BIT Numerical Mathematics (2022) 62:1673–1701
https://doi.org/10.1007/s10543-022-00927-x

BIT

Using the Dafermos entropy rate criterion in numerical schemes

Simon-Christian Klein[1]

Received: 4 November 2021 / Accepted: 28 April 2022 / Published online: 1 June 2022
© The Author(s) 2022

Abstract
The following work concerns the construction of an entropy dissipative finite volume solver based on the convex combination of an entropy conservative and an entropy dissipative flux. We aim to construct a semidiscrete scheme that is entropy stable in the sense of the entropy criterion of Dafermos as well as in the classical sense entropy dissipative. The proposed semidiscrete scheme shows nice properties like $2p$ order accuracy in smooth regions as well as a non-oscillatory behavior around shocks.

Keywords Entropy stability · High-order methods · Finite-volume methods

Mathematics Subject Classification 35L03 · 35L45 · 35L65 · 35L67 · 65M08 · 65M12 · 65M20 · 76L05

1 Introduction

The robustness of numerical methods for hyperbolic conservation laws of the form

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f \circ u(x,t)}{\partial x} = 0 \quad \text{for} \quad u(x,t) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^m \quad \text{with} \quad f : \mathbb{R}^m \to \mathbb{R}^m \tag{1.1}$$

is greatly enhanced by numerical methods that do not only approximate (1.1) but also satisfy entropy inequalities

$$\frac{\partial U \circ u}{\partial t} + \frac{\partial F \circ u}{\partial x} \leq 0. \tag{1.2}$$

Communicated by Jan Nordström.

✉ Simon-Christian Klein
simon-christian.klein@tu-bs.de

[1] Institute for Partial Differential Equations, TU Braunschweig, Braunschweig, Germany

These are used to select one weak solution out of many possible weak solutions. One could further assume that the error for fixed grid size could be reduced by adhering to entropy inequalities. A scheme has to satisfy (1.2) in a discrete sense

$$\frac{U_k^{n+1} - U_k^n}{\Delta t} + \frac{F_{k+\frac{1}{2}}^n - F_{k-\frac{1}{2}}^n}{\Delta x} \leq 0$$

as proposed in [20, 30, 31] for all or at least one entropy pair (U, F). If solutions of a scheme satisfy all of these inequalities it is called an entropy stable scheme and entropy dissipative if only one entropy inequality is satisfied. Examples of schemes constructed with the aim of being entropy dissipative are for example given in [1, 2, 4, 13, 14, 19]. While the objective of this work is also centered around entropy dissipative schemes the motivation stems from an alternative entropy criterion by Dafermos. A second distinction lies in the fact that most of the aforementioned authors construct generalizations of finite element methods while this work is based on classical finite volume methods. We will first look at some numerical artifacts that can still occur with entropy dissipative schemes. Afterwards, a scheme will be constructed that is entropy dissipative and at least approximately satisfies the entropy condition of Dafermos [5] and some numerical tests using this scheme will be carried out. Dafermos defined a different entropy criterion using the total entropy in the domain

$$E_u(t) = \int U \circ u(x,t) \mathrm{d}x.$$

A Dafermos entropy solution u is a weak solution that satisfies

$$\forall t > 0: \quad \frac{\mathrm{d}E_u(t)}{\mathrm{d}t} \leq \frac{\mathrm{d}E_{\tilde{u}}(t)}{\mathrm{d}t}$$

compared to all other weak solutions $\tilde{u}$ of the conservation law (1.1). In essence entropy of the solution decreases faster than the entropy of all other solutions.

2 Comparing schemes by their entropy dissipation

Stable high order schemes are often constructed by the addition of suitable dissipation to an at least entropy conservative base scheme, e.g. [29]. The equation approximated by the resulting scheme is typically of the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon \frac{\partial^{2s} u}{\partial x^{2s}}.$$

The amount of dissipation ε has a strong influence on the resulting errors. To much dissipation results in a simulation of a diffusion (heat) equation while too small amounts of dissipation are responsible for oscillations and can lead to instabilities and order losses. It arises the question of why schemes with low entropy dissipation show bad

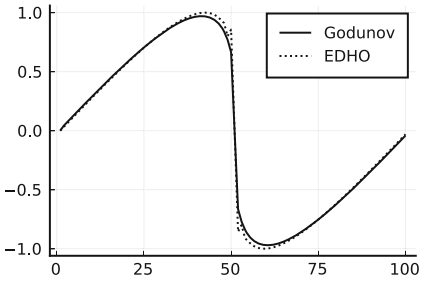
behavior even if they are formally of high order and entropy dissipative, as they fulfil the entropy inequality for at least one entropy. We would like to shed some light on the connection between the correct amount of entropy dissipation defined by the Dafermos criterion using the following numerical experiments. It should be noted that the Dafermos criterion was designed for solutions to conservation laws and not their numerical approximations; this means we will look at some numerical solutions and make some assumptions about their limit solutions and their behavior.

Numerical experiment 1 (Comparing two schemes by the Dafermos criterion) *A simulation of the Burgers equation*

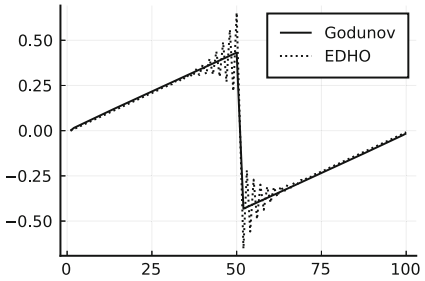
$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0$$

with $u_0(x) = \sin(\pi x)$ *was carried out on a periodic domain* $\Omega = [0, 2)$ *for* $t \in [0, 2]$. *The entropy conservative flux of order 4 from* [22, 32] *was used with a dissipation operator due to* [8] *and* $\varepsilon = 0.5$ *as dissipation coefficient. The numerical solution was compared to a solution calculated by a Godunov scheme. The solution and graphs of the complete entropy in the domain for the quadratic entropy can be seen in Fig.* 1. *Several times a new simulation was started using the solution of the entropy conservative fluxes in conjunction with dissipation as a starting point and the Godunov method as solver. The corresponding total entropy was also plotted in the total entropy diagrams. As we would like to be sure that our conclusions do not depend on the number of points in the domain the simulation was carried out once more with 3000 instead of 100 cells. We can clearly see that the entropy dissipative method produces bad results, because oscillations appear around the shock. We can also see that the Godunov scheme dissipates more entropy than the other scheme. The simulations which where carried out by the Godunov method with the solution of the high order method at different times as a starting point are especially interesting. These show a strong reduction of the total entropy until the total entropy of the solution calculated by the Godunov method from the beginning is reached. The Dafermos entropy criterion is only partially applicable in this case as the solutions are approximate solutions. It states in this case that the solution of the high order solver is not the entropy solution, although the solver is technically entropy dissipative, because the negative derivative of the total entropy can even be more negative. It should be noted that the Godunov method on the other hand dissipates entropy even for smooth solutions. This opposes the known theory of hyperbolic conservation laws, as smooth solutions satisfy an entropy equality referred to as an additional conservation law* [5, 20]. *The Godunov method satisfies the entropy equality only approximately as the entropy dissipation is small compared to the entropy dissipation after the onset of the shock, but not zero. The Godunov method is still the best possible three point first order method as it is the method with the least possible dissipation that converges to the entropy solution* [30, 31].

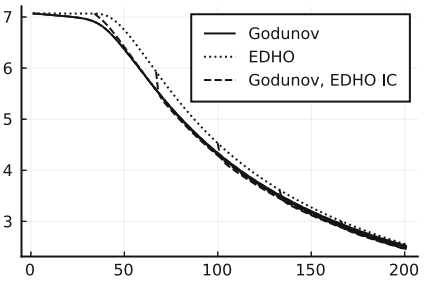
Numerical experiment 2 (Per cell dissipation of the Godunov method) *As we saw in the last example the Godunov method leads to a significantly higher total entropy*



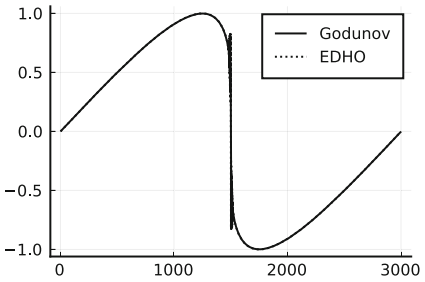
(a) Solution at $t = 0.34$ using $N = 100$ points



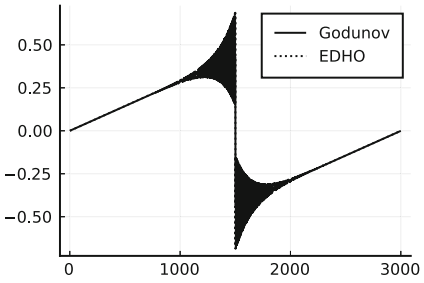
(b) Solution at $t = 1.99$ using $N = 100$ points



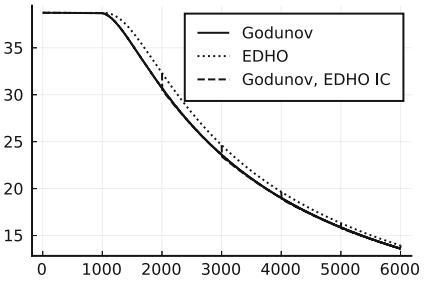
(c) Total entropies over time using $N = 100$ points



(d) Solution at $t = 0.33$ using $N = 3000$ points

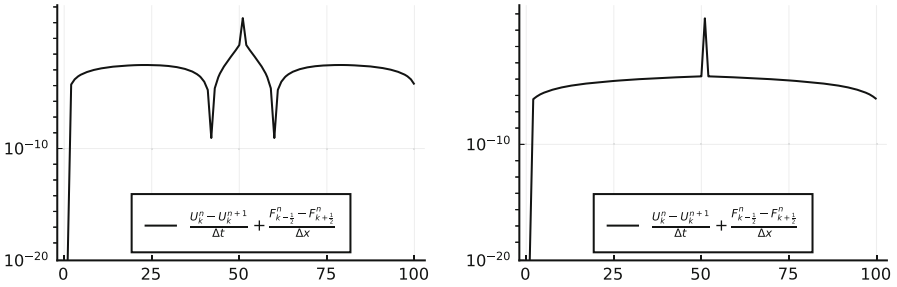


(e) Solution at $t = 2.0$ using $N = 3000$ points



(f) Total entropies over time using $N = 3000$ points

Fig. 1 Solution to $u_0(x) = \sin(\pi x)$ for the Burgers equation with $N = 100$ cells in the first 3 Graphs and $N = 3000$ cells in the last 3 Graphs. A simulation with the Godunov method was started with the solution of the entropy dissipative high order (EDHO) scheme as a starting point at different times.The Godunov method is the basic Godunov scheme with the exact Riemann solver for the Burgers equation and without any reconstruction. Time integration was carried out using a CFL number of $\lambda = 0.5$ and the SSPRK104 scheme.The high order scheme is composed of an entropy stable flux and a dissipation operator. The fourth order entropy conservative flux constructed out of Tadmors entropy conservative flux [32] for the Burgers equation and the linear combination developed by LeFloch, Mercier and Rhode [22] is used. A periodic fourth order dissipation operator with the coefficients given in [8] was used as a dissipation operator with strength $\varepsilon = 0.5$. Time integration was, as in the case before, done using the SSPRK104 method



(a) Entropy equality violation at $t = 0.34$ **(b)** Entropy equality violation at $t = 1.99$

Fig. 2 Per cell entropy inequality for the Godunov scheme. The same Godunov method was used as in Fig. 1 with $N = 100$ cells

reduction than our high order method, which leads to the question of where this dissipation occurs. This is why the violation of the entropy equation was plotted for the aforementioned numerical experiment for the Godunov method in Fig. 2. We can see that a small amount of entropy dissipation occurs during the simulation of a smooth solution while a much bigger amount of entropy dissipation occurs centered around a shock, if present, in compliance with the entropy inequality for shocks. This knowledge was already put to work in [34] *using edge sensors.*

The last two numerical experiments lead to a new design philosophy for numerical schemes. A good numerical scheme should not only be entropy dissipative in the sense of the entropy inequality. It should also dissipate the correct amount of entropy. This can be governed by the entropy equality for smooth areas, the entropy inequality for shocks and the Dafermos entropy criterion. The Godunov method violates this philosophy by dissipating entropy in smooth areas, while the aforementioned high order method dissipates less entropy than needed and possible around shocks, which violates the Dafermos criterion. It should be noted at the same time that schemes can also dissipate too much entropy in the vicinity of a shock or a maximum. Our proposed scheme will be built out of the following components.

- Let the scheme decide if the entropy equality or the entropy inequality holds in an area - this is equivalent to the presence of a shock.
- Use an entropy conservative flux if the entropy equality holds.
- Dissipate entropy with correct rate in the other case by the use of a dissipative first order flux.

Deciding which amount of entropy dissipation is the correct amount is a non-trivial sub-problem. It is not wise to aim for unconstrained maximized entropy dissipation in a numerical method as given by the Dafermos criterion. The reason for this is that the conservation law works as a constraint for the variational formulation of entropy dissipation. Numerical solvers can violate this constraint to some extent and dissipate even more entropy at the cost of higher approximation errors, as even more dissipation leads to larger approximation errors. This is why the highest amount of entropy loss that does not sacrifice low approximation errors is needed.

It is difficult to find a definition for a suitable amount of entropy loss that does not sacrifice low approximation errors. Godunov's method dissipates the least amount of entropy possible for single conservation laws of all E-fluxes [30, 31], and is thereby a natural candidate. Especially as a high order approximation makes no sense for a region of discontinuity.

Remark 2.1 One could ask why the Godunov and not an even more dissipative flux like the local Lax-Friedrichs flux should be selected. We are interested in the highest amount of dissipation that does not lower the accuracy, in the sense of the error between numerical approximation and exact (entropy) solution for fixed grid size. While the Lax-Friedrichs method has the same formal order of accuracy the method is less accurate for a fixed grid than the Godunov method and therefore violates our additional constraint. Another perspective can be that a higher entropy dissipation rate than the Godunov method has to be also higher than the entropy rate of the exact solution as the Godunov method uses averages of exact solutions. One could conjecture that such a high dissipation is not possible for any exact weak solution. The Lax-Friedrichs method will be still used in some of the following numerical tests to avoid solving Riemann problems for the Euler equations as the error between exact solution and LF method vanishes for growing grid sizes.

The following chapter is devoted to the construction of the aforementioned solver that tries to satisfy these requirements and uses the Godunov flux as a guide for the correct amount of entropy to dissipate. For simplicity this is done by using an entropy stable first order flux in this case for the dissipation and Tadmor's high order flux in entropy conservative areas of the domain. This scheme thereby should be a numerical scheme that at least approximately satisfies the Dafermos entropy criterion.

3 The best of both worlds

We will use entropy conservative fluxes as pioneered in [32]. A flux f will be termed entropy conservative if it satisfies a semidiscrete entropy equality

$$\frac{\mathrm{d}U(u_k(t))}{\mathrm{d}t} = \frac{F(u_{k-p-1},\ldots,u_{k+p}) - F(u_{k-p},\ldots,u_{k+p+1})}{\Delta x}.$$

Definition 3.1 *(Convex combination flux)* We define a new numerical flux by

$$f_\alpha^{GT}(u_i,u_{i+1}) = \alpha f^G(u_i,u_{i+1}) + (1-\alpha) f^T(u_i,u_{i+1})$$

where $\alpha \in [0,1]$ is a parameter controlling a convex combination between the Godunov flux presented in [11, 23] and the entropy conservative flux given in [32]. The value of

$$\alpha = \alpha(u_{i-p+1},\ldots,u_{i+p})$$

will in general depend on u_i and therefore the properties of the flux will depend on the selected function $\alpha(u_{i-p+1},\ldots,u_{i+p})$.

It should be clear that this construction does not depend on the use of the Godunov flux. In fact any other numerical flux function could be used, and we will refer to a flux constructed this way using the Lax-Friedrichs scheme as the LFT-Flux and to a flux constructed using the Godunov scheme as the GT flux. Several other entropy conservative fluxes [18, 26] have been constructed for some conservation laws and these can be also substituted for the basic Tadmor entropy conservative flux.

Lemma 3.1 *The GT-Flux is a consistent and local Lipschitz continuous numerical flux.*

Proof Consistency can be proved by direct insertion.

$$f_\alpha(u,u) = \alpha f^G(u,u) + (1-\alpha) f^T(u,u) = \alpha f(u) + (1-\alpha) f(u) = f(u)$$

We will interpret the arguments of the numerical fluxes as tuples $a = (u_i, u_{i+1})$ during the rest of the proof. The Godunov and Tadmor fluxes are Lipschitz continuous with the constants L_G and L_T,

$$\begin{aligned} \left| f_G(a) - f_g(b) \right| &\leq L_G \left\| a - b \right\| \\ \left| f_T(a) - f_T(b) \right| &\leq L_T \left\| a - b \right\|. \end{aligned}$$

We can conclude using the triangle inequality that the fluxes are also bounded for any bounded subset $U \subset \mathbb{R}^{2p\times m}$

$$\begin{aligned} \forall a \in U,\ \forall I \in \{G,T\}: \quad \left| f_I(a) \right| &\leq \left| f_I(a) - f_I(a_0) \right| + \left| f_I(a_0) \right| \\ &\leq \left| f_I(a_0) \right| + L_I \left\| a - a_0 \right\| \\ &\leq \left| f_I(a_0) \right| + L_I M_{a_0} = M_I, \end{aligned}$$

where $M_{a_0} > 0$ is any bound that satisfies $\forall a \in U : \|a - a_0\| \leq M_{a_0}$ and $a_0 \in U$ is an arbitrary point. Another calculation shows that

$$f_\alpha(a) = F(\alpha, a) : [0,1] \times \mathbb{R}^{2p\times m} \to \mathbb{R}^m$$

is a local Lipschitz continuous function

$$\begin{aligned} \left| f_\alpha(a) - f_\beta(b) \right| &= \left| \alpha f_G(a) + (1-\alpha) f_T(a) - \beta f_G(b) - (1-\beta) f_T(b) \right| \\ &= \left| \alpha f_G(a) - \beta f_G(b) + (1-\alpha) f_T(a) - (1-\beta) f_T(b) \right| \\ &= \left| \alpha f_G(a) - \beta f_G(a) + \beta f_G(a) - \beta f_G(b) \right. \\ &\quad \left. + (1-\alpha) f_T(a) - (1-\beta) f_T(a) + (1-\beta) f_T(a) - (1-\beta) f_T(b) \right| \\ &\leq |\alpha - \beta| \left| f_G(a) \right| + |\beta| \left| f_G(a) - f_G(b) \right| \\ &\quad + |\beta - \alpha| \left| f_T(a) \right| + |1-\beta| \left| f_T(a) - f_T(b) \right| \\ &\leq |\alpha - \beta| (M_G + M_T) + \left\| a - b \right\| (L_G + L_T). \end{aligned}$$

□

Using the previous lemma proving that $f_{\alpha(u_{i-p+1},\ldots,u_{i+p})}(u_i, u_{i+1})$ is a local Lipschitz continuous flux boils down to proving that $\alpha : \mathbb{R}^{2p\times m} \to [0, 1]$ is local Lipschitz continuous.

We will now prove that this new flux satisfies a semidiscrete entropy inequality at least if there is a cell boundary where $\alpha_{k+\frac{1}{2}} \neq 0$ holds. The proof is based on cell subdivision, averaging and the convexity of the entropy as already used in [31].

Theorem 3.1 *The GT flux satisfies the semidiscrete cell entropy inequality*

$$\frac{\mathrm{d}U \circ u_k}{\mathrm{d}t} \leq \frac{F^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-1}, u_k) - F^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+1})}{\Delta x}$$

with the numerical entropy Flux

$$F^{GT}_{\alpha}(u_l, u_r) = \alpha F^G(u_l, u_r) + (1-\alpha)F^T(u_l, u_r)$$

where $F^G(u_l, u_r) = F(u_R(0, u_l, u_r))$ *and*

$$F^T(u_l, u_r) = \frac{\langle \frac{\partial U}{\partial u}(u_l) + \frac{\partial U}{\partial u}(u_r), f^T(u_l, u_r)\rangle + F(u_l) + F(u_r) - \langle \frac{\partial U}{\partial u}(u_l), f(u_l)\rangle - \langle \frac{\partial U}{\partial u}(u_r), f(u_r)\rangle}{2}$$

are the respective entropy fluxes of the Godunov [31] *and Tadmor fluxes* [32].

Proof We begin our proof by deriving a semidiscrete cell entropy inequality from the discrete cell entropy inequality for the Godunov flux by going over to the limit $\Delta t \to 0$

$$\begin{aligned}
0 &\geq \lim_{\Delta t\to 0} \frac{U(u_k^{n+1}) - U(u_k^n)}{\Delta t} - \frac{F^G(u_{k-1}^n, u_k^n) - F^G(u_k^n, u_{k+1}^n)}{\Delta x} \\
&= \frac{\mathrm{d}U \circ u}{\mathrm{d}t} - \frac{F^G(u_{k-1}, u_k) - F^G(u_k, u_{k+1})}{\Delta x} \\
&= \left\langle v_k, \frac{\mathrm{d}u_k}{\mathrm{d}t}\right\rangle - \frac{F^G(u_{k-1}, u_k) - F^G(u_k, u_{k+1})}{\Delta x} \\
&= \left\langle v_k, \frac{f^G(u_{k-1}, u_k) - f^G(u_k, u_{k+1})}{\Delta x}\right\rangle - \frac{F^G(u_{k-1}, u_k) - F^G(u_k, u_{k+1})}{\Delta x}
\end{aligned}$$

The same holds in the sense of an equality also for the product of the entropy variable with the Tadmor flux. We first look at the special case $\alpha_{k-\frac{1}{2}} = \alpha = \alpha_{k+\frac{1}{2}}$ and use the entropy variable $v_k = \frac{\partial U \circ u}{\partial u}|_{u_k}$ to find

$$\begin{aligned}
\frac{\mathrm{d}U \circ u_k}{\mathrm{d}t} &= \left\langle v_k, \frac{\mathrm{d}u}{\mathrm{d}t} \right\rangle \\
&= \left\langle v_k, \frac{\alpha f^G(u_{k-1}, u_k) + (1-\alpha) f^T(u_{k-1}, u_k) - \alpha f^T(u_k, u_{k+1}) - (1-\alpha) f^G(u_k, u_{k+1})}{\Delta x} \right\rangle \\
&= \alpha \left\langle v_k, \frac{f^G(u_{k-1}, u_k) - f^G(u_k, u_{k+1})}{\Delta x} \right\rangle + (1-\alpha) \left\langle v_k, \frac{f^T(u_{k-1}, u_k) - f^T(u_k, u_{k+1})}{\Delta x} \right\rangle \\
&\leq \alpha \frac{F^G(u_{k-1}, u_k) - F^G(u_k, u_{k+1})}{\Delta x} + (1-\alpha) \frac{F^T(u_{k-1}, u_k) - F^T(u_k, u_{k+1})}{\Delta x} \\
&= \frac{F_\alpha^{GT}(u_{k-1}, u_k) - F_\alpha^{GT}(u_k, u_{k+1})}{\Delta x}.
\end{aligned}$$

Furthermore, we now consider the general case $\alpha_{k-\frac{1}{2}} \neq \alpha_{k+\frac{1}{2}}$ under usage of the first case. The derivative of the average u_k can be rewritten as the average of two schemes for the averages $u_{k-\frac{1}{4}}$ and $u_{k+\frac{1}{4}}$

$$\begin{aligned}
\frac{\mathrm{d}u_{k-\frac{1}{4}}}{\mathrm{d}t} &= \frac{f_{\alpha_{k-\frac{1}{2}}}(u_{k-1}, u_k) - f_{\alpha_{k-\frac{1}{2}}}(u_k, u_k)}{\Delta x/2} \\
\frac{\mathrm{d}u_{k+\frac{1}{4}}}{\mathrm{d}t} &= \frac{f_{\alpha_{k+\frac{1}{2}}}(u_k, u_k) - f_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+1})}{\Delta x/2},
\end{aligned}$$

that can be thought of as the cell subdivision in Fig. 3

$$\begin{aligned}
\frac{\mathrm{d}u_k}{\mathrm{d}t} &= \frac{f_{\alpha_{k-\frac{1}{2}}}(u_{k-1}, u_k) - f_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+1})}{\Delta x} \\
&= \frac{1}{2} \left(\frac{f_{\alpha_{k-\frac{1}{2}}}(u_{k-1}, u_k) - f_{\alpha_{k-\frac{1}{2}}}(u_k, u_k)}{\Delta x/2} + \frac{f_{\alpha_{k+\frac{1}{2}}}(u_k, u_k) - f_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+1})}{\Delta x/2} \right) \\
&= \frac{\frac{\mathrm{d}u_{k-\frac{1}{4}}}{\mathrm{d}t} + \frac{\mathrm{d}u_{k+\frac{1}{4}}}{\mathrm{d}t}}{2}.
\end{aligned}$$

This is the semidiscrete equivalent of the cell division usually employed to make use of the convexity of the entropy. In our case this allows us to change from $\alpha_{k-\frac{1}{2}}$ to

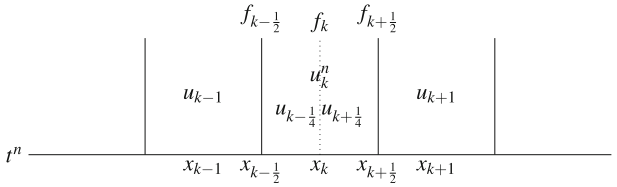


Fig. 3 The subdivision of a cell in space, initialized with the mean value of the old cell

$\alpha_{k+\frac{1}{2}}$ as our fluxes are consistent, namely $f^{GT}_{\alpha_{k-\frac{1}{2}}}(u,u) = f(u) = f^{GT}_{\alpha_{k-\frac{1}{2}}}(u,u)$. This implies together with the consistency of the entropy fluxes

$$\begin{aligned}\frac{\mathrm{d}U\circ u_k}{\mathrm{d}t} &= \left\langle v_k, \frac{\mathrm{d}u_k}{\mathrm{d}t}\right\rangle = \frac{1}{2}\left(\left\langle v_k, \frac{\mathrm{d}u_{k-\frac{1}{4}}}{\mathrm{d}t}\right\rangle + \left\langle v_k, \frac{\mathrm{d}u_{k+\frac{1}{4}}}{\mathrm{d}t}\right\rangle\right)\\ &\leq \frac{1}{2}\left(\frac{F^{GT}_{\alpha-\frac{1}{2}}(u_{k-1},u_k) - F^{GT}_{\alpha-\frac{1}{2}}(u_k,u_k)}{\Delta x/2} + \frac{F^{GT}_{\alpha+\frac{1}{2}}(u_k,u_k) - F^{GT}_{\alpha+\frac{1}{2}}(u_k,u_{k+1})}{\Delta x/2}\right)\\ &= \frac{F^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-1},u_k) - F^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k,u_{k+1})}{\Delta x}\end{aligned} \tag{3.1}$$

and completes the Proof. □

The aforementioned arguments show that our flux is entropy dissipative in the usual sense if α is chosen to be nonzero.

In [22, Sect. 4.1] the entropy conservative flux of Tadmor was extended via the usage of linear combinations into an entropy conservative flux of order $2p$. We will also use this idea on our flux. As we have already used α as a parameter for convex combinations we will use c_p^r instead to denote the coefficients in the linear combination.

Definition 3.2 We define the high order LMRGT flux of order $2p$ as

$$f_\alpha^{LMRGT}(u_{k-p+1},\ldots,u_{k+p}) = \sum_{r=1}^{p} c_p^r (f_\alpha^{GT}(u_k,u_{k+r}) + \cdots + f_\alpha^{GT}(u_{k-r+1},u_{k+1})).$$

It follows from the definition that this flux is of order $2p$ for $\alpha_k = 0$ i.e. when the entropy equality holds. One further deduces from

$$\begin{aligned}&f_\alpha^{LMRGT}(u_{k-p+1},\ldots,u_{k+p}) - f_0^{LMRGT}(u_{k-p+1},\ldots,u_{k+p})\\ &= \alpha\sum_{r=1}^{p} c_p^r((f^G(u_k,u_{k+r}) - f^T(u_k,u_{k+r}) + \cdots + f^G(u_{k-r+1},u_{k+1}) - f^T(u_{k-r+1},u_{k+1}))\\ &= \alpha\,\mathcal{O}(\Delta x),\end{aligned}$$

that the scheme is even of order $2p$ if the linear combination for an $2p$ order accurate entropy conservative flux is used in the construction and $\alpha(u_{k-q+1}, \ldots, u_{k+q}) = \mathcal{O}((\Delta x)^{2p-1})$ holds.

While one aims for a discrete entropy inequality we are at least able to proof a semidiscrete cell entropy inequality for this flux

Corollary 3.1 *The semidiscrete scheme*

$$\begin{aligned}\frac{\mathrm{d}u_k}{\mathrm{d}t} &= \frac{f^{LMRGT}_{\alpha_{k-\frac{1}{2}}}(u_{k-p}, \ldots, u_{k+p-1}) - f^{LMRGT}_{\alpha_{k+\frac{1}{2}}}(u_{k-p+1}, u_{k+p})}{\Delta x} \\ &= -\sum_{r=1}^{p} c_p^r \frac{f^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+r}) - f^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-r}, u_k)}{\Delta x}\end{aligned}$$

satisfies a semidiscrete entropy inequality

$$\frac{\mathrm{d}U \circ u_k}{\mathrm{d}t} \leq \frac{F^{LMRGT}_{\alpha_{k-\frac{1}{2}}}(u_{k-p}, \ldots, u_{k+p-1}) - F^{LMRGT}_{\alpha_{k+\frac{1}{2}}}(u_{k-p+1}, \ldots, u_{k+p})}{\Delta x}$$

with an consistent numerical entropy flux given by

$$F^{LMRGT}_{\alpha}(u_{k-p+1}, \ldots, u_{k+p}) = \sum_{r=1}^{p} c_p^r (F^{GT}_{\alpha}(u_k, u_{k+r}) + \cdots + F^{GT}_{\alpha}(u_{k-r+1}, u_{k+1}))$$

if $\forall k : \alpha_{k+\frac{1}{2}} \in (0, 1]$ *holds.*

Proof We follow the proof of the semidiscrete entropy inequality from [22, Sect. 4.1] and multiply the definition of the scheme by the entropy variable v_k to find

$$\begin{aligned}\frac{\mathrm{d}U \circ u_k}{\mathrm{d}t} = \langle v_k, \frac{\mathrm{d}u_k}{\mathrm{d}t} \rangle &= \left\langle v_k, \sum_{r=1}^{p} c_p^r \frac{f^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-r}, u_k) - f^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+r})}{\Delta x} \right\rangle \\ &= \sum_{r=1}^{p} c_p^r \left\langle v_k, \frac{f^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-r}, u_k) - f^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+r})}{\Delta x} \right\rangle \\ &\overset{(3.1)}{\leq} \sum_{r=1}^{p} c_p^r \frac{F^{GT}_{\alpha_{k-\frac{1}{2}}}(u_{k-r}, u_k) - F^{GT}_{\alpha_{k+\frac{1}{2}}}(u_k, u_{k+r})}{\Delta x} \\ &= \frac{F^{GTLMR}_{\alpha_{k-\frac{1}{2}}}(u_{k-p}, u_{k+p-1}) - F^{GTLMR}_{\alpha_{k+\frac{1}{2}}}(u_{k-p+1}, u_{k+p})}{\Delta x}.\end{aligned}$$

□

Our first numerical experiment showed that an entropy dissipative scheme alone is not enough to guarantee good approximate solutions. This is why we will now construct an algorithm to find values for α to control our flux according to the Dafermos entropy criterion.

Definition 3.3 We call $\alpha : \mathbb{R}^{2p\times m} \to [0,1]$ an entropy inequality predictor with a $(2p)$ point stencil if

$$\lim_{h\to 0} \alpha(u_{i-p+1},\dots,u_{i+p}) = \begin{cases} 0 & \exists x \in [x_i-(p-1)\Delta x, x_i+p\Delta x] : \frac{\partial U\circ u}{\partial t} + \frac{\partial F\circ u}{\partial x} < 0 \\ 1 & \forall x \in [x_i-(p-1)\Delta x, x_i+p\Delta x] : \frac{\partial U\circ u}{\partial t} + \frac{\partial F\circ u}{\partial x} = 0 \end{cases}$$

holds for the complete stencil. The input values $u_{i-p+1},\dots,u_{i+p}$ shall be the mean values of the solution in the respective cells as present in a Finite Volume solver. We will call the entropy inequality predictor slope limited if

$$|\alpha_i - \alpha_{i+1}| < M \quad \text{with} \quad \alpha_i = \alpha(u_{i-p+1},\dots,u_{i+p})$$

holds for some $M < 1$ and all i.

The slope limiting property was inspired by the idea to limit the slope of α with respect to the grid index i. This should not be mixed up with a bound on the slope of α with respect to x. Such a bound would be scaling with the distance between x_i and x_{i+1} as present in the usual definition of a difference quotient. This ensures that α switches between 0 and 1 over several mesh points while the size of this switch is scaled down with respect to the physical scale for a finer grid. The switch needs at least $\lfloor 1/M \rfloor$ points.

Lemma 3.2 (Smoothstep [24]) *The function*

$$H_{sm}(x) = \begin{cases} 0 & x \le 0 \\ 6x^5 - 15x^4 + 10x^3 & 0 \le x \le 1 \\ 1 & 1 \le x \end{cases}$$

is a $\mathcal{C}^2$ function with zero first and second derivatives at $x = 0$ and $x = 1$.

We will need some special operations on functions for the construction of our predictor which are motivated by mollification. The convolution [21, p. 216] of f and g is defined as

$$f * g(x) = \int_\Omega f(y)g(x-y)\mathrm{d}y$$

for suitable f and g. Please note that this can be interpreted as the integral of all combinations of $f(\cdot)$ with $g(-\cdot)$ multiplied over $\mathbb{R}$ and indexed by their respective

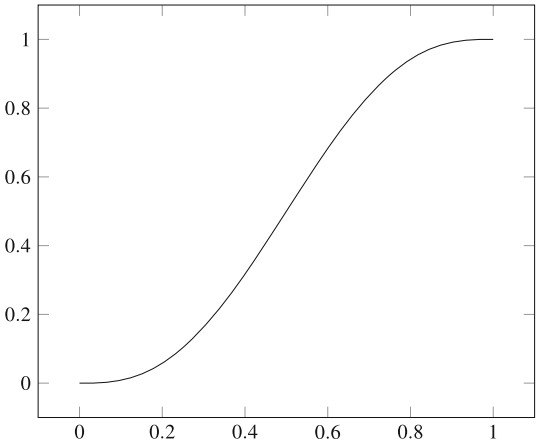


Fig. 4 Plot of the smooth step function

shift between the argument of f and g. A mollification, first defined in [9], is a convolution of a function f with a suitable g giving a smoother function as f. The result [21, p. 216]

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1 ,$$

relates the norm of the mollified function $f * g$ to the original function. As convolution is coupled to (Lebesgue)-integration which in turn provides the Lebesgue norms, one can ask if also an equivalent of convolution for other norms exists. We will find such an equivalent with interesting properties for our application related to the uniform norm $\|\cdot\|_\infty$.

Definition 3.4 *(Minkowsky sum and Minkowsky product)* Given two sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$ we define the Minkowsky sum and Minkoswky product as

$$A \oplus B = \{a + b \mid a \in A, b \in B\}$$

and

$$A \odot B = \{ab \mid a \in A, b \in B\}.$$

Lemma 3.3 (Special properties of the Minkowsky product and sum) *Let $A, B \subset \mathbb{R}$ be two sets. In this case*

$$\sup(A \oplus B) = \sup A + \sup B$$

holds. If additionally $\forall a \in A : a \geq 0$ and $\forall b \in B : b \geq 0$ hold the equality

$$\sup(A \odot B) = \sup A \cdot \sup B$$

is also satisfied. In other words the supremum is additive and positively homogeneous for sets.

Definition 3.5 *(Minkowsky product of functions)* Let $f, g : \mathbb{R} \to \mathbb{R}$. Their Minkowsky product is a map $f \odot g : \mathbb{R} \to \mathbb{R}$, defined by

$$(f \odot g)(x) = \{f(x - y)g(y) \mid y \in \mathbb{R}\}.$$

In other words the Minkowsky product of f and g at x is the set of all values of f multiplied by g so that the argument added together gives x. Compare this to the integrand of the convolution.

Definition 3.6 Let $f, g : \mathbb{R} \to \mathbb{R}$ be bounded real maps. The sup-mollification of f and g is defined as

$$f \circledast g(x) = \sup f \odot g(x) = \sup_{y \in \mathbb{R}} (f(x - y)g(x)).$$

Please note that $f \odot g$ is a set depending on x and the supremum is not taken over x, but over the set at the point x.

We will use the defined sup-mollification operator to ensure the slope limiting property of our entropy inequality predictor. We will now prove some useful lemmas that will also show that our entropy inequality predictor is Lipschitz continuous and keeps α up at one in a region around an entropy dissipating shock.

Lemma 3.4 *Let $f, g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be bounded functions. In this case*

$$\sup_x f \circledast g(x) = \sup_z f(z) \sup_z g(z)$$

holds.

Proof An easy calculation shows

$$\begin{aligned} \sup_x f \circledast g &= \sup_x \sup_y f(y)g(x - y) = \sup_{(x,y) \in \mathbb{R}^2} f(y)g(x - y) \\ &= \sup_{(x,y) \in \mathbb{R}^2} f(y)g(x) = \sup \operatorname{ran} f \odot \operatorname{ran} g \\ &= (\sup \operatorname{ran} f) \cdot (\sup \operatorname{ran} g). \end{aligned}$$

□

Lemma 3.5 *Let $f, g : \mathbb{R} \to \mathbb{R}$ be bounded functions. Then*

$$\left| \sup_x f(x) - \sup_y g(y) \right| \leq \sup_x |f(x) - g(x)|$$

holds.

Proof We start by stating that

$$\sup_x f(x) = \sup_x g(x) + f(x) - g(x) \leq \sup_y g(y) + \sup_x f(x) - g(x)$$

holds. This can be rearranged and bounded so that

$$\sup_x f(x) - \sup_y g(y) \leq \sup_x f(x) - g(x) \leq \sup_x |f(x) - g(x)|$$

holds. As this also holds if the roles of f, g are swapped, it follows

$$\left| \sup_x f(x) - \sup_y g(y) \right| \leq \sup_x |f(x) - g(x)| .$$

□

Lemma 3.6 (Sup-mollification is a Lipschitz continuous operator) *For bounded* $f_1, f_2, g : \mathbb{R} \to \mathbb{R}$ *it holds*

$$\|f_1 \circledast g - f_2 \circledast g\|_\infty \leq \|g\|_\infty \|f_1 - f_2\|_\infty$$

Proof We use lemma 3.5 and 3.4 to prove

$$\begin{aligned}
\|f_1 \circledast g - f_2 \circledast g\|_\infty &= \sup_x \left| \sup_y f_1(y) g(x-y) - \sup_y f_2(y) g(x-y) \right| \\
&\overset{\text{lem } 3.5}{\leq} \sup_x \sup_y |f_1(y) g(x-y) - f_2(y) g(x-y)| \\
&= \sup_x \sup_y |f_1(y) - f_2(y)| \, |g(x-y)| \\
&\overset{\text{lem } 3.4}{=} \sup_x |f_1(x) - f_2(x)| \sup_y |g(y)| \\
&= \|g\|_\infty \|f_1 - f_2\|_\infty .
\end{aligned}$$

□

Lemma 3.7 (Slope condition inequality) *Let* $f, g : \mathbb{R} \to \mathbb{R}$ *be bounded functions. If* g *satisfies for a fixed* $h \in \mathbb{R}$

$$\exists M \in \mathbb{R} : \sup_{x \in \mathbb{R}} |g(x+h) - g(x)| \leq M,$$

then the sup-mollification $f \circledast g$ *full fills*

$$|f \circledast g(x+h) - f \circledast g(x)| \leq M \cdot \sup |f(y)| .$$

Proof We again use lemma 3.5 to prove

$$\begin{aligned}
|f \circledast g(x+h) - f \circledast g(x)| &= \left| \sup_y f(y)g(x+h-y) - \sup f(y)g(x-y) \right| \\
&\leq \sup_y |f(y)g(x+h-y) - f(y)g(x-y)| \\
&= \sup_y |f(y)|\,|g(x+h-y) - g(x-y)| \\
&\leq \sup_y |f(y)| \sup_y |g(x+h-y) - g(x-y)| \\
&\leq M \cdot \sup |f|\,.
\end{aligned}$$

□

Lemma 3.8 (Plateau condition inequality) *Let* $f, g : \mathbb{R} \to \mathbb{R}$ *be bounded,* $x_0 \in \mathbb{R}$ *and* $\varepsilon > 0$. *Then*

$$\forall x \in [-\varepsilon, \varepsilon] : g(x) > c \in \mathbb{R}$$

implies

$$\forall x \in [x_0 - \varepsilon, x_0 + \varepsilon] : f \circledast g(x) \geq c f(x_0).$$

Proof Let $x \in [x_0 - \varepsilon, x_0 + \varepsilon]$. If we set $y = x_0$ it follows $x - y \in [-\varepsilon, \varepsilon]$ and

$$\begin{aligned}
& f(y)g(x-y) = f(x_0)g(x-y) \geq f(x_0)c \\
\implies\ & f(x_0)c \leq \sup_y f(y)g(x-y) = f \circledast g(x).
\end{aligned}$$

□

Definition 3.7 *(Discrete sup-mollification)* Let for $n \in \mathbb{N}$ be the vector space of step functions on [0, 1] denoted as

$$S_n = \{f : [0,1] \to \mathbb{R} \mid \forall i = 0, \ldots, n-1 : f|_{[i/n,(i+1)/n]} = f_i \in \mathbb{R}\}.$$

We can further define an embedding $Z : S_n \to S$ of this space into the step functions over $\mathbb{R}$, denoted as S, by

$$Z : S_n \to S,\ f \mapsto Zf,\ Zf(x) = \begin{cases} f(x) & x \in [0,1] \\ 0 & \text{else} \end{cases}.$$

These two definitions allow us to define the discrete sup-mollification of $f, g \in S_n$ as

$$(f \circledast g)|_{[i/n,(i+1)/n]}$$

$$= (Zf \circledast Zg)|_{[i/n,(i+1)/n]} = \max_{j\in\{0,\dots,n-1\}} \tilde{f}_j \tilde{g}_{i-j}. \quad \text{for } i = 0,\dots,n-1.$$

The values $\tilde{f}_i \in \mathbb{R}$ and $\tilde{g}_i \in \mathbb{R}$ relate to f_i and g_i as

$$\tilde{f}_i = \begin{cases} f_i & i \in \{0,\dots,n-1\} \\ 0 & \text{else} \end{cases} \quad \tilde{g}_i = \begin{cases} g_i & i \in \{0,\dots,n-1\} \\ 0 & \text{else} \end{cases}.$$

The sup mollfication for step functions on the interval $[a, b]$ shall be defined using the coordinate transform $\varphi(x) = \frac{x-a}{b-a}$ and the corresponding inverse $\varphi^{-1}(y) = a + y(b-a)$. Using this transform yields

$$(f \circledast g)(x) = (Z(f \circ \varphi) \circledast Z(g \circ \varphi)) \circ \varphi^{-1}(x).$$

It is also possible to exactly sup-mollify piecewise linear functions.

Example 3.1 (The Godunov Flux entropy inequality predictor) A suitable entropy inequality predictor can be constructed from the Godunov flux by looking at it's entropy dissipation

$$s_k^n = \frac{F(u_{k+1}^n, u_k^n) - F(u_k^n, u_{k-1}^n)}{\Delta x} + \frac{U(u_k^{n+1}) - U(u_k^n)}{\Delta t} \leq 0$$

as given in [30, 31]. As the Godunov scheme is entropy stable $s_k^n \leq 0$ holds $\forall n, k$. We can use this value to predict if the entropy equality holds - or the inequality. A problem occurs as in fact $s_k^n < 0$ is even true for smooth initial conditions like $u_0(x) = \sin(\pi x)$ in the first time step.

Similar Problems appear in the context of edge sensors and local viscosity [3, 34] and are usually solved by a thresholding process. In our case this threshold will be carried out by the smooth step function H_{sm} from lemma 3.2 to ensure a Lipschitz-continuous transition. As there are in fact two free parameters in this approach, one for determining the lower threshold, and another to control the width of the smooth step, it is imperative to find parameters that are at least independent of the used grid. We therefore define

$$u^+ = \max_k u_k^n \quad u^- = \min_k u_k^n$$

for single conservation laws and

$$u^+ = u(\arg\max_x U \circ u(x,t), t) \quad u^- = u(\arg\min_x U \circ u(x,t), t)$$

as the cell values having maximum and minimum entropy in the domain for systems of conservation laws. These values can be afterwards used to construct the Riemann

problem with the initial conditions

$$u_1(x,0) = \begin{cases} u^- \\ u^+ \end{cases} \qquad u_2(x,0) = \begin{cases} u^+ & x < 0 \\ u^- & x \geq 0 \end{cases}.$$

By looking at their entropy dissipation s_k^n when the Godunov scheme is applied one finds a reference

$$s^{\text{ref}} = \min\left(\min_k s_k^n(u_1(\cdot,0)), \min_k s_k^n(u_2(\cdot,0))\right)$$

for the entropy dissipation of a strong shock that could be present in the solution. While this approach involves a lot of hand waving the numerical results are quite satisfactory and further research could be centered around this issue. The values

$$r_k^n = H_{sm}\left(\frac{\frac{s_k^n}{s^{\text{ref}}} - a}{b}\right)$$

depend smoothly on s_k^n but can still have extremely localized spikes as wide as only a few cells. The parameter $a \in \mathbb{R}$ is a threshold under which the result of the entropy inequality predictor should be thought of as zero, while $b \in \mathbb{R}$ corresponds to a typical amplitude of a spike in the entropy dissipation indicating a shock. Numerical tests indicate that a instantaneous switching between fluxes leads to undesired oscillations around their interface. Furthermore, the stencil of the high order Tadmor flux is wider than the stencil of the Godunov scheme and the derivation of the high order Tadmor scheme assumes an entropy conservative solution in its derivation, which will be violated by an entropy dissipating discontinuity in the solution. These two problems are considered in the definition of the entropy inequality predictor. Responsible are the slope limiting property and the definition as a scale for the violation of the entropy equality on the entire stencil of a scheme. In this case the wider stencil of the high order modified Tadmor scheme is relevant. We will satisfy these requirements using sup-mollification of r_k^n, i.e. its associated piecewise constant function, and a suitable kernel. We chose the cut hat function

$$h(x) = \max(0, \min(1, 2x + 2, -2x + 2))$$

in a properly rescaled fashion for this purpose. and define

$$\alpha = r \circledast h.$$

Other choices are possible, and it is not clear yet if a smoother mollifier improves the scheme.

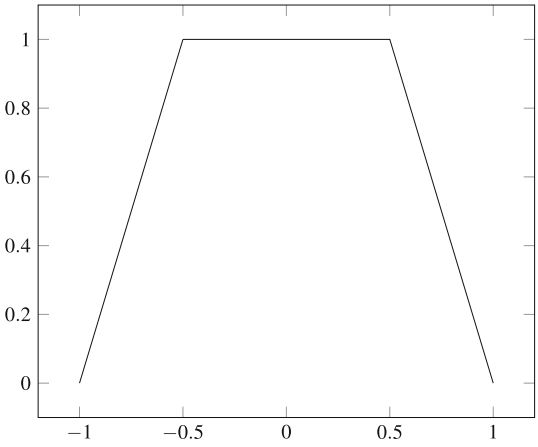


Fig. 5 Plot of a cut hat function $h(x)$

Lemma 3.9 *The Godunov Flux inequality predictor*

$$\alpha^n = H_{sm}\left(\frac{\frac{s_k^n}{s^{ref}} - a}{b}\right) \circledast h$$

is slope limited.

Proof This follows from the fact that

$$\text{ran } H_{sm}\left(\frac{\frac{s_k^n}{s^{\text{ref}}} - a}{b}\right) \subset [0, 1]$$

holds and the cut hat function has limited slope using the slope condition inequality.□

While the aforementioned entropy inequality predictor is able to deliver satisfactory results the solution of a Riemann problem, needed to calculate the Godunov flux, is not always easily obtained. This is why two other entropy inequality predictors, one of theoretical and one also of practical value, were constructed.

Example 3.2 (The Lax-Friedrichs entropy inequality predictor) The aforementioned construction can be also applied to the Lax-Friedrichs scheme, and it's corresponding entropy inequality and entropy flux, proved in [20, 30, 31]. The entropy flux is given by

$$F(u_l, u_r) = \frac{F(u_l) + F(u_r)}{2} + \frac{U(u_l) - U(u_r)}{2\lambda},$$

leading to the entropy production

$$s_k^n = \frac{F(u_{k+1}^n, u_k^n) - F(u_k^n, u_{k-1}^n)}{\Delta x} + \frac{U(u_k^{n+1}) - U(u_k^n)}{\Delta t} \leq 0.$$

Entering these results leads to the entropy inequality predictor

$$\alpha^n = H_{sm}\left(\frac{\frac{s_k^n}{s^{\text{ref}}} - a}{b}\right) \circledast h.$$

Sadly, while this predictor has a rigorous provable background from [20], its practical use is complicated. The line between an area of entropy conservation and an entropy dissipating shock is blurred by the big amount of dissipation present in the Lax-Friedrichs scheme that also happens in smooth areas. This is why a second entropy inequality predictor was constructed with reduced dissipation. This reduction is based on linear reconstruction in an ENO type fashion [16].

Example 3.3 (The ENO2 Lax Friedrichs entropy inequality predictor) Given an piecewise constant solution u_k^n one first calculates the piecewise linear reconstructions

$$\tilde{u}_k^n(x) = \begin{cases} u_k^n + a_l(x - x_k) & a_l < a_r \\ u_k^n + a_r(x - x_k) & a_r \geq a_l \end{cases} \quad a_l = \frac{u_k^n - u_{k-1}^n}{x_k - x_{k-1}} \quad a_r = \frac{u_{k+1}^n - u_k^n}{x_{k+1} - x_k}.$$

Using this reconstruction directly in a finite volume entropy inequality is not possible, as the scheme

$$u_k^{n+1} = u_k^n + \lambda\left(f\left(\tilde{u}_{k-1}^n\left(x_{k-\frac{1}{2}}\right), \tilde{u}_k^n\left(x_{k-\frac{1}{2}}\right)\right) - f\left(\tilde{u}_k^n\left(x_{k+\frac{1}{2}}\right), \tilde{u}_{k+1}^n\left(x_{k+\frac{1}{2}}\right)\right)\right)$$

has to the authors knowledge no known entropy fluxes. We instead seek to calculate an approximation of the entropy dissipation of this reconstructed solution by using it as the initial condition for a first order Lax-Friedrichs solver. It is sufficient to use this solver at points of discontinuity as the entropy equality holds for the smooth areas of the solution. We therefore use the subdivision of our primary cells sketched in Fig. 6 to start the Lax-Friedrichs method. Let $x_{k+1/2}$ be the cell boundary between the cell around x_k and x_{k+1} and $\frac{\Delta x}{6} \geq \varepsilon > 0$ an arbitrary parameter for a sub-cell size. We introduce new cell boundaries at

$$x_l^- = x_{k+\frac{1}{2}} - 3\varepsilon \quad x_l^+ = x_m^- = x_{k+\frac{1}{2}} - \varepsilon \quad x_m^+ = x_{k+\frac{1}{2}} + \varepsilon = x_r^- \quad x_r^+ = x_{k+\frac{1}{2}} + 3\varepsilon$$

to form new cells around

$$x_l = x_{k+\frac{1}{2}} - 2\varepsilon \quad x_m = x_{k+\frac{1}{2}} \quad x_r = x_{k+\frac{1}{2}} + 2\varepsilon.$$

These cells are initialized with the mean values of $\tilde{u}^n(x)$ in these cells

$$v_l = \frac{1}{2\varepsilon}\int_{x_l^-}^{x_l^+} \tilde{u}^n(x)\mathrm{d}x \quad v_m = \frac{1}{2\varepsilon}\int_{x_m^-}^{x_m^+} \tilde{u}^n(x)\mathrm{d}x \quad v_r = \frac{1}{2\varepsilon}\int_{x_r^-}^{x_r^+} \tilde{u}^n(x)\mathrm{d}x.$$

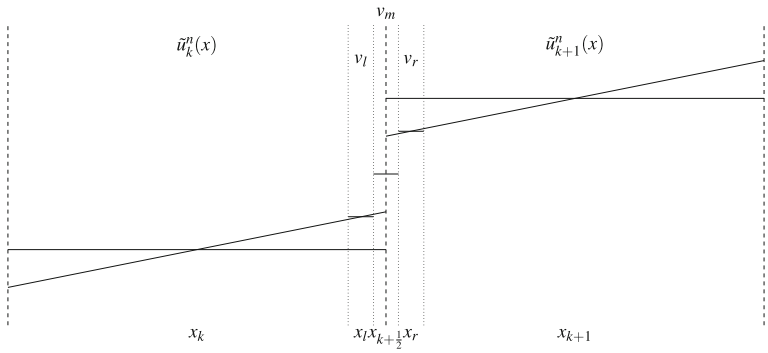


Fig. 6 Subdivision and averaging after an ENO reconstruction

After one step of calculations we can take the entropy dissipation of the Lax-Friedrichs Scheme in the middle cell

$$s_{k+\frac{1}{2}} = U\left(\frac{v_l + v_r}{2} + \lambda\frac{f(v_l) - f(v_r)}{2}\right) - \frac{U(v_l) + U(v_r))}{2} + \lambda\frac{F(v_l) - F(v_r)}{2}$$

as an approximation of the true entropy dissipation at this edge. The value of ε is not critical in this calculation, and we can pass to the limit $\varepsilon \to 0$ to find

$$s_{k+\frac{1}{2}} = U\left(\frac{\tilde{u}_k^n\left(x_{k+\frac{1}{2}}\right) + \tilde{u}_{k+1}^n\left(x_{k+\frac{1}{2}}\right)}{2} + \lambda\frac{f\left(\tilde{u}_k^n\left(x_{k+\frac{1}{2}}\right)\right) - f\left(\tilde{u}_{k+1}^n\left(x_{k+\frac{1}{2}}\right)\right)}{2}\right)$$
$$- \frac{U\left(\tilde{u}_k^n\left(x_{k+\frac{1}{2}}\right)\right) + U\left(\tilde{u}_{k+1}^n\left(x_{k+\frac{1}{2}}\right)\right)}{2} + \lambda\frac{F\left(\tilde{u}_k^n\left(x_{k+\frac{1}{2}}\right)\right) - F\left(\tilde{u}_{k+1}^n\left(x_{k+\frac{1}{2}}\right)\right)}{2}$$

and proceed as before with the usual stepping and sup-mollification operators. One should note that as λ is constant the time step used for this calculation tends to zero and hence this entropy inequality is of no use for the ENO-LxF scheme and only gives an estimate for the entropy dissipation.

Remark 3.1 The aforementioned method is easily generalized to several space dimensions using a grid with tensor product structure and application of the presented methods in every direction, as also pointed out for correction procedure via reconstruction (CPR) Methods in their summation-by-parts (SBP) interpretation in [27].

Remark 3.2 The entropy inequality predictors are not discontinuity sensors. The function α should sense positions where entropy dissipation takes place. The entropy equality dictates that there has to be in fact a discontinuity if entropy is dissipated. The opposite implication does not hold. A discontinuity can be present in the solution but still no entropy is dissipated. An example of such behavior is the contact discontinuity present in some solutions to Riemann problems for the Euler equations. Therefore, a different approach, not equivalent to sensing discontinuities, is the aim of the scheme.

4 Numerical tests

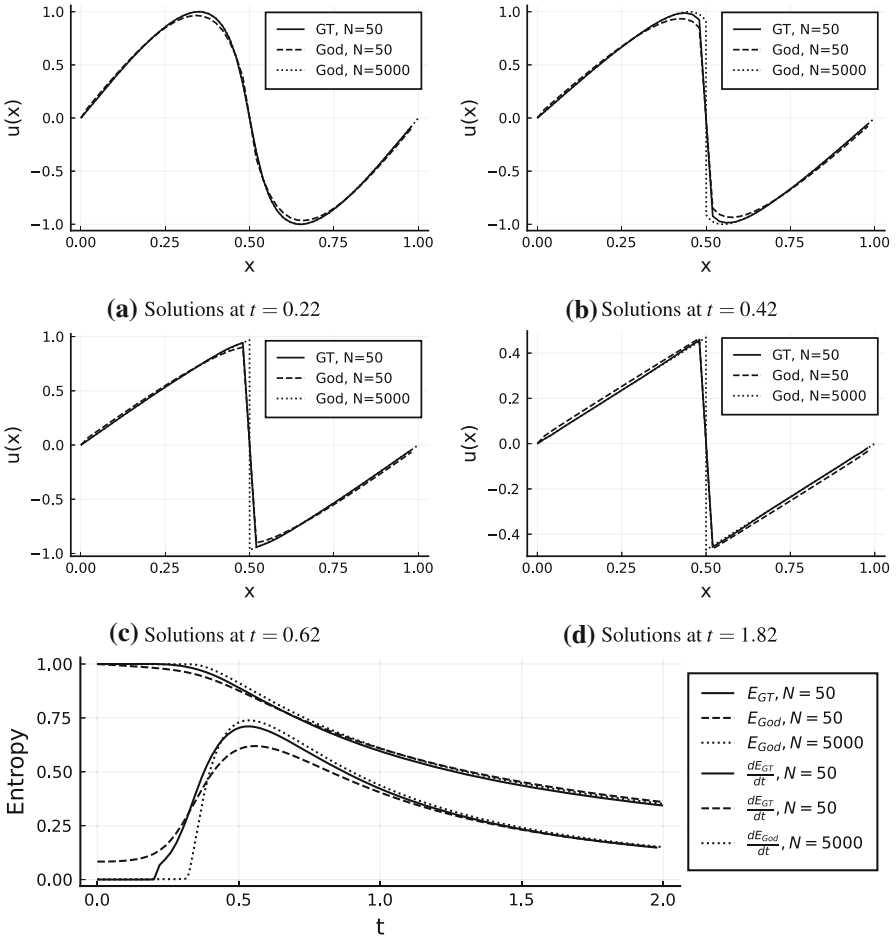
4.1 Numerical tests for the Burgers equation

Numerical tests were carried out for the new numerical flux composed of the entropy inequality predictor coupled to the convex combination flux. Sadly our scheme is not free of open parameters. The parameters a, b were chosen as $a = 1/20, b = 1/100$ after some experiments. Wrong selection of a results in a late or early detection of needed entropy dissipation. Those problems vanish for finer grids, as the entropy inequality predictor gives a refined distinction between conservation and entropy dissipation in this case. Still the optimal values for a are only distributed over one order of magnitude. To high values of b result in difficulties during time integration as this yields big Lipschitz constants for the resulting flux, while to small values result in a slow switching of the scheme between entropy conservation and extremal entropy dissipation. The values for b are not as critical as values for a and equally acceptable values span several orders of magnitude. A second decision which had to be made concerned the entropy pair. The pair $U(u) = u^2/2$, $F(u) = u^3/3$ was chosen for this purpose. We compare the new scheme directly to the Godunov scheme as the new scheme uses the Godunov scheme for dissipative regions. The Burgers equation was solved for $N = 50$ cells and periodic boundary conditions. A known good solution was calculated by a Godunov scheme with $N_{control} = 5000$ cells. Time integration was carried out using the SSPRK104 algorithm [12].

Our numerical tests were carried out to test two assumptions.

- The total entropy of the numerical solution of the GT scheme is a (good) approximation of the total entropy of the true solution.
- The norm $\|u(\cdot, t) - u_{numeric}(\cdot, t)\|$ is improved by our scheme over the error one gets from the Godunov scheme.

The first assumption seems to be true. By looking at Fig. 10a the entropy of the GT scheme is, by construction, constant as long as u is smooth. The behavior is also desirable for non-smooth solutions as the numerical derivative of the total entropy approximates the exact derivative quite well. Interestingly the total entropy of the less dissipative GT scheme is smaller than the total entropy of the Godunov method for large times, which is the same for the exact solution. Assumptions on the quality of the solution can be made from the solution plots in 7. The smooth solutions show good correspondence between exact solution and the GT solution. In the discontinuous case the solution at the discontinuity corresponds to the solution of the Godunov method but is still significantly more exact in smooth areas. After these qualitative assumptions some quantitative measurements were carried out in form of norms of the errors. While the L_1 norm of the error was reduced for smooth and non-smooth solutions the L_2 norm error for non-smooth solutions was only improved by a small amount as the shock is not better resolved than by the Godunov scheme. Several upwind schemes show glitches concerning rarefaction waves [35]. The scheme was tested for this deficiency using a Riemann problem with $u_l = -1.0$ and $u_r = 1.0$ as initial condition and the results are shown in Fig. 9. One could imagine that the sonic glitch, clearly present in the solution calculated by the Godunov method, will be also part of the solution



(a) Solutions at $t = 0.22$

(b) Solutions at $t = 0.42$

(c) Solutions at $t = 0.62$

(d) Solutions at $t = 1.82$

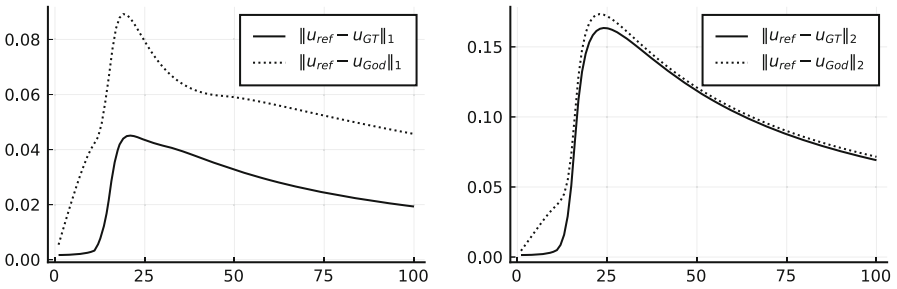
(e) Total entropies and their derivatives with respect to time for all 3 solvers.

Fig. 7 Numerical experiment with the GT scheme. The entropy conservative flux is the eight order flux [22, 32] while the classic Godunov scheme with exact Riemann solver was used as entropy dissipative flux. Time integration was carried out using the SSPRK104 method and a CFL number of $\lambda = 0.5$. The parameters of the Godunov entropy inequality predictor were $a = 1/20$, $b = 1/100$. The cutted hat function used in the sup mollification was rescaled to fit the support of the hat into a $2p + 1$ wide stencil with $p = 8$, i.e. to fit the stencil of the high order flux

calculated by the GT scheme. This is only partly the case. The strength of the sonic glitch is significantly reduced compared to the Godunov scheme.

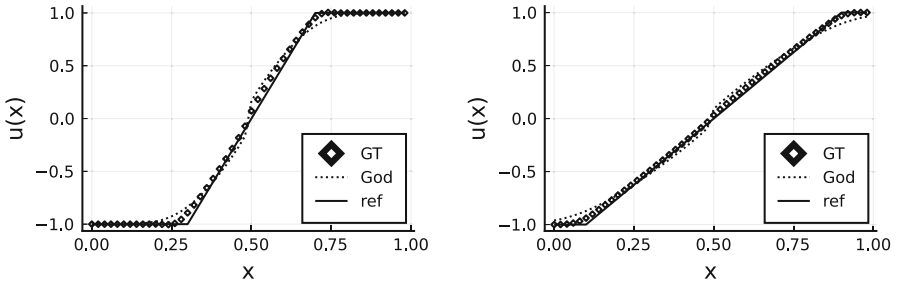
4.2 Numerical tests for the Euler equations of gas-dynamics

After these promising results for the Burgers equation numerical tests were carried out for the Euler equations of gas dynamics



(a) L_1 norm **(b)** L_2 norm

Fig. 8 Error norms over time for the solutions in Fig. 7 over time



(a) $t = 0.4$ **(b)** $t = 0.8$

Fig. 9 Results for a Riemann Problem given by the initial condition $u_l = -1.0, u_r = 1.0$. A grid consisting of 50 cells was used in conjunction with a CFL number of $\lambda = 0.5$ and the SSPRK104 time integration method [12]. Parameters where the same as in Fig. 7. The solution of the basic Godunov scheme was plotted as a reference for the possible sonic point glitch. The used GT scheme, based on the eight order entropy conservative flux, uses the same Godunov method as a low order flux. The sonic glitch is significantly reduced by the application of the GT scheme

$$u = (\rho, \rho v, E) \quad f(\rho, \rho v, E) = \begin{bmatrix} \rho v \\ \rho v^2 + p \\ v(E + p) \end{bmatrix} \quad P = (\gamma - 1)\left(E - \frac{1}{2}\rho v^2\right)$$

in conjunction with the LxFRI scheme and the ENO2LxF entropy inequality predictor. The physical entropy [15, 33]

$$U(\rho, \rho v, E) = -\rho S \quad F(\rho, \rho v, E) = -\rho v S \quad S = \ln(p\rho^{-\gamma})$$

was used in the entropy inequality predictor whereas the entropy conservative flux

$$f^R(u_l, u_r) = \begin{pmatrix} \hat{\rho}\hat{u} \\ \hat{\rho}\hat{u}^2 + \hat{p}_1 \\ \hat{\rho}\hat{u}\hat{H} \end{pmatrix} \quad z = \sqrt{\frac{\rho}{p}} \begin{pmatrix} 1 \\ u \\ p \end{pmatrix}$$

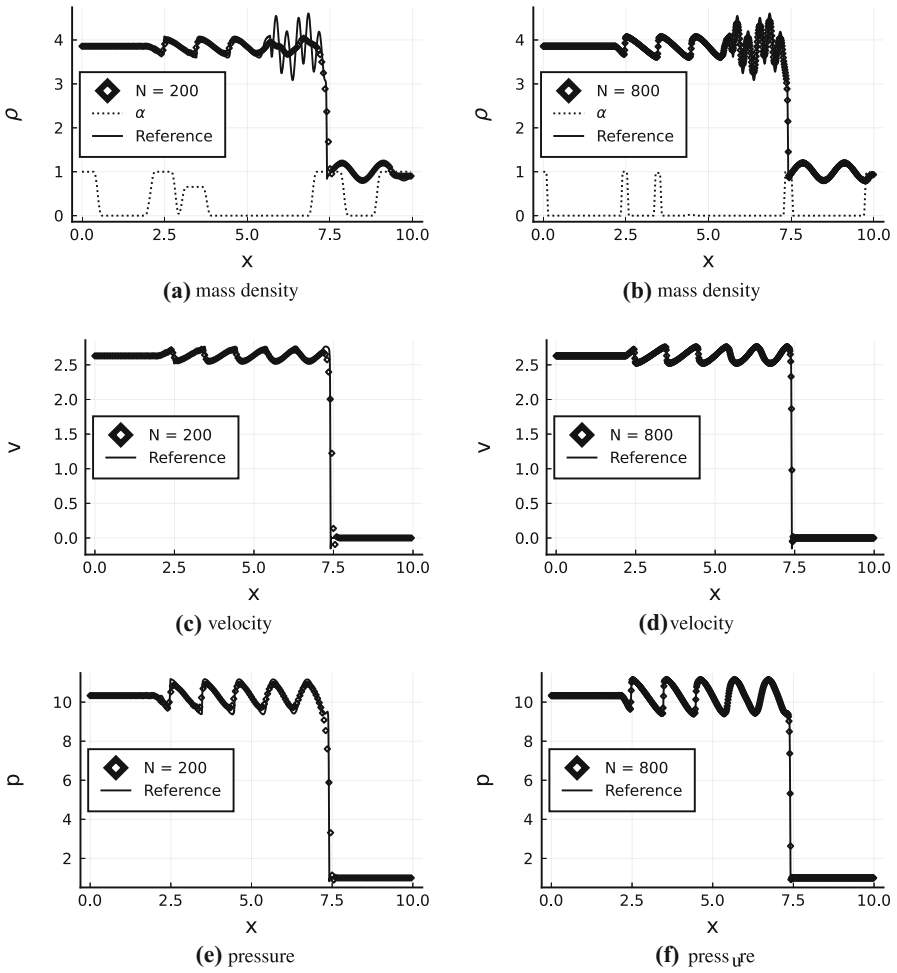
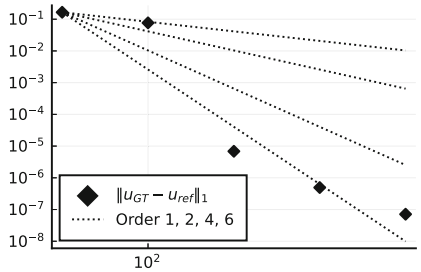


Fig. 10 Shu-Osher testcase at $t = 1.8$. LMRLxFRI scheme of Order 6. The entropy conservative flux is the entropy conservative flux from [18] while the Lax-Friedrich flux was used as entropy dissipative flux. Time integration was carried out using the SSPRK104 method and a CFL number of $\lambda = 0.1$. The parameters of the ENO2-Lax-Friedrichs entropy inequality predictor were $a = 1/1000, b = 1/1000$. The cutted hat function used in the sup mollification was rescaled to fit the support of the hat into a $2p + 1$ wide stencil with $p = 6$, i.e. to fit the stencil of the high order flux. The values of $\alpha_{k+\frac{1}{2}}$ were also plotted

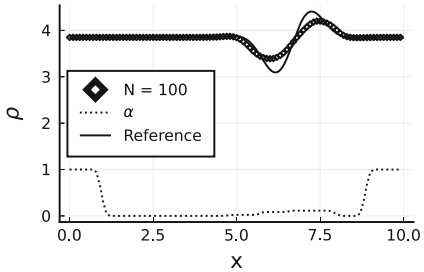
$$\hat{\rho} = \overline{z}_1 z_3^{\ln} \quad \hat{p}_1 = \frac{\overline{z}_3}{\overline{z}_1} \quad \hat{p}_2 = \frac{\gamma+1}{2\gamma}\frac{z_3^{\ln}}{z_1^{\ln}} + \frac{\gamma-1}{2\gamma}\frac{\overline{z}_3}{\overline{z}_1}$$

$$\hat{a} = \sqrt{\frac{\gamma \hat{p}_2}{\hat{\rho}}} \quad \hat{H} = \frac{\hat{a}^2}{\gamma-1} + \frac{\hat{u}^2}{2}$$

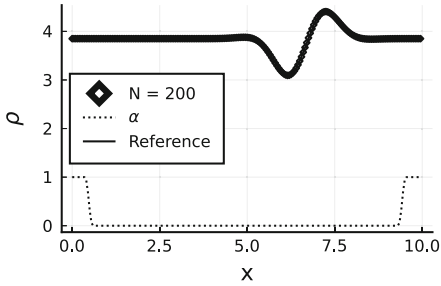
developed by Ismail and Roe in [18] that conserves the selected entropy was used for the entropy conservative part of the scheme. This flux was selected as it is also used



(a) L_1 Error for $N \in \{50, 100, 200, 400, 800\}$ Points. Please note that the fourth order strong stability preserving Runge-Kutta method used for time-integration limits the achievable order to 4.



(b) Solution for $N = 100$ points. The Entropy inequality predictor (miss) detects an entropy dissipating shock and partially activates the flux to dissipate entropy.



(c) Solution for $N = 200$ points. The higher resolution deactivates the entropy inequality predictor and the entropy conservative high order flux reduces the entropy dissipation to zero.

Fig. 11 convergence analysis for the Euler equations. LMRLxFRI Scheme of order 6. The same parameters and fluxes as in Fig. 10 were used

in several other publications [6, 7]. Other options, including fluxes that also conserves the kinetic energy, are possible [25]. The parameters $a = 1/1000$ and $b = 1/1000$, that were determined experimentally as before, were used. A new set of parameters is needed as a different entropy inequality predictor is used whose typical amplitudes and offset are different. Different entropies can also influence these parameters and an analysis giving explicit formulas is planed for a future publication. Reference solutions were calculated by the LMRLxFRI scheme with $N = 1600$ points of order 6 and using SSPRK104 for time integration. First the ability of the scheme to resolve shocks was tested by the Shu-Osher test case number 6 from [28] given by the initial conditions

$$\rho_0(x,0) = \begin{cases} 3.857153 \\ 1 + \varepsilon \sin(5x) \end{cases} \quad v_0(x,0) = \begin{cases} 2.629 \\ 0 \end{cases} \quad p_0(x,0) = \begin{cases} 10.333 & x < 1 \\ 1 & x \geq 1 \end{cases}$$

The results can examined in Fig. 10. A second experiment was carried out to demonstrate the ability of the scheme to achieve high order in smooth areas. The initial condition

$$\rho_0(x,0) = 3.857153 + \varepsilon \sin(2x) \quad v_0(x,0) = 2.0 \quad p_0(x,0) = 10.33333.$$

is a density variation that is carried downstream to the right and the results from the convergence analysis are shown in Fig. 11.

5 Conclusion

We first looked at some numerical solutions to hyperbolic conservation laws and saw that the entropy inequality is not enough to guarantee high quality solutions. Afterwards a new philosophy for the construction of schemes was proposed as they should satisfy the Dafermos entropy criterion and the entropy equality for smooth solutions. We then constructed such a solver by the hybrid usage of entropy conservative and entropy dissipative fluxes. Numerical experiments showed that having no entropy dissipation for smooth solutions, as motivated by the entropy equality, and enough entropy dissipation in discontinuous areas by the Godunov or respective LxF scheme provides a scheme with improved accuracy in smooth areas over the Godunov scheme and an accuracy not worse than the Godunov or respective LxF scheme in non-smooth areas. This can be seen as an improvement over prior attempts of using the Dafermos criterion for numerical schemes as in [26, Chap. 9.2] where excessive dissipation in smooth areas lead to bad solutions. The primary difference being that the stencil selector proposed in [26, Chap. 9.2] also tried to dissipate the maximum amount of entropy in smooth areas while in fact the analytic theory in form of the entropy equality dictates the conservation of entropy as the maximum allowable reduction of entropy in this case. Research is ongoing concerning the improvement of stencil selection algorithms in reconstruction based methods by taking into account not only the maximum entropy dissipation but also the entropy equality for smooth areas. The methods that were constructed to calculate the coefficient α could be used also in methods based on steered dissipation as for example in [10, Chap. 11]. Better α distributions could on the other hand greatly enhance the abilities of the constructed schemes. An algorithm based on artificial intelligence has been tested by the author and a preprint [17] concerning several other algorithms to calculate α is available. Practical applications of finite volume methods are often multidimensional problems therefore a future publication concerning this scheme will generalize the presented method to multiple space dimensions on unstructured grids.

Acknowledgements Simon Klein would like to thank Thomas Sonar and Marko Stautz for their support during the preparation of this manuscript. Mr. Klein would further like to thank the anonymous reviewer for his comments on the first draft of the manuscript, Hendrik Ranocha and Philipp Öffner for interesting discussions on entropy dissipative schemes, leading to an improved presentation of the material. The author was partially supported by the German Science Foundation (DFG) under the Grant SO 363/14-1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abgrall, R., Nordström, J., Öffner, P., Tokareva, S.: Analysis of the SBP-SAT stabilization for finite element methods part I: Linear problems. Commun. Appl. Math. Comput (2020). https://doi.org/10.1007/s42967-020-00086-2
2. Abgrall, R., Nordström, J., Öffner, P., Tokareva, S.: Analysis of the SBP-SAT stabilization for finite element methods part II: entropy stability. Commun. Appl. Math. Comput. (2021). https://doi.org/10.1007/s42967-020-00086-2
3. Archibald, R., Gelb, A., Yoon, J.: Polynomial fitting for edge detection in irregularly sampled signals and images. J. Numer. Anal. **43**, 259–279 (2005)
4. Chen, T., Shu, C.W.: Review of entropy stable discontinuous Galerkin methods for systems of conservation laws on unstructured simplex meshes. Trans. Appl. Math. **1**, 1–52 (2020)
5. Dafermos, C.M.: The entropy rate admissibility criterion for solutions of hyperbolic conservation laws. J. Differ. Equ. **14**, 202–212 (1972)
6. Fisher, T., Carpenter, M.: High-order entropy stable finite difference schemes for nonlinear conservation laws. finite domains. *NASA Technical Report*, (2013). https://doi.org/10.1016/j.jcp.2013.06.014
7. Fjordholm, U.S., Mishra, S., Tadmor, E.: Arbitrarily high order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. SIAM J. Numer. Anal. **50**, 544–573 (2012)
8. Fornberg, B.: Calculation of weights in finite difference formulas. *SIAM Rev.*, 40(3):685–691, (1998). URL http://www.jstor.org/stable/2653239
9. Friedrichs, K.O.: The identity of weak and strong extensions of differential operators. Trans. Am. Math. Soc. **55**, 132–151 (1944). https://doi.org/10.2307/1990143
10. Glaubitz, J.: *Shock capturing and high-order methods for hyperbolic conservation laws*. PhD thesis, TU Braunschweig, (2020)
11. Godunov, C.K.: A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations. Matematicheskii Sbornik **47**, 271–306 (1959)
12. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**, 05 (2001). https://doi.org/10.1137/S003614450036757X
13. Guermond, J.-L., Pasquetti, R., Popov, B.: Entropy viscosity method for nonlinear conservation laws. J. Comput. Phys. **230**, 4248–4267 (2011)
14. Guermond, J.-L., Nazarov, M., Popov, B., Tomas, I.: Second-order invariant domain preserving approximation of the Euler equations using convex limiting. J. Sci. Comput. **40**, 3211–3239 (2018)
15. Harten, A.: On the symmetric form of systems of conservation laws with entropy. J. Comput. Phys. **49**, 151–164 (1983)
16. Harten, A., Enquist, B., Osher, S., Chakravarthy, S.R.: Uniformly high order accurate essentially non-oscillatory schemes III. J. Comput. Phys. **71**, 231–303 (1987)
17. Hillebrand, D., Klein, S.C., Öffner, P.: Comparison to control oscillations in high-order Finite Volume schemes via physical constraint limiters, neural networks and polynomial annihilation. *arXiv e-prints*, art. arXiv:2203.00297, (March 2022)
18. Ismail, F., Roe, P.L.: Affordable, entropy-consistent flux functions II: entropy production at shocks. J. Comput. Phys. **228**, 5410–5436 (2009)
19. Kuzmin, D., Hajduk, H., Rupp, A.: Limiter-based entropy stabilization of semi-discrete and fully discrete schemes for nonlinear hyperbolic problems. Comput. Methods Appl. Mech. Eng. **389**, 114428 (2021)
20. Lax, P.D.: Shock waves and entropy. In: Contributions to Nonlinear Functional Analysis, pp. 603–634. Academic Press, Cambridge (1971)

21. Lax, P.D.: Functional Analysis. Wiley Interscience, New York (2002)
22. LeFloch, P.G., Mercier, J.M., Rohde, C.: Fully discret, entropy conservative schemes of arbitrary order. SIAM J. Numer. Anal. **40**, 1968–1992 (2002)
23. Osher, S.: Riemann solvers, the entropy condition, and difference appsoximations. SIAM J. Numer. Anal. **21**, 217–235 (1984)
24. Perlin, K.: *Texturing and Modeling: A Procedural Approach*. Morgan Kaufmann, (2002)
25. Ranocha, H.: Comparison of some entropy conservative numerical fluxes for the Euler equations. J. Sci. Comput. **76**(1), 216–242 (2018)
26. Ranocha, H.: *Generalised Summation-by-Parts Operators and Entropy Stability of Numerical Methods for Hyperbolic Balance Laws*. PhD thesis, TU Braunschweig, (02 2018)
27. Ranocha, H., Öffner, P., Sonar, T.: Summation-by-parts operators for correction procedure via reconstruction. J. Comput. Phys. **311**, 299–328 (2016). https://doi.org/10.1016/j.jcp.2016.02.009
28. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes II. J. Comput. Phys. **83**, 439–471 (1989)
29. Sonar, T., Öffner, P., Glaubitz, J., Ranocha, H.: Stability of artificial dissipation and modal filtering for flux reconstruction schemes using summation-by-parts operators. Appl. Numer. Math. **128**, 1–23 (2018)
30. Tadmor, E.: The large-time behavior of the scalar, genuinely nonlienar Lax-Friedrichs scheme. Math. Comput. **43**, 353–368 (1984)
31. Tadmor, E.: Numerical viscosity and the entropy condition for conservative difference schemes. Math. Comput. **43**, 369–381 (1984)
32. Tadmor, E.: The numerical viscosity of entropy stable schemes for systems of conservation laws. Math. Comput. **49**, 91–103 (1987)
33. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time dependent problems. Acta Numer **12**, 451–512 (2003)
34. Tadmor, E., Waagan, K.: Adaptive spectral viscosity for hyperbolic conservation laws. J. Sci. Comput. **34**, 993–1009 (2012)
35. Tang, H.: On the sonic point glitch. J. Comput. Phys. **202**, 507–532 (2005). https://doi.org/10.1016/j.jcp.2004.07.013

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.