SIAM J. NUMER. ANAL.
Vol. 55, No. 6, pp. 3120–3146

INVARIANT DOMAINS AND SECOND-ORDER CONTINUOUS FINITE ELEMENT APPROXIMATION FOR SCALAR CONSERVATION EQUATIONS[*]

JEAN-LUC GUERMOND[†] AND BOJAN POPOV[†]

Abstract. We propose a technique for approximating nonlinear scalar conservation equations that uses continuous finite elements and is formally (at least) second-order accurate in space and maximum principle preserving. The method is explicit in time, uses unstructured continuous finite elements in space, and works in any space dimension. The stability and accuracy of the method are achieved by adapting the artificial viscosity.

Key words. conservation equations, second-order accuracy, maximum principle, entropy-viscosity, finite element method, limiters

AMS subject classifications. 65M60, 65M10, 65M15, 35L65

DOI. 10.1137/16M1106560

1. Introduction.

In this paper we are concerned with the approximation of nonlinear scalar conservation equations in arbitrary space dimension using explicit time stepping and unstructured continuous finite elements in space. The objective is to introduce a continuous finite element method that is formally second-order accurate in space, locally preserves the maximum principle, and is observed to converge to the entropy solution. Throughout the paper we consider the scalar conservation equation

$$(1)\quad \partial_t u + \nabla\!\cdot\! \boldsymbol{f}(u) = 0 \text{ for a.e. } (\boldsymbol{x},t)\in\mathbb{R}^d\!\times\!\mathbb{R}_+, \qquad u(\boldsymbol{x},0)=u_0(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x}\in\mathbb{R}^d,$$

with Lipschitz flux $\boldsymbol{f}:\mathbb{R}\longrightarrow\mathbb{R}^d$. To simplify questions regarding boundary conditions, we assume that either the initial data is constant outside a compact set or periodic boundary conditions are enforced. The spatial domain where the approximation is constructed is denoted by D in both cases. In the case of the Cauchy problem, D is a bounded polygonal open set of $\mathbb{R}^d$ large enough that the domain of influence of $\boldsymbol{u}_0$ is always included in D over the entire duration of the simulation.

In the finite volume literature there are many second- or higher-order nonoscillatory schemes based on minmod or other limiters that are considered to be numerically robust. The use of nonlinear limiters such as minmod, generalized minmod, superbee, uniformly nonoscillatory (UNO), essentially nonoscillatory (ENO), and others is essential in the construction of maximum principle preserving schemes, and this is typically done via either slope or flux limiting; see, for example, Harten and Osher [22], Harten et al. [23], Sweby [37], Colella and Glaz [10], van Leer [39], and the references therein. We also refer the reader to the series of papers by Zhang and Shu [42] (and references therein) where limiting is applied in the context of the discontinuous Galerkin method.

[*] Received by the editors December 5, 2016; accepted for publication (in revised form) October 2, 2017; published electronically December 19, 2017.

http://www.siam.org/journals/sinum/55-6/M110656.html

Funding: The work of the authors was supported in part by National Science Foundation grants DMS-1620058 and DMS-1619892, by the Air Force Office of Scientific Research, USAF, under grant/contract FA9550-15-1-0257, and by the Army Research Office under grant/contract W911NF-15-1-0517.

[†] Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843 (guermond@math.tamu.edu, popov@math.tamu.edu).

In the wake of the work of Burman and Ern [7], significant progress on the theory of the maximum principle has been made in the continuous finite element literature on the linear advection diffusion equation. But the general case of nonlinear conservation equations is not well developed, and with the exception of the so-called flux corrected transport method (FCT) of Zalesak [41], there is no real equivalent to simple flux limiting such as minmod and variations thereof (see also Kuzmin and Turek [31] and Kuzmin, Löhner, and Turek [32] for reviews of FCT).

There are three objectives in this paper. The first is to identify a good first-order finite element method that is consistent with the maximum principle and with all the entropy inequalities. This objective is achieved in section 3 by adapting some of the tools of the finite volume literature to the continuous finite element setting recently developed in [17]. The method introduced in [17] is a first-order invariant domain preserving scheme for general nonlinear hyperbolic systems in arbitrary space dimension; the method is based on a guaranteed bound on the local maximum wave speed and is henceforth referred to as the GMS scheme (guaranteed maximum speed). In passing we produce a counterexample establishing that the maximum principle preserving method usually referred to in the literature as local extrema diminishing (LED) does not converge to the entropy solution (see Lemma 3.2), thereby showing that it is necessary to have a good control on the maximum wave speed to be convergent. The second objective is to propose various maximum principle preserving extensions of the GMS scheme that are formally second-order accurate in space. This is done in section 4, where we investigate the following methods: (i) a smoothness-based technique inspired by Jameson, Schmidt, and Turkel [28, eq. (12)], Jameson [27, p. 6], and Burman [6, Thm. 4.1]; (ii) a greedy viscosity; (iii) an FCT-based viscosity. The originality of the present work is that the L^∞-stability and second-order accuracy are obtained by adapting the artificial viscosity of the GMS scheme instead of limiting slopes or interface fluxes as is usually done in the finite volume literature. In section 4 we also show that removing viscosity in a blind way like FCT limiting may lead to serious convergence problems (although the maximum principle holds). More specifically, we give a simple counterexample showing that the combination of the Galerkin method and FCT limiting fails to converge properly; see Lemma 4.6. The last objective of the present work is to propose a method that, in addition to being maximum principle preserving, is also formally entropy consistent and potentially high-order for any polynomial degree. (Note that the methods proposed in section 4 are only second-order accurate in space and may fail to be entropy consistent.) This method, which we call entropy viscosity [18], is presented in section 5. It is made maximum principle preserving by using FCT limiting. The low-order method is GMS, and the high-order viscosity is based on an entropy residual. This combination is shown to be the most robust one in the numerical section, section 6. In particular, it is shown that when the flux is not genuinely nonlinear, this combination is the only one that survives without adding any ad hoc entropy fixes. The other three methods proposed in section 4 are good alternatives to the entropy-viscosity method only when composite waves are not present (convex flux).

The paper is organized as follows. The notation and definitions are introduced in section 2. The GMS first-order method is discussed in section 3. The content of this section is not new but facilitates the reading of the following sections. We discuss various techniques to make the artificial viscosity formally high-order in section 4. An entropy-viscosity method using the techniques introduced in sections 3 and 4 as a "low-order" method is proposed in section 5. The methods proposed in sections 4 and 5 are illustrated numerically in section 6. The convergence properties of these methods are verified on a sequence of linear and nonlinear benchmark examples. Concluding remarks are reported in section 7.

2. Preliminaries.

We introduce the notation and definitions in this section. We adopt the same notation as in [17]. The reader familiar with the content of [17] is invited to jump to section 3.

2.1. The finite element space.

In order to approximate the solution of (1) with continuous finite elements, we consider a shape-regular sequence of matching meshes $(\mathcal{T}_h)_{h>0}$. To be general, we assume that the elements in the mesh sequence are generated from a finite number of reference elements denoted by $\widehat{K}_1,\dots,\widehat{K}_\varpi$. For example, in two space dimensions the mesh $\mathcal{T}_h$ could be composed of a combination of parallelograms and triangles; in this case $\varpi=2$. In three space dimensions, $\mathcal{T}_h$ could be composed of a combination of tetrahedra, parallelepipeds, and triangular prisms; in this case $\varpi=3$. We denote by $\boldsymbol{T}_K:\widehat{K}_r\longrightarrow K$ the diffeomorphism mapping $\widehat{K}_r$ to an arbitrary element $K\in\mathcal{T}_h$. We now introduce a set of reference finite elements $\{(\widehat{K}_r,\widehat{P}_r,\widehat{\Sigma}_r)\}_{1\le r\le\varpi}$ (the index $r\in\{1{:}\varpi\}$ will be omitted in the rest of the paper to alleviate the notation), and we define the scalar-valued finite element space

$$P(\mathcal{T}_h)=\{v\in\mathcal{C}^0(D;\mathbb{R})\mid v_{|K}\circ\boldsymbol{T}_K\in\widehat{P}\ \forall K\in\mathcal{T}_h\},\tag{2}$$

where $\widehat{P}$ is the reference space (note that the index r has been omitted). Letting $n_{\rm sh}:=\dim\widehat{P}$, the shape functions on the reference element are denoted by $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$. We assume that the basis $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$ has the partition of unity property:

$$\sum_{i\in\{1:n_{\rm sh}\}}\widehat{\theta}_i(\widehat{\boldsymbol{x}})=1\quad\forall\widehat{\boldsymbol{x}}\in\widehat{K}.\tag{3}$$

The global shape functions are denoted by $\{\varphi_i\}_{i\in\{1:I\}}$. They form a basis of $P(\mathcal{T}_h)$, and the partition of unity property implies that $\sum_{i\in\{1:I\}}\varphi_i(\boldsymbol{x})=1$ for all $\boldsymbol{x}\in D$. The support of φ_i is denoted by S_i, $i\in\{1{:}I\}$. Let E be a union of cells in $\mathcal{T}_h$; we denote by $|E|$ the measure of E. The set of integers that contains the indices of all the shape functions whose support on E is of nonzero measure is denoted $\mathcal{I}(E):=\{j\in\{1{:}I\}\mid |S_j\cap E|\neq 0\}$.

The matrix with entries $m_{ij}:=\int_D\varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $i,j\in\{1{:}I\}$, is called the consistent mass matrix and is denoted by $\mathcal{M}\in\mathbb{R}^{I\times I}$. The diagonal matrix with entries equal to $m_i:=\int_{S_i}\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ is called the lumped mass matrix and is denoted by $\mathcal{M}^L$. The partition of unity property implies that $\sum_{j\in\mathcal{I}(S_i)}m_{ij}=m_i$. One key assumption that is used in the rest of the paper is that

$$m_i>0\quad\forall i\in\{1{:}I\}.\tag{4}$$

The assumptions (3) and (4) hold for many Lagrange elements and for Bernstein–Bezier finite elements of any polynomial degree.

2.2. Generic form of the algorithm.

Since, as demonstrated in [20], it is impossible to construct an explicit continuous finite element method that is stabilized with artificial viscosity and satisfies the maximum principle if the time derivative is approximated with the consistent mass matrix, we use instead the lumped mass matrix. That is to say, denoting by $u_h^n=\sum_{i\in\{1:I\}}\mathsf{U}_i^n\varphi_i\in P(\mathcal{T}_h)$ the approximation of u at time t^n, the time derivative $\int_D\partial_t u_h\varphi_i\,\mathrm{d}\boldsymbol{x}$ is approximated by

$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau}$, where $\tau := t^{n+1}-t^n$ is the time step. It is also shown in [17] that the Galerkin approximation of the flux term $\int_D \nabla\cdot(\boldsymbol{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}\boldsymbol{x}$ can be approximated by $\sum_{i\in\{1:I\}} \boldsymbol{f}(\mathsf{U}_j^n)\cdot \int_D \varphi_i \nabla\varphi_j \,\mathrm{d}\boldsymbol{x}$. Upon introducing the coefficients $\boldsymbol{c}_{ij}\in\mathbb{R}^d$,

$$\boldsymbol{c}_{ij} := \int_D \varphi_i \nabla\varphi_j \,\mathrm{d}\boldsymbol{x}, \tag{5}$$

the Galerkin approximation of the flux term has the following alternative representation: $\sum_{j\in\mathcal{I}(S_i)} \big(\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)\big)\cdot\boldsymbol{c}_{ij}$; notice that $\sum_{j\in\mathcal{I}(S_i)} \boldsymbol{c}_{ij} = 0$ owing to the partition of unity property. In the rest of the paper we consider the following generic form for the finite element approximation of (1):

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(S_i)} \big((\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n))\cdot\boldsymbol{c}_{ij} - d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)\big) = 0, \tag{6}$$

where the artificial viscosity coefficients d_{ij}^n are yet to be specified but are systematically assumed to satisfy the following properties:

$$0\le d_{ij}^n, \qquad d_{ij}^n = d_{ji}^n, \qquad \sum_{j\in\mathcal{I}(S_i)} d_{ij}^n = 0. \tag{7}$$

The objective of the paper is to investigate various definitions of d_{ij}^n to make the method formally second-order accurate in space and maximum principle preserving. Moreover, the resulting method should be verified, numerically or theoretically, to converge to the entropy solution.

Remark 2.1 (conservation). The symmetry assumption on the artificial viscosity coefficients implies that $\sum_{i,j\in\{1:I\}} d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n) = 0$. Then, upon observing that $\int_D u_h^n \,\mathrm{d}\boldsymbol{x} = \sum_{i\in\{1:I\}} m_i \boldsymbol{\mathsf{U}}_i^n$, we infer that (6) implies conservation:

$$\int_D \boldsymbol{u}_h^{n+1}\,\mathrm{d}\boldsymbol{x} = \int_D \boldsymbol{u}_h^n\,\mathrm{d}\boldsymbol{x} - \tau\int_D \nabla\cdot\Big(\sum_{j\in\{1:I\}} \boldsymbol{f}(\boldsymbol{\mathsf{U}}_j^n)\varphi_j\Big)\,\mathrm{d}\boldsymbol{x} \qquad \forall n\ge 0. \tag{8}$$

3. Choice of the first-order viscosity.

We show in this section that, contrary to appearances, choosing a first-order viscous method is not a trivial matter. We illustrate our point by comparing two maximum principle preserving first-order methods. Both are local maximum principle preserving, but one violates entropy inequalities because it does not correctly estimate the local maximum wave speed, whereas the other satisfies a local entropy inequality for every convex entropy.

3.1. Local extrema diminishing (LED) schemes.

Let us first consider an approach known in the literature as local extrema diminishing (LED) due to Roe [35, p. 361], Jameson [26, section 2.1], and others; see, e.g., Kuzmin, Löhner, and Turek [32, p. 163] and Kuzmin and Turek [31, eq. (32)–(33)]. The technique consists of rewriting (6) as follows:

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau} = -\sum_{i\ne j\in\mathcal{I}(S_i)} \frac{\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_j^n-\mathsf{U}_i^n}\cdot\boldsymbol{c}_{ij}(\mathsf{U}_j^n-\mathsf{U}_i^n) + \sum_{j\in\mathcal{I}(S_i)} d_{ij}^n\mathsf{U}_j^n. \tag{9}$$

Let $\boldsymbol{n}_{ij} = \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}$ and let $A_{ij}(\mathsf{U}_L,\mathsf{U}_R) := \boldsymbol{n}_{ij}\cdot\frac{\boldsymbol{f}(\mathsf{U}_L)-\boldsymbol{f}(\mathsf{U}_R)}{\mathsf{U}_L-\mathsf{U}_R}$ if $\mathsf{U}_L\ne\mathsf{U}_R$ and $A_{ij}(\mathsf{U},\mathsf{U}) := \boldsymbol{f}'(\mathsf{U})\cdot\boldsymbol{c}_{ij}$ otherwise. Denoting $k_{ij}^n := A_{ij}(\mathsf{U}_i^n,\mathsf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}$, we have

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n\Big(1-\frac{\tau}{m_i}\sum_{i\ne j\in\mathcal{I}(S_i)}(-k_{ij}^n+d_{ij}^n)\Big) + \sum_{i\ne j\in\mathcal{I}(S_i)} \frac{\tau}{m_i}(-k_{ij}^n+d_{ij}^n)\mathsf{U}_j^n, \tag{10}$$

where we have used $d_{ii} := -\sum_{i\neq j\in\mathcal{I}(S_i)} d_{ij}^n$. Following [26, 31, 32], we now set

$$d_{ij}^n := \max(0, k_{ij}^n, k_{ji}^n), \qquad i \neq j. \tag{11}$$

Lemma 3.1 (maximum principle). *Assume* (11) *and assume that* τ *is small enough that* $1 - \frac{\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(S_i)}(-k_{ij}^n + d_{ij}^n) \geq 0$ *for all* $i \in \{1{:}I\}$. *Then* $\mathsf{U}_i^{n+1} \in \text{conv}\{\mathsf{U}_j^n,\ j \in \mathcal{I}(S_i)\}$ *for all* $i \in \{1{:}I\}$.

Proof. The definition (11) implies that $-k_{ij}^n + d_{ij}^n \geq 0$ for all $i \in \{1{:}I\}$, $j \in \mathcal{I}(S_i)$. The assumed CFL condition and (10) imply that $\mathsf{U}_i^{n+1} \in \text{conv}\{\mathsf{U}_j^n,\ j \in \mathcal{I}(S_i)\}$. □

Note that owing to the assumed boundary conditions, the definition of $\boldsymbol{c}_{ij}$ implies that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$, i.e., $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ij}$. The definition of k_{ij}^n in turn implies $k_{ij}^n = -k_{ij}^n$; hence we can reformulate the definition of d_{ij}^n in (11) as follows:

$$d_{ij}^n := |A_{ij}(\mathsf{U}_i^n, \mathsf{U}_j^n)|\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}, \qquad i \neq j. \tag{12}$$

In conclusion, d_{ij}^n is proportional to the speed $A_{ij}(\mathsf{U}_i^n, \mathsf{U}_j^n)$. Although the above technique preserves the maximum principle locally and looks reasonable a priori (and is used a lot in the literature), it turns out that it is not diffusive enough to make the method convergent; see [16, section 3.3] and [17, section 5.1]. The convergence result established in [16] requires an estimation of the wave speed that is more accurate than just the average speed $|A_{ij}(\mathsf{U}_i^n, \mathsf{U}_j^n)| := |\boldsymbol{n}_{ij}\cdot\frac{\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_j^n-\mathsf{U}_i^n}|$, which is invoked in (12). If the Riemann problem with data $(\mathsf{U}_i^n, \mathsf{U}_j^n)$ is a simple shock, the above definition of the wave speed is correct since it is the speed given by the Rankin–Hugoniot formula; but it may not be sufficient if the Riemann solution is an expansion or a composite wave.

Lemma 3.2 (LED counterexample). *There exist* C^∞ *fluxes and piecewise smooth initial data such that under the CFL condition stated in Lemma* 3.1, *the approximate sequence given by* (9) *and* (11) *(or* (12)*) with continuous piecewise linear approximation does not converge to the unique entropy solution of* (1).

Proof. Let us consider Burgers's equation in one space dimension, $\boldsymbol{f}(u) := f(u)\boldsymbol{e}_x$, $f(u) := \frac{1}{2}u^2$, $D := (-1, 1)$, with data $u_0(x) := -1$ if $x \leq 0$ and $u_0(x) := 1$ otherwise. Let $N \in \mathbb{N}\backslash\{0\}$, and let us consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, where the nodes x_i, $i \in \{0{:}2N\}$, are such that $-1 = x_0 < x_1 < \cdots < x_{2N} = 1$ and $x_N \leq 0 < x_{N+1}$. The mesh may not be uniform; we denote $h_i = x_i - x_{i-1}$, $1 \leq i \leq 2N$. Let $P(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$. We have $\boldsymbol{c}_{ii-1} = -\frac{1}{2}\boldsymbol{e}_x$, $\boldsymbol{c}_{ii} = \boldsymbol{0}$, and $\boldsymbol{c}_{ii+1} = \frac{1}{2}\boldsymbol{e}_x$, $m_i = \frac{h_i+h_{i+1}}{2}$. Let us consider the approximate initial data $u_h^0 = \sum_{i\in\{0:2N\}} \mathsf{U}_i^0\varphi_i(x)$ with $\mathsf{U}_i^0 = -1$ if $i \leq N$ and $\mathsf{U}_i^0 = 1$ if $i \geq N+1$. The definition for the update U_i^{n+1}, $n \geq 0$, is

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n + \frac{\tau}{2m_i}(f(\mathsf{U}_{i-1}^n) - f(\mathsf{U}_{i+1}^n)) + \frac{\tau}{m_i}d_{ii-1}^n(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{m_i}d_{ii+1}^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$

with the convention that $\mathsf{U}_{-1}^n = -1$ and $\mathsf{U}_{2N+1}^n = 1$. Clearly, $\mathsf{U}_i^1 = \mathsf{U}_i^0$ for $i \leq N-1$ and $N+2 \leq i$. For $i = N$ we have $\mathsf{U}_{N-1}^0 = -1$, $\mathsf{U}_N^0 = -1$, and $\mathsf{U}_{N+1}^0 = 1$, giving $f(\mathsf{U}_{N+1}^0) - f(\mathsf{U}_{N-1}^0) = \frac{1}{2}(1-1) = 0$, and $d_{NN-1}^0 = \frac{1}{2}$, $d_{NN+1}^0 = 0$; hence $\mathsf{U}_N^1 = \mathsf{U}_N^0$. Similarly for $i = N+1$, we have $\mathsf{U}_N^0 = -1$, $\mathsf{U}_{N+1}^0 = 1$, and $\mathsf{U}_{N+2}^0 = 1$, giving $f(\mathsf{U}_{N+2}^0) - f(\mathsf{U}_N^0) = \frac{1}{2}(1-1) = 0$, and $d_{N+1N}^0 = 0$, $d_{N+1N+2}^0 = \frac{1}{2}$; hence

$\mathsf{U}^1_{N+1} = \mathsf{U}^0_{N+1}$. In conclusion, $u^1_h = u^0_h$, i.e., $u^n_h = u^0_h$ for every $n \geq 0$. This proves that the solution is a stationary discontinuity, whereas it should be an approximation of an expansion wave; hence the method does not converge to the entropy solution. □

Remark 3.3 (Roe's average). The above arguments generalize to hyperbolic systems with flux $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^{m\times d}$. Consider the average Jacobian matrix defined by

$$A_{ij}(\mathbf{U}_L, \mathbf{U}_R) := \boldsymbol{n}_{ij}\cdot \int_0^1 D_{\boldsymbol{u}}\boldsymbol{f}(\mathbf{U}_R + \theta(\mathbf{U}_L - \mathbf{U}_R))\,\mathrm{d}\theta. \tag{13}$$

This definition implies that $(\boldsymbol{f}(\mathbf{U}_L) - \boldsymbol{f}(\mathbf{U}_R))\cdot\boldsymbol{n}_{ij} = A_{ij}(\mathbf{U}_L, \mathbf{U}_R)(\mathbf{U}_L - \mathbf{U}_R)$. By definition of hyperbolicity, the $m\times m$ matrix $D_{\boldsymbol{u}}\boldsymbol{f}\cdot\boldsymbol{n}_{ij}$ is diagonalizable with real eigenvalues for any unit vector $\boldsymbol{n}_{ij}$ in $\mathbb{R}^d$, but it may not be the case of $A_{ij}(\mathbf{U}_L, \mathbf{U}_R)$. Anyway, *if* the two states $\mathbf{U}_L, \mathbf{U}_R$ are close enough so that $A_{ij}(\mathbf{U}_L, \mathbf{U}_R)$ is diagonalizable with real eigenvalues, and *if* the Riemann problem with left and right states $\mathbf{U}_L, \mathbf{U}_R$ has a solution consisting of a single discontinuity (shock or contact for the Euler equations), then the wave speed of the discontinuity is one of the eigenvalues of $A(\mathbf{U}_L, \mathbf{U}_R)$; see, e.g., Bressan [5, section 5.2]. In that case, the spectral radius of $A(\mathbf{U}_L, \mathbf{U}_R)$ is a guaranteed upper bound of the maximum wave speed. This observation is at the origin of the popularity of the so-called Roe's average. But the above argument relies on two *ifs*, and in general there is no guarantee at all that the spectral radius of $A(\mathbf{U}_L, \mathbf{U}_R)$ is an upper bound of the maximum wave speed in the Riemann problem, as clearly demonstrated in Lemma 3.2.

Remark 3.4 (entropy glitch). The phenomenon at the origin of the above counterexample is known in the literature as the "entropy glitch" and various "fixes" are available; see, e.g., Harten and Hyman [21]. It turns out that the correction proposed in [21] amounts to computing an actual upper bound of the maximum wave speed of the exact solution of the Riemann problem (at least for scalar conservation equations); that is, the correction proposed in [21] consists of *not using* the LED method but using instead a method that correctly estimates the maximum wave speed from above. It is shown in the next section that estimating from above the maximum wave speed of the exact solution of the Riemann problem is the only thing that really matters.

3.2. Guaranteed maximum speed (GMS) schemes. This section addresses the shortcomings of the LED method unveiled in the previous section. Following an idea by Hoff [24, 25], it is shown in [17] that (6) can be reformulated as follows:

$$\mathsf{U}^{n+1}_i = \mathsf{U}^n_i\Big(1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d^n_{ij}}{m_i}\Big) + \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d^n_{ij}}{m_i}\overline{\mathsf{U}}^n_{ij}, \tag{14}$$

with the auxiliary quantities $\overline{\mathsf{U}}^n_{ij}$ defined by

$$\overline{\mathsf{U}}^n_{ij} := \frac{1}{2}(\mathsf{U}^n_j + \mathsf{U}^n_i) - (\boldsymbol{f}(\mathsf{U}^n_j) - \boldsymbol{f}(\mathsf{U}^n_i))\cdot\frac{\boldsymbol{c}_{ij}}{2d^n_{ij}}. \tag{15}$$

Let $\boldsymbol{n}$ be a unit vector in $\mathbb{R}^d$. The key observation made in [17] consists of introducing the following one-dimensional Riemann problem:

$$\partial_t u + \partial_x(\boldsymbol{f}(u)\cdot\boldsymbol{n}) = 0, \quad (x,t)\in\mathbb{R}\times\mathbb{R}_+, \qquad u(x,0) = \begin{cases} u_L & \text{if } x < 0,\\ u_R & \text{if } x > 0.\end{cases} \tag{16}$$

Let $u(\boldsymbol{n}, u_L, u_R)(x,t)$ be the unique solution to (16). Let $\lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ be the maximum wave speed in (16), i.e., $u(\boldsymbol{n}, u_L, u_R)(x,t) = u_L$ for all $x < -|\lambda_{\max}|t$ and $u(\boldsymbol{n}, u_L, u_R)(x,t) = u_R$ for all $|\lambda_{\max}|t < x$; see Lemma 3.8. Upon denoting $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$ and introducing the fake time $t := \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^n$, we realize that

$$\overline{\mathsf{U}}_{ij}^n = \int_{-\frac{1}{2}}^{\frac{1}{2}} u(\boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)(x,t)\,\mathrm{d}x \tag{17}$$

provided $t\lambda_{\max}(\boldsymbol{n}_{ij}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$. An immediate consequence of this observation is that $\overline{\mathsf{U}}_{ij}^{n+1} \in \operatorname{conv}(\mathsf{U}_i^n, \mathsf{U}_j^n)$, i.e., $\overline{\mathsf{U}}_{ij}^{n+1}$ satisfies the maximum principle. The above arguments then motivate the following definition for the viscosity coefficients d_{ij}^n:

$$d_{ij}^n := \max(\lambda_{\max}(\boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji}, \mathsf{U}_j^n, \mathsf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}). \tag{18}$$

The following results are proved in [17].

THEOREM 3.5 (local invariance and entropy inequality). *Let the finite element setting satisfy the structural hypotheses detailed in section* 2.1. *Let* $n \geq 0$, *and let* $i \in \{1{:}I\}$. *Assume that* τ *is small enough that* $1 + 2\tau\frac{d_{ii}^n}{m_i} \geq 0$. *Then*
(i) $\mathsf{U}_i^{n+1} \in \operatorname{conv}\{\mathsf{U}_j^n \mid j \in \mathcal{I}(S_i)\}$.
(ii) *The following local entropy inequality holds for every entropy pair* $(\eta, \boldsymbol{q})$*:*

$$m_i\frac{(\eta(\mathsf{U}_i^{n+1}) - \eta(\mathsf{U}_i^n))}{\tau} + \int_D \nabla{\cdot}\Bigg(\sum_{j\in\mathcal{I}(S_i)} \boldsymbol{q}(\mathsf{U}_j^n)\varphi_j\Bigg)\varphi_i\,\mathrm{d}\boldsymbol{x} \;-\sum_{j\in\mathcal{I}(S_i)} d_{ij}^n\eta(\mathsf{U}_j^n) \leq 0.$$

COROLLARY 3.6 (global invariance 1). *Let the finite element setting satisfy the hypotheses detailed in section* 2.1. *Let* $n \in \mathbb{N}$. *Assume that* τ *is small enough that* $\min_{i\in\{1:I\}}\left(1 + 2\tau\frac{d_{ii}^n}{m_i}\right) \geq 0$. *Then* $\operatorname{conv}\{\mathsf{U}_i^{n+1} \mid i \in \{1{:}I\}\} \subset \operatorname{conv}\{\mathsf{U}_i^n \mid i \in \{1{:}I\}\}$.

The above results say that the maximum principle holds for the coordinate vector U^{n+1}, but they do not say whether this property holds for $\boldsymbol{u}_h^{n+1}$. To answer this question, we now assume that the reference shape functions $\{\widehat{\theta}_i\}_{i\in\{1:n_{\text{sh}}\}}$ are nonnegative:

$$\widehat{\theta}_i(\boldsymbol{x}) \geq 0 \quad \forall\widehat{\boldsymbol{x}} \in \widehat{K}. \tag{19}$$

This property holds for linear Lagrange elements on simplices, quadrangular elements, and hexahedra in two and three dimensions. It also holds for first-order prismatic elements in three dimensions. It is true also for Bernstein–Bezier finite elements of any polynomial degree. This assumption implies that $\varphi_i(\boldsymbol{x}) \geq 0$ for all $i \in \{1{:}I\}$ and all $\boldsymbol{x} \in D$. This, together with the partition of unity property (3) and the identity $\boldsymbol{u}_h^{n+1}(\boldsymbol{x}) = \sum_{i\in\{1:I\}} \mathsf{U}_i^{n+1}\varphi_i(\boldsymbol{x})$, in turn implies that $u_h^{n+1}(\boldsymbol{x}) \in \operatorname{conv}\{\mathsf{U}_i^{n+1} \mid i \in \{1{:}I\}\}$. Note finally that (4) is now just a consequence of (19).

COROLLARY 3.7 (global invariance 2). *Assume that the hypotheses detailed in section* 2.1 *and* (19) *hold. Let* $B := [\min_{\boldsymbol{x}\in D} u_0(\boldsymbol{x}), \max_{\boldsymbol{x}\in D} u_0(\boldsymbol{x})] \cup \operatorname{conv}\{\mathsf{U}_i^0 \mid i \in \{1{:}I\}\}$. *Let* $N \in \mathbb{N}$. *Suppose that* τ *is small enough that* $1 + 2\tau\frac{d_{ii}^n}{m_i} \geq 0$ *for all* $i \in \{1{:}I\}$ *and all* $n \in \{0{:}N\}$. *Then* $\operatorname{conv}\{\mathsf{U}_i^n \mid i \in \{1{:}I\}\} \subset B$ *and* $\boldsymbol{u}_h^n \in B$ *for all* $n \in \{0{:}N+1\}$.

We finish this section by recalling a standard result on the maximum wave speed.

Lemma 3.8. *Let $g \in \mathrm{Lip}(\mathbb{R};\mathbb{R})$, and let v be the entropy solution of*

$$\partial_t v + \partial_x g(v) = 0, \quad (x,t) \in \mathbb{R}\times\mathbb{R}_+, \qquad v(x,0) = \begin{cases} v_L & \textit{if } x < 0, \\ v_R & \textit{if } x > 0. \end{cases} \tag{20}$$

Then the maximum wave speed $\lambda_{\max}(g, v_L, v_R)$ in (20) *is given by*

$$\begin{cases} \max\left(\left| \inf_{v_L<y\le v_R} \frac{g(v_L)-g(y)}{v_L-y} \right|, \left| \sup_{v_L\le y<v_R} \frac{g(v_R)-g(y)}{v_R-y} \right| \right) & \textit{if } v_L < v_R, \\ \max\left(\left| \sup_{v_R<y\le v_L} \frac{g(v_R)-g(y)}{v_R-y} \right|, \left| \inf_{v_R\le y<v_L} \frac{g(v_L)-g(y)}{v_L-y} \right| \right) & \textit{if } v_L > v_R, \end{cases} \tag{21}$$

and we always have $\lambda_{\max}(g, v_L, v_R) \ge |\frac{g(v_R)-g(v_L)}{v_R-v_L}|$.

Proof. The construction of the exact solution of the Riemann problem for general (nonconvex) flux was first established for piecewise linear flux in Dafermos [11, Lemma 3.1]. The general case follows by density arguments using a perturbation result from Bouchut and Perthame [4, Thm. 3.1(iii)]. ∎

4. Higher-order viscosity.

We now explore the possibility of adding less viscosity than we did in section 3.2 in order to make the method more accurate in space. We change notation and denote by $d_{ij}^{\mathrm{V},n}$ the viscosity defined in (18) and denote by $\mathsf{U}_i^{\mathrm{V},n+1}$ the solution given by (14) with (18); i.e., we have added the index v.

4.1. Heuristic motivations.

The idea is to reduce the viscosity in regions where it is not needed and keep it first-order in regions where entropy production is essential. We then introduce a vector $\psi^n \in \mathbb{R}^I$ with the property that $0 \le \psi_i^n \le 1$ for all $i \in \{1{:}I\}$ and all $n \ge 0$, and we define the (hopefully) high-order viscosity

$$d_{ij}^n := d_{ij}^{\mathrm{V},n} \max(\psi_i^n, \psi_j^n) \qquad \forall i \ne j \in \{1{:}I\}, \tag{22}$$

with the convention $d_{ii}^n := -\sum_{i\ne j\in\mathcal{I}(S_i)} d_{ij}^n$. Then U_i^{n+1} is defined by

$$m_i \frac{\mathsf{U}_i^{n+1} - \mathsf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(S_i)} \left((\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n)) \cdot \boldsymbol{c}_{ij} - d_{ij}^n(\mathsf{U}_j^n - \mathsf{U}_i^n) \right) = 0. \tag{23}$$

Traditional flux limiting techniques can be reformulated in this setting. More precisely let us introduce the "high-order flux" $F_{ij}^{\mathrm{H},n} := \frac{1}{2}(\boldsymbol{f}(\mathsf{U}_j^n) + \boldsymbol{f}(\mathsf{U}_i^n))\cdot(2\boldsymbol{c}_{ij})$ which corresponds to the centered (Galerkin, or nonviscous) flux in the finite volume lingua; similarly, let us introduce the "low-order flux" $F_{ij}^{\mathrm{L},n} := F_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{V},n}\mathsf{U}_i^n$. (Note in passing that we have $2\|\boldsymbol{c}_{ij}\|_{\ell^2} = 1$ in dimension one, which justifies the factor 2 in the above formula.) The conservation properties $\sum_{j\in\mathcal{I}(S_i)} \boldsymbol{c}_{ij} = 0$, $\sum_{j\in\mathcal{I}(S_i)} d_{ij}^n = 0$, and $\sum_{j\in\mathcal{I}(S_i)} d_{ij}^{\mathrm{V},n} = 0$ imply that (23) can be rewritten as

$$m_i \frac{\mathsf{U}_i^{n+1} - \mathsf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(S_i)} F_{ij}^{\mathrm{L},n} + (1 - \max(\psi_i^n, \psi_j^n))(F_{ij}^{\mathrm{H},n} - F_{ij}^{\mathrm{L},n}) = 0, \tag{24}$$

which, after introducing the so-called limiting function $\phi_{ij}^n := 1 - \max(\psi_i^n, \psi_j^n)$, corresponds exactly to the traditional interpretation of flux limiting in the finite volume literature. (The function ϕ_{ij}^n is often denoted by $\mathcal{L}_{ij}^n$ in the FCT literature.) There is a large body of finite volume literature dedicated to the estimation of ϕ_{ij} by measuring the local smoothness of the approximate solution; see, for example, Harten and

Osher [22], Harten et al. [23], Sweby [37], Colella and Glaz [10], van Leer [39], and the references therein. We also refer the reader to Zhang and Shu [42] and Persson and Peraire [34], where this idea is exploited in the context of discontinuous Galerkin techniques.

In sections 4.3 and 4.4 we propose two definitions for the limiting function ψ_i^n that preserve the maximum principle and that we conjecture yield convergent methods, at least when $\boldsymbol{f}\cdot\boldsymbol{n}$ is convex or concave for all unit vectors $\boldsymbol{n}$; see sections 4.3 and 4.4. We recall the basic principles of the FCT technique in section 4.5 and show that the method can be recast into the form (23) with a viscosity defined like in (22).

4.2. Preliminary lemma. We start with a lemma which will be useful in designing various limiting functions. Given $i \in \{1{:}I\}$ and $n \geq 0$, we start by introducing the maximum and the minimum of U^n in the support of the shape function φ_i:

$$\mathsf{U}_i^{\mathrm{M},n} := \max_{j\in\mathcal{I}(S_i)} \mathsf{U}_j^n, \qquad \mathsf{U}_i^{\mathrm{m},n} := \min_{j\in\mathcal{I}(S_i)} \mathsf{U}_j^n. \tag{25}$$

Our objective is to tune ψ_i^n so that $\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. We introduce the parameter

$$\theta_i^n := \begin{cases} \frac{\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}} & \text{if } \mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0, \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{26}$$

We define $\mathcal{I}(S_i^+) = \{j \in \mathcal{I}(S_i) \mid \mathsf{U}_i^n < \mathsf{U}_j^n\}$ and $\mathcal{I}(S_i^-) = \{j \in \mathcal{I}(S_i) \mid \mathsf{U}_j^n < \mathsf{U}_i^n\}$, and we decompose the local viscous CFL number, $\gamma_i^n := -\frac{2\tau d_{ii}^{\mathrm{V},n}}{m_i}$, as follows:

$$\gamma_i^{+,n} := \frac{2\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^+)} d_{ij}^{\mathrm{V},n}, \qquad \gamma_i^{-,n} := \frac{2\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^-)} d_{ij}^{\mathrm{V},n}. \tag{27}$$

The following lemma is the key result of this section.

Lemma 4.1. *Assume that* $\gamma_i^n < 1$ *and* $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$, $n \geq 0$, $i \in \{1{:}I\}$*; then*

$$\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\mathrm{M},n} - (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{-,n}\right), \tag{28}$$

$$\mathsf{U}_i^{n+1} \geq \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\left(\theta_i^n(1-\gamma_i^n) - (1-\theta_i^n)(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{+,n}\right). \tag{29}$$

Proof. We obtain $\mathsf{U}_i^{n+1} = \mathsf{U}_i^{\mathrm{V},n+1} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(S_i)}(d_{ij}^n - d_{ij}^{\mathrm{V},n})(\mathsf{U}_j^n - \mathsf{U}_i^n)$ by subtracting (6) from (23). Upon setting $\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_j^n + \mathsf{U}_i^n) - (\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{V},n}}$, the identity (14) gives $\mathsf{U}_i^{\mathrm{V},n+1} = \mathsf{U}_i^n(1 - \sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^{\mathrm{V},n}}{m_i}) + \sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^{\mathrm{V},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$. An important property of the auxiliary states is that $\mathsf{U}_i^{\mathrm{m},n} \leq \overline{\mathsf{U}}_{ij}^n \leq \mathsf{U}_i^{\mathrm{M},n}$ (see (17)). Owing to the definition of γ_i^n and $d_{ii}^{\mathrm{V},n}$ we have $\gamma_i^n := \sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^{\mathrm{V},n}}{m_i}$, which implies that the quantity $\mathsf{U}_i^{*,n} := \frac{1}{\gamma_i^n}\sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^{\mathrm{V},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$ is a convex combination of $\{\overline{\mathsf{U}}_{ij}^n\}_{i\in\mathcal{I}(S_i)}$, that is to say $\mathsf{U}_i^{\mathrm{m},n} \leq \mathsf{U}_{ij}^{*,n} \leq \mathsf{U}_i^{\mathrm{M},n}$. Then $\mathsf{U}_i^{\mathrm{V},n+1}$ can be rewritten as follows: $\mathsf{U}_i^{\mathrm{V},n+1} = \mathsf{U}_i^n(1-\gamma_i^n) + \gamma_i^n\mathsf{U}_i^{*,n}$. Using that $\mathsf{U}_{ij}^{*,n} \leq \mathsf{U}_i^{\mathrm{M},n}$, we infer that

$$\begin{aligned} \mathsf{U}_i^{n+1} &= \mathsf{U}_i^n(1-\gamma_i^n) + \gamma_i^n\mathsf{U}_i^{*,n} + \frac{\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(S_i)}(d_{ij}^n - d_{ij}^{\mathrm{V},n})(\mathsf{U}_j^n - \mathsf{U}_i^n) \\ &\leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(S_i)}(d_{ij}^n - d_{ij}^{\mathrm{V},n})(\mathsf{U}_j^n - \mathsf{U}_i^n). \end{aligned}$$

Then using that $d_{ij}^n \le d_{ij}^{\mathrm{V},n}$ by definition, the above inequality gives

$$\begin{aligned}\mathsf{U}_i^{n+1} &\le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^-)} (d_{ij}^{\mathrm{V},n} - d_{ij}^n)(\mathsf{U}_i^n - \mathsf{U}_j^n)\\ &\le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^-)} (d_{ij}^{\mathrm{V},n} - d_{ij}^n)(\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}).\end{aligned}$$

Now using that $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$ and, by definition of $\mathsf{U}_i^{\mathrm{M},n}$ and $\mathsf{U}_i^{\mathrm{m},n}$, U_i^n is in the convex hull of $\mathsf{U}_i^{\mathrm{M},n}$ and $\mathsf{U}_i^{\mathrm{m},n}$, we have $\mathsf{U}_i^n = \theta_i^n \mathsf{U}_i^{\mathrm{M},n} + (1-\theta_i^n)\mathsf{U}_i^{\mathrm{m},n}$, where $\theta_i^n \in [0,1]$ is as defined in (26). Hence $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n} = -\theta_i^n(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$ and $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n} = (1-\theta_i^n)(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$. The above inequality can then be rewritten as

$$\mathsf{U}_i^{n+1} \le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^-)} (d_{ij}^{\mathrm{V},n} - d_{ij}^n)\right).$$

Using that $d_{ij}^n \ge d_{ij}^{\mathrm{V},n}\psi_i^n$ and $\psi_i^n \ge 0$, we infer that $-d_{ij}^n \le -d_{ij}^{\mathrm{V},n}\psi_i^n$, which in turn implies the following inequalities:

$$\begin{aligned}\mathsf{U}_i^{n+1} &\le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(S_i^-)} d_{ij}^{\mathrm{V},n}\right)\\ &\le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{-,n}\right).\end{aligned}$$

The other estimate is obtained similarly; i.e., using that $\mathsf{U}_{ij}^{*,n} \ge \mathsf{U}_i^{\mathrm{m},n}$, we infer that

$$\begin{aligned}\mathsf{U}_i^{n+1} &\ge \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(S_i^+)} (d_{ij}^n - d_{ij}^{\mathrm{V},n})(\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\\ &\ge \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n})\left(\theta_i^n(1-\gamma_i^n) - (1-\psi_i^n)(1-\theta_i^n)\tfrac{1}{2}\gamma_i^{+,n}\right),\end{aligned}$$

which completes the proof. □

4.3. Smoothness-based viscosity. We start by introducing a limiting technique based on a measure of the local smoothness of the solution in the spirit of the finite volume literature (see, e.g., Jameson, Schmidt, and Turkel [28, eq. (12)] and Jameson [27, p. 6]). Assuming that $\mathsf{U}_i^{\mathrm{m},n} \neq \mathsf{U}_i^{\mathrm{M},n}$, we introduce the quantity

$$\alpha_i^n := \frac{\left|\sum_{j\in\mathcal{I}(S_i)} \beta_{ij}(\mathsf{U}_j^n - \mathsf{U}_i^n)\right|}{\sum_{j\in\mathcal{I}(S_i)} \beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|}, \tag{30}$$

where the real numbers β_{ij} are assumed to be positive. One can use the parameters β_{ij} to make $\alpha_i^n = 0$ if $\boldsymbol{u}_h^n$ is linear on the support of the shape function φ_i, which then makes the method linearity preserving (see Berger, Aftosmis, and Murman [2] for a review on linearity preserving limiters in the finite volume literature). Note that $\alpha_i^n \in [0,1]$ for all $i \in \{1{:}I\}$ and all $n \ge 0$. The first motivation for the above definition is that α_i^n is equal to 1 if U_i^n is a local extrema, and in this case one wants to set $d_{ij}^n = d_{ij}^{\mathrm{V},n}$, i.e., $\psi_i^n = 1$; hence setting $\psi_i^n = \alpha_i^n$ could be a good idea. The second motivation is that if the coefficients β_{ij} are defined so that $\alpha_i^n = 0$ if u_h^n is linear on S_i, then the numerator of (30) behaves like $h^2\|D^2u(\boldsymbol{\xi},t^n)\|_{\ell^2(\mathbb{R}^{d\times d})}$ at some point $\boldsymbol{\xi}$, whereas the denominator behaves like $h\|\nabla u(\boldsymbol{\zeta},t^n)\|_{\ell^2(\mathbb{R}^d)}$ at some point $\boldsymbol{\zeta}$. In these conditions $\alpha_i^n \approx h\|D^2u(\boldsymbol{\xi},t^n)\|_{\ell^2(\mathbb{R}^{d\times d})}/\|\nabla u(\boldsymbol{\zeta},t^n)\|_{\ell^2(\mathbb{R}^d)}$, that is to say α_i^n is of order h in smooth regions away from extrema, which makes the method formally second-order consistent.

Let us denote $\beta_i^m := \min_{j\in\mathcal{I}(S_i)}\beta_{ij}$ and $\beta_i^M := \max_{j\in\mathcal{I}(S_i)}\beta_{ij}$, and suppose that there exists $\beta^\sharp < \infty$, uniform with respect to $(\mathcal{T}_h)_{h>0}$, such that

$$0 < \beta_{ij} \quad \forall i \in \{1{:}I\},\ \forall j \in \mathcal{I}(S_i), \qquad \max_{i\in\{1:I\}} \frac{\beta_i^M}{\beta_i^m} < \beta^\sharp. \tag{31}$$

THEOREM 4.2. *Let $\psi \in C^{0,1}([0,1];[0,1])$ be any positive function such that $\psi(1) = 1$. Let k_ψ be the Lipschitz constant of ψ. The scheme* (23) *with the diffusion matrix defined in* (22)*, with the assumptions* (31)*, and the definition*

$$\psi_i^n := \psi(\alpha_i^n), \tag{32}$$

is locally maximum principle preserving under the local CFL condition $\gamma_i^n \le \frac{1}{1+k_\psi c_\sharp}$, where $c_\sharp = \beta^\sharp \max_{i\in\{1:I\}} \operatorname{card}(\mathcal{I}(S_i))$ (this number is uniformly bounded with respect to the mesh sequence).

Proof. Note first that if $\mathsf{U}_i^{\mathrm{M},n} = \mathsf{U}_i^{\mathrm{m},n}$, then $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ irrespective of the value of d_{ij}^n, which proves the statement. Let us assume now that $\mathsf{U}_i^{\mathrm{M},n} \ne \mathsf{U}_i^{\mathrm{m},n}$. If $\theta_i^n \in \{0,1\}$, where we recall that $\theta_i^n := \frac{\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}}$, then $\alpha_i^n = 1$, which implies that $\psi(\alpha_i^n) = 1$; as a result, $d_{ij}^n = d_{ij}^{\mathrm{V},n}\max(1,\psi(\alpha_j)) = d_{ij}^{\mathrm{V},n}$, which again implies that $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. Finally, let us assume that $0 < \theta_i^n < 1$. Observing that $||y| - |x|| = \max(-|x| + |y|, |x| - |y|)$, we infer that $-||y| - |x|| \le |y| - |x|$ for all $x, y \in \mathbb{R}$. This inequality in turn implies that

$$\begin{aligned}
1 - \alpha_i^n &= 1 - \frac{\left|\sum_{j\in\mathcal{I}(S_i^+)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n| - \sum_{j\in\mathcal{I}(S_i^-)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|\right|}{\sum_{j\in\mathcal{I}(S_i)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|} \\
&\le \frac{\sum_{j\in\mathcal{I}(S_i)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n| + \sum_{j\in\mathcal{I}(S_i^+)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n| - \sum_{j\in\mathcal{I}(S_i^-)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|}{\sum_{j\in\mathcal{I}(S_i)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|} \\
&\le 2\frac{\sum_{j\in\mathcal{I}(S_i^+)}\beta_{ij}(\mathsf{U}_j^n - \mathsf{U}_i^n)}{\sum_{j\in\mathcal{I}(S_i)}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|} \le 2\frac{\sum_{j\in\mathcal{I}(S_i^+)}\beta_{ij}(\mathsf{U}_j^{\mathrm{M},n} - \mathsf{U}_i^n)}{\beta_i^m|\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n| + \beta_i^m|\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^n|} \\
&\le 2\frac{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}}\frac{\beta_i^M}{\beta_i^m}\operatorname{card}(\mathcal{I}(S_i^+)) \le 2c_\sharp(1 - \theta_i^n),
\end{aligned}$$

where $c_\sharp = \beta^\sharp \max_{i\in\{1:I\}} \operatorname{card}(\mathcal{I}(S_i))$ is a number uniformly bounded with respect to the mesh sequence. Likewise we have $1 - \alpha_i^n \le 2c_\sharp\theta_i^n$. Let k_ψ be the Lipschitz constant of ψ; then $1 - \psi(\alpha_i^n) = \psi(1) - \psi(\alpha_i^n) \le k_\psi(1 - \alpha_i^n)$. This in turn implies that

$$\begin{aligned}
(1 - \theta_i^n)(1 - \gamma_i^n) - \theta_i^n(1 - \psi(\alpha_i^n))\tfrac{1}{2}\gamma_i^{-,n} &\ge (1 - \theta_i^n)(1 - \gamma_i^n) - k_\psi c_\sharp \theta_i^n(1 - \theta_i^n)\gamma_i^n \\
&\ge (1 - \theta_i^n)(1 - (1 + k_\psi c_\sharp \theta_i^n)\gamma_i^n) \ge 0,
\end{aligned}$$

provided $\gamma_i^n \le \frac{1}{1+k_\psi c_\sharp}$. Similarly we have

$$\begin{aligned}
\theta_i^n(1 - \gamma_i^n) - (1 - \theta_i^n)(1 - \psi(\alpha_i^n))\tfrac{1}{2}\gamma_i^{+,n} &\ge \theta_i^n(1 - \gamma_i^n) - k_\psi c_\sharp \theta_i^n(1 - \theta_i^n)\gamma_i^n \\
&\ge \theta_i^n(1 - (1 + k_\psi c_\sharp(1 - \theta_i^n))\gamma_i^n) \ge 0,
\end{aligned}$$

provided again that $\gamma_i^n \le \frac{1}{1+k_\psi c_\sharp}$. The conclusion follows by invoking Lemma 4.1. □

We finish this section by discussing possible constructions of the coefficients β_{ij}, although a detailed analysis of this construction is irrelevant for the rest of the paper. Actually, all the tests reported in section 6 are done with $\beta_{ij} = \frac{1}{\text{card}(\mathcal{I}(S_i))-1}$ for all i, j.

Let us start by considering piecewise linear Lagrange finite elements on a one-dimensional nonuniform grid. Consider two consecutive cells $[x_{i-1}, x_i]$, $[x_i, x_{i+1}]$, and let $h_i = x_i - x_{i-1}$, $h_{i+1} = x_{i+1} - x_i$. If u_h is linear over $[x_{i-1}, x_{i+1}]$, then $\mathsf{U}_{i+1}\frac{x_i - x_{i-1}}{x_{i+1}-x_{i-1}} + \mathsf{U}_{i-1}\frac{x_i - x_{i+1}}{x_{i-1}-x_{i+1}} - \mathsf{U}_i$ should be equal to zero. This quantity can also be rewritten: $(\mathsf{U}_{i+1} - \mathsf{U}_i)\frac{h_i}{h_i+h_{i+1}} + (\mathsf{U}_{i-1} - \mathsf{U}_i)\frac{h_{i+1}}{h_i+h_{i+1}}$. Hence, in one dimension it is natural to take $\beta_{i,i-1} = \frac{h_i}{h_i+h_{i+1}}$ and $\beta_{i,i+1} = \frac{h_{i+1}}{h_i+h_{i+1}}$. The above argument generalizes in higher space dimension if one assumes that the support of φ_i is convex. In that case it is possible to find generalized barycentric coordinates, $(\omega_{ij})_{i\neq j\in\mathcal{I}(S_i)}$, so that at any point $\boldsymbol{x}$ in $\text{conv}\{\boldsymbol{a}_j \mid i \neq j \in \mathcal{I}(S_i)\}$ one has $\boldsymbol{x} = \sum_{i\neq j\in\mathcal{I}(S_i)} \omega_{ij}(\boldsymbol{x})\boldsymbol{a}_j$, with $\sum_{i\neq j\in\mathcal{I}(S_i)} \omega_{ij}(\boldsymbol{x}) = 1$, $\omega_{ij(\boldsymbol{x})\geq 0}$. For instance, one can use the so-called Wachspress coordinates in two dimensions. In higher dimensions one can use the technique described in Warren et al. [40]. We refer the reader to Floater [12] for a review on generalized barycentric coordinates. If u_h is linear over S_i, then $u_h(\boldsymbol{a}_i)$ is equal to $\sum_{i\neq j\in\mathcal{I}(S_i)} \omega_{ij}(\boldsymbol{a}_i)u_h(\boldsymbol{a}_j)$; hence the quantity $\sum_{i\neq j\in\mathcal{I}(S_i)} \omega_{ij}(\boldsymbol{a}_i)(\mathsf{U}_j - \mathsf{U}_i)$ should be zero. This argument shows that in this case it is natural to take $\beta_{ij} = \omega_{ij}(\boldsymbol{a}_i)$.

Remark 4.3 (literature). During the preparation of this paper we were made aware by G. Barrenechea of a technique proposed in Burman [6, Thm. 4.1] and Barrenechea, Burman, and Karakatsani [1, eqs. (2.4)–(2.5)] that resembles the one proposed here. The quantity α_i^p is used in [6] to construct a nonlinear viscosity that yields the maximum principle and convergence to the entropy solution for the one-dimensional Burgers equation. It is used in [1] for solving scalar-valued linear convection-diffusion equations.

4.4. Greedy viscosity. We continue with a limiting technique entirely based on the observations made in Lemma 4.1 irrespective of any smoothness considerations.

THEOREM 4.4. *Let* θ_n^n, $\gamma_i^{-,n}$, *and* $\gamma_i^{+,n}$ *be the quantities defined in* (26)–(27) *for all* $i \in \{1{:}I\}$. *The scheme* (23) *with the diffusion matrix defined in* (22) *with*

$$\psi_i^n := \max\left(1 - 2(1-\gamma_i^n)\min\left(\frac{1}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{(1-\theta_i^n)}\right), 0\right) \tag{33}$$

and the convention $\psi_i^n = 1$ *if* $\theta_i^n \in \{0, 1\}$ *is locally maximum principle preserving under the same CFL condition as in Theorem* 3.5, *i.e.*, $1 \geq \gamma_i^n$.

Proof. Note first that if $\mathsf{U}_i^{\text{M},n} = \mathsf{U}_i^{\text{m},n}$, then $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\text{m},n}, \mathsf{U}_i^{\text{M},n}]$ irrespective of the value of d_{ij}^n, which proves the statement. If $\theta_i^n \in \{0, 1\}$, then $\psi_i^n = 1$ implies that $d_{ij}^n = d_{ij}^{\text{V},n}\max(1, \psi_j^n) = d_{ij}^{\text{V},n}$, which again implies that $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\text{m},n}, \mathsf{U}_i^{\text{M},n}]$. Finally, let us assume that $0 < \theta_i^n < 1$. The definition of ψ_i^n in (33) implies that $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}$, which in turn gives $\theta_i^n(\psi_i^n - 1)\frac{1}{2}\gamma_i^{-,n} + (1-\gamma_i^n)(1-\theta_i^n) \geq 0$. This is the condition in Lemma 4.1 that shows that $\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\text{M},n}$; see (28). Similarly, we have $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{+,n}}\frac{\theta_i^n}{1-\theta_i^n}$, which gives $(\psi_i^n - 1)(1-\theta_i^n)\frac{1}{2}\gamma_i^{+,n} + (1-\gamma_i^n)\theta_i^n \geq 0$. This is the condition in Lemma 4.1 that shows that $\mathsf{U}_i^{\text{m},n} \leq \mathsf{U}_i^{n+1}$; see (29). ◻

Remark 4.5 (small CFL number). Note that when the local CFL number γ_i^n is small and U_i^n is not a local extremum, the definition (33) implies that ψ_i^n is almost equal to 1 irrespective of the smoothness of the solution. More precisely, assume that we work in one dimension on a uniform mesh and the solution is locally linear. Then $\theta_i = \frac{1}{2}$, irrespective of the local slope of the solution. As a result, $\min(\frac{1}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{(1-\theta_i^n)}) \geq \frac{1}{\gamma_i^n}$ and $\psi_i^n \leq \max(1-2\frac{1-\gamma_i^n}{\gamma_i^n}, 0)$; hence $\psi_i^n = 0$ if $\gamma_i^n \leq \frac{2}{3}$. This shows that the smaller the CFL, the greedier the method.

4.5. Flux corrected transport (FCT) viscosity. We finish by showing that the FCT limitation technique of Zalesak [41] (see also Boris and Book [3]) can be used to construct a second-order viscosity as described in section 4.1. We first recall the basic principle of the FCT method for completeness and refer the reader to Kuzmin and Turek [31] for a comprehensive review of FCT-related methods.

The FCT theory is algebraic. Let $\mathsf{U}_L^{n+1} \in \mathbb{R}^I$ and $\mathsf{U}_H^{n+1} \in \mathbb{R}^I$ be two coordinate vectors which we refer to as low- and high-order solutions, respectively. We assume that the low-order solution satisfies some type of minimum and maximum principle; say there are two vectors $\mathsf{U}^{\min} \in \mathbb{R}^I$ and $\mathsf{U}^{\max} \in \mathbb{R}^I$ such that $\mathsf{U}_i^{\min} \leq \mathsf{U}_{L,i}^{n+1} \leq \mathsf{U}_i^{\max}$ for all $i \in \{1{:}I\}$. We furthermore assume that U_L^{n+1} and U_H^{n+1} are related as follows:

$$\frac{m_i}{\tau}\mathsf{U}_{H,i}^{n+1} = \frac{m_i}{\tau}\mathsf{U}_{L,i}^{n+1} + \sum_{j\in\mathcal{I}(S_i)} a_{ij}^n, \tag{34}$$

where the coefficients a_{ij}^n are skew-symmetric: $a_{ij}^n = -a_{ji}^n$ for all $1 \leq i,j \leq I$. Note that this property implies conservation, i.e., $\sum_{i\in\{1:I\}} m_i\mathsf{U}_{H,i}^{n+1} = \sum_{i\in\{1:I\}} m_i\mathsf{U}_{L,i}^{n+1}$. Since it is not guaranteed that U_H^{n+1} satisfies the minimum and the maximum principle, the FCT strategy consists of modifying the high-order solution as follows:

$$\frac{m_i}{\tau}\mathsf{U}_{i}^{n+1} = \frac{m_i}{\tau}\mathsf{U}_{L,i}^{n+1} + \sum_{j\in\mathcal{I}(S_i)} \mathcal{L}_{ij}^n a_{ij}^n, \tag{35}$$

where, inspired by Zalesak [41] (see equations (10)–(13) in [41]), the coefficients $\mathcal{L}_{ij}^n$ are computed as follows. First, compute P_i^+, P_i^-, Q_i^+, Q_i^-, R_i^+, and R_i^-, $1 \leq i \leq N$:

$$P_i^+ := \sum_{j\in\mathcal{S}(i)} \max\{0, a_{ij}^n\}, \qquad P_i^- := \sum_{j\in\mathcal{S}(i)} \min\{0, a_{ij}^n\}, \tag{36}$$

$$Q_i^+ := \frac{m_i}{\tau}(\mathsf{U}_i^{\max} - \mathsf{U}_{L,i}^{n+1}), \qquad Q_i^- := \frac{m_i}{\tau}(\mathsf{U}_i^{\min} - \mathsf{U}_{L,i}^{n+1}), \tag{37}$$

$$R_i^+ := \begin{cases} \min\{1, \frac{Q_i^+}{P_i^+}\} & \text{if } P_i^+ \neq 0,\\ 1 & \text{otherwise,}\end{cases} \qquad R_i^- := \begin{cases} \min\{1, \frac{Q_i^-}{P_i^-}\} & \text{if } P_i^- \neq 0,\\ 1 & \text{otherwise.}\end{cases} \tag{38}$$

Then define the limiting coefficients $\mathcal{L}_{ij}^n$ by

$$\mathcal{L}_{ij}^n := \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } a_{ij} \geq 0,\\ \min\{R_i^-, R_j^+\} & \text{otherwise.}\end{cases} \tag{39}$$

Note that this definition implies that $\mathcal{L}_{ij}^n = \mathcal{L}_{ji}^n$, which is essential to guarantee conservation, i.e., $\sum_{i\in\{1:I\}} m_i\mathsf{U}_{i}^{n+1} = \sum_{i\in\{1:I\}} m_i\mathsf{U}_{L,i}^{n+1}$. Note also that $\mathcal{L}_{ij}^n \in [0,1]$. Let us mention in passing that, upon setting $\phi_i^n := \min\{R_i^+, R_i^-\}$, an alternative definition of $\mathcal{L}_{ij}^n$ could be $\mathcal{L}_{ij}^n = \min(\phi_i^n, \phi_i^n)$. One can verify that this definition also yields the local maximum principle. It is mainly the definition (39), which is slightly greedier than the above alternative, that is used in the literature.

The FCT viscosity is defined as follows: (i) Let U_H^{n+1} be the Galerkin approximation; i.e., U_H^{n+1} is obtained by setting $d_{ij}^n = 0$ in (23). (ii) The low-order solution is defined by $\mathsf{U}_L^{n+1} := \mathsf{U}^{\mathrm{V},n+1}$. Then $\frac{m_i}{\tau}\mathsf{U}_{H,i}^{n+1} = \frac{m_i}{\tau}\mathsf{U}_{L,i}^{n+1} + \sum_{j\in\mathcal{I}(S_i)} d_{ij}^{\mathrm{V},n}(\mathsf{U}_j^n - \mathsf{U}_i^n)$. (iii) Compute $\mathcal{L}_{ij}^n$ with the above algorithm with $a_{ij}^n := -d_{ij}^{\mathrm{V},n}(\mathsf{U}_j^n - \mathsf{U}_i^n)$, $\mathsf{U}_i^{\min} = \min_{j\in\mathcal{I}(S_i)} \mathsf{U}_i^n$, and $\mathsf{U}_i^{\max} = \max_{j\in\mathcal{I}(S_i)} \mathsf{U}_i^n$ (this is legitimate since U_L^{n+1} satisfies the local maximum principle). After inserting the definition of $\mathsf{U}_{L,i}^{n+1}$ into (35), we obtain

$$(40)\quad m_i \frac{\mathsf{U}_i^{n+1} - \mathsf{U}_i^n}{\tau} = \sum_{j\in\mathcal{I}(S_i)} \boldsymbol{c}_{ij}\cdot(\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n)) + \sum_{j\in\mathcal{I}(S_i)} (1-\mathcal{L}_{ij}^n) d_{ij}^{\mathrm{V},n}(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

In conclusion, FCT limitation is equivalent to solving (6) or (23) with

$$(41)\qquad d_{ij}^n = d_{ij}^{\mathrm{V},n}(1-\mathcal{L}_{ij}^n).$$

The above definition of d_{ij}^n is similar to (22); here $1-\mathcal{L}_{ij}^n$ plays the role of $\max(\psi_i^n, \psi_j^n)$. The FCT viscosity is greedier than the techniques proposed in sections 4.3 and 4.4, but being too greedy may have disastrous consequences as shown in the following result.

LEMMA 4.6 (FCT counterexample). *There exist C^∞ fluxes and initial data such that, under the CFL condition $1+2\tau\frac{d_{ii}^{\mathrm{V},n}}{m_i} \geq 0$, the approximate sequence given by* (23) *with the FCT viscosity* (41) *does not converge to the entropy solution of* (1).

Proof. We use the same counterexample as in Lemma 3.2. We consider Burgers's equation in one space dimension, $\boldsymbol{f}(u) := f(u)\boldsymbol{e}_x$, $f(u) := \frac{1}{2}u^2$, $D := (-1,1)$, with initial data $u_0(x) := -1$ if $x \leq 0$ and $u_0(x) := 1$ otherwise. Let $N \in \mathbb{N}\backslash\{0\}$, and let us consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$ where the nodes x_i, $i \in \{0{:}2N\}$, are such that $-1 = x_0 < x_1 < \cdots < x_{2N} = 1$ and $x_N \leq 0 < x_{N+1}$. Let $P(\mathcal{T}_h)$ be the space composed of continuous piecewise linear functions on $\mathcal{T}_h$. We have $\boldsymbol{c}_{ii-1} = -\frac{1}{2}\boldsymbol{e}_x$, $\boldsymbol{c}_{ii} = \boldsymbol{0}$ and $\boldsymbol{c}_{ii+1} = \frac{1}{2}\boldsymbol{e}_x$, $m_i = \frac{h_i+h_{i+1}}{2}$. The equation for U_i^{n+1} is

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n + \frac{\tau}{2m_i}(f(\mathsf{U}_{i-1}^n) - f(\mathsf{U}_{i+1}^n)) + \frac{\tau}{m_i} d_{ii-1}^n(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{m_i} d_{ii+1}^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$

where we have defined by convention that $\mathsf{U}_{-1}^n := -1$ and $\mathsf{U}_{2N+1}^n := 1$. Let us consider the approximate initial data $u_h^0 = \sum_{i\in\{0:2N\}} \mathsf{U}_i^0\varphi_i(x)$ with $\mathsf{U}_i^0 = -1$ if $i \leq N$ and $\mathsf{U}_i^0 = 1$ if $i \geq N+1$. Let U_H^1 be the Galerkin solution at $t^1 := \tau$, which we recall is by definition obtained by solving the above equation with $d_{ij}^1 = 0$. Since $f(\mathsf{U}_{i-1}^0) - f(\mathsf{U}_{i+1}^0) = 0$ for all $i \in \{0{:}2N\}$, we then obtain $\mathsf{U}_H^1 = \mathsf{U}^0$. Let us now estimate the low-order solution U_L^1. It is clear that $\mathsf{U}_{L,i}^1 = \mathsf{U}_i^0$ for all $i < N$ and all $N+1 < i$. For $i = N$ we have $\mathsf{U}_{N-1}^0 = -1$, $\mathsf{U}_N^0 = -1$, $\mathsf{U}_{N+1}^0 = 1$, $f(\mathsf{U}_{N+1}^0) - f(\mathsf{U}_{N-1}^0) = 0$. Note that $d_{N+1N}^{v,0} = d_{NN+1}^{v,0} = \frac{1}{2}$ since the maximum wave speed in the Riemann problem with the data $(-1,1)$ is 1. Hence,

$$\mathsf{U}_{L,N}^1 = \mathsf{U}_N^0 + \frac{\tau}{m_N} d_{NN+1}(\mathsf{U}_{N+1}^0 - \mathsf{U}_N^0) = -1 + \frac{\tau}{m_N},$$
$$\mathsf{U}_{L,N+1}^1 = \mathsf{U}_{N+1}^0 + \frac{\tau}{m_{N+1}} d_{N+1N}(\mathsf{U}_N^0 - \mathsf{U}_{N+1}^0) = 1 - \frac{\tau}{m_{N+1}}.$$

In the FCT notation, we have

$$m_N\mathsf{U}_{H,N}^1 = m_N\mathsf{U}_{L,N}^1 - \frac{\tau}{2}(\mathsf{U}_{N+1}^0 - \mathsf{U}_N^0),$$
$$m_{N+1}\mathsf{U}_{H,N+1}^1 = m_{N+1}\mathsf{U}_{L,N+1}^1 - \frac{\tau}{2}(\mathsf{U}_N^0 - \mathsf{U}_{N+1}^0).$$

This means that $a_{N,N+1} = -\frac{\tau}{2}(\mathsf{U}^0_{N+1} - \mathsf{U}^0_N) = -\tau$ and $a_{N+1,N} = -\frac{\tau}{2}(\mathsf{U}^0_N - \mathsf{U}^0_{N+1}) = \tau$. Let us now compute the limiter coefficient $\mathcal{L}_{N,N+1}$ with $\mathsf{U}^{\max}_N = \mathsf{U}^{\max}_{N+1} = 1$ and $\mathsf{U}^{\min}_N = \mathsf{U}^{\min}_{N+1} = -1$. We now compute the FCT coefficients,

$$\begin{aligned}
&P^+_N = 0, && P^-_N = -\tau, && P^+_{N+1} = \tau, && P^-_{N+1} = 0,\\
&Q^+_N = 2m_N - \tau, && Q^-_N = m_N(-\tfrac{\tau}{m_N}), && Q^+_{N+1} = m_{N+1}(\tfrac{\tau}{m_{N+1}}), && Q^-_{N+1} = -2m_{N+1} + \tau,\\
&R^+_N = 1, && R^-_N = 1, && R^+_{N+1} = 1, && R^-_{N+1} = 1,
\end{aligned}$$

which gives $\mathcal{L}_{N,N+1} = 1$. Hence $\mathsf{U}^1 = \mathsf{U}^1_H$, which gives $\mathsf{U}^1 = \mathsf{U}^0$ since $\mathsf{U}^1_H = \mathsf{U}^0$. In conclusion, $u^1_h = u^0_h$; i.e., $u^n_h = u^0_h$ for every $n \geq 0$. This proves that the numerical solution is a stationary discontinuity, whereas the entropy solution of the problem is an expansion wave; hence the method does not converge to the entropy solution. □

5. Entropy-viscosity and dispersion correction.

We describe in this section two techniques that improve the entropy consistency and the convergence rates of the methods presented in section 4: the first technique is the entropy-viscosity, and the second consists of using the consistent mass matrix. Each of these techniques violates the maximum principle, but this problem is corrected by using the FCT technique to postprocess the solution.

5.1. Entropy-viscosity.

We start by recalling the entropy-viscosity technique introduced in [18] and further refined in [14, 19]. (It seems though that the terminology "entropy-viscosity" was actually first proposed in Harten and Hyman [21, Appendix A] in an effort to fix the deficiencies of Roe's average; see section 3.1. We were not aware of this terminology when we "introduced" it in [18].) The idea proposed in [18] consists of estimating the artificial viscosity by measuring entropy residuals. One of the deficiencies of the early attempt made in [18] is that the construction of the viscosity requires a measure of the local mesh size, which may not be available on anisotropic meshes. Moreover, the dissipation is introduced by invoking the weak form of the Laplacian operator $-\nabla{\cdot}(\nu\nabla\psi)$. But the maximum principle is not robust with respect to the shape of the cells when using the bilinear form $(\psi, \varphi) \longrightarrow \int_K \nabla\psi{\cdot}\nabla\varphi \,\mathrm{d}\boldsymbol{x}$; see, e.g., Burman and Ern [7] and [14, 19]. More specifically, the convex combination argument that is invoked to prove the maximum principle works only if $\int_D \nabla\varphi_i{\cdot}\nabla\varphi_j \,\mathrm{d}\boldsymbol{x} < 0$ for all pairs of shape functions, φ_i, φ_j, with common support of nonzero measure. This is the well-known acute angle condition assumption which a priori excludes many meshes, in particular in three space dimensions. Finally, the scaling of the entropy proposed in [14, 19] is not natural since it is global. We now propose reformulating the entropy viscosity method to address the above problems.

The key idea, as suggested in [14], is to get rid of the Laplacian-based stabilization, to use a graph Laplacian instead, and to construct a viscosity matrix $\{d^n_{ij}\}_{i,j\in\{1:I\}}$ based on the estimation on a nondimensional entropy residual. Let u^n_h be the approximate solution at t^n, $n \geq 0$. Let $u^{\mathrm{G},n}_h$ be the Galerkin solution obtained by

$$\int_D \left(\frac{u^{\mathrm{G},n}_h - u^n_h}{\tau} + \nabla{\cdot}(\boldsymbol{f}(u^n_h)) \right) v \,\mathrm{d}x = 0 \qquad \forall v \in P(\mathcal{T}_h). \tag{42}$$

Note that a linear system involving the consistent mass matrix has to be solved here. Let η be an entropy. Let us denote $\eta^{\max,n}_i := \max_{j\in\mathcal{I}(S_i)} |\eta(\mathsf{U}^n_j)|$ and $\eta^{\min,n}_i :=$

$\min_{j\in\mathcal{I}(S_i)} |\eta(\mathsf{U}_j^n)|$. Then we define the entropy residual by setting

$$R_i^n(u_h^n) := \frac{2}{\eta_i^{\max,n} - \eta_i^{\min,n}} \int_D \left(\frac{u_h^{\mathrm{G},n} - u_h^n}{\tau} + \boldsymbol{f}'(u_h^n)\cdot\nabla(u_h^n) \right) \eta'(u_h^n(\boldsymbol{x}))\varphi_i \,\mathrm{d}x. \tag{43}$$

The entropy-viscosity is defined by

$$d_{ij}^n := \min(d_{ij}^{\mathrm{L},n}, \max(|R_i^n(u_h^n)|, |R_j^n(u_h^n)|)), \tag{44}$$

where $d_{ij}^{\mathrm{L},n}$ is a maximum principle preserving low-order viscosity. For instance, one can set $d_{ij}^{\mathrm{L},n} = d_{ij}^{\mathrm{V},n}$, where $d_{ij}^{\mathrm{V},n}$ is defined in (18), or one can choose $d_{ij}^{\mathrm{L},n}$ to be any of the high-order viscosities defined in section 4. Observe that d_{ij}^n can be put into the form (22) by setting $\psi_i^n = \min(1, \frac{|R_i^n|}{d_{ij}^{\mathrm{L},n}})$. The viscosity thus defined does not make the scheme maximum principle preserving in general, but ample numerical evidence shows that it is high-order and the viscosity is large enough to select the entropy solution if the entropy is nonlinear enough. Note in particular that $R_i^n(u_h^n) = 0$ and $d_{ij}^n = 0$ if $\eta(v) = v$; hence the entropy must be nonlinear. Our experience is that the method works well with entropies like $\eta(v) = |v|^p$ for any $p \in \mathbb{N}$, $p \geq 2$, or e^v. Note also that, contrary to the over avatars of the entropy viscosity method presented in our previous works, [18, 19], the definition of the residual (43) is invariant by any scaling and translation on the entropy; that is, replacing $\eta(v)$ by $\lambda_i\eta(v) + \mu_i$ for any $\lambda_i \in \mathbb{R}\backslash\{0\}$, $\mu_i \in \mathbb{R}$ does not change $R_i^n(u_h^n)$.

To summarize, the entropy viscosity solution u_h^{n+1} is obtained by solving

$$(\mathcal{M}\mathsf{U}^{n+1})_i = \sum_{j\in\mathcal{I}(S_i)} m_{ij}\mathsf{U}_j^n - \tau\left((\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\boldsymbol{c}_{ij} - d_{ij}^n(\mathsf{U}_j^n - \mathsf{U}_i^n)\right), \tag{45}$$

where $\mathcal{M}$ is the consistent mass matrix and d_{ij}^n is computed as in (42)–(43)–(44). The rationale for using the consistent mass matrix is explained in section 5.2. Let us finally mention again that there is no guarantee that the above algorithm is maximum principle preserving. This problem is addressed in section 5.3.

5.2. Dispersion correction.

5.2. Dispersion correction. Note that all the methods listed in sections 3 and 4 assume that the mass matrix is lumped. This is an important deficiency, at least for piecewise linear approximation, since it is well known that lumping the mass matrix induces dispersion errors that have adverse effects when solving transport-like equations with nonsmooth initial data. For instance, it is shown in Christon, Martinez, and Voth [8], Gresho, Sani, and Engelman [13], Guermond and Pasquetti [15], and Thompson [38] that the consistent mass matrix automatically corrects the dispersion error (at least for piecewise linear approximation). The beneficial effects of the consistent mass matrix are particularly visible when solving problems with nonsmooth solutions; see, e.g., [15, Fig. 5.5].

Unfortunately, the price to pay to eliminate dispersion errors is violations of the maximum principle. For instance, it is proved in [20] that the continuous finite element method based on artificial viscosity in space and explicit time stepping cannot satisfy the maximum principle when using the consistent mass matrix. More precisely, setting $n = 0$ in (45), it is possible to construct an initial datum u_h^0 such that for any viscosity distribution d_{ij}^0 it is impossible for the update (45) to be maximum principle preserving. In conclusion, to benefit fully from the antidispersive properties of the mass matrix, some limitation must be applied. This question is addressed in section 5.3.

5.3. FCT postprocessing. We finish this section by explaining how FCT is used to correct the violations of the maximum principle induced by the use of the entropy viscosity introduced in section 5.1 or the use of the consistent mass matrix motivated in section 5.2. The technique described in the rest of this section is henceforth referred to as "FCT postprocessing" to distinguish it from the technique described in section 4.5. We now essentially paraphrase Kuzmin, Löhner, and Turek [32, section 6.1] and do not claim originality here.

Let $d_{ij}^{\mathrm{H},n}$ be the entropy viscosity defined in (44) and $u_{H,h}^{n+1} = \sum_{i\in\mathcal{I}(S_i)} \mathsf{U}_{H,i}^{n+1}\varphi_i$ be the solution obtained by solving (45). Let $d_{ij}^{\mathrm{L},n}$ be one of the maximum principle satisfying viscosities defined in sections 3 and 4, and let $u_{L,h}^{n+1} = \sum_{i\in\mathcal{I}(S_i)} \mathsf{U}_{L,i}^{n+1}\varphi_i$ be the solution obtained by solving (6). By construction $u_{L,h}^{n+1}$ satisfies the local maximum principle, and in matrix form we have

$$\mathcal{M}\frac{\mathsf{U}_H^{n+1}-\mathsf{U}^n}{\tau} - d^{\mathrm{H},n}\mathsf{U}_H^n = \mathcal{M}^L\frac{\mathsf{U}_L^{n+1}-\mathsf{U}^n}{\tau} - d^{\mathrm{L},n}\mathsf{U}^n. \tag{46}$$

This identity is rewritten in the following form better suited for FCT postprocessing:

$$\mathcal{M}^L\frac{\mathsf{U}_H^{n+1}-\mathsf{U}_L^{n+1}}{\tau} = (\mathcal{M}^L-\mathcal{M})\frac{\mathsf{U}_H^{n+1}-\mathsf{U}^n}{\tau} + (d^{\mathrm{H},n}-d^{\mathrm{L},n})\mathsf{U}^n. \tag{47}$$

Since by definition $\sum_{j\in\mathcal{I}(S_i)}(m_i\delta_{ij}-m_{ij}) = 0$ and $\sum_{j\in\mathcal{I}(S_i)} d_{ij}^{\mathrm{H},n} = 0 = \sum_{j\in\mathcal{I}(S_i)} d_{ij}^{\mathrm{L},n}$, the above identity can be rewritten as $\frac{m_i}{\tau}(\mathsf{U}_{H,i}^{n+1}-\mathsf{U}_{L,i}^{n+1}) = \sum_{j\in\mathcal{I}(S_i)} a_{ij}^n$ with

$$a_{ij}^n := \tfrac{(m_i\delta_{ij}-m_{ij})}{\tau}\big(\mathsf{U}_{H,j}^{n+1}-\mathsf{U}_j^n-(\mathsf{U}_{H,i}^{n+1}-\mathsf{U}_i^n)\big) - (d_{ij}^{\mathrm{H},n}-d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n-\mathsf{U}_i^n). \tag{48}$$

This is exactly the structure that is needed to apply FCT as described in section 4.5.

In conclusion, the technique that we henceforth refer to as "entropy-viscosity with FCT postprocessing" is the solution obtained from the identity

$$\frac{m_i}{\tau}\mathsf{U}_i^{n+1} = \frac{m_i}{\tau}\mathsf{U}_{L,i}^{n+1} + \sum_{j\in\mathcal{I}(S_i)}\mathcal{L}_{ij}^n a_{ij}^n, \tag{49}$$

where the coefficients $\mathcal{L}_{ij}^n$ are obtained by using the FCT algorithm (36)–(39), $d_{ij}^{\mathrm{H},n}$ is the entropy viscosity defined in (44), $d_{ij}^{\mathrm{L},n}$ is one of the maximum principle satisfying viscosities defined in sections 3 and 4, $u_{H,h}^{n+1}$ is obtained by solving (45), and $u_{L,h}^{n+1}$ is obtained by solving (6).

6. Numerical tests. We illustrate the performances of the techniques introduced in sections 3, 4, and 5. All the tests reported in this section use continuous $\mathbb{P}_1$ finite elements on unstructured triangulations. The time stepping is done with the SSP RK(3,3) method (strong stability preserving Runge–Kutta, three stages, third-order); see Shu and Osher [36, eq. (2.18)]. All the L^p-norms of the errors reported in this section are relative to the L^p-norm of the exact solution, $p\in\{1,2,\infty\}$. The L^p-norms are estimated by high-order quadratures.

6.1. Illustration of Lemma 3.2. Although the LED method is not really the topic of the paper, since it is a first-order method, we start by illustrating the negative result of Lemma 3.2. Recall that the proof of the lemma is one-dimensional and based on the existence of a stationary sonic point (a state v such that $\|\boldsymbol{f}'(v)\|_{\ell^2} = 0$ is called a sonic point). We now show numerically that the LED technique also does not converge in two dimensions on nonuniform grids even when the expansion waves are not symmetric with respect to the sonic points. We consider the domain $D = (-0.25, 1.75)^2$ and Burgers's equation

$$\partial_t u + \nabla\cdot(\boldsymbol{f}(u)) = 0, \qquad \boldsymbol{f}(u) = \tfrac{1}{2}(u^2, u^2)^\mathsf{T}, \qquad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \text{ a.e. } \boldsymbol{x} \in D, \tag{50}$$

where

$$u_0(\boldsymbol{x}) = \begin{cases} 1 & \text{if } |x_1 - \frac{1}{2}| \leq 1 \text{ and } |x_2 - \frac{1}{2}| \leq 1, \\ -a & \text{otherwise.} \end{cases} \tag{51}$$

The tests are done with $a = 0.75$. The exact solution is computed as follows. With the convention $\boldsymbol{x} = (x_1, x_2)^\mathsf{T}$, assume first that $x_2 \leq x_1$. Then set $z_1 := x_1 - \frac{1}{2}$, $z_2 := x_2 - \frac{1}{2}$, and $\alpha := z_1 - z_2$. There are three cases depending on the value of α:

$$\text{if } \alpha \leq 1 - \tfrac{t(1+a)}{2}, \quad u(\boldsymbol{x}, t) = \begin{cases} \frac{z_2}{t} & \text{if } -at \leq z_2 < t, \\ 1 & \text{if } t \leq z_2 < 1 - a + (1-a)\frac{t}{2}, \\ -a & \text{otherwise;} \end{cases} \tag{52}$$

$$\text{if } 1 - \tfrac{t(1+a)}{2} < \alpha \leq 1, \quad u(\boldsymbol{x}, t) = \begin{cases} \frac{z_2}{t} & \text{if } -at \leq z_2 < \sqrt{2(1+a)t(1-\alpha)} - at, \\ -a & \text{otherwise;} \end{cases} \tag{53}$$

if $1 < \alpha$, then $u(\boldsymbol{x}, t) = -a$. Finally, we set $u((x_1, x_2), t) = u((x_2, x_1), t)$ if $x_1 \leq x_2$.

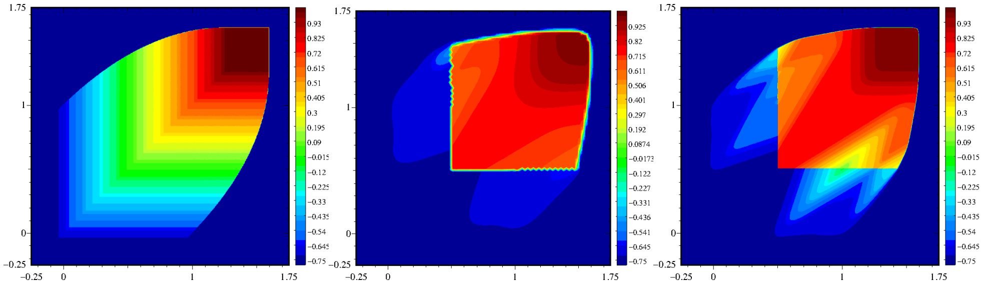

FIG. 1. *Burgers's equation* (50). *Left:* $\mathbb{P}_1$ *interpolant of the exact solution at* $t = 0.75$. *Center and right: Piecewise linear approximation of the solution using the first-order LED scheme described in section* 3.1 *with* 7543 *grid points (center) and* 474189 *grid points (right).*

We show in the left panel of Figure 1 the $\mathbb{P}_1$ interpolant of the exact solution at $t = 0.75$ on an unstructured triangulation composed of 474189 grid points. The center and right panels show the piecewise linear approximation of the solution at $t = 0.75$ using the LED viscosity described in section 3.1. The computation is done on two nonuniform triangulations composed of 7543 and 474189 grid points, respectively. Actually, the number of grid points used for this computation is irrelevant since the approximate solution changes very little as the mesh is refined and does not converge to the entropy solution. This example shows that the negative result stated in Lemma 3.2 in dimension 1 about the LED method actually holds in higher dimensions.

6.2. Linear transport, smooth solution. We test the convergence orders of the limiting techniques proposed in the paper on the linear transport equation

$\partial_t u + \nabla\cdot \boldsymbol{f}(u) = 0$ in the domain $D = (0,1)^2$ with the flux $\boldsymbol{f}(u) := \boldsymbol{\beta}(\boldsymbol{x},t)u$, where $\boldsymbol{\beta}(\cdot,t)$ is divergence free and

$$\boldsymbol{\beta}(\boldsymbol{x},t) := \begin{pmatrix} -2\sin(\pi x_2)\cos(\pi x_2)\sin^2(\pi x_1)\cos(\pi t) \\ 2\sin(\pi x_1)\cos(\pi x_1)\sin^2(\pi x_2)\cos(\pi t) \end{pmatrix}, \tag{54}$$

where we have set $\boldsymbol{x} := (x_1, x_2)$. This is a swirling deformation flow. We use the initial data $u_0(\boldsymbol{x}) = \sin(2\pi x_1)\sin(2\pi x_2)$. The motion is periodic in time with period 2; the solution is smooth with respect to time and space and returns to the initial data at every half period, i.e., $u(\boldsymbol{x}, k) = u_0(\boldsymbol{x})$ for every $k \in \mathbb{N}$.

6.2.1. High-order viscosities from section 4. We show in Table 1 the L^1-norm, L^2-norm, and L^∞-norm of the error at $t = 1$ for the three techniques described in section 4. The first one is the smoothness-based viscosity using $\psi(\alpha) = \alpha^2$ in (32). The second one is the greedy viscosity defined in (33). The third one is the FCT viscosity defined in (41). We use six nonnested, nonuniform meshes of approximate mesh size $\frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}, \frac{1}{320}, \frac{1}{640}$. More precisely, the total number of vertices, I, in each mesh is $507, 1927, 7545, 29870, 118851, 474186$.

TABLE 1
Swirling flow problem (54) *at* $t = 1$, $CFL = 0.25$, L^1*-norm (rows* 2–7*),* L^2*-norm (rows* 8–13*), and* L^∞*-norm of the error (rows* 14–19*). Viscosities are* $\psi(\alpha) = \alpha^2$ *(columns* 3–4*); greedy viscosity (columns* 5–6*); FCT viscosity (columns* 7–8*).*

Norm	I	$\psi(\alpha) = \alpha^2$ visc.		Greedy visc.		FCT visc.	
L^1	507	1.41E-01	Rate	6.71E-02	Rate	1.50E-02	Rate
	1927	4.32E-02	1.77	1.28E-02	2.48	3.71E-03	2.09
	7545	8.81E-03	2.33	2.16E-03	2.61	9.02E-04	2.07
	29870	1.65E-03	2.43	4.18E-04	2.39	2.18E-04	2.06
	118851	3.18E-04	2.50	8.69E-05	2.38	5.34E-05	2.13
	474186	6.58E-05	2.18	1.90E-05	2.11	1.31E-05	1.94
L^2	507	1.70E-01	–	8.97E-02	–	2.00E-02	–
	1927	5.67E-02	1.64	2.29E-02	2.04	4.94E-03	2.08
	7545	1.51E-02	1.94	4.82E-03	2.28	1.13E-03	2.16
	29870	4.02E-03	1.93	1.09E-03	2.16	2.60E-04	2.13
	118851	1.15E-03	1.89	2.52E-04	2.22	6.16E-05	2.18
	474186	3.64E-04	1.60	5.90E-05	2.01	1.46E-05	1.99
L^∞	507	2.71E-01	–	1.76E-01	–	5.81E-02	–
	1927	1.18E-01	1.25	6.83E-02	1.42	2.18E-02	1.47
	7545	4.62E-02	1.38	2.36E-02	1.56	7.03E-03	1.66
	29870	1.88E-02	1.31	8.47E-03	1.49	2.25E-03	1.66
	118851	9.29E-03	1.06	3.12E-03	1.52	7.83E-04	1.60
	474186	4.62E-03	0.97	1.15E-03	1.38	2.59E-04	1.53

We observe that the rate of convergence in the L^1-norm of the three methods is second-order, as expected. There is a slight loss of accuracy in the L^2-norm for the smoothness-based technique; the rate is around 1.90. All three methods are suboptimal in the L^∞-norm. The rate for the smoothness-based technique is between 1.2 and 1.0, and the rate of the other two methods is about 1.5. These rates are compatible with other limiting techniques reported in the finite volume literature; see, for instance, the so-called minimum angle plane reconstruction limiter used in Christov and Popov [9] and the minmod-type limiting from Kurganov and Tadmor [29].

6.2.2. Entropy-viscosity. We test the entropy-viscosity (EV) technique as defined in section 5.1 with the entropy $\eta(v) = v^4$. Tests not reported here show that

Table 2

Swirling deformation flow problem (54) *at* $t = 1$, $CFL = 0.25$. L^∞*-norm of the error for EV solution (columns* 3–4*); EV solution + loc. FCT postprocessing (columns* 5–6*); EV solution + glob. FCT postprocessing (columns* 7–8*). "Low-order" viscosities are* $\psi(\alpha) = \alpha^2$ *(rows* 2–7*); greedy viscosity (rows* 8–13*); FCT viscosity (rows* 14–19*).*

	I	EV alone		EV + loc. FCT		EV + glob. FCT	
$\psi(\alpha) = \alpha^2$	507	1.18E-01	Rate	1.18E-01	Rate	8.62E-02	Rate
	1927	4.49E-02	1.44	4.46E-02	1.45	1.44E-02	2.68
	7545	1.62E-02	1.49	1.63E-02	1.48	1.97E-03	2.91
	29870	5.39E-03	1.60	5.41E-03	1.60	3.20E-04	2.64
	118851	1.67E-03	1.77	1.68E-03	1.77	9.14E-05	1.90
	474186	4.96E-04	1.68	6.04E-04	1.42	2.13E-05	2.01
Greedy visc.	507	8.75E-02	–	9.87E-02	–	9.96E-02	–
	1927	1.07E-02	3.15	2.50E-02	2.06	1.87E-02	2.51
	7545	1.88E-03	2.45	7.95E-03	1.68	1.99E-03	3.28
	29870	8.98E-04	1.07	2.76E-03	1.54	3.08E-04	2.71
	118851	4.10E-04	1.19	1.13E-03	1.35	7.42E-05	2.16
	474186	1.80E-04	1.14	4.25E-04	1.36	1.79E-05	1.97
FCT visc.	507	5.20E-02	–	7.26E-02	–	1.80E-02	–
	1927	5.63E-03	3.33	2.59E-02	1.54	4.51E-03	2.07
	7545	1.15E-03	2.32	8.83E-02	1.57	1.14E-03	2.01
	29870	3.07E-04	1.91	2.77E-03	1.69	2.83E-04	2.02
	118851	9.88E-05	1.72	9.37E-04	1.64	6.99E-05	2.12
	474186	2.48E-05	1.91	3.33E-04	1.43	1.78E-05	1.90

the method performs well with any entropy of the type $\eta(v) = |v|^p$, $p \in \mathbb{N}$, $p \geq 2$, or e^v. We use (45) with the consistent mass matrix $\mathcal{M}$, and the entropy viscosity d_{ij}^n is computed by using (42)–(43)–(44).

We report in the column labeled "EV alone" in Table 2 the L^∞-norm of the error in the three cases considered above. (We focus our attention on the L^∞-norm of the error, since it is in this norm that full second-order is difficult to obtain.) In the first case (rows 2–7) the "low-order" viscosity is the smoothness-based viscosity using $\psi(\alpha) = \alpha^2$ in (32). In the second case (rows 8–13) the "low-order" viscosity is the greedy viscosity defined in (33). In the third case (rows 14–19) the "low-order" viscosity is the FCT viscosity defined in (41). The rates of convergence in the L^1-norm and L^2-norm (not reported for brevity) are second-order. We observe that the convergence rates of the entropy viscosity solution in the column labeled "EV alone" in Table 2 are better than those reported in Table 1, thereby confirming that the entropy viscosity technique is formally at least second-order accurate. Note though that the convergence rate in the L^∞-norm is slightly less than 2.

We now test the entropy viscosity technique as defined in section 5.1 with the same "low-order" viscosities as above, but the maximum principle is now ensured by applying the FCT postprocessing as explained in section 5.3. For each "low-order" viscosity we run two simulations. In the first case (labeled "loc. FCT"), we define the local minimum $\mathsf{U}_i^{\min}$ and the local maximum $\mathsf{U}_i^{\max}$ by $\mathsf{U}_i^{\min} = \min_{i \in \mathcal{I}(S_i)}\{\mathsf{U}_i^n\}$ and $\mathsf{U}_i^{\max} = \max_{i \in \mathcal{I}(S_i)}\{\mathsf{U}_i^n\}$. In the second case (labeled "glob. FCT"), we a priori set $\mathsf{U}_i^{\min} = -1$ and $\mathsf{U}_i^{\max} = 1$. The approximate solution satisfies the local maximum principle in the first case and the global maximum principle in the second case. The results are shown in the columns labeled "EV + loc. FCT" and "EV + glob. FCT" in Table 2.

When comparing the columns "EV alone" and "EV + loc. FCT" in Table 2, we observe that the local FCT postprocessing reduces the accuracy of the entropy-viscosity. The origin of this accuracy loss is the well-known clipping effect of FCT. The results in the column "glob. FCT" show that the entropy-viscosity solution enjoys full second-order in the L^∞-norm when the global maximum principle is enforced. Note finally that the errors and the rates in the column "glob. FCT" are almost independent of the "low-order" viscosity.

6.3. Linear transport with nonsmooth solution. We test in this section the convergence rate of the second-order viscosities of section 4 and the entropy-viscosity on the transport equation $\partial_t u + \nabla\cdot\boldsymbol{f}(u) = 0$ in $D = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\|_{\ell^2} < 1\}$ with the flux $\boldsymbol{f}(u) := \boldsymbol{\beta}(\boldsymbol{x},t)u$, where $\boldsymbol{\beta}(\cdot,t) = 2\pi(-x_2,x_1)^\mathsf{T}$. The initial data is defined by

$$u_0(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_d\|_{\ell^2} \le r_0 \text{ and } (|x_1| \ge 0.05 \text{ or } x_2 \ge 0.7),\\ 1 - \frac{\|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell 2}}{r_0} & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell^2} \le r_0,\\ g(\|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2}) & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2} \le r_0,\\ 0 & \text{otherwise,} \end{cases} \tag{55}$$

where $r_0 = 0.3$, $g(r) := \frac{1}{4}\left[1+\cos\left(\pi\frac{r}{r_0}\right)\right]$, $\boldsymbol{x}_d = (0,0.5)$, $\boldsymbol{x}_c = (0,-0.5)$, $\boldsymbol{x}_h = (-0.5,0)$. The graph of u_0 consists of three solids: a slotted cylinder of height 1, a smooth hump of height $\frac{1}{2}$, and a cone of height 1; see Leveque [33] and Zalesak [41] for more details.

We use five nonnested, nonuniform meshes of approximate mesh size $\frac{1}{20}$, $\frac{1}{40}$, $\frac{1}{80}$, $\frac{1}{160}$, $\frac{1}{320}$. More precisely, the total number of vertices, I, in each mesh is 1605, 6561, 29870, 98648, 389860. Two series of convergence tests are performed; the results are reported in Table 3. In the first series of tests (columns 2–3) we use the second-order viscosities of section 4 but the update is done with the consistent mass matrix and postprocessed with FCT as explained in section 5.3 (with $d^{\mathrm{H},n} = d^{\mathrm{L},n}$); the second series of tests is done with the entropy viscosity plus global FCT postprocessing with $\mathsf{U}^{\min} = 0$ and $\mathsf{U}^{\max} = 1$ as explained in section 6.2.2. We observe that the accuracy of the entropy-viscosity solution is always better than the "low-order" solution; the convergence rates are also slightly higher when using the greedy or the FCT "low-order" viscosities in the definition of the entropy viscosity; see (44). These results are compatible with (or slightly better than) those reported in Table 4 in [19].

Table 3

Three solids problem (55) *at* $t = 1$, $CFL = 0.25$. L^1*-norm of the error for "low-order" solution (columns* 3–4*); EV solution + glob. FCT postprocessing (columns* 5–6*). "Low-order" viscosities are* $\psi(\alpha) = \alpha^2$ *(rows* 2–6*); greedy viscosity (rows* 7–11*); FCT viscosity (rows* 12–16*).*

	I	"Low-order" sol.		EV + glob. FCT	
$\psi(\alpha) = \alpha^2$	1605	7.68E-01	Rate	4.47E-01	Rate
	6561	4.68E-01	0.77	2.83E-01	0.65
	24917	2.63E-01	0.87	1.57E-01	0.88
	98648	1.49E-01	0.82	9.59E-02	0.72
	389860	8.69E-02	0.78	5.58E-02	0.79
Greedy	1605	4.74E-01	–	3.67E-01	–
	6561	2.34E-01	1.00	2.21E-01	0.72
	24917	1.26E-01	0.93	1.14E-01	0.99
	98648	7.83E-02	0.69	7.06E-02	0.70
	389860	4.90E-02	0.68	4.02E-02	0.82
FCT	1605	3.81E-01	–	3.14E-01	–
	6561	1.86E-01	1.01	1.63E-01	0.93
	24917	1.05E-01	0.86	8.96E-02	0.90
	98648	5.95E-02	0.82	5.17E-02	0.80
	389860	3.60E-02	0.73	2.91E-02	0.84

To illustrate the performance of the entropy-viscosity method, we show in Figure 2

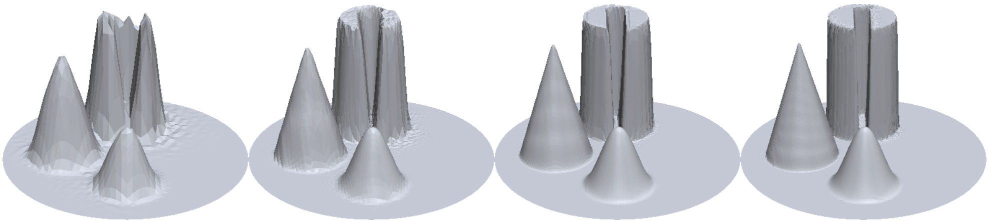

FIG. 2. *Three solids problem* (55) *at* $t = 1$*, EV solution (with FCT-viscosity)+ glob. FCT postprocessing* ([0, 1])*,* $CFL = 0.25$*. From left to right:* $I = 1605$*;* $I = 6561$*;* $I = 24917$*;* $I = 98648$*.*

TABLE 4

Burgers's equation (50)*,* $CFL = 0.25$*.* L^1*-norm of the error at* $t = 0.75$*. Viscosities are* $\psi(\alpha) = \alpha^2$ *(columns* 2–3*); greedy viscosity (columns* 4–5*); FCT viscosity (columns* 6–7*).*

I	$\psi(\alpha) = \alpha^2$ visc.		Greedy visc.		FCT visc.	
507	1.34E-01	–	2.06E-01	–	3.61E-01	–
1927	6.71E-02	1.04	1.31E-01	0.66	3.18E-01	0.19
7545	3.71E-02	0.87	9.27E-02	0.50	3.09E-01	0.04
29870	2.11E-02	0.82	5.61E-02	0.72	3.02E-01	0.03
118851	1.04E-02	1.02	3.22E-02	0.80	2.98E-01	0.02
474189	5.35E-03	0.96	1.81E-02	0.83	2.94E-01	0.02

the graph of the solutions computed with the entropy-viscosity using FCT viscosity as "low-order" viscosity. The global FCT postprocessing is used with $\mathsf{U}^{\min} = 0$ and $\mathsf{U}^{\max} = 1$. We show from left to right the solutions obtained with $I = 1605, 6561, 24917, 98648$.

6.4. Burgers's equation. We consider again Burgers's equation in the setting described in section 6.1, but this time we test the three high-order viscosities described in section 4. The first is the smoothness-based viscosity using $\psi(\alpha) = \alpha^2$ in (32). The second is the greedy viscosity defined in (33). The third is the FCT viscosity defined in (41). The computational domain is $D = (-0.25, 1.75)^2$. We use six nonnested, nonuniform meshes of approximate mesh size $\frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}, \frac{1}{320}$. More precisely, the total number of vertices, I, in each mesh is $507, 1927, 7545, 29870, 118851, 474189$. The computations are run at $\text{CFL} = 0.25$ up to $t = 0.75$. We show in Table 4 the L^1-norm of the error at $t = 0.75$ for the three methods. The smoothness-based correction converges with an order close to 1, which is optimal. The greedy viscosity is slightly suboptimal. The convergence rate seems to grow to 1 as the mesh is refined, though. Inspection of the solution shows that the approximate solution produces a small shock in the sonic point region, but the amplitude of this spurious shock vanishes as the mesh is refined. The most striking result, though, is that the method using the "FCT viscosity" does not converge. The approximate solution forms a shock around the sonic points that never disappears. This result shows that the negative result stated in Lemma 4.6 in dimension 1 actually holds true in higher dimensions.

The behavior of the $\psi(\alpha) = \alpha^2$ viscosity is robust with respect to the CFL (tests not reported), but this is not the case of the greedy viscosity. To further investigate the behavior of the greedy viscosity with respect to the CFL, we report in Table 5 tests done with three different CFL numbers: 0.25, 0.125, 0.0625. The table shows that when the CFL is fixed, the convergence order increases as the mesh is refined and it seems to converge to 1 (this conjecture has been verified on one-dimensional tests not reported). We also see that when the meshsize is fixed, the convergence order decreases

TABLE 5
Burgers's equation (50), *L^1-norm of the error at $t = 0.75$. Greedy viscosity versus CFL.*

	I	CFL=0.25		CFL=0.125		CFL=0.0625	
Greedy visc.	507	2.06E-01	–	2.67E-01	–	3.08E-01	–
	1927	1.31E-01	0.66	1.89E-01	0.52	2.35E-01	0.41
	7545	9.27E-02	0.50	1.47E-01	0.37	1.92E-01	0.29
	29870	5.61E-02	0.72	1.01E-01	0.54	1.42E-01	0.43
	118851	3.22E-02	0.80	6.51E-02	0.64	9.83E-02	0.54
	474189	1.81E-02	0.83	4.10E-02	0.67	6.54E-02	0.59

with the CFL. Inspection of the solution shows that the amplitude of the spurious shock forming in the vicinity of the sonic points increases as the CFL decreases. This is the well-known stair-casing effect that is usually observed with compressive limiters, and it is fully consistent with the analysis reported in Remark 4.5.

We finish this section by investigating the behavior of the entropy viscosity technique combined with the "low-order" methods described in section 4. The L^1-norm of the error in the three cases is shown in Table 6. It is striking that the three "low-order" methods behave similarly when associated with the entropy viscosity; the convergence order is close to 1 in the three cases. This test shows that the entropy viscosity algorithm is capable of fixing the nonconvergence issue of the "FCT viscosity." We have verified (tests not shown here) that the behavior of the greedy viscosity is robust with respect to the CFL, that is to say, below the stability threshold, the convergence rate is independent of the CFL.

TABLE 6
Burgers's equation (50), *CFL* $= 0.25$. *L^1-norm of the error at $t = 0.75$ for the entropy viscosity solution. The "low-order" viscosities are $\psi(\alpha) = \alpha^2$ (columns 3–4); greedy viscosity (columns 5–6); FCT viscosity (columns 7–8).*

	I	$\psi(\alpha) = \alpha^2$ visc.		Greedy visc.		FCT visc.	
EV	507	1.20E-01	–	1.04E-01	–	1.11E-01	–
	1927	5.56E-02	1.15	4.82E-02	1.11	5.51E-02	1.04
	7545	3.02E-02	0.89	2.50E-02	0.95	2.68E-02	1.06
	29870	1.76E-02	0.79	1.51E-02	0.73	1.61E-02	0.74
	118851	8.35E-02	1.08	7.16E-03	1.07	7.46E-03	1.10
	474189	4.47E-03	0.91	3.89E-03	0.88	4.00E-03	0.90

6.5. Nonconvex flux. We finish by solving a two-dimensional scalar conservation equation with a nonconvex flux originally proposed in Kurganov, Petrova, and Popov [30]:

$$\partial_t u + \nabla \cdot \boldsymbol{f}(u) = 0, \quad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) = \begin{cases} \frac{14\pi}{4} & \text{if } \sqrt{x^2+y^2} \le 1, \\ \frac{\pi}{4} & \text{otherwise,} \end{cases} \tag{56}$$

with $\boldsymbol{f}(u) := (\sin u, \cos u)^{\mathsf{T}}$, and the computational domain is $D = [-2, 2] \times [-2.5, 1.5]$. The solution has a two-dimensional composite wave structure which high-order numerical schemes have difficulty capturing correctly. Actually, many high-order schemes have a tendency to produce shocks where one should have expansions; see, e.g., [30].

The maximum wave speed that is used in (18) is estimated from above as follows. Let θ_{ij} be such that $\boldsymbol{n}_{ij} = (\cos(\theta_{ij}), \sin(\theta_{ij}))^{\mathsf{T}}$; then $\boldsymbol{f}(u) \cdot \boldsymbol{n}_{ij} = \sin(u + \theta_{ij})$. If $\lfloor \frac{\mathsf{U}_i^n + \theta_{ij}}{\pi} \rfloor \neq \lfloor \frac{\mathsf{U}_j^n + \theta_{ij}}{\pi} \rfloor$ (where $\lfloor \cdot \rfloor$ is the floor function), then $\boldsymbol{f}(u) \cdot \boldsymbol{n}_{ij}$ is not convex on the interval $[\min(\mathsf{U}_i^n, \mathsf{U}_j^n), \max(\mathsf{U}_i^n, \mathsf{U}_j^n)]$, and we take $\lambda_{\max} = 1$ in (18). If instead

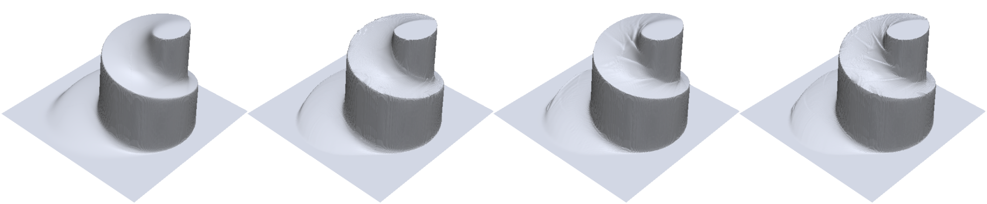

FIG. 3. *KPP problem* (56) *at* $t = 1$, $I = 118850$. *From left to right: GMS solution (first-order viscosity); EV + GMS viscosity + FCT postprocessing;* $\psi(\alpha) = \alpha^2$ *viscosity; EV +* $\psi(\alpha) = \alpha^2$ *viscosity + FCT postprocessing.*

$\lfloor \frac{\mathsf{U}_i^n+\theta_{ij}}{\pi} \rfloor = \lfloor \frac{\mathsf{U}_j^n+\theta_{ij}}{\pi} \rfloor$, then $\boldsymbol{f}(u)\cdot\boldsymbol{n}_{ij}$ is convex on the interval $[\min(\mathsf{U}_i^n, \mathsf{U}_j^n), \max(\mathsf{U}_i^n, \mathsf{U}_j^n)]$. If $\boldsymbol{f}(\mathsf{U}_i^n)\cdot\boldsymbol{n}_{ij} \le \boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{n}_{ij}$, the Riemann solution is an expansion wave, and we take $\lambda_{\max} = \max(|\boldsymbol{f}(\mathsf{U}_i^n)\cdot\boldsymbol{n}_{ij}|, |\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{n}_{ij}|)$; otherwise we have a shock, and we take $\lambda_{\max} = |(\boldsymbol{f}(\mathsf{U}_i^n) - \boldsymbol{f}(\mathsf{U}_j^n))\cdot\boldsymbol{n}_{ij}/(\mathsf{U}_i^n - \mathsf{U}_j^n)|$. Of course, one could take $\lambda_{\max} = 1$ in all the cases since this is also a guaranteed upper bound.

We show in Figure 3 four simulations done at $t = 1$ on a nonuniform mesh with 118850 $\mathbb{P}_1$ nodes. The leftmost panel in Figure 3 shows the graph of the first-order solution obtained with the GMS scheme (i.e., the viscosity defined in (18) with $\lambda_{\max}$ computed as above). The second panel shows the entropy-viscosity solution and FCT postprocessing with the low-order viscosity being the GMS viscosity. The helicoidal composite wave is clearly visible. In the eyeball norm, the entropy-viscosity solution is clearly sharper than the first-order solution and does not look spurious.

The third panel in Figure 3 shows the solution obtained with the smoothness-based viscosity and $\psi(\alpha) = \alpha^2$ with the little modification in (22) consisting of setting $\psi_i^n = \psi_j^n = 1$ if $\boldsymbol{f}(u)\cdot\boldsymbol{n}_{ij}$ is not convex on the interval $[\min(\mathsf{U}_i^n, \mathsf{U}_j^n), \max(\mathsf{U}_i^n, \mathsf{U}_j^n)]$. This modification is similar to the one proposed in Kurganov, Petrova, and Popov [30, section 4]. We have verified on manufactured solutions (test not reported here) that this does not change the convergence order of the method; the reason is that the number of pairs of degrees of freedom where convexity is lost becomes negligible as the mesh is refined. The method fails to converge to the entropy solution if this correction is not applied. The rightmost panel shows the graph of the solution obtained with the entropy viscosity and FCT postprocessing, the "low-order" viscosity being the smoothness-based viscosity described above (with the convexity correction). Similar results are obtained with the greedy viscosity or the FCT viscosity as the low-order method: they both fail to converge to the entropy solution if the convexity correction is not applied (tests not shown). Even with the convexity fix, the graph of the approximate solution obtained with these methods shows artifacts similar to (or worse than) those that are visible in the third and fourth panels of Figure 3. In conclusion, the most robust method is the entropy-viscosity with FCT postprocessing, the low-order viscosity being GMS.

7. Conclusions. Various high-order maximum principle preserving techniques for nonlinear scalar conservation equations have been investigated. It has been proved that the so-called local extrema diminishing (LED) technique (also known as "discrete upwinding" in the literature) is actually equivalent to constructing an artificial viscosity where the maximum wave speed is estimated by Roe's average. As a result, this type of method cannot be convergent when used to solve nonlinear conservation equations. A counterexample is produced in Lemma 3.2. Three families of second-order viscosities that are maximum principle preserving have been investigated in section 4. The first one is based on a measure of the local smoothness of the solution. The second is based on the bounds established in Lemma 4.1. The third, which we have called "FCT viscosity," consists of applying the FCT technique to the Galerkin solution. It is established in Lemma 4.6 that the method using the "FCT viscosity" has convergence problems when applied to nonlinear conservation equations, thereby casting some doubt on the blind use of the FCT technique in the literature. The entropy viscosity idea has been recast into the framework of the GMS scheme in section 5. All the above methods have been tested in section 6. The key conclusions from these tests are as follows: (i) Among the three second-order maximum principle preserving viscosities considered in this paper (smoothness-based, greedy, and "FCT viscosity"), the smoothness-based viscosity is slightly less accurate than the other two methods, but it is the most robust one when used alone. (ii) The greedy viscosity is a bit more accurate, but it lacks robustness with respect to the CFL number when solving problems with sonic points as shown in section 6.4. (iii) The "FCT viscosity" is the most accurate, but it is not robust (it is actually unreliable; see rightmost column in Table 4). (iv) Finally, when combined with the entropy viscosity, all the above viscosities (smoothness-based, greedy, and FCT) perform extremely well, and all three combinations are very robust when the flux is convex or concave (i.e., $\boldsymbol{f}(v){\cdot}\boldsymbol{n}$ is convex or concave for all unit vectors $\boldsymbol{n}$). The entropy-viscosity is actually at least second-order accurate in the L^∞-norm when tested on the linear transport equation and the FCT postprocessing is done with global bounds (see rightmost column in Table 2). It will be shown in a forthcoming paper that full second-order accuracy in the maximum norm can be achieved with local bounds provided that these bounds are relaxed with second-order tolerance.

REFERENCES

[1] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes*, Numer. Math., 135 (2017), pp. 521–545.

[2] M. Berger, M. J. Aftosmis, and S. M. Murman, *Analysis of Slope Limiters on Irregular Grids*, AIAA paper 2005-0490, American Institute for Aeronautics and Astronautics, Reno, NV, 2005. Also NASA TM NAS-05-007.

[3] J. P. Boris and D. L. Book, *Flux-corrected transport.* I. *SHASTA, a fluid transport algorithm that works [J. Comput. Phys.,* 11 (1973)*, pp.* 38–69*]*, J. Comput. Phys., 135 (1997), pp. 170–186.

[4] F. Bouchut and B. Perthame, *Kružkov's estimates for scalar conservation laws revisited*, Trans. Amer. Math. Soc., 350 (1998), pp. 2847–2870.

[5] A. Bressan, *Hyperbolic Systems of Conservation Laws: The One-Dimensional Cauchy Problem*, Oxford Lect. Ser. Math. Appl. 20, Oxford University Press, Oxford, UK, 2000.

[6] E. Burman, *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws*, BIT, 47 (2007), pp. 715–733.

[7] E. Burman and A. Ern, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652.

[8] M. A. Christon, M. J. Martinez, and T. E. Voth, *Generalized Fourier analyses of the advection-diffusion equation, part* I*: One-dimensional domains*, Internat. J. Numer. Methods Fluids, 45 (2004), pp. 839–887.

[9] I. Christov and B. Popov, *New non-oscillatory central schemes on unstructured triangulations for hyperbolic systems of conservation laws*, J. Comput. Phys., 227 (2008), pp. 5736–5757.

[10] P. Colella and H. M. Glaz, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys., 54 (1984), pp. 174–201.

[11] C. M. Dafermos, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[12] M. S. Floater, *Generalized barycentric coordinates and applications*, Acta Numer., 24 (2015), pp. 161–214.

[13] P. Gresho, R. Sani, and M. Engelman, *Incompressible Flow and the Finite Element Method: Advection-Diffusion and Isothermal Laminar Flow*, Wiley, New York, 1998.

[14] J.-L. Guermond and M. Nazarov, *A maximum-principle preserving C^0 finite element method for scalar conservation equations*, Comput. Methods Appl. Mech. Engrg., 272 (2013), pp. 198–213.

[15] J.-L. Guermond and R. Pasquetti, *A correction technique for the dispersive effects of mass lumping for transport problems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 186–198.

[16] J.-L. Guermond and B. Popov, *Error estimates of a first-order Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 54 (2016), pp. 57–85, https://doi.org/10.1137/140990863.

[17] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016) pp. 2466–2489, https://doi.org/10.1137/16M1074291.

[18] J.-L. Guermond, R. Pasquetti, and B. Popov, *Entropy viscosity method for nonlinear conservation laws*, J. Comput. Phys., 230 (2011), pp. 4248–4267.

[19] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 52 (2014), pp. 2163–2182, https://doi.org/10.1137/130950240.

[20] J.-L. Guermond, B. Popov, and Y. Yang, *The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations*, J. Sci. Comput., 70 (2017), pp. 1358–1366.

[21] A. Harten and J. M. Hyman, *Self-adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comput. Phys., 50 (1983), pp. 235–269.

[22] A. Harten and S. Osher, *Uniformly high-order accurate nonoscillatory schemes* I, SIAM J. Numer. Anal., 24 (1987), pp. 279–309, https://doi.org/10.1137/0724022.

[23] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high-order accurate essentially nonoscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.

[24] D. Hoff, *A finite difference scheme for a system of two conservation laws with artificial viscosity*, Math. Comp., 33 (1979), pp. 1171–1193.

[25] D. Hoff, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610.

[26] A. Jameson, *Positive schemes and shock modelling for compressible flows*, Internat. J. Numer. Methods Fluids, 20 (1995), pp. 743–776.

[27] A. Jameson, *Origins and further development of the Jameson-Schmidt-Turkel scheme*, AIAA J., 55 (2017), pp. 1487–1510.

[28] A. Jameson, W. Schmidt, and E. Turkel, *Numerical solution of the Euler equations by finite volume. Methods using Runge-Kutta time-stepping schemes*, in 14th AIAA Fluid and Plasma Dynamics Conference, 1981, AIAA paper 1981–1259.

[29] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.

[30] A. Kurganov, G. Petrova, and B. Popov, *Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws*, SIAM J. Sci. Comput., 29 (2007), pp. 2381–2401, https://doi.org/10.1137/040614189.

[31] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558.

[32] D. Kuzmin, R. Löhner, and S. Turek, *Flux-Corrected Transport*, Scientific Computation, Springer, Dordrecht, The Netherlands, 2005.

[33] R. J. LeVeque, *High-resolution conservative algorithms for advection in incompressible flow*, SIAM J. Numer. Anal., 33 (1996), pp. 627–665, https://doi.org/10.1137/0733033.

[34] P.-O. Persson and J. Peraire, *Sub-cell shock capturing for discontinuous Galerkin methods*, in 44th AIAA Aerospace Sciences Meeting and Exhibit, AIAA paper 2015-2006-112, 2006.

[35] P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.

[36] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[37] P. K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., 21 (1984), pp. 995–1011, https://doi.org/10.1137/0721062.

[38] T. Thompson, *A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation*, J. Comput. Appl. Math., 303 (2016), pp. 229–248.

[39] B. van Leer, *Towards the ultimate conservative difference scheme.* II. *Monotonicity and conservation combined in a second-order scheme*, J. Comput. Phys., 14 (1974), pp. 335–362.

[40] J. Warren, S. Schaefer, A. N. Hirani, and M. Desbrun, *Barycentric coordinates for convex sets*, Adv. Comput. Math., 27 (2007), pp. 319–338.

[41] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.

[42] X. Zhang and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 467 (2011), pp. 2752–2776.