_{j,+}^n$ depends on u_{j-1}^n, u_j^n and u_{j+1}^n, while $u_{j+1,-}$ depends on u_j^n, u_{j+1}^n and u_{j+2}^n. Thus (35) is a 4-points flux with $s_L = s_R = 2$.

In the case of hyperbolic system with $p > 1$, this strategy is mimicked componentwise. For the Euler equation (4) we followed the strategy of [1]. The piecewise linear 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 entropic variables $\left(\rho, v, s = \dfrac{p}{\rho^\gamma}\right)$ and to deduce the values $(\rho v)_{j,\pm}^n = \rho_{j,\pm}^n v_{j,\pm}^n$ with

$$E_{j,\pm}^n = \frac{p_{j,\pm}^n}{\gamma - 1} + \frac{\rho_{j,\pm}^n \left(v_{j,\pm}^n\right)^2}{2} \quad \text{or} \quad E_{j,\pm}^n = \frac{(\rho_{j,\pm}^n)^\gamma s_{j,\pm}^n}{\gamma - 1} + \frac{\rho_{j,\pm}^n \left(v_{j,\pm}^n\right)^2}{2}.$$

A variation of the discrete entropy inequality (7) is obtained in [1] when $\mathcal{F}$ is an entropy satisfying first order flux. In the case of the limitation of conservative variables it has the form

$$\eta\left(u_j^{n+1}\right) \le \frac{\eta\left(u_{j,-}^n\right) + \eta\left(u_{j,+}^n\right)}{2} - \frac{\Delta t}{\Delta x}\left(\mathcal{G}\left(u_{j,+}^n, u_{j+1,-}^n\right) - \mathcal{G}\left(u_{j-1,+}^n, u_{j,-}^n\right)\right)$$

which does not imply the original (7) since the reconstruction step increases the entropy:

$$\eta\left(u_j^n\right) \le \frac{\eta\left(u_{j,-}^n\right) + \eta\left(u_{j,+}^n\right)}{2}$$

The restriction on the time step encompasses the reconstructed states $u_{j,\pm}^n$ in the computation of the maximal wavespeed.

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. We first compute

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

In a second step we compute

$$\bar{\bar{u}}_j^n = \bar{u}_j^n - \frac{\Delta t}{\Delta x}\left(\mathcal{F}\left(\bar{u}_{j-s_L+1}^n, \dots, \bar{u}_{j+s_R}^n\right)\mathcal{F}\left(\bar{u}_{j-s_L}^n, \dots, \bar{u}_{j+s_R-1}^n\right)\right)$$

The final update is

$$u_j^{n+1} = \frac{u_j^n + \bar{\bar{u}}_j^n}{2}.$$

The whole procedure rewrites in a compact form as

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(\bar{\mathcal{F}}\left(u_{j-2s_L+1}^n, \dots, u_{j+2s_R}^n\right) - \bar{\mathcal{F}}\left(u_{j-2s_L}^n, \dots, u_{j+2s_R-1}^n\right)\right)$$

with

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

This is indeed a numerical flux with $2s_L$ points to the left and $2s_R$ points to the right since the computations of $\bar{u}_{j-s_L+1}^n$ to $\bar{u}_{j+s_R}^n$ uses the values from $\bar{u}_{j-2s_L+1}^n$ to $\bar{u}_{j+2s_R}^n$.

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

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

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 [1, Equation (2.12)], with the slight modification that Δt_2 cannot exceed Δt_1.

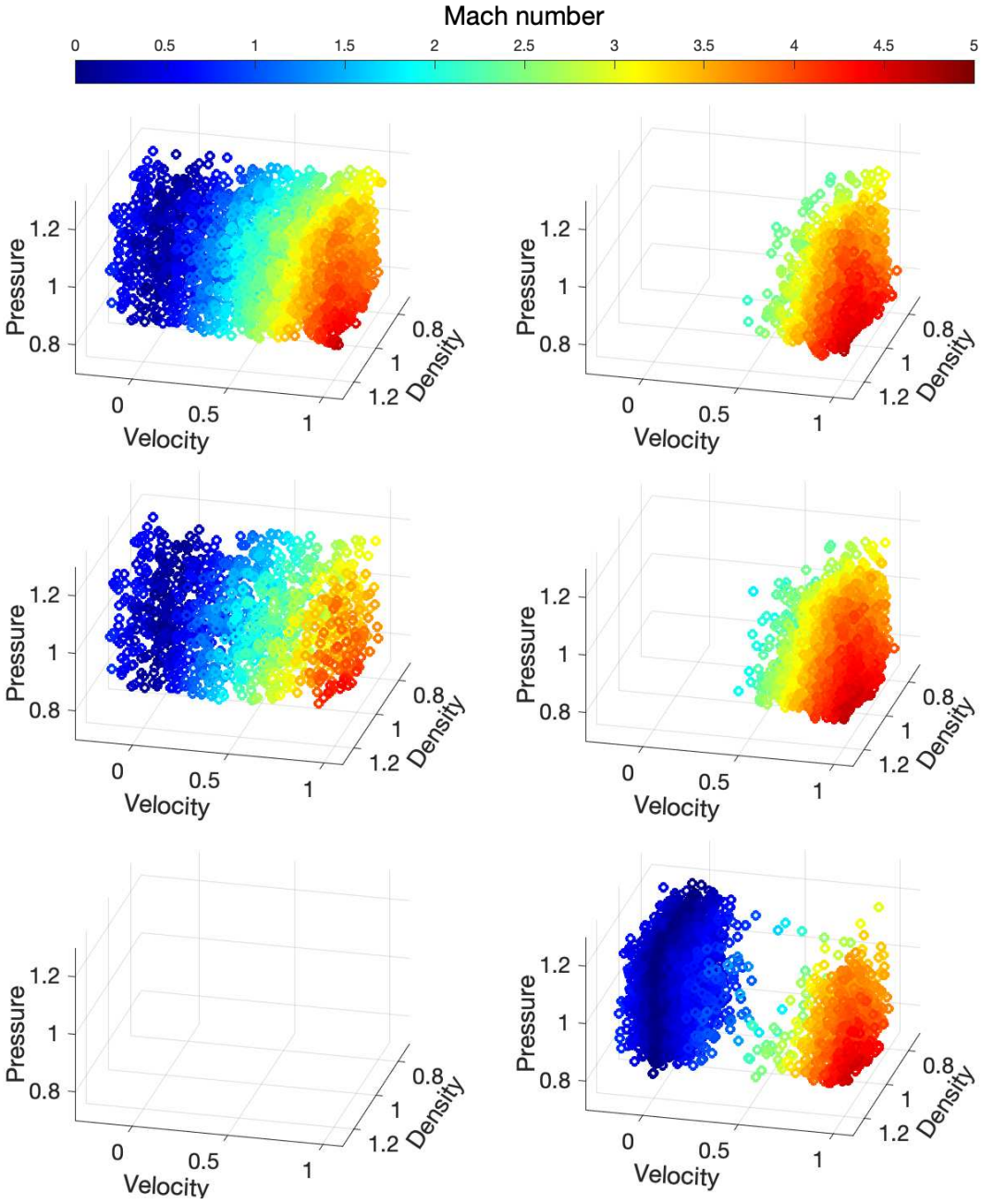

FIGURE 11. Zoom on a the region $[0.8, 1.2] \times [-0.2, 1] \times [0.8, 1.2]$.

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

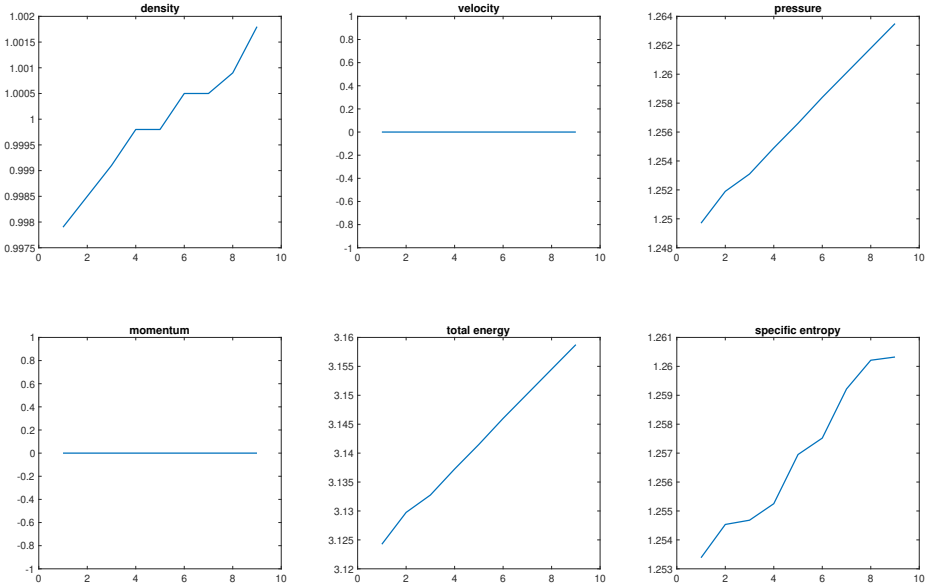

FIGURE 12. The MUSCL+RK2 scheme with this initial data is non entropy satisfying, with an HLL first order scheme