Comput. Methods Appl. Mech. Engrg. 272 (2014) 198–213

Contents lists available at ScienceDirect

Comput. Methods Appl. Mech. Engrg.

journal homepage: www.elsevier.com/locate/cma

# A maximum-principle preserving $C^0$ finite element method for scalar conservation equations ☆

Jean-Luc Guermond *,[1], Murtazo Nazarov

[a] Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843, USA

ARTICLE INFO

*Article history:*
Received 8 October 2013
Received in revised form 20 December 2013
Accepted 30 December 2013
Available online 23 January 2014

*AMS subject classifications:*
65M12
65N12
35L65

*Keywords:*
Conservation equations
Parabolic regularization
Upwinding
First-order viscosity
Entropy solutions

ABSTRACT

This paper introduces a first-order viscosity method for the explicit approximation of scalar conservation equations with Lipschitz fluxes using continuous finite elements on arbitrary grids in any space dimension. Provided the lumped mass matrix is positive definite, the method is shown to satisfy the local maximum principle under a usual CFL condition. The method is independent of the cell type; for instance, the mesh can be a combination of tetrahedra, hexahedra, and prisms in three space dimensions.



## 1. Introduction

The maximum principle is a key property for the analysis of nonlinear scalar conservation equations. Reproducing this property at the discrete level is a feature that is often desired in applications and, from the theoretical point of view, greatly facilitates the convergence analysis of algorithms.

Enforcing the maximum principle using discontinuous finite elements or finite volumes with piece-wise constant approximation is a problem that has been solved since the early work of Lax [21] or possibly earlier; the key is to use the upwind flux. Although it has been shown by Godunov that monotonicity preserving linear methods cannot be second-order accurate, it is possible to construct higher-order accurate discontinuous Galerkin and finite volume methods by making use of limiting techniques as demonstrated in [25,22,17] for finite volumes and [32–35] for the discontinuous Galerkin method.

☆ This material is based upon work supported in part by the National Science Foundation Grants DMS-1015984, and DMS-1217262, by the Air Force Office of Scientific Research, USAF, under Grant/Contract number FA9550-09-1-0424, FA99550-12-0358, and by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

* Corresponding author.
*E-mail address:* guermond@math.tamu.edu (J.-L. Guermond).
[1] On leave from CNRS, France.

0045-7825/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cma.2013.12.015

It seems that the success achieved by finite volume methods and discontinuous Galerkin methods have not yet been matched by continuous finite elements. The first barrier to overcome is the design of a robust maximum-principle satisfying first-order continuous finite element method. To the best of our knowledge, we are not aware of the existence in the literature of an explicit method for continuous piecewise linear finite elements that can be proven to satisfy the maximum principle on any grid, irrespective of any angle condition, and for any Lipschitz continuous flux, (although solutions to this problem have been given for the linear transport equation on particular grids, see e.g., the work of Tabata [26] and Codina [8], and nonlinear viscosities have been shown to yield the desired result on strictly acute meshes in [3,4]). The primary objective of the present paper is to propose such a method. The technique that we propose is the first building block on the road leading to higher-order continuous finite element methods using limiting techniques in the spirit of [25,32] or that of the flux corrected methodology of [2] (see also the book of [23,20, Chap. 9]).

This paper is organized as follows: The formulation of the problem and introductory material are presented in Section 2. Some key shortcomings of the standard artificial viscosity based on the operator $-\nabla\cdot(\nu_h\nabla)$, where $\nu_h$ is scalar-valued, are listed in Section 2.4. A tensor-valued viscosity is proposed in Section 3 in the particular case of simplicial meshes. The method is shown to satisfy the maximum principle on any meshes composed of simplices. The key definition of this section is (3.8) and the main result is Theorem 3.2. The method is extended to general meshes in Section 4. Using the heuristic developed for simplicial meshes in Section 3, we propose an expression for the artificial viscosity (4.11) for which we prove the maximum principle. The key result of this section is Theorem 4.2. A high-order extension of the method using the notion of entropy viscosity is also proposed in this section. The method and its high-order extension are finally illustrated numerically in Section 5.

## 2. Preliminaries

We formulate the problem in this section and recall details on the standard approach based on the notion of isotropic artificial viscosity which we show cannot be extended to general meshes. The material presented in this section is by no means original; it is, however, useful to better appreciate the material presented in Sections 3 and 4.

### 2.1. Formulation of the problem

Let $\Omega$ be an open polyhedral domain in $\mathbb{R}^d$, $d$ is the space dimension. Let $\boldsymbol{f}\in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ be the flux, and let $u_0\in L^\infty(\Omega)$ be some initial data. We consider the scalar-valued conservation equations

$$\partial_t u+\nabla\cdot\boldsymbol{f}(u)=0,\quad u(\boldsymbol{x},0)=u_0(\boldsymbol{x}),\ \ (\boldsymbol{x},t)\in\Omega\times\mathbb{R}_+. \tag{2.1}$$

To simplify questions regarding boundary conditions, we assume that either periodic boundary conditions are enforced, or the initial data is compactly supported and we are interested in the solution before the domain of influence of $u_0$ reaches the boundary of $\Omega$. This problem has a unique entropy solution satisfying the additional entropy inequalities $\partial_t E(u)+\nabla\cdot\boldsymbol{F}(u)\leq 0$ for all convex entropy $E\in\mathrm{Lip}(\mathbb{R};\mathbb{R})$ and associated entropy flux $\boldsymbol{F}$ with $\boldsymbol{F}_i'(u)=\int_0^u E'(v)\boldsymbol{f}_i'(v)\,\mathrm{d}v$, $1\leqslant i\leqslant d$ (see [18,1]).

### 2.2. The mesh

Let $\{\mathcal{K}_h\}_{h>0}$ be a mesh family that we assume to be conforming (no hanging nodes) and shape-regular in the sense of Ciarlet. By convention, the elements in $\mathcal{K}_h$ are closed in $\mathbb{R}^d$. Let $\{(\widehat{K},\widehat{P},\widehat{\Sigma})\}$ be a finite family of reference Lagrange finite elements in the sense of Ciarlet. The map between $\widehat{K}$ and an arbitrary element $K\in\mathcal{K}_h$ is denoted $\Phi_K:\widehat{K}\longrightarrow K$. For the sake of simplicity we assume that $\Phi_K$ is affine.

Our objective is to approximate the entropy solution of (2.1) with $H^1$-conforming Lagrange finite elements. To this end we define the scalar-valued finite element approximation space

$$X_h=\{v\in\mathcal{C}^0(\Omega;\mathbb{R});\ v|_K\circ\Phi_K\in\widehat{P},\ \forall K\in\mathcal{K}_h\}, \tag{2.2}$$

where $\widehat{P}$ is the reference polynomial space associated with $K\in\mathcal{K}_h$.

Let $\{\varphi_1,\dots,\varphi_N\}$ be the nodal Lagrange basis associated with the vertices of the mesh $\mathcal{K}_h$, say $\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_N\}$, i.e., $\varphi_i(\boldsymbol{a}_j)=\delta_{ij}$. We denote by $S_i$ the support of $\varphi_i$ and by $|S_i|$ the measure of $S_i$, $i=1,\dots,N$. We also define $S_{ij}:=S_i\cap S_j$ the intersection of the two supports $S_i$ and $S_j$. Let $E$ be a union of cells in $\mathcal{K}_h$; we define $\mathcal{I}(E):=\{j\in\{1,\dots,N\};|S_j\cap E|\neq 0\}$. The set $\mathcal{I}(E)$ contains the indices of all the shape functions whose support on $E$ is non-empty.

Although the notion of local meshsize is not relevant for the method to be presented in the remainder of the paper, it is useful to define a CFL number. We then define the so-called local minimum meshsize, say $h_K$, of a cell $K\in\mathcal{K}_h$ as follows:

$$h_K:=\frac{1}{\max\limits_{i\in\mathcal{I}(K)}\|\nabla\varphi_i\|_{L^\infty(K)}}, \tag{2.3}$$

and the global minimum mesh size as $h:=\min_{K\in\mathcal{K}_h}h_K$; this parameter is solely used to define the CFL number.

### 2.3. The finite volume/DG0 approaches

Let us restrict ourselves for the time being to one space dimension and assume that the mesh is uniform with mesh size $h$ and time step $\Delta t$. The flux in (2.1) is rewritten $\boldsymbol{f}(u)=f(u)\boldsymbol{e}_x$ where $\boldsymbol{e}_x$ is the unit vector on the real line pointing towards $+\infty$. Let us consider finite volumes or equivalently the piecewise constant Discontinuous Galerkin approximation. Denote by $\{U_i^k\}_{i=1,\cdots,N}$ the piecewise constant approximation of $u$ at time $t^k$. Let $U_{i-1}^k$, $U_i^k$ and $U_{i+1}^k$ be the three approximate values of $u$ over cells $[x_{i-\frac{3}{2}},x_{i-\frac{1}{2}}]$, $[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$ and $[x_{i+\frac{1}{2}},x_{i+\frac{3}{2}}]$ at time $t^k$. Let $\Gamma_{i-\frac{1}{2}}$ and $\Gamma_{i+\frac{1}{2}}$ be the two interfaces separating the three above cells. The Lax–Friedrichs scheme for (2.1) consists of setting

$$h\frac{U_i^{k+1}-U_i^k}{\Delta t}=-(\boldsymbol{n}_{\Gamma_{i-\frac{1}{2}}}\cdot\widehat{\boldsymbol{f}}(U_i,U_{i-1})+\boldsymbol{n}_{\Gamma_{i+\frac{1}{2}}}\cdot\widehat{\boldsymbol{f}}(U_i,U_{i+1})), \tag{2.4}$$

where $\boldsymbol{n}_\Gamma$ is the unit normal vector at the interface pointing from the interior to the exterior of the cell; i.e., $\boldsymbol{n}_{\Gamma_{i-\frac{1}{2}}}=-\boldsymbol{e}_x$ and $\boldsymbol{n}_{\Gamma_{i+\frac{1}{2}}}=\boldsymbol{e}_x$. The numerical flux $\widehat{\boldsymbol{f}}(V^i,V^e)$ chosen is the Lax–Friedrichs flux

$$\widehat{\boldsymbol{f}}(V^i,V^e)=\frac{1}{2}(\boldsymbol{f}(V^i)+\boldsymbol{f}(V^e))+\frac{1}{2}|\beta|(V^i-V^e)\boldsymbol{n}_\Gamma, \tag{2.5}$$

where $V^i$ and $V^e$ are the interior and exterior values. The quantity $|\beta|$ is the maximum wave speed, i.e., $|\beta|:=\|f'\|_{L^\infty(\mathbb{R})}$. It is then possible to recast (2.4) into the following form

$$\frac{U_i^{k+1}-U_i^k}{\Delta t}=-\frac{(f(U_{i+1}^k)-f(U_{i-1}^k))}{2h}+\frac{1}{2}|\beta|h\frac{(U_{i+1}^k-2U_i^k+U_{i-1}^k)}{h^2}. \tag{2.6}$$

**Remark 2.1** (*Upwinding*). Note in passing that the so-called Lax–Friedrichs flux (2.5) reduces to the upwind flux in the case of linear transport. For instance assuming that $\boldsymbol{f}=\beta u\boldsymbol{e}_x$, we obtain $\widehat{\boldsymbol{f}}(V^i,V^e)=V_i\frac{1}{2}(\beta\boldsymbol{e}_x+|\beta|\boldsymbol{n}_\Gamma)+V_e\frac{1}{2}(\beta\boldsymbol{e}_x-|\beta|\boldsymbol{n}_\Gamma)$. Then $\widehat{\boldsymbol{f}}(V^i,V^e)=V^e\beta\boldsymbol{e}_x=\boldsymbol{f}(V^e)$ if $\beta\boldsymbol{e}_x\cdot\boldsymbol{n}_\Gamma<0$ (i.e., if the flow enters the cell), and $\widehat{\boldsymbol{f}}(V^i,V^e)=V_i\beta\boldsymbol{e}_x=\boldsymbol{f}(V^i)$ otherwise (i.e., if the flow exists the cell).

**Remark 2.2** (*Lax–Friedrichs scheme*). The so-called Lax–Friedrichs as originally introduced in [21, p. 163] is as follows:

$$\frac{U_i^{k+1}-U_i^k}{\Delta t}=-\frac{(f(U_{i+1}^k)-f(U_{i-1}^k))}{2h}+\frac{1}{\mathrm{cfl}}\frac{1}{2}|\beta|h\frac{(U_{i+1}^k-2U_i^k+U_{i-1}^k)}{h^2}, \tag{2.7}$$

where $\mathrm{cfl}:=|\beta|\Delta t h^{-1}$ is the Courant–Friedrichs–Levy number. This scheme is more dissipative than (2.6) due to the presence of the CFL number in the denominator of the artificial viscosity. It is an accepted practice to refer to both (2.6) and (2.7) as Lax–Friedrichs schemes, although this may sometimes lead to confusions.

A key property of the scheme (2.6) is that it satisfies the local discrete maximum principle (formulated in [21, p. 190] for (2.7)).

**Theorem 2.1.** *Assume that* $f\in\mathrm{Lip}(\mathbb{R};\mathbb{R})$ *and* $-\infty<u_{\min}:=\min_{x\in\mathbb{R}}u_0(x)\leqslant\min_iU_i^0\leqslant\max_iU_i^0\leqslant\max_{x\in\mathbb{R}}u_0(x):=u_{\max}<\infty$. *Let* $|\beta|=\sup_{v\in[u_{\min},u_{\max}]}|f'(v)|$. *Assume that* $|\beta|\Delta t\leqslant h$, *then both algorithms (2.6) and (2.7) satisfy the local discrete maximum principle, i.e.,* $u_{\min}\leqslant\min(U_{i-1}^k,U_i^k,U_{i+1}^k)\leqslant U_i^{k+1}\leqslant\max(U_{i-1}^k,U_i^k,U_{i+1}^k)\leqslant u_{\max}$.

**Proof.** Although the argument is standard and can be found in many textbooks, we give the proof for completeness since we are going to reuse the argument latter. The argument proceeds by induction. At $t^0$ we have $u_{\min}\leqslant\min_iU_i^0\leqslant\max_iU_i^0\leqslant u_{\max}$ by assumption. Let us now assume that this property still holds at time $t^k$. The key is to construct a convex combination of $U_{i-1}^k,U_i^k,U_{i+1}^k$. Using the mean-value theorem, there is $V$ between $U_{i-1}$ and $U_{i+1}$ so that $f(U_{i+1}^k)-f(U_{i-1}^k)=f'(V)(U_{i+1}^k-U_{i-1}^k)$. Then

$$U_i^{k+1}=U_i^k\left(1-\frac{|\beta|\Delta t}{h}\right)+U_{i+1}^k\frac{\Delta t}{2h}(|\beta|-f'(V))+U_{i-1}^k\frac{\Delta t}{2h}(|\beta|+f'(V)).$$

The induction assumption implies that $V\in[u_{\min},u_{\max}]$ which in turn implies that $|f'(V)|\leqslant|\beta|$ from the assumptions on $f$. The assumption on the CFL number finally implies that $U_i^{k+1}=a_{i-1}U_{i-1}^k+a_iU_i^k+a_{i+1}U_{i+1}^k$ with $\sum_ja_j=1$ and $a_j\geqslant0$, for $j=i-i,i,i+1$, i.e., $U_i^{k+1}$ is a convex combination of $U_{i-1}^k,U_i^k,U_{i+1}^k$. This implies in particular that $\min(U_{i-1}^k,U_i^k,U_{i+1}^k)\leqslant U_i^{k+1}\leqslant\max(U_{i-1}^k,U_i^k,U_{i+1}^k)$, which in turn proves that the induction hypothesis holds at time $t^{k+1}$. This concludes the proof. □

The key conclusion of this section is that formula (2.6) can be reinterpreted as a continuous piecewise linear finite element approximation of (2.1) with an artificial viscosity equal to $\frac{1}{2}|\beta|h$. This observation is at the origin of a large body of research in the continuous finite element literature trying to reproduce the stabilizing properties of the upwind flux of the DG approximation by augmenting the Galerkin formulation with semi-linear forms like $\int_\Omega\nu_h(u)\nabla u\cdot\nabla v\,\mathrm{d}\boldsymbol{x}$ where $\nu_h(u)$ is some nonlinear artificial viscosity scaling like $\frac{1}{2}|\beta|h$ in regions of large gradients.

### 2.4. Some shortcomings of scalar viscosities

We review in this section some intrinsic shortcomings of the semi-linear form $\int_\Omega v_h(u)\nabla u\cdot\nabla v\,d\boldsymbol{x}$, where $v_h(u)$ is scalar-valued. Let us adopt this definition for the time being and, to simplify, assume that $v_h(u)\sim|\beta|h$, i.e., we are using linear first-order artificial viscosity.

The first obstacle one runs into when using continuous finite elements is that of a proper definition of the meshsize $h$ on non-uniform anisotropic meshes. Although many clever and reasonably well justified ideas have been proposed to address this non-trivial issue, see e.g., [28,27,5,9], to the best of our knowledge, none of them have yet lead to a provable maximum principle holding for every nonlinear flux $\boldsymbol{f}\in\mathrm{Lip}(\mathbb{R},\mathbb{R}^d)$ and every mesh, assuming piecewise linear approximation of course, i.e., assuming that (3.1) holds.

Once some meshsize $h$ has been chosen, one then runs into the problem of choosing the constant to multiply $|\beta|h$ to form the viscosity. Although $\frac{1}{2}$ seems to be a reasonable choice justified by the one-dimensional analysis on uniform grids, we do not know of any rational for tuning this constant in two and three space dimensions on arbitrary grids besides heuristic arguments and trial and errors tests. Again, even-though it is possible to establish well-founded heuristic arguments to tune the constant (see e.g., [28,27,15]), we do not know of any argument yielding a provable maximum principle for every nonlinear flux $\boldsymbol{f}$ and every mesh.

The last argument that finally lead us to revisit the first-order theory is that it is not robust with respect to the shape of the cells. When trying to reproduce the one-dimensional argument in arbitrary space dimension with continuous finite elements one observes that the convex combination argument in the proof of Theorem 2.1 can be made to work only if $\int_{S_{ij}}\nabla\varphi_i\cdot\nabla\varphi_j\,d\boldsymbol{x}<0$ for all pairs of shape functions, $\varphi_i,\varphi_j$, with common support of nonzero measure. This is the well-known acute angle condition assumption, see e.g., [24, p. 182], [8, Eq. (35)], see also [30, Eq. (2.5)], [4, Eq. (8)] for a slightly weaker version of this condition. For instance, one easily verifies that the acute angle condition assumption fails on the grid shown in Fig. 2.1.

**Remark 2.3** (*Re-orientation of the gradients*). We show in the next section that the acute angle assumption can be avoided on simplicial meshes by realizing that, since the viscosity is artificial and has no particular physical meaning, one can reorient $\nabla\varphi_i$ and $\nabla\varphi_i$ at will to make the scalar product of the two new vectors negative and bounded away from zero uniformly. This observation, which makes the viscosity tensor-valued, is the key to the entire paper.

## 3. A tensor-valued viscosity for simplices

We restrict ourselves in this section to meshes composed of simplices only and we assume that the shape functions are piecewise $\mathbb{P}_1$ (i.e., multivariate polynomials of total degree at most 1). General meshes (quadrangular, hexahedral) and higher-order polynomial approximations are considered in Section 4. The objective of this section is to introduce the definitions (3.3) and (3.4), which are the inspiration for the definitions (4.9) and (4.11) that we are going to use in the general case presented in Section 4.

### 3.1. Some geometry

We assume in this entire section that $\widehat{K}$ is the regular simplex whose edges all have length 1, i.e., $\widehat{K}$ is the equilateral triangle of side 1 in two space dimension, and $\widehat{K}$ is the regular tetrahedron (all four faces are equilateral triangles) in three space dimensions. Since $X_h$ is composed of piecewise linear functions, we have

$$\min_{\ell\in\mathcal{I}(K)}v(\boldsymbol{a}_\ell)\leqslant v(\boldsymbol{x})\leqslant\max_{\ell\in\mathcal{I}(K)}v(\boldsymbol{a}_\ell),\quad\forall v\in X_h,\ \forall\boldsymbol{x}\in K,\ \forall K\in\mathcal{K}_h. \tag{3.1}$$

This property holds on simplicial meshes with $\widehat{P}=\mathbb{P}_1$ (set of multivariate polynomials of total degree at most 1). It holds also on quadrilateral and hexahedral meshes with $\widehat{P}=\mathbb{Q}_1$ (set of multivariate polynomials of partial degree at most 1). It also works with prismatic elements.

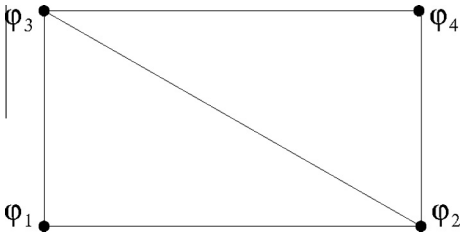


**Fig. 2.1.** Example of a mesh violating the minimal angle condition: $\int_{S_{23}}\nabla\varphi_2\cdot\nabla\varphi_3\,d\boldsymbol{x}=0$.

Let $K$ be an arbitrary cell in the mesh. Denoting by $\Phi_K : \widehat{K} \longrightarrow K$ the affine mapping that transforms $\widehat{K}$ to $K$, $\mathbb{J}_K$ the Jacobian matrix of $\Phi_K$, and $\mathbb{J}_K^{\intercal}$ its transpose, the chain rule implies that

$$\nabla(u \circ \Phi_K) = \mathbb{J}_K^{\intercal}(\nabla u)(\Phi_K), \tag{3.2}$$

for all weakly differentiable function $u$ defined over $K$.

Let $\varphi_i \neq \varphi_j$ be two shape functions with non-empty support in $K$. We now claim that $\mathbb{J}_K^{\intercal}(\nabla\varphi_i)\cdot\mathbb{J}_K^{\intercal}(\nabla\varphi_j)$ has exactly the right property we are looking for.

**Lemma 3.1.** *Under the assumptions on the mesh family $\{\mathcal{K}_h\}_{h>0}$ and approximation spaces $\{X_h\}_{h>0}$ stated in Section 2.2, there is are constants $\alpha := \frac{1}{\widehat{d}h^2} > 0$, $\gamma := \frac{1}{\widehat{h}^2} > 0$, where $\widehat{h} = \sqrt{\frac{d+1}{2d}}$ is the height of $\widehat{K}$, so that*

$$\int_K (\mathbb{J}_K^{\intercal}\nabla\varphi_i)\cdot(\mathbb{J}_K^{\intercal}\nabla\varphi_j)\,\mathrm{d}\boldsymbol{x} = -\alpha|K|, \quad \forall K \in \mathcal{K}_h,\ \ \forall i, \forall j \in S(K), j \neq i, \tag{3.3}$$

$$\int_K \|\mathbb{J}_K^{\intercal}\nabla\varphi_i\|^2\,\mathrm{d}\boldsymbol{x} = \gamma|K|, \quad \forall K \in \mathcal{K}_h,\ \ \forall i \in S(K). \tag{3.4}$$

**Proof.** Let $\mathcal{K}_h$ be a mesh and $K$ be a mesh cell in $\mathcal{K}_h$. Let us define $\widehat{\varphi}_l := \varphi_l \circ \Phi_K$ for all $l \in \mathcal{I}(K)$. Let $i \neq j \in \mathcal{I}(K)$. Making use of (3.2) and the symmetry properties of $\widehat{K}$, we obtain

$$\begin{aligned}\int_K (\mathbb{J}_K^{\intercal}\nabla\varphi_i)\cdot(\mathbb{J}_K^{\intercal}\nabla\varphi_j)\frac{\mathrm{d}\boldsymbol{x}}{|\det(\mathbb{J}_K)|} &= \int_{\widehat{K}} \nabla\widehat{\varphi}_i\cdot\nabla\widehat{\varphi}_j\,\mathrm{d}\widehat{\boldsymbol{x}} = \frac{1}{d}\sum_{\mathcal{I}(K)\ni l\neq i}\int_{\widehat{K}}\nabla\widehat{\varphi}_i\cdot\nabla\widehat{\varphi}_l\,\mathrm{d}\widehat{\boldsymbol{x}} = -\frac{1}{d}\int_{\widehat{K}}\nabla\widehat{\varphi}_i\cdot\nabla\widehat{\varphi}_i\,\mathrm{d}\widehat{\boldsymbol{x}}\\ &= -\frac{1}{d}\int_{\widehat{K}}\|\nabla\widehat{\varphi}_i\|^2\,\mathrm{d}\widehat{\boldsymbol{x}} < 0.\end{aligned}$$

Note that the constant $\alpha := \frac{1}{d|\widehat{K}|}\int_{\widehat{K}}\|\nabla\widehat{\varphi}_i\|^2\,\mathrm{d}\widehat{\boldsymbol{x}}$ is actually independent of $i$ owing to the symmetry properties of $\widehat{K}$. The definition $\widehat{\varphi}_i := \varphi_i \circ \Phi_K$ implies that $\|\nabla\widehat{\varphi}_i\| = \widehat{h}^{-1}$, meaning that (3.3) holds with $\alpha := \frac{1}{\widehat{d}h^2} > 0$, since $\det(\mathbb{J}_K) = |K|/|\widehat{K}|$. □

The above argument shows that $\alpha = \frac{2}{d+1}$ and $\gamma = \frac{2d}{d+1}$, i.e.,

$$\int_K (\mathbb{J}_K^{\intercal}\nabla\varphi_i)\cdot(\mathbb{J}_K^{\intercal}\nabla\varphi_j)\,\mathrm{d}\boldsymbol{x} = -|K|\begin{cases}\frac{2}{3} & \text{in 2D},\\ \frac{1}{2} & \text{in 3D}.\end{cases} \tag{3.5}$$

$$\int_K \|\mathbb{J}_K^{\intercal}\nabla\varphi_i\|^2\,\mathrm{d}\boldsymbol{x} = |K|\begin{cases}\frac{4}{3} & \text{in 2D},\\ \frac{3}{2} & \text{in 3D}.\end{cases} \tag{3.6}$$

### 3.2. The tensor-valued viscosity

The argumentation in the above section suggests that a proper way to reformulate the action of the artificial viscosity consists of using the following bilinear form: $\sum_{K\in\mathcal{K}_h}\int_K \nu_K(\mathbb{J}_K^{\intercal}\nabla u)\cdot(\mathbb{J}_K^{\intercal}\nabla v)\,\mathrm{d}\boldsymbol{x}$.

Let $u_{0h} \in X_h$ be an approximation of $u_0$. Let us denote by $u_h^k := \sum_{j=1}^N U_j^k\varphi_j \in X_h$ an approximation of $u(\cdot, t^k)$ where $t^k \geqslant 0$ is some time. Let $\Delta t^k$ be the next time step so that $t^{k+1} = t^k + \Delta t^k$. We construct $u_h^{k+1} = \sum_{j=1}^N U_j^{k+1}\varphi_j \in X_h$ to be such that

$$U_i^{k+1} = U_i^k - \Delta t^k m_i^{-1}\sum_{K\subset S_i}\int_K \left(\nu_K^k(\mathbb{J}_K^{\intercal}\nabla u_h^k)\cdot(\mathbb{J}_K^{\intercal}\nabla\varphi_i) + \nabla\cdot(\boldsymbol{f}(u_h^k))\varphi_i\right)\mathrm{d}\boldsymbol{x}. \tag{3.7}$$

Note that the mass matrix has been lumped so that $\int_{S_i} u_h\varphi_i\,\mathrm{d}\boldsymbol{x}$ has been replaced by $U_i m_i$ where $m_i := \int_{S_i}\varphi_i\,\mathrm{d}\boldsymbol{x}$. The above definition can be recast into the following more algebraic form:

$$\begin{aligned}U_i^{k+1} = U_i^k\Bigg(1 - \Delta t^k m_i^{-1}\sum_{K\subset S_i}\int_K\left(\nu_K^k\|\mathbb{J}_K^{\intercal}\nabla\varphi_i\|^2 + (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_i)\varphi_i\right)\mathrm{d}\boldsymbol{x}\Bigg)\\ - \Delta t^k m_i^{-1}\sum_{\mathcal{I}(S_i)\ni j\neq i} U_j^k\sum_{K\subset S_{ij}}\int_K\left(\nu_K^k(\mathbb{J}_K^{\intercal}\nabla\varphi_j)\cdot(\mathbb{J}_K^{\intercal}\nabla\varphi_i) + (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i\right)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

This equation can also be formally put into the form of linear combination as follows:

$$U_i^{k+1} = \sum_{j\in\mathcal{I}(S_i)} a_{ij}U_j^k, \quad i \in \{1,\dots,N\}.$$

Note that owing to the fact that $\sum_{j\in\mathcal{I}(S_i)}\varphi_j|_{S_i} = 1$, we have $\sum_{j\in\mathcal{I}(S_i)} a_{ij} = 1$, since

$$1 = 1 - \Delta t^k m_i^{-1} \sum_{K\subset S_i} \int_K \left( \nu_K^k (\mathbb{J}_K^{\mathsf{T}} \nabla \sum_{j\in\mathcal{I}(S_i)} \varphi_j) \cdot (\mathbb{J}_K^{\mathsf{T}} \nabla \varphi_i) + (\boldsymbol{f}\,'(u_h^k) \cdot \nabla \sum_{j\in\mathcal{I}(S_i)} \varphi_j)\varphi_i \right) \mathrm{d}\boldsymbol{x} = \sum_{j\in\mathcal{I}(S_i)} a_{ij}.$$

At this point, a reasonable definition of the artificial viscosity $\nu_K^k$ becomes clear. The maximum principle will be satisfied if the above combination is convex, i.e., $a_{ij} \geqslant 0$. A sufficient condition for this property to hold is that

$$\nu_K^k = \max_{i\neq j\in\mathcal{I}(K)} \frac{\left| \int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k) \cdot \nabla\varphi_j)\varphi_i \,\mathrm{d}\boldsymbol{x} \right|}{-\int_{S_{ij}} (\mathbb{J}^{\mathsf{T}} \nabla\varphi_j) \cdot (\mathbb{J}^{\mathsf{T}} \nabla\varphi_i)\,\mathrm{d}\boldsymbol{x}}, \tag{3.8}$$

where we defined $\mathbb{J}$ so that $\mathbb{J}|_K = \mathbb{J}_K$ for all $K \in \mathcal{K}_h$.

**Remark 3.1.** (*Anisotropic viscosity*). Note that the above definition makes sense since the denominator in (3.8) is negative. Actually $\int_{S_{ij}} (\mathbb{J}^{\mathsf{T}} \nabla\varphi_j) \cdot (\mathbb{J}^{\mathsf{T}} \nabla\varphi_i)\,\mathrm{d}\boldsymbol{x} = -\alpha|S_{ij}|$, i.e., $\nu_K^k = \frac{1}{\alpha} \max_{i\neq j\in\mathcal{I}(K)} \frac{1}{|S_{ij}|} \left| \int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k) \cdot \nabla\varphi_j)\varphi_i \,\mathrm{d}\boldsymbol{x} \right|$. We have thus defined a viscous-tensor $\mathbb{M}_K^k := \nu_K^k \mathbb{J}_K \mathbb{J}_K^{\mathsf{T}}$. The corresponding artificial viscosity operator is $-\nabla \cdot (\mathbb{M}_K^k \cdot \nabla u_h)$.

**Remark 3.2** (*Scaling*). Note that the piecewise constant scalar-valued function $\nu_K^k$ scales like a wave speed times $h_K^{-1}$ whereas the Frobenius norm of $\mathbb{M}_K^k$ scales like a wave speed times $h_K$.

### 3.3. Maximum principle

We establish the maximum principle for the scheme (3.7) in this section.

**Theorem 3.2.** *Assume that* $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ *and let* $u_{\min} := \inf_{\boldsymbol{x}\in\mathbb{R}^d} u_0(\boldsymbol{x})$, $u_{\max} := \sup_{\boldsymbol{x}\in\mathbb{R}^d} u_0(\boldsymbol{x})$, *and* $\beta = \sup_{v\in[u_{\min},u_{\max}]} \|\boldsymbol{f}\,'(v)\|$. *Assume that* $u_{\min} \leqslant U_i^0 \leqslant u_{\max}$, *for all* $i = 1,\dots,N$, *and* $\beta\Delta t^k h^{-1} \leqslant 1/(1+d)$. *Then the solution to* (3.7) *satisfies the local discrete maximum principle, i.e.,* $u_{\min} \leqslant \min_{j\in\mathcal{I}(S_i)} U_j^k \leqslant U_i^{k+1} \leqslant \max_{j\in\mathcal{I}(S_i)} U_j^k \leqslant u_{\max}$ *for all* $k \geqslant 0$.

**Proof.** We proceed by induction. Let us assume that $u_{\min} \leqslant U_i^k \leqslant u_{\max}$ for some $k \geqslant 0$ and for all $i = 1,\dots,N$. Note that the induction assumption holds for $k = 0$ by definition. Note that $u_h^k(\boldsymbol{x}) \in [u_{\min}, u_{\max}]$ for all $\boldsymbol{x} \in \Omega$, since the approximation space satisfies the convexity property (3.1) and the nodal values of $u_h^k$, $\{U_i^k\}_{i=1,\dots,N}$, satisfy the induction assumption. The definition of $h_K$, (2.3), implies that $\|\nabla\varphi_j\|_{\boldsymbol{L}^\infty(K)} \leqslant h_K^{-1}$. This bound together with the definition of $\nu_K^k$ in (3.8) and the identity $\int_K |\varphi_j|\,\mathrm{d}\boldsymbol{x} = \int_K \varphi_j\,\mathrm{d}\boldsymbol{x} = (d+1)^{-1}|K|$ implies that

$$\nu_K^k \leqslant \max_{v\in[u_{\min},u_{\max}]} \|\boldsymbol{f}\,'(v)\| \max_{i\neq j\in\mathcal{I}(K)} \frac{|\int_{S_{ij}} \|\nabla\varphi_j\|\varphi_i\,\mathrm{d}\boldsymbol{x}}{\alpha|S_{ij}|} \leqslant \alpha^{-1}\beta h^{-1}(d+1)^{-1}.$$

Let us now evaluate $a_{ii}$,

$$\begin{aligned} a_{ii} &:= 1 - \Delta t^k m_i^{-1} \sum_{K\subset S_i} \int_K \left( \nu_K^k \|J_K^{\mathsf{T}} \nabla\varphi_i\|^2 + (\boldsymbol{f}\,'(u_h^k) \cdot \nabla\varphi_i)\varphi_i \right) \mathrm{d}\boldsymbol{x} \\ &\geqslant 1 - \Delta t^k m_i^{-1} \sum_{K\subset S_i} (\alpha^{-1}\beta h^{-1}\gamma|K| + \beta h^{-1}|K|)(d+1)^{-1} \\ &\geqslant 1 - \beta\Delta t^k h^{-1}(1+\gamma\alpha^{-1})|S_i|(d+1)^{-1} m_i^{-1} \geqslant 1 - \beta\Delta t^k h^{-1}(1+d), \end{aligned}$$

where we used that $m_i = (d+1)^{-1}|S_i|$ and $\gamma\alpha^{-1} = d$. This implies that $a_{ii} \geqslant 0$ since $|\beta|\Delta t^k h^{-1} \leqslant 1/(1+d)$. The definition (3.8) implies that

$$a_{ij} := -\Delta t^k m_i^{-1} \sum_{K\subset S_{ij}} \int_K \left( \nu_K^k (\mathbb{J}_K^{\mathsf{T}} \nabla\varphi_j) \cdot (\mathbb{J}_K^{\mathsf{T}} \nabla\varphi_i) + (\boldsymbol{f}\,'(u_h^k) \cdot \nabla\varphi_j)\varphi_i \right) \mathrm{d}\boldsymbol{x} \geqslant 0.$$

Finally the property $\sum_{j\in\mathcal{I}(S_i)} a_{ij} = 1$ implies that $U_i^{k+1}$ is a convex combination of $\{U_j\}_{j\in\mathcal{I}(S_i)}$. This proves the local discrete maximum principle, which in turns implies that the induction hypothesis holds for $k+1$. □

**Remark 3.3** (*SSP extension*). The result of Theorem (3.2) can be directly extended to any higher-order Strong Stability Preserving time stepping (see e.g., [10] for a review), since these schemes construct higher-order accurate approximations in time by making convex combination of solutions of forward Euler sub-steps.

**Remark 3.4** (*CFL condition*). The CFL condition in Theorem 3.2 is slightly suboptimal by the factor $d/(d+1)$. For instance we obtain CFL $\leqslant \frac{1}{2}$ in one space dimension on uniform grids instead of the standard result CFL $\leqslant 1$. The reason for this is that the expression $\int_{S_i} \nabla \cdot (\boldsymbol{f}(u_h^k))\varphi_i\,\mathrm{d}\boldsymbol{x}$ has not been integrated by parts; as a result there is the term $\int_{S_i} (\boldsymbol{f}\,'(u_h^k) \cdot \nabla\varphi_i)\varphi_i\,\mathrm{d}\boldsymbol{x}$ in the definition of $a_{ii}$ that inflates the evaluation of the CFL number. This extra term is actually zero for linear transport, and in this particular case we obtain CFL $\leqslant \frac{1}{d}$, which is optimal.

### 3.4. Examples

We apply in this section the above theory to the linear transport equation on a uniform Cartesian mesh composed of triangles to give a better intuition of the action of the viscosity defined in (3.8). We want in particular to illustrate the fact that the definition (3.8) puts the right amount of viscosity in each direction of the mesh, independently of the anisotropy ratio.

Consider the uniform Cartesian mesh shown in Fig. 3.1 and denote by $h_x, h_y$ the mesh size in direction $x$ and $y$, respectively. Assume that $f(u) = \boldsymbol{\beta} u$ where $\boldsymbol{\beta} = (\beta_x, \beta_y)$ is a constant vector. Owing to the symmetries of the mesh, definition (3.8) gives

$$\nu_K^k = \max_{i\neq j\in\mathcal{I}(K)} \frac{\left|\int_{S_{ij}} (\boldsymbol{\beta}\cdot\nabla\phi_j)\varphi_i \,\mathrm{d}\boldsymbol{x}\right|}{\frac{4}{3}|K|} = \frac{1}{4}\max\left(2\frac{|\beta_x|}{h_x}+\frac{|\beta_y|}{h_y}, \frac{|\beta_x|}{h_x}+2\frac{|\beta_y|}{h_y}\right), \tag{3.9}$$

i.e., the viscosity coefficient $\nu_K^k$ is constant over the entire mesh and scales like a speed over a distance (see Remark 3.2). Upon setting $B = \max(2\frac{|\beta_x|}{h_x}+\frac{|\beta_y|}{h_y}, \frac{|\beta_x|}{h_x}+2\frac{|\beta_y|}{h_y})$, (3.7) is recast into the following form:

$$\begin{aligned} U_5^{k+1} = {} & U_5^k(1-\Delta t^k B) + U_6^k\frac{\Delta t^k}{6}\left(-2\frac{\beta_x}{h_x}-\frac{\beta_y}{h_y}+B\right) + U_8^k\frac{\Delta t^k}{6}\left(-\frac{\beta_x}{h_x}-2\frac{\beta_y}{h_y}+B\right) + U_4^k\frac{\Delta t^k}{6}\left(2\frac{\beta_x}{h_x}+\frac{\beta_y}{h_y}+B\right) \\ & + U_2^k\frac{\Delta t^k}{6}\left(\frac{\beta_x}{h_x}+2\frac{\beta_y}{h_y}+B\right) + U_3^k\frac{\Delta t^k}{6}\left(-\frac{\beta_x}{h_x}+\frac{\beta_y}{h_y}+B\right) + U_7^k\frac{\Delta t^k}{6}\left(\frac{\beta_x}{h_x}-\frac{\beta_y}{h_y}+B\right), \end{aligned} \tag{3.10}$$

which is clearly a convex combination of $U_2^k, \ldots, U_8^k$. We can also rewrite this equation as follows:

$$\begin{aligned} & \frac{U_5^{k+1}-U_5^k}{\Delta t^k} + \frac{1}{3}\left(2\beta_x+\beta_y\frac{h_x}{h_y}\right)\frac{U_6^k-U_4^k}{2h_x} + \frac{1}{3}\left(2\beta_y+\beta_x\frac{h_y}{h_x}\right)\frac{U_8^k-U_2^k}{2h_y} + \frac{1}{3}\left(\frac{\beta_x}{h_x}-\frac{\beta_y}{h_y}\right)\sqrt{h_x^2+h_y^2}\,\frac{U_3^k-U_7^k}{2\sqrt{h_x^2+h_y^2}} \\ & -\frac{Bh_x^2}{6}\frac{U_6^k-2U_5^k+U_4^k}{h_x^2} - \frac{Bh_y^2}{6}\frac{U_8^k-2U_5^k+U_2^k}{h_y^2} - \frac{B(h_x^2+h_y^2)}{6}\frac{U_3^k-2U_5^k+U_7^k}{h_x^2+h_y^2} = 0. \end{aligned} \tag{3.11}$$

This computation is summarized in Table 3.1. The representation (3.11) shows that the effective wave speed in the directions $\boldsymbol{a}_4\boldsymbol{a}_6, \boldsymbol{a}_2\boldsymbol{a}_8$, and $\boldsymbol{a}_7\boldsymbol{a}_3$ is $\frac{1}{3}(2\beta_x+\beta_y\frac{h_x}{h_y}), \frac{1}{3}(2\beta_y+\beta_x\frac{h_y}{h_x})$, and $\frac{1}{3}(\frac{\beta_x}{h_x}-\frac{\beta_y}{h_y})\sqrt{h_x^2+h_y^2}$, respectively. In the direction $\boldsymbol{a}_4\boldsymbol{a}_6$, the viscosity, $\frac{Bh_x^2}{6}$, is larger than $\frac{1}{3}(2|\beta_x|+|\beta_y|\frac{h_x}{h_y})\frac{1}{2}h_x$; in the direction $\boldsymbol{a}_2\boldsymbol{a}_8$, the viscosity, $\frac{Bh_y^2}{6}$, is larger than $\frac{1}{3}(2|\beta_y|+|\beta_x|\frac{h_y}{h_x})\frac{1}{2}h_y$; and in the direction $\boldsymbol{a}_7\boldsymbol{a}_3$, the viscosity, $\frac{1}{6}B(h_x^2+h_y^2)$, is larger than $\frac{1}{3}(\frac{|\beta_x|}{h_x}+\frac{|\beta_y|}{h_y})\sqrt{h_x^2+h_y^2}\frac{1}{2}\sqrt{h_x^2+h_y^2}$. In conclusion, the scheme puts exactly the right amount of viscosity in each direction irrespective of the anisotropy ratio of the mesh.

## 4. General meshes and higher-order extension

We revisit the above theory in this section and adapt it to general meshes and higher-order approximation settings that may not satisfy (3.1). We also extend the method to the entropy viscosity framework to make it higher-order.

### 4.1. The cubic obstruction

We show in this section that definitions (4.1) and (4.2) are not appropriate to prove the maximum principle on hexahedral meshes.

Let us assume that the mesh $\mathcal{K}_h$ is composed of quadrangles in two space dimensions or hexahedra in three space dimensions. Surprisingly, the direct extension of the above theory to this type of cells is not trivial. Let us assume that $\widehat{K}$ is the unit square or unit cube with side of unit length. Let $K$ be an arbitrary cell in $\mathcal{K}_h$ and let $\Phi_K : \widehat{K} \longrightarrow K$ be the $\mathbb{Q}_1$ mapping that

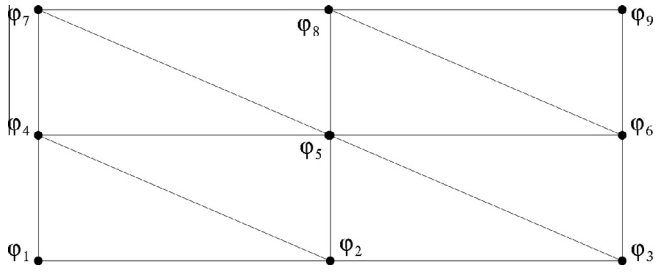


**Fig. 3.1.** Example of an anisotropic mesh.

**Table 3.1**
Artificial viscosity for example (3.10).

| Dir. | $h$ | Speed | $\frac{1}{2}\lvert\text{speed}\rvert \times h$ | Viscosity |
|---|---|---|---|---|
| $\boldsymbol{a}_4\boldsymbol{a}_6$ | $h_x$ | $\frac{1}{3}(2\beta_x+\beta_y\frac{h_x}{h_y})$ | $\frac{1}{6}(2\frac{\lvert\beta_x\rvert}{h_x}+\frac{\lvert\beta_y\rvert}{h_y})h_x^2$ | $\frac{Bh_x^2}{6}$ |
| $\boldsymbol{a}_2\boldsymbol{a}_8$ | $h_y$ | $\frac{1}{3}(2\beta_y+\beta_x\frac{h_y}{h_x})$ | $\frac{1}{6}(2\frac{\lvert\beta_y\rvert}{h_y}+\frac{\lvert\beta_x\rvert}{h_x})h_y^2$ | $\frac{Bh_y^2}{6}$ |
| $\boldsymbol{a}_7\boldsymbol{a}_3$ | $\sqrt{h_x^2+h_y^2}$ | $\frac{1}{3}(\frac{\beta_x}{h_x}-\frac{\beta_y}{h_y})\sqrt{h_x^2+h_y^2}$ | $\frac{1}{6}(\frac{\lvert\beta_x\rvert}{h_x}+\frac{\lvert\beta_y\rvert}{h_y})(h_x^2+h_y^2)$ | $\frac{B(h_x^2+h_y^2)}{6}$ |

transforms $\widehat{K}$ to $K$. Let $\mathbb{J}_K$ be the Jacobian matrix of $\Phi_K$. A seemingly reasonable definition of the first-order artificial viscosity is as follows:

$$v_K^k = \max_{i\neq j\in\mathcal{I}(K)} \frac{\left|\int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i\,\mathrm{d}\boldsymbol{x}\right|}{-\sum_{K\subset S_{ij}}\int_K(\mathbb{J}^{\mathsf{T}}\nabla\varphi_j)\cdot(\mathbb{J}^{\mathsf{T}}\nabla\varphi_i)|K||\det\mathbb{J}_K^{-1}|\,\mathrm{d}\boldsymbol{x}}. \tag{4.1}$$

Note that if the mapping $\Phi_K$ is affine, i.e., the mesh is composed of parallelograms or parallelepipeds, $|K||\det\mathbb{J}_K^{-1}|=1$, and the above definition is identical to (3.8). The artificial viscosity bilinear form is then

$$(u_h^k, v_h)\mapsto\sum_{K\in\mathcal{K}_h}\int_K v_K^k(\mathbb{J}^{\mathsf{T}}\nabla u_h^k)\cdot(\mathbb{J}^{\mathsf{T}}\nabla v_h)|K||\det\mathbb{J}_K^{-1}|\,\mathrm{d}\boldsymbol{x}. \tag{4.2}$$

One then runs in a major obstruction in three space dimension. To illustrate the problem we are facing, assume that the mesh is composed of identical cubes with sides of length $h$. Let $\varphi_i$ and $\varphi_j$ be two shape functions whose nodes are located at the two extremities of one edge of a cell $K$ in the mesh. In that case $\mathbb{J}_K=h\mathbb{I}$, where $\mathbb{I}$ is the identity matrix, and $|K||\det\mathbb{J}_K^{-1}|=1$. A simple computation shows that $\int_K(\mathbb{J}_K^{\mathsf{T}}\nabla\varphi_i)\cdot(\mathbb{J}_K^{\mathsf{T}}\nabla\varphi_j)\,\mathrm{d}\boldsymbol{x}=0$. This observation shows that definitions (4.1) and (4.2) are not sufficient to prove the maximum principle on hexahedral meshes.

### 4.2. General case

The obstruction that we have identified on uniform cubic meshes shows that simple geometric arguments are not enough to construct a general theory of artificial viscosity. We now propose to change the point of view and adopt a graph theoretic perspective.

#### 4.2.1. Mesh considerations

We do not make any specific assumption on the shape of the cells composing $\mathcal{K}_h$, and the degree of the reference polynomial spaces in the family $\{\widehat{P}\}$ is unspecified. Note that $\mathcal{K}_h$ could be composed of a mixture of triangles and quadrangles in two space dimensions, or composed of a mixture of tetrahedra, hexahedra, and prims in three space dimensions. At this point we do not require that $X_h$ satisfy (3.1). It is essential though that $X_h$ be such that the mass matrix can be lumped and be positive definite. We formalize this hypothesis by introducing the following notation and assumption for all $K\in\mathcal{K}_h$:

$$0<\mu_K^{\min}:=\min_{i\in\mathcal{I}(K)}\frac{1}{|K|}\int_K\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\quad \mu_K^{\max}:=\max_{i\in\mathcal{I}(K)}\frac{1}{|K|}\int_K|\varphi_i(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}. \tag{4.3}$$

**Lemma 4.1.** *The following inequalities hold:*

$$0<\mu_K^{\min}|S_i|\leqslant m_i\leqslant\mu_K^{\max}|S_i|. \tag{4.4}$$

**Proof.** Let $\varphi_i$ be a shape function, $1\leqslant i\leqslant N$. The definition of $m_i$ implies that

$$m_i=\int_{S_i}\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\sum_{K\subset S_i}|K|\frac{1}{|K|}\int_K\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\geqslant\mu_K^{\min}|S_i|>0.$$

The upper bound in (4.4) is derived similarly. □

Let $K$ be a cell in $\mathcal{K}$ and let $n_K$ be the number of vertices in $K$, i.e., $n_K=\mathrm{card}(\mathcal{I}(K))$. Owing to the mesh being affine, the quantities $\mu_K^{\min}$, $\mu_K^{\max}$, and $n_K$ only depend on $\widehat{K}$. Since the number of reference elements defining the mesh family $\{\mathcal{K}_h\}_{h>0}$ is finite, we now define

$$\lambda := \max_{\mathcal{K}_h} \max_{K\in\mathcal{K}_h} \frac{\mu_K^{\max}}{\mu_K^{\min}} < +\infty, \quad \rho := \min_{\mathcal{K}_h} \min_{K\in\mathcal{K}_h} \frac{1}{n_K - 1} > 0. \tag{4.5}$$

**Remark 4.1** (*Higher-order polynomials*). Note that $\lambda = 1$ and the assumption $\mu_K^{\min} > 0$ holds if $X_h$ is such that the convexity assumption (3.1) holds. For instance $\rho = \frac{1}{d}$ and $\mu_K^{\min} = \mu_K^{\max} = (d+1)^{-1}$ for $\mathbb{P}_1$ approximation on simplices, and $\rho = \frac{1}{d}$ and $\mu_K^{\min} = \mu_K^{\max} = 2^{-d}$ for $\mathbb{Q}_1$ approximation on parallelotopes. Note that $\mu_K^{\min} = 0$ for $\mathbb{P}_2$ approximation. The assumption $\mu_K^{\min} > 0$ does hold though for $\mathbb{P}_3$ approximation on simplices, at least in $\mathbb{R}^2$, when the reference nodes on $\widehat{K}$ are either equidistributed on a uniform lattice or are the Fekete points. More generally, $\mathbb{Q}_k$ finite elements with tensor-product Gauss–Lobatto nodes are such that $\mu_K^{\min} > 0$.

We now estimate the maximum wave speed to define a CFL number. If (3.1) holds, we set

$$\beta := \sup_{v\in[u_{\min},u_{\max}]} \|\boldsymbol{f}'(v)\|. \tag{4.6}$$

Otherwise we define $B_h := \{v_h \in X_h \mid u_{\min} \leqslant \min_{1\leqslant i\leqslant N} v_h(\boldsymbol{a}_i) \leqslant \max_{1\leqslant i\leqslant N} v_h(\boldsymbol{a}_i) \leqslant u_{\max}\} \subset X_h$. Note that $B_h$ is not a vector space. It can be shown that

$$u_{\min}^* := \inf_{\{\mathcal{K}_h\}_{h>0}} \inf_{v_h\in B_h,\, \boldsymbol{x}\in\Omega} v_h(\boldsymbol{x}) \quad \text{and} \quad u_{\max}^* := \sup_{\{\mathcal{K}_h\}_{h>0}} \sup_{v_h\in B_h} \sup_{\boldsymbol{x}\in\Omega} v_h(\boldsymbol{x}), \tag{4.7}$$

are two finite numbers; in particular, there are constants $c_{\min}$ and $c_{\max}$, that depend only on the collection of reference finite elements $\{(\widehat{K}, \widehat{P}, \widehat{\Sigma})\}$, such that $c_{\min}u_{\min} \leqslant u_{\min}^*$ and $u_{\max}^* \leqslant c_{\max}u_{\max}$. We then define the maximum wave speed to be

$$\beta := \sup_{v\in[u_{\min}^*,u_{\max}^*]} \|\boldsymbol{f}'(v)\|. \tag{4.8}$$

*4.2.2. Viscous bilinear form*

Taking inspiration from (3.3) and (3.4), we define the local bilinear form $b_K$ so that

$$b_K(\varphi_j,\varphi_i) = \begin{cases} -\frac{1}{n_K-1}|K| & \text{if } i\neq j,\ \ i,j\in\mathcal{I}(K),\\ |K| & \text{if } i=j,\ \ i,j\in\mathcal{I}(K),\\ 0 & \text{if } i\notin\mathcal{I}(K) \text{ or } j\notin\mathcal{I}(K). \end{cases} \tag{4.9}$$

The global artificial viscosity bilinear form at time $t^k$ is then defined as follows:

$$b(u_h^k, v_h) = \sum_{K\in\mathcal{K}_h} \sum_{i,j\in\mathcal{I}(K)} \nu_K^k U_j^k V_i b_K(\varphi_j,\varphi_i), \tag{4.10}$$

where $u_h^k = \sum_{i=1}^N U_i^k\varphi_i$, $v_h = \sum_{i=1}^N V_i\varphi_i$. Taking inspiration again from (3.8), the local artificial coefficient $\nu_K^k$ is defined as follows:

$$\nu_K^k = \max_{i\neq j\in\mathcal{I}(K)} \frac{\left|\int_{S_{ij}} (\boldsymbol{f}'(u_h^k)\cdot\nabla\varphi_j)\varphi_i \,\mathrm{d}\boldsymbol{x}\right|}{-\sum_{T\subset S_{ij}} b_T(\varphi_j,\varphi_i)}. \tag{4.11}$$

*4.2.3. Algorithm*

The time stepping is done by proceeding like in (3.7). Let $u_{0h} \in X_h$ be an approximation of $u_0$ that satisfies the discrete maximum principle,

$$u_{\min} := \inf_{\boldsymbol{x}\in\Omega} u_0(\boldsymbol{x}) \leqslant \min_{1\leqslant i\leqslant N} u_{0h}(\boldsymbol{a}_i) \leqslant \max_{1\leqslant i\leqslant N} u_{0h}(\boldsymbol{a}_i) \leqslant \sup_{\boldsymbol{x}\in\Omega} u_0(\boldsymbol{x}) := u_{\max}. \tag{4.12}$$

Let $\Delta t^k$ be the next time step so that $t^{k+1} = t^k + \Delta t^k$. The nodal values of the solution $u_h^{k+1} = \sum_{j=1}^N U_j^{k+1}\varphi_j \in X_h$ at time $t^{k+1}$ is evaluated as follows:

$$U_i^{k+1} = U_i^k - \Delta t^k m_i^{-1} \sum_{K\subset S_i} \left( \nu_K^k b_K(u_h^k,\varphi_i) + \int_K \nabla\cdot(\boldsymbol{f}(u_h^k))\varphi_i \,\mathrm{d}\boldsymbol{x} \right). \tag{4.13}$$

Note again that the mass matrix has been lumped and $m_i := \int_{S_i} \varphi_i \,\mathrm{d}\boldsymbol{x}$ is positive owing to (4.4).

**Theorem 4.2** (*Discrete maximum principle*). *Assume that the CFL number is small enough, i.e.,* $\beta\Delta t^k h^{-1} \leqslant 1/(\lambda(1+\rho^{-1}))$. *Then the solution to (4.13) satisfies the local discrete maximum principle, i.e.,* $u_{\min} \leqslant \min_{j\in\mathcal{I}(S_i)} U_j^k \leqslant U_i^{k+1} \leqslant \max_{j\in\mathcal{I}(S_i)} U_j^k \leqslant u_{\max}$ *for all* $k \geqslant 0$.

**Proof.** The proof is similar to that of Theorem 3.2, i.e., we proceed by induction. Let $k \geqslant 0$ and assume that $u_{\min} \leqslant U_i^k \leqslant u_{\max}$ for all $i = 1, \ldots, N$. Note that the induction assumption holds for $k = 0$. The definition of $v_K^k$ in (4.11) together with the definition of $\rho$ in (4.5), the inequality $\|\nabla\varphi_j\|_{\boldsymbol{L}^\infty(K)} \leqslant h_K^{-1}$ and the inequality $\int_K |\varphi_j| \mathrm{d}\boldsymbol{x} \leqslant \mu_K^{\max}|K|$ (see (4.3)) implies that

$$v_K^k \leqslant \beta \max_{i\neq j\in\mathcal{I}(K)} \frac{|\int_{S_{ij}} \|\nabla\varphi_j\|\varphi_i \mathrm{d}\boldsymbol{x}}{\rho|S_{ij}|} \leqslant \rho^{-1}\beta h^{-1}\mu_K^{\max}.$$

The expression (4.13) can be recast into the following more algebraic form:

$$\begin{aligned} U_i^{k+1} = U_i^k\left(1 - \Delta t^k m_i^{-1}\sum_{K\subset S_i}\left(v_K^k b_K(\varphi_i,\varphi_i) + \int_K (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_i)\varphi_i \mathrm{d}\boldsymbol{x}\right)\right) \\ - \Delta t^k m_i^{-1}\sum_{\mathcal{I}(S_i)\ni j\neq i} U_j^k \sum_{K\subset S_{ij}}\left(v_K^k b_K(\varphi_j,\varphi_i) + \int_K (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i \mathrm{d}\boldsymbol{x}\right), \end{aligned}$$

that we formally rewrite as $U_i^{k+1} = \sum_{j\in\mathcal{I}(S_i)} a_{ij} U_j^k$. First, we observe that

$$\sum_{j\in\mathcal{I}(S_i)} a_{ij} = 1 - \Delta t^k m_i^{-1}\sum_{K\subset S_i}\left(v_K^k b_K\left(\sum_{j\in\mathcal{I}(S_i)}\varphi_j,\varphi_i\right) + \int_K (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\sum_{j\in\mathcal{I}(S_i)}\varphi_j)\varphi_i \mathrm{d}\boldsymbol{x}\right) = 1$$

since $\sum_{j\in\mathcal{I}(S_i)}\varphi_j|_{S_i} = 1$ and $b_K(\sum_{j\in\mathcal{I}(S_i)}\varphi_j,\varphi_i) = b_K(\sum_{j\in\mathcal{I}(K)}\varphi_j,\varphi_i) = 0$ owing to the definition (4.9). Second, we evaluate a bound from below for $a_{ii}$,

$$\begin{aligned} a_{ii} &:= 1 - \Delta t^k m_i^{-1}\sum_{K\subset S_i}\left(v_K^k b_K(\varphi_i,\varphi_i) + \int_K (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_i)\varphi_i \mathrm{d}\boldsymbol{x}\right) \\ &\geqslant 1 - \Delta t^k m_i^{-1}\sum_{K\subset S_i}(\rho^{-1}\beta h^{-1}|K| + \beta h^{-1}|K|)\mu_K^{\max} \\ &\geqslant 1 - \beta\Delta t^k h^{-1}(1+\rho^{-1})|S_i|\mu_K^{\max}m_i^{-1} \geqslant 1 - \beta\Delta t^k h^{-1}(1+\rho^{-1})\lambda, \end{aligned}$$

where we used that $\mu_K^{\max}|S_i|m_i^{-1} \leqslant \mu_K^{\max}/\mu_K^{\min} \leqslant \lambda$ (see (4.4)). This implies that $a_{ii} \geqslant 0$ since $|\beta|\Delta t^k h^{-1} \leqslant 1/(\lambda(1+\rho^{-1}))$. Third, we evaluate a bound from below for $a_{ij}$, with $i \neq j$. Observing that $b_K(\varphi_j,\varphi_i) \leqslant 0$. The definition (4.11) implies that

$$-\sum_{K\subset S_{ij}} v_K^k b_K(\varphi_j,\varphi_i) \geqslant -\sum_{K\subset S_{ij}} \frac{\left|\int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i \mathrm{d}\boldsymbol{x}\right|}{-\sum_{T\subset S_{ij}} b_T(\varphi_j,\varphi_i)} b_K(\varphi_j,\varphi_i) \geqslant |\int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i \mathrm{d}\boldsymbol{x}|,$$

which gives

$$a_{ij} := \Delta t^k m_i^{-1}\left(\sum_{K\subset S_{ij}} - v_K^k b_K(\varphi_j,\varphi_i) - \int_{S_{ij}} (\boldsymbol{f}\,'(u_h^k)\cdot\nabla\varphi_j)\varphi_i \mathrm{d}\boldsymbol{x}\right) \geqslant 0.$$

The above argument shows that $U_i^{k+1}$ is a convex combination of $\{U_j\}_{j\in\mathcal{I}(S_i)}$. This proves the local discrete maximum principle, which in turns implies that the induction hypothesis holds for $k+1$. □

**Corollary 4.3.** *Under the assumptions of Theorem 4.2, the solution to (4.13) satisfies the following $L^\infty$-estimates*

$$c_{\min}u_{\min} \leqslant u_h^n(\boldsymbol{x}) \leqslant c_{\max}u_{\max}, \quad \forall n \geqslant 0, \ \forall \boldsymbol{x}\in\Omega. \tag{4.14}$$

*Moreover the following holds for all $K \in \mathcal{K}_h$ if the space $X_h$ satisfies (3.1):*

$$\min_{i\in\mathcal{I}(K)}\min_{j\in\mathcal{I}(S_i)} U_j^n \leqslant u_h^{n+1}(\boldsymbol{x}) \leqslant \max_{i\in\mathcal{I}(K)}\max_{j\in\mathcal{I}(S_i)} U_j^n, \quad \forall n \geqslant 0. \tag{4.15}$$

**Proof.** The estimate (4.14) is a consequence of Theorem 4.2 and (4.7). The second estimate is a consequence of Theorem 4.2 and the convexity assumption (3.1). □

**Remark 4.2** (*Graph Laplacian*). The definition (4.10) is somewhat reminiscent of that of a graph Laplacian, (see e.g., [30, Eq. (2.3)] where a graph Laplacian is used to stabilized an advection diffusion equation).

**Table 5.1**
Burger's equation: convergence tests for $\mathbb{P}_1$ entropy viscosity and first-order viscosity.

| $h$ | $\mathbb{P}_1$ Entropy viscosity | | | | $\mathbb{P}_1$ First-order viscosity | | | |
|---|---|---|---|---|---|---|---|---|
| | $L^1$ | Rate | $L^2$ | Rate | $L^1$ | Rate | $L^2$ | Rate |
| 1/25 | 3.95E−02 | – | 1.16E−01 | – | 4.33E−02 | – | 1.21E−01 | – |
| 1/50 | 2.43E−02 | 0.70 | 1.01E−01 | 0.20 | 2.87E−02 | 0.59 | 1.07E−01 | 0.17 |
| 1/100 | 1.19E−02 | 1.00 | 6.85E−02 | 0.56 | 1.52E−02 | 0.92 | 7.27E−02 | 0.56 |
| 1/200 | 5.94E−03 | 1.03 | 4.87E−02 | 0.49 | 8.28E−03 | 0.87 | 5.31E−02 | 0.45 |
| 1/400 | 3.09E−03 | 0.94 | 3.64E−02 | 0.42 | 4.57E−03 | 0.86 | 4.04E−02 | 0.40 |

**Remark 4.3** (*CFL condition*). Note that the CFL condition in Theorem 4.2 is exactly the same as in Theorem 3.2 for simplicial meshes, since in that case $\rho^{-1} = n_K - 1 = d$ and $\lambda = 1$. It is remarkable that the CFL number scales like $1/(\lambda(1+\rho^{-1}))$ on every meshes; see also Remark 3.4.

**Remark 4.4** (*Convergence*). We expect that the solution of (4.13) converges to the entropy solution, since the method is essentially a first-order viscosity method. This result will be established in a forthcoming paper.

### 4.3. Extension to the entropy viscosity method

We now propose a variation of the above technique to make it higher-order using the notion of entropy viscosity introduced in [13,15]. The method consists of using a SSP method to step in time where each Runge–Kutta sub-step is of the form (4.13) with a higher-order viscosity.

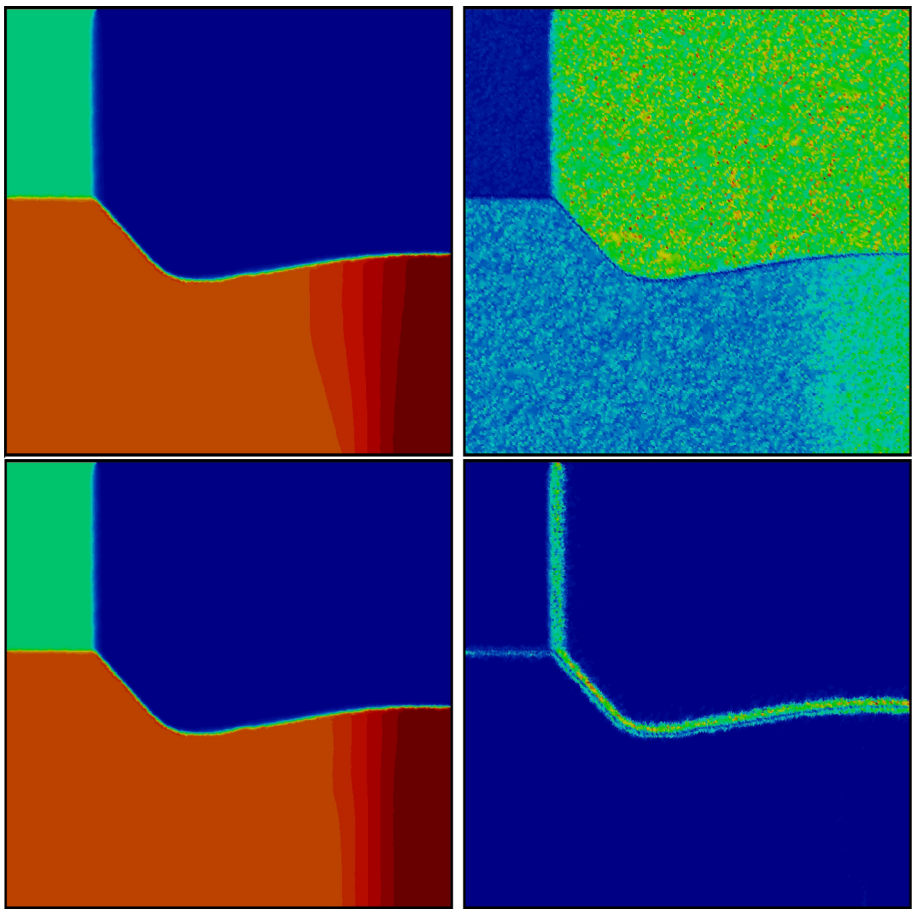

**Fig. 5.1.** Burger's equation at time $t = 0.5$ on an unstructured mesh composed of 36836 $\mathbb{P}1$ nodes. First row: first-order approximation; solution $-1.0 \leqslant u \leqslant 0.8$; viscosity coefficient $9.5 \leqslant v \leqslant 209.14$. Second row: entropy viscosity approximation; solution $-1.002 \leqslant u \leqslant 0.802$; viscosity coefficient $0 \leqslant v \leqslant 176.04$; $c_E = 1$, $c_J = 4$.

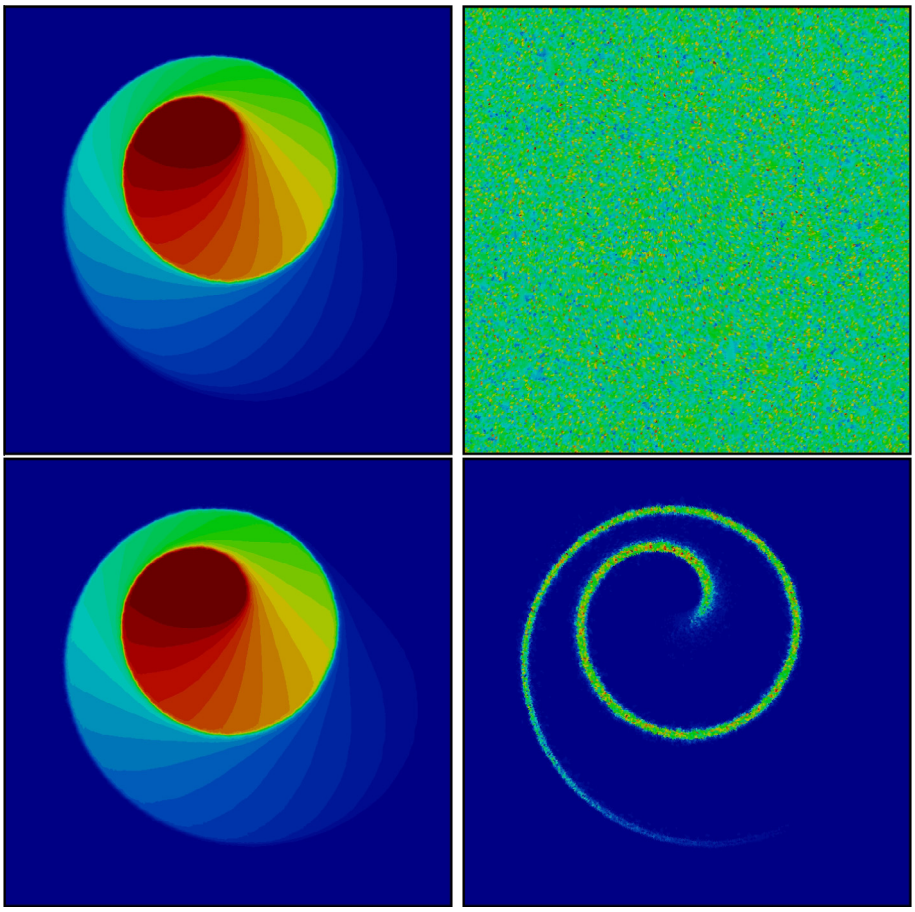

**Fig. 5.2.** KPP problem at time $t = 1$ on an unstructured mesh composed of 52963 $\mathbb{P}_1$ nodes. First row: first-order approximation; solution $\frac{\pi}{4} \leqslant u \leqslant \frac{14\pi}{4}$; viscosity coefficient $10.18 \leqslant \nu_K \leqslant 44.62$. Second row: entropy viscosity approximation; solution $0.67 \leqslant u \leqslant 11.02$; viscosity coefficient $0 \leqslant \nu_K \leqslant 42.02$; $c_E = 1$, $c_J = 4$.

The higher-order viscosity is defined to be the minimum of the first-order viscosity defined in (4.11) and an entropy residual. Let $E \in \mathrm{Lip}(\mathbb{R};\mathbb{R})$ be a convex entropy. Let $K$ be a cell in the mesh $\mathcal{K}_h$. We define an entropy residual over $K$ as follows:

$$R_K^k(u_h^k, u_h^{k-1}) = \left\| \frac{1}{\Delta t^{k-1}} (E(u_h^k) - E(u_h^{k-1})) + \boldsymbol{f}\,'(u_h^k) \cdot \nabla E(u_h^k) \right\|_{L^\infty(K)}. \tag{4.16}$$

The $L^\infty$-norm over $K$ is estimated by evaluating the residual at the quadrature points. A first-order approximation of the time derivative $\partial_t E(u)$ is used here, but a second-order or a higher-order approximation of the derivative of the entropy can easily be constructed. For instance, the evaluation of the entropy residual can be embedded within Runge–Kutta sub-steps; we omit the details for simplicity. Let $F$ be a face of $K$ and assume that $F$ is an interface, i.e., $F$ is not a boundary face. It is useful to evaluate the entropy jump across the cell interfaces

$$J_F^k(u_h^k) = \| \boldsymbol{f}\,'(u_h^k) \cdot \boldsymbol{n} [\![ \partial_n E(u_h^k) ]\!] \|_{L^\infty(F)}. \tag{4.17}$$

Let $\nu_K^{v,k}$ be the first-order viscosity defined in (4.11). The so-called entropy viscosity is constructed as follows:

$$\nu_K^k = \min \left( \nu_K^{v,k}, \frac{c_E R_K^k(u_h^k, u_h^{k-1}) + c_J \max_{F \in \partial K} J_F^k(u_h^k)}{\| E(u_h^k) - \overline{E}(u_h^k) \|_{L^\infty(\Omega)}} \right), \tag{4.18}$$

where $c_E$ and $c_J$ are user-defined parameters of order unity and $\overline{E}(u_h^k)$ is the mean of $E(u_h^k)$ over $\Omega$. The key improvement over earlier versions of the method, see e.g., [13,15], is that the definition (4.18) does not require a notion of mesh-size. This sim-

plifies significantly the implementation of the method and makes it better suited for realistic computations on arbitrary or/and anisotropic meshes. All the quantities $v_K^{v,k}, R_K^k, J_K^k$ scale like the inverse of the distance to the square times a wave speed. The method is not yet parameter free due to the presence of $c_E$ and $c_J$. We have observed that $c_E = 1$ and $c_J = 4$ makes the method work satisfactorily in all the applications we have investigated.

Let us finally mention that the method is implemented without lumping the mass matrix in the time stepping. Although, this strategy makes the method to lose the maximum principle property, it is nevertheless more accurate to work with the consistent mass matrix, since lumping the mass matrix induces high dispersion errors as shown in e.g., [6,7,11,29,14].

The resulting method is not maximum preserving, but it has been established in [12] that it can be made so after some limiting and still be conservative and be formally second-order. The key to the argument is to combine a conservative treatment of the mass matrix proposed in [14] with the flux corrected methodology of Boris–Book–Zalesak [2,31] (see also [23,20, Chap. 9]).

## 5. Numerical illustrations

To illustrate numerically the methods presented in the previous sections, we now solve three different scalar conservation equations in two space dimensions on nonuniform triangular grids using $\mathbb{P}_1$ approximation. All the results presented below are done using SSP RK3 to step in time. For the first-order algorithm we use the artificial viscosity defined in (4.9)–(4.11) and each RK3 substep is done using (4.13). For the higher-order version of the algorithm we use the definition (4.18) for the viscosity and the variant of (4.13) with the consistent mass matrix.

### 5.1. 2D Burgers

Consider the two dimensional Burger's equation in $\mathbb{R}^2$

$$\partial_t u + \nabla \cdot \left(\frac{1}{2}\boldsymbol{\beta} u^2\right) = 0, \qquad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \tag{5.1}$$

where $\boldsymbol{\beta} = (1, 1)$ is a constant vector field, and the initial condition is

$$u_0 = \begin{cases} -0.2, & \text{if } x < 0.5 \text{ and } y > 0.5, \\ -1, & \text{if } x > 0.5 \text{ and } y > 0.5, \\ 0.5, & \text{if } x < 0.5 \text{ and } y < 0.5, \\ 0.8, & \text{if } x > 0.5 \text{ and } y < 0.5. \end{cases} \tag{5.2}$$

This exact solution to this problem is as follows:

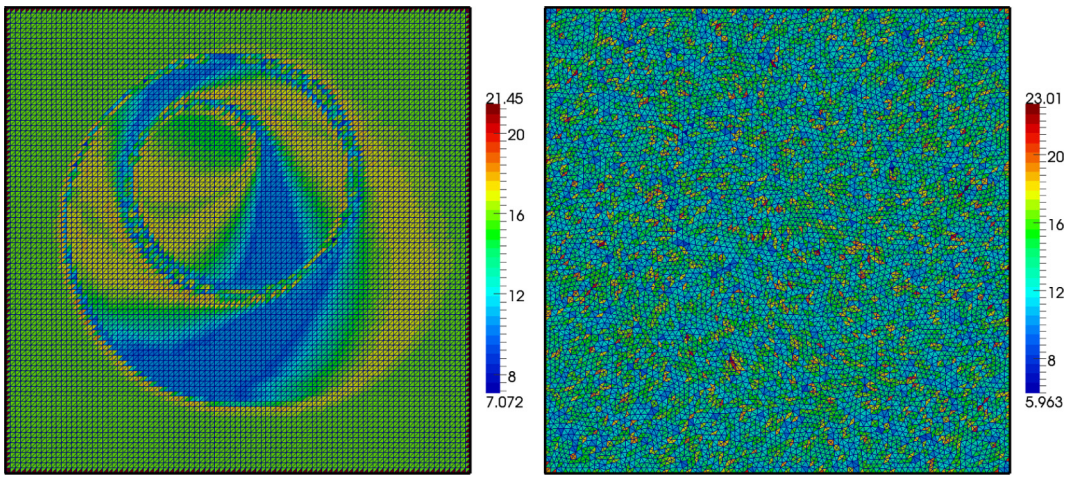


**Fig. 5.3.** First-order viscosity given by (4.11) for the KPP problem at $t = 1$. Left panel: uniform grid. Right panel: unstructured grid.

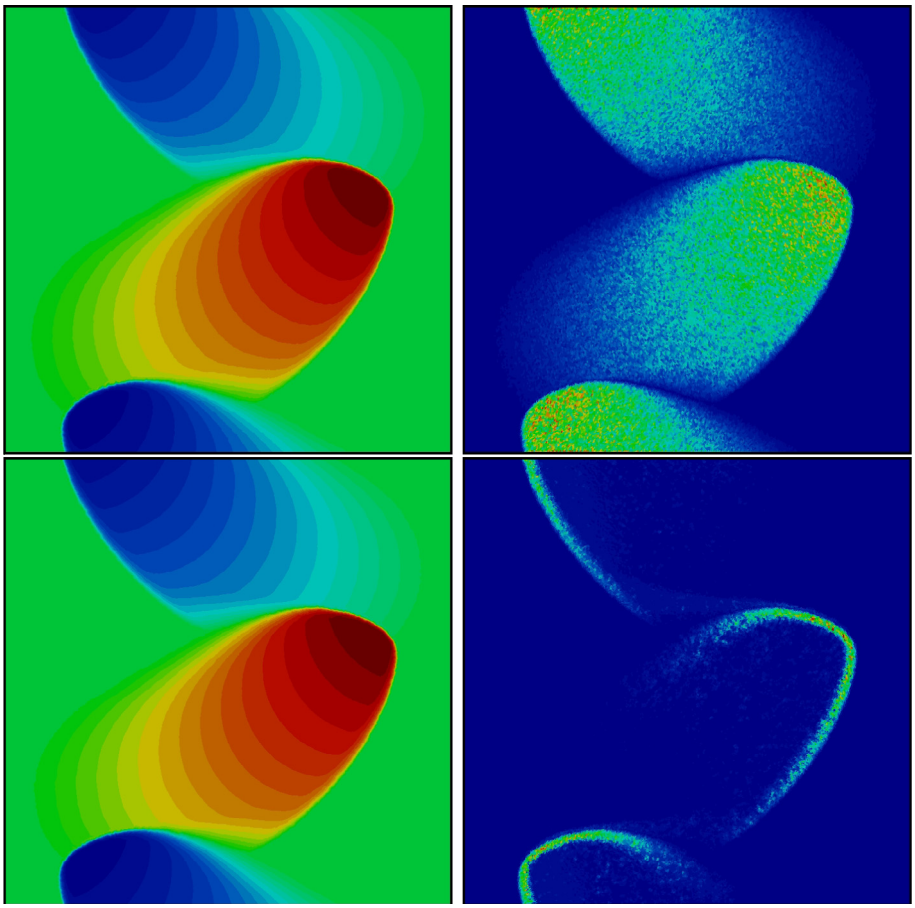

**Fig. 5.4.** Two wave problem at time $t = 2$ on an unstructured triangular grid composed 64720 $\mathbb{P}_1$ nodes. First row: first-order approximation, solution $-0.69 \leqslant u \leqslant 0.69$, viscosity coefficient $0 \leqslant v \leqslant 79.89$. Second row: entropy viscosity approximation; solution $-0.712 \leqslant u \leqslant 0.711$; viscosity coefficient $0 \leqslant v \leqslant 71.62$, $c_E = 1$, $c_J = 4$.

$$u(x,y,t) = \begin{cases} -0.2 & \text{if } x < \frac{1}{2} - \frac{3t}{5} \text{ and } \begin{cases} y > \frac{1}{2} + \frac{3t}{20}, \\ \text{otherwise}, \end{cases} \\ 0.5 & \\ -1 & \text{if } \frac{1}{2} - \frac{3t}{5} < x < \frac{1}{2} - \frac{t}{4} \text{ and } \begin{cases} y > -\frac{8x}{7} + \frac{15}{14} - \frac{15t}{28}, \\ \text{otherwise}, \end{cases} \\ 0.5 & \\ -1 & \text{if } \frac{1}{2} - \frac{t}{4} < x < \frac{1}{2} + \frac{t}{2} \text{ and } \begin{cases} y > \frac{x}{6} + \frac{5}{12} - \frac{5t}{24}, \\ \text{otherwise}, \end{cases} \\ 0.5 & \\ -1 & \text{if } \frac{1}{2} + \frac{t}{2} < x < \frac{1}{2} + \frac{4t}{5} \text{ and } \begin{cases} y > x - \frac{5}{18t}\left(x + t - \frac{1}{2}\right)^2, \\ \text{otherwise}, \end{cases} \\ \frac{2x-1}{2t} & \\ -1 & \text{if } x > \frac{1}{2} + \frac{4t}{5} \text{ and } \begin{cases} y > \frac{1}{2} - \frac{t}{10}, \\ \text{otherwise}. \end{cases} \\ 0.8 & \end{cases} \tag{5.3}$$

Convergence tests are presented in Table 5.1. We compare the convergence rates in the $L^1$- and $L^2$-norms for the entropy viscosity and first-order methods in the left and right panels of the table, respectively. Note that the rate of convergence in the $L^1$-norm is close to optimality for the entropy viscosity method, whereas it is suboptimal for the fist-order viscosity, as expected. We have verified numerically that the first-order viscosity method satisfies the local maximum principle up to round off errors at every time step.

Fig. 5.1 presents the solution (left panels) and the viscosity coefficient $\nu_K$ (right panel) at the final time. The figures in the top row are the results of the first-order viscosity method, and the those in the bottom row are the results of the entropy viscosity method using definition (4.18). The entropy for this problem is chosen to be: $E(u) = \frac{1}{2}u^2$. The control parameters are $c_E = 1$, $c_J = 4$. The computations are done with CFL = 0.2 on an unstructured triangular mesh consisting of 36826 $\mathbb{P}_1$ nodes. We observe that the entropy viscosity technique adds dissipation only in the shock regions.

### 5.2. 2D KPP rotating wave

In this section we solve so called KPP rotating wave problem, a 2D scalar nonlinear conservation laws

$$\partial_t u + \nabla\cdot \boldsymbol{f}(u) = 0, \quad u(\boldsymbol{x},0) = u_0(\boldsymbol{x}) = \begin{cases} \frac{14\pi}{4}, & \text{if } \sqrt{x^2+y^2} \leqslant 1, \\ \frac{\pi}{4}, & \text{otherwise.} \end{cases}, \tag{5.4}$$

where $\boldsymbol{f}(u) = (\sin u, \cos u)$. This test was originally proposed in [19]. It is challenging to many high-order numerical schemes because the solution has a two-dimensional composite wave structure. For example central-upwind schemes based on WENO5, Minmod 2 and SuperBee reconstructions converge to non-entropic solutions; see [19] for details.

The computation is done in the square $[-2,2]\times[-2.5,1.5]$, with CFL $= 0.2$. Fig. 5.2 shows the solution (left panels) computed on an unstructured triangular grid composed of 52964 $\mathbb{P}_1$ nodes. The magnitude of the viscosity coefficient is shown in the right panels. The figures in the top row are the results of the first-order viscosity method, and the those in the bottom row are the results of the entropy viscosity method using definition (4.18) and the entropy $E(u) = \frac{1}{2}u^2$. The parameters for this computations are $c_E = 1$ and $c_J = 4$. We have verified numerically that the first-order viscosity method satisfies the local maximum principle up to round off errors provided the integral $\int_{S_{ij}} (\boldsymbol{f}'(u_h^k)\cdot\nabla\varphi_j)\,d\boldsymbol{x}$ in the definition of the viscosity (see (4.11)) is evaluated exactly. In particular, replacing $\boldsymbol{f}'(u_h^k)$ by its Lagrange interpolant leads to slight violations of the local maximum principle.

The focusing of the entropy viscosity in the shock region is striking. Observe also that using an unstructured grid makes the first-order viscosity coefficient $\nu_K$ to appear like a random field (see top right panel in Fig. 5.2). To emphasize that the viscosity coefficient $\nu_K$ given in (4.11) does indeed encode the meshsize information, we have redone the above computation using a structured grid composed of $91\times 91$ square cells divided into two triangles and an unstructured Delaunay grid composed of triangles of similar sizes. The structure grid has 8464 $\mathbb{P}_1$ nodes and the unstructured grid has 8560 $\mathbb{P}_1$ nodes. The resulting viscosity fields $\nu_K$ are shown in the Fig. 5.3. The differences between the viscosity fields in the left and right panels are just due to the differences in the mesh structure.

### 5.3. Two wave test

We finally consider a two-dimensional scalar conservation equation studied in [16]:

$$\partial_t u + \partial_x\left(\frac{1}{2}u^2\right) + \partial_y\left(\frac{1}{3}u^3\right) = 0, \quad u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), \tag{5.5}$$

with periodic boundary conditions and the following initial condition:

$$u_0 = \begin{cases} -1, & \text{if } (x-0.5)^2 + (y-0.5)^2 < 0.16, \\ 1, & \text{if } (x+0.5)^2 + (y+0.5)^2 < 0.16, \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

This problem has a fully two dimensional structure; there are two shock waves traveling in different directions. Fig. 5.4 shows the solution (left panels) computed on an unstructured triangular grid composed of 64720 $\mathbb{P}_1$ nodes. The magnitude of the viscosity coefficient is shown in the right panels. The figures in the top row are the results of the first-order viscosity method, and those in the bottom row are the results of the entropy viscosity method using definition (4.18) and the entropy $E(u) = \frac{1}{2}u^2$. The parameters for this computations are $c_E = 1$ and $c_J = 4$. The CFL number is 0.3.

## Acknowledgments

The authors acknowledge fruitful discussions with Wolfgang Bangerth, Bojan Popov, and Raytcho Lazarov.

## References

[1] C. Bardos, A.Y. le Roux, J.-C. Nédélec, First order quasilinear equations with boundary conditions, Commun. Partial Diff. Equat. 4 (9) (1979) 1017–1034.
[2] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. 11 (1973), no. 1, 38–69]. J. Comput. Phys. 135 (2) (1997) 170–186. With an introduction by Steven T. Zalesak, Commemoration of the 30th anniversary of J. Comput. Phys.
[3] E. Burman, A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion-reaction equation, Comput. Methods Appl. Mech. Engrg. 191 (35) (2002) 3833–3855.
[4] E. Burman, A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence, Math. Comput. 74 (252) (2005) 1637–1652 (electronic).
[5] J.C. Campbell, M.J. Shashkov, A tensor artificial viscosity using a mimetic finite difference algorithm, J. Comput. Phys. 172 (2) (2001) 739–765.
[6] M.A. Christon, The influence of the mass matrix on the dispersive nature of the semi-discrete, second-order wave equation, Comput. Methods Appl. Mech. Engrg. 173 (12) (1999) 147–166.
[7] M.A. Christon, M.J. Martinez, T.E. Voth, Generalized fourier analyses of the advection-diffusion equation-part i: one-dimensional domains, Int. J. Numer. Methods Fluids 45 (8) (2004) 839–887.
[8] R. Codina, A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation, Comput. Methods Appl. Mech. Engrg. 110 (3–4) (1993) 325–342.

[9] V.A. Dobrev, T.V. Kolev, R.N. Rieben, High-order curvilinear finite element methods for Lagrangian hydrodynamics, SIAM J. Sci. Comput. 34 (5) (2012) B606–B641.
[10] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM Rev. 43 (1) (2001) 89–112 (electronic).
[11] P. Gresho, R. Sani, M. Engelman, Incompressible flow and the finite element method: advection-diffusion and isothermal laminar flow, Incompressible Flow & the Finite Element Method, Wiley, 1998.
[12] J.-L. Guermond, M. Nazarov, B. Popov, Y. Yang, A second-order maximum principle preserving lagrange finite element technique for nonlinear scalar conservation equations, SIAM J. Numer. Anal., 2013, submitted for publication.
[13] J.-L. Guermond, R. Pasquetti, Entropy-based nonlinear viscosity for Fourier approximations of conservation laws, C.R. Math. Acad. Sci. Paris 346 (13–14) (2008) 801–806.
[14] J.-L. Guermond, R. Pasquetti, A correction technique for the dispersive effects of mass lumping for transport problems, Comput. Methods Appl. Mech. Engrg. 253 (2013) 186–198.
[15] J.-L. Guermond, R. Pasquetti, B. Popov, Entropy viscosity method for nonlinear conservation laws, J. Comput. Phys. 230 (11) (2011) 4248–4267.
[16] H. Holden, K.H. Karlsen, K.-A. Lie, N.H. Risebro, Splitting methods for partial differential equations with rough solutions, EMS Series of Lectures in Mathematics, European Mathematical Society (EMS), Zürich, 2010. Analysis and MATLAB programs.
[17] G.-S. Jiang, E. Tadmor, Nonoscillatory central schemes for multidimensional hyperbolic conservation laws, SIAM J. Sci. Comput. 19 (6) (1998) 1892–1917 (electronic).
[18] S.N. Kružkov, First order quasilinear equations with several independent variables, Mat. Sb. (N.S.) 81 (123) (1970) 228–255.
[19] A. Kurganov, G. Petrova, B. Popov, Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws, SIAM J. Scient. Comput. 29 (6) (2007) 2381–2401.
[20] D. Kuzmin, On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection, J. Comput. Phys. 219 (2) (2006) 513–531.
[21] P.D. Lax, Weak solutions of nonlinear hyperbolic equations and their numerical computation, Commun. Pure Appl. Math. 7 (1954) 159–193.
[22] X.-D. Liu, A maximum principle satisfying modification of triangle based adapative stencils for the solution of scalar hyperbolic conservation laws, SIAM J. Numer. Anal. 30 (3) (1993) 701–716.
[23] R. Löhner, Applied CFD Techniques, second ed., J. Wiley & Sons, 2008.
[24] A. Mizukami, T.J.R. Hughes, A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle, Comput. Methods Appl. Mech. Engrg. 50 (2) (1985) 181–193.
[25] R. Sanders, A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws, Math. Comput. 51 (184) (1988) 535–558.
[26] M. Tabata, A theoretical and computational study of upwind-type finite element methods, in: Patterns and waves, vol. 18 of Stud. Math. Appl., North-Holland, Amsterdam, 1986, pp. 319–356.
[27] T. Tezduyar, S. Sathe, Stabilization parameters in SUPG and PSPG formulations, J. Comput. Appl. Mech. 4 (1) (2003) 71–88 (electronic).
[28] T.E. Tezduyar, Y. Osawa, Finite element stabilization parameters computed from element matrices and vectors, Comput. Methods Appl. Mech. Engrg. 190 (3 4) (2000) 411–430.
[29] T.E. Voth, M.J. Martinez, M.A. Christon, Generalized fourier analyses of the advection diffusion equation part ii: two-dimensional domains, Int. J. Numer. Methods Fluids 45 (8) (2004) 889–920.
[30] J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comput. 68 (228) (1999) 1429–1446.
[31] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, J. Comput. Phys. 31 (3) (1979) 335–362.
[32] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. 229 (9) (2010) 3091–3120.
[33] X. Zhang, 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 (2134) (2011) 2752–2776.
[34] X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes, J. Sci. Comput. 50 (1) (2012) 29–62.
[35] Y. Zhang, X. Zhang, C.-W. Shu, Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes, J. Comput. Phys. 234 (2013) 295–316.