SIAM J. NUMER. ANAL.
Vol. 54, No. 4, pp. 2466–2489

INVARIANT DOMAINS AND FIRST-ORDER CONTINUOUS FINITE ELEMENT APPROXIMATION FOR HYPERBOLIC SYSTEMS*

JEAN-LUC GUERMOND† AND BOJAN POPOV†

Abstract. We propose a numerical method for solving general hyperbolic systems in any space dimension using forward Euler time stepping and continuous finite elements on nonuniform grids. The properties of the method are based on the introduction of an artificial dissipation that is defined so that any convex invariant set containing the initial data is an invariant domain for the method. The invariant domain property is proved for any hyperbolic system provided a CFL condition holds. The solution is also shown to satisfy a discrete entropy inequality for every admissible entropy of the system. The method is formally first-order accurate in space and can be made high-order in time by using strong stability preserving algorithms. This technique extends to continuous finite elements the work of [D. Hoff, *Math. Comp.,* 33 (1979), pp. 1171–1193], [D. Hoff, *Trans. Amer. Math. Soc.,* 289 (1985), pp. 591–610], and [H. Frid, *Arch. Ration. Mech. Anal.,* 160 (2001), pp. 245–269].

Key words. conservation equations, hyperbolic systems, parabolic regularization, invariant domain, first-order method, finite element method

AMS subject classifications. 65M60, 65M12, 35L45, 35L65

DOI. 10.1137/16M1074291

1. Introduction.

The objective of this paper is to propose a first-order approximation technique for general nonlinear hyperbolic systems using continuous finite elements of arbitrary polynomial degree and explicit time stepping on nonuniform meshes in any space dimension. The method is invariant domain preserving and satisfies a local entropy inequality for every admissible entropy pair for any hyperbolic system on any unstructured meshes.

Consider the following hyperbolic system in conservation form:

$$\begin{cases} \partial_t \boldsymbol{u} + \nabla\cdot\boldsymbol{f}(\boldsymbol{u}) = 0 \quad \text{for } (\boldsymbol{x},t) \in \mathbb{R}^d\times\mathbb{R}_+, \\ \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}) \quad \text{for } \boldsymbol{x}\in\mathbb{R}^d, \end{cases} \tag{1.1}$$

where the dependent variable $\boldsymbol{u}$ takes values in $\mathbb{R}^m$ and the flux $\boldsymbol{f}$ takes values in $(\mathbb{R}^m)^d$. In this paper $\boldsymbol{u}$ is considered as a column vector $\boldsymbol{u}=(u_1,\dots,u_m)^{\mathsf{T}}$. The flux is a matrix with entries $f_{ij}(\boldsymbol{u})$, $1\le i\le m$, $1\le j\le d$, and $\nabla\cdot\boldsymbol{f}$ is a column vector with entries $(\nabla\cdot\boldsymbol{f})_i=\sum_{1\le j\le d}\partial_{x_j}f_{ij}$. For any $\boldsymbol{n}=(n_1,\dots,n_d)^{\mathsf{T}}\in\mathbb{R}^d$, we denote by $\boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n}$ the column vector with entries $\sum_{1\le l\le d}n_l f_{il}(\boldsymbol{u})$, where $i\in\{1{:}m\}$. The unit sphere in $\mathbb{R}^d$ centered at 0 is denoted by $S^{d-1}(\mathbf{0},1)$.

To simplify questions regarding boundary conditions, we assume that either periodic boundary conditions are enforced, or the initial data is compactly supported or constant outside a compact set. In both cases we denote by D the spatial domain where the approximation is constructed. The domain D is the d-torus in the case of periodic boundary conditions. In the case of the Cauchy problem, D is a compact,

*Received by the editors September 25, 2015; accepted for publication (in revised form) May 23, 2016; published electronically August 16, 2016.
http://www.siam.org/journals/sinum/54-4/M107429.html
Funding: The research of the authors was supported in part by National Science Foundation grant DMS-1217262; by Air Force Office of Scientific Research, USAF, grant/contract FA99550-12-0358; and by Army Research Office 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).

polygonal portion of $\mathbb{R}^d$ large enough so that the domain of influence of $\boldsymbol{u}_0$ is always included in D over the entire duration of the simulation.

The method that we propose is explicit in time and uses continuous finite elements on nonuniform grids in any space dimension. The algorithm is described in section 3.2; see (3.9) with definitions (3.5)-(3.8)-(3.16). It is a somewhat loose adaptation of the nonstaggered Lax–Friedrichs scheme to continuous finite elements. The key results of the paper are Theorems 4.1 and 4.7. It is shown in Theorem 4.1 that the proposed scheme preserves all the convex invariant sets as defined in Definition 2.3, and it is shown in Theorem 4.7 that the approximate solution satisfies a discrete entropy inequality for every entropy pair of the hyperbolic system. Similar results have been established by Hoff [13, 14], Perthame and Shu [27], and Frid [9] for the compressible Euler equations and the p-system using various finite volumes schemes based on piecewise constant approximation. Our scheme has no restriction on the nature of the hyperbolic system, besides the speed of propagation being finite, and can use continuous finite elements of arbitrary polynomial degree on unstructured meshes in any space dimension. To the best of our knowledge, we are not aware of any similar scheme in the continuous finite element literature. The proposed method is meant to be a stepping stone for the construction of higher-order continuous finite element methods by using, for instance, the flux transport correction methodology à la Boris, Book, and Zalesak, or any generalization thereof, to implement limitation on the artificial viscosity. None of these generalizations is discussed in this paper. Our sole objective is to prepare the groundwork for future research on limitation by giving a firm, indisputable, theoretical footing for first-order, invariant-domain preserving methods using continuous finite elements on unstructured grids in arbitrary space dimension for every hyperbolic system.

The paper is organized as follows. The notions of invariant sets and invariant domains with various examples and other preliminaries are introduced in section 2. The method is introduced in section 3. Stability properties of the algorithm are analyzed in section 4. Numerical illustrations and comparisons with existing first-order methods are presented in section 5.

2. Preliminaries. The objective of this section is to introduce notation and preliminary results that will be useful in the rest of the paper. We mostly use the notation and terminology of Chueh, Conley, and Smoller [5], Hoff [13, 14], and Frid [9]. The reader who is familiar with the notions of invariant domains and Riemann problems may skip this section and go directly to section 3, although the reader should be aware that our definitions of invariant sets and domains are slightly different from those of [5, 13, 14, 9].

2.1. Riemann problem. We assume that (1.1) is such that there is a clear notion for the solution of the Riemann problem. That is, there exists a (nonempty) admissible set $\mathcal{A} \subset \mathbb{R}^m$ such that for any pair of states $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A}\times\mathcal{A}$ and any unit vector $\boldsymbol{n} \in S^{d-1}(\mathbf{0}, 1)$, the one-dimensional Riemann problem

$$\partial_t \boldsymbol{u} + \partial_x(\boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = 0, \quad (x,t) \in \mathbb{R}\times\mathbb{R}_+, \qquad \boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_L & \text{if } x < 0, \\ \boldsymbol{u}_R & \text{if } x > 0 \end{cases} \tag{2.1}$$

has a unique (physical) solution, which we henceforth denote $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$.

The theory of the Riemann problem for general nonlinear hyperbolic systems with data far apart is an open problem. Moreover, it is unrealistic to expect a general theory for every system with arbitrary initial data. However, when the system is strictly

hyperbolic with smooth flux and all the characteristic fields are either genuinely nonlinear or linearly degenerate, it is possible to show that there exists $\delta > 0$ such that the Riemann problem has a unique self-similar weak solution in Lax's form for any initial data such that $\|\boldsymbol{u}_L - \boldsymbol{u}_R\|_{\ell^2} \le \delta$; see Lax [21] and Bressan [3, Thm. 5.3]. In particular, there are $2m$ numbers

$$\lambda_1^- \le \lambda_1^+ \le \lambda_2^- \le \lambda_2^+ \le \cdots \le \lambda_m^- \le \lambda_m^+ \tag{2.2}$$

defining up to $2m+1$ sectors (some could be empty) in the (x,t) plane:

$$\frac{x}{t} \in (-\infty, \lambda_1^-), \quad \frac{x}{t} \in (\lambda_1^-, \lambda_1^+), \ldots, \quad \frac{x}{t} \in (\lambda_m^-, \lambda_m^+), \quad \frac{x}{t} \in (\lambda_m^+, \infty). \tag{2.3}$$

The Riemann solution is $\boldsymbol{u}_L$ in the sector $\frac{x}{t} \in (-\infty, \lambda_1^-)$, and $\boldsymbol{u}_R$ in the last sector $\frac{x}{t} \in (\lambda_m^+, \infty)$. The solution in the other sectors is either a constant state or an expansion; see Bressan [3, Chap. 5]. The sector $\lambda_1^- t < x < \lambda_m^+ t$, $0 < t$, is henceforth referred to as the Riemann fan. The key result that we are going to use is that there is a maximum speed of propagation $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \max(|\lambda_1^-|, |\lambda_m^+|)$ such that for $t \ge 0$ we have

$$\boldsymbol{u}(x,t) = \begin{cases} \boldsymbol{u}_L & \text{if } x \le -t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R), \\ \boldsymbol{u}_R & \text{if } x \ge t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R). \end{cases} \tag{2.4}$$

Actually, even if the above structure of the Riemann solution is not available or valid, we henceforth make the following assumption:

(2.5) The unique solution of (2.1) has a finite speed of propagation for any $\boldsymbol{n}$; i.e., there is $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ such that (2.4) holds.

For instance, this is the case for strictly hyperbolic systems that may have characteristic families that are either not genuinely nonlinear or not linearly degenerate; see, e.g., Liu [24, Thm. 1.2] and Dafermos [8, Thm. 9.5.1]. We refer the reader to Osher [26, Thm. 1] for the theory of the Riemann problem for scalar conservation equations with nonconvex fluxes. In the case of general hyperbolic systems, we refer the reader to Bianchini and Bressan [2, section 14] for characterizations of the Riemann solution using viscosity regularization. We also refer the reader to Young [30, Thm. 2] for the theory of the Riemann problem for the p-system with arbitrary data (i.e., with possible formation of vacuum).

The following elementary result is an important, well-known consequence of (2.4); i.e., the Riemann solution is equal to $\boldsymbol{u}_L$ for $x \in (-\infty, \lambda_1^- t)$ and equal to $\boldsymbol{u}_R$ for $x \in (\lambda_m^+ t, \infty)$.

LEMMA 2.1. *Let $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathcal{A}$, let $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ be the Riemann solution to* (2.1)*, let $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x$, and assume that $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \frac{1}{2}$; then*

$$\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) = \frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R) - t\big(\boldsymbol{f}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{f}(\boldsymbol{u}_L)\cdot\boldsymbol{n}\big). \tag{2.6}$$

If the system (1.1) has an entropy pair $(\eta, \boldsymbol{q})$, and if the Riemann solution is defined to be entropy satisfying, i.e., if the following holds:

$$\partial_t \eta(\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)) + \partial_x\big(\boldsymbol{q}(\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R))\cdot\boldsymbol{n}\big) \le 0 \tag{2.7}$$

in some appropriate sense (distribution sense, measure sense, etc.), then we have the following additional result.

LEMMA 2.2. *Let $(\eta, \boldsymbol{q})$ be an entropy pair for* (1.1)*, and assume that* (2.7) *holds. Let $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathcal{A}$, and let $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ be the Riemann solution to* (2.1)*. Assume that $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \frac{1}{2}$. Then*

$$\eta(\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)) \le \frac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) - t(\boldsymbol{q}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{u}_L)\cdot\boldsymbol{n}). \tag{2.8}$$

Proof. Under the CFL assumption $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \frac{1}{2}$, the inequality (2.7) implies that

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R))(x,t)\,\mathrm{d}x \le \frac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) - t(\boldsymbol{q}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{u}_L)\cdot\boldsymbol{n}). \tag{2.9}$$

Jensen's inequality $\eta(\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)) \le \int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t))\,\mathrm{d}x$ then implies the desired result. ∎

2.2. Invariant sets and domains. We introduce in this section the notions of invariant sets and invariant domains. Our definitions are slightly different from those in Chueh, Conley, and Smoller [5], Hoff [14], Smoller [28], and Frid [9]. We will associate invariant sets only with solutions of Riemann problems and define invariant domains only for an approximation process.

DEFINITION 2.3 (invariant set). *We say that a set $B \subset \mathcal{A} \subset \mathbb{R}^m$ is invariant for* (1.1) *if for any pair $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in B \times B$, any unit vector $\boldsymbol{n} \in \mathcal{S}^{d-1}(\boldsymbol{0}, 1)$, and any $t > 0$, the average of the entropy solution of the Riemann problem* (2.1) *over the Riemann fan, say, $\frac{1}{t(\lambda_m^+ - \lambda_1^-)} \int_{\lambda_1^- t}^{\lambda_m^+ t} \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x$, remains in B.*

Note that if in addition we assume that B is convex, then the above definition implies that given $t > 0$ and any interval I such that $(\lambda_1^- t, \lambda_m^+ t) \subset I$, we have that $\frac{1}{I} \int_I \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x \in B$. Note also that most of the time expansion waves and shocks are not invariant sets.

We now introduce the notion of invariant domain for an approximation process. Let $\boldsymbol{X}_h \subset L^1(\mathbb{R}^d; \mathbb{R}^m)$ be a finite-dimensional approximation space, and let $S_h : \boldsymbol{X}_h \ni \boldsymbol{u}_h \longmapsto S_h(\boldsymbol{u}_h) \in \boldsymbol{X}_h$ be a discrete process over $\boldsymbol{X}_h$. Henceforth we abuse notation by saying that a member of $\boldsymbol{X}_h$, say $\boldsymbol{u}_h$, is in the set $B \subset \mathbb{R}^m$ when actually we mean that $\{\boldsymbol{u}_h(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\} \subset B$.

DEFINITION 2.4 (invariant domain). *A convex invariant set $B \subset \mathcal{A} \subset \mathbb{R}^m$ is said to be an invariant domain for the process S_h if and only if for any state $\boldsymbol{u}_h$ in B, the state $S_h(\boldsymbol{u}_h)$ is also in B.*

For scalar conservation equations the notions of invariant sets and invariant domains are closely related to the maximum principle; see Example 1 in section 2.3. In the case of nonlinear systems, the notion of maximum principle does not apply and must be replaced by the notion of invariant domain. To the best of our knowledge, the definition of invariant sets for the Riemann problem was introduced in Nishida [25], and the general theory of positively invariant regions was developed in Chueh, Conley, and Smoller [5]. Applications and extensions to numerical methods were developed in Hoff [13, 14] and Frid [9].

The invariant domain theory when $m = 2$ and $d = 1$ relies on the existence of global Riemann invariants; the best known examples are the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian form; see Example 2 in section 2.4 and Lions, Perthame, and Souganidis [22]. For results on general hyperbolic systems,

we refer the reader to Frid [9], where a characterization of invariant domains for the Lax–Friedrichs scheme and some flux splitting schemes is given. In particular the existence of invariant domains is established for the above-mentioned schemes for the compressible Euler equations in the general case $m = d+2$ (positive density, internal energy, and minimum principle on the specific entropy); see Frid [9, Thms. 7 and 8]. Similar results have been established for various finite volume schemes in two-space dimension for the Euler equations in Perthame and Shu [27, Thm. 3].

The objective of this paper is to propose an explicit numerical method based on continuous finite elements to approximate (1.1) such that any convex invariant set of (1.1) is an invariant domain for the process generated by the said numerical method.

To facilitate the reading of the paper we now illustrate the abstract notions of invariant sets and invariant domains with some examples.

2.3. Example 1: Scalar equations. Assume that $m = 1$ and d is arbitrary; i.e., (1.1) is a scalar conservation equation. Provided $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$, any bounded interval is an admissible set for (1.1). For any Riemann data u_L, u_R, the maximum speed of propagation in (2.4) is bounded by $\lambda_{\max}(u_L,u_R) := \|\boldsymbol{f}\cdot\boldsymbol{n}\|_{\mathrm{Lip}(u_{\min},u_{\max})}$, where $u_{\min} = \min(u_L,u_R)$, $u_{\max} = \max(u_L,u_R)$. If $\boldsymbol{f}$ is convex and is of class C^1, we have $\lambda_{\max}(u_L,u_R) = \max(|\boldsymbol{n}\cdot\boldsymbol{f}'(u_L)|,|\boldsymbol{n}\cdot\boldsymbol{f}'(u_R)|)$ if $\boldsymbol{n}\cdot\boldsymbol{f}'(u_L) \le \boldsymbol{n}\cdot\boldsymbol{f}'(u_R)$ and $\lambda_{\max}(u_L,u_R) = \boldsymbol{n}\cdot(\boldsymbol{f}(u_L)-\boldsymbol{f}(u_R))/(u_L-u_R)$ otherwise. Any interval $[a,b] \subset \mathbb{R}$ is admissible and is an invariant set for (1.1), i.e., if $u_R, u_L \in [a,b]$, then $a \le u(\boldsymbol{n},u_L,u_R) \le b$ for all times; this is the maximum principle. For any $a \le b \in \mathbb{R}$, the interval $[a,b]$ is an invariant domain for any maximum principle satisfying numerical schemes. Note that the maximum principle can be established for a large number of numerical methods (whether monotone or not); see, for example, Crandall and Majda [7].

2.4. Example 2: p-system. The one-dimensional motion of an isentropic gas is modeled by the so-called p-system, and in Lagrangian coordinates the system is written as follows:

$$\begin{cases} \partial_t v + \partial_x u = 0, \\ \partial_t u + \partial_x p(v) = 0 \quad \text{for } (x,t) \in \mathbb{R}\times\mathbb{R}_+. \end{cases} \tag{2.10}$$

Here $d = 1$ and $m = 2$. The dependent variables are the velocity u and the specific volume v, i.e., the reciprocal of density. The mapping $v \mapsto p(v)$ is the pressure and assumed to be of class $C^2(\mathbb{R}_+;\mathbb{R})$ and to satisfy

$$p' < 0, \qquad 0 < p''. \tag{2.11}$$

A typical example is the so-called gamma-law, $p(v) = rv^{-\gamma}$, where $r > 0$ and $\gamma \ge 1$. Using the notation $\boldsymbol{u} = (v,u)^{\mathsf{T}}$, any set $\mathcal{A}$ in $(0,\infty)\times\mathbb{R}$ is admissible.

Using the notation $\mathrm{d}\mu := \sqrt{-p'(s)}\,\mathrm{d}s$, and assuming $\int_1^\infty \mathrm{d}\mu < \infty$, the system has two families of global Riemann invariants:

$$w_1(\boldsymbol{u}) = u + \int_v^\infty \mathrm{d}\mu \quad \text{and} \quad w_2(\boldsymbol{u}) = u - \int_v^\infty \mathrm{d}\mu. \tag{2.12}$$

Note that $\int_1^\infty \mathrm{d}\mu < \infty$ if $\gamma > 1$. If $\gamma = 1$, we can use $w_1(\boldsymbol{u}) = u - \sqrt{r}\log v$ and $w_2(\boldsymbol{u}) = u + \sqrt{r}\log v$. Let $a,b \in \mathbb{R}$; then it can be shown that any set $A_{ab} \in \mathbb{R}_+\times\mathbb{R}$ of the form

$$A_{ab} := \{\boldsymbol{u} \in \mathbb{R}_+\times\mathbb{R} \mid a \le w_2(\boldsymbol{u}),\ w_1(\boldsymbol{u}) \le b\} \tag{2.13}$$

is an invariant set for the system (2.10) for $\gamma \geq 1$; see Hoff [14, Exp. 3.5, p. 597] for a proof in the context of parabolic regularization, or use the results from Young [30] for a direct proof. Moreover, A_{ab} is an invariant domain for the Lax–Friedrichs scheme; see Hoff [13, Thm. 2.1] and Hoff [14, Thm. 4.1].

Since in the rest of the paper the maximum wave speed is the only information we are going to need from the Riemann solution, we give the following result.

LEMMA 2.5. *Let* $(v_L, u_L), (v_R, u_R) \in \mathbb{R}_+ \times \mathbb{R}$ *with* $v_R, v_L < \infty$. *Then*

$$\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) = \begin{cases} \sqrt{-p'(\min(v_L, v_R))} & \text{if } u_L - u_R > \sqrt{(v_L - v_R)(p(v_R) - p(v_L))}, \\ \sqrt{-p'(v^*)} & \text{otherwise,} \end{cases}$$

where v^* *is the unique solution of* $\phi(v) := f_L(v) + f_R(v) + u_L - u_R = 0$ *and*

$$f_Z(v) := \begin{cases} -\sqrt{(p(v) - p(v_Z)(v_Z - v)} & \text{if } v \leq v_Z, \\ \displaystyle\int_{v_Z}^{v} \mathrm{d}\mu & \text{if } v > v_Z. \end{cases}$$

Upon setting $w_1^{\max} := \max(w_1(\boldsymbol{u}_L), w_1(\boldsymbol{u}_R))$ *and* $w_2^{\min} := \min(w_2(\boldsymbol{u}_L), w_2(\boldsymbol{u}_R))$ *we also have* $v^0 \leq \min(v_L, v_R, v^*)$, *i.e.,* $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) \leq \sqrt{-p'(v^0)}$, *where*

$$v^0 := (\gamma r)^{\frac{1}{\gamma-1}} \left(\frac{4}{(\gamma-1)(w_1^{\max} - w_2^{\min})} \right)^{\frac{2}{(\gamma-1)}}.$$

Proof. It is well known that the solution of the Riemann problem consists of three constant states $\boldsymbol{u}_L$, $\boldsymbol{u}^*$, and $\boldsymbol{u}_R$ connected by two waves: a 1-wave connects $\boldsymbol{u}_L$ and $\boldsymbol{u}^*$, and a 2-wave connects $\boldsymbol{u}^*$ and $\boldsymbol{u}_R$. Moreover, a vacuum forms if and only if $\lim_{v\to+\infty} \phi(v) \geq 0$; see Young [30] for details. In the presence of vacuum, the equation $\phi(v) = 0$ has no solutions, and in this case we conventionally set $v^* := +\infty$ and $\sqrt{-p'(v^*)} := 0$. Note that since ϕ is an increasing and concave function with $\lim_{v\to 0+} \phi(v) = -\infty$, the solution v^* is unique. We also have that the maximum speed of the exact solution is $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) = \max(\sqrt{-p'(v_L)}, \sqrt{-p'(v^*)}, \sqrt{-p'(v_R)})$. The only possibility for $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) = \sqrt{-p'(v^*)}$ is if $v^* \leq \min(v_L, v_R)$, i.e., the solution contains two shock waves, which is equivalent to $\phi(\min(v_L, v_R)) \geq 0$. Using the definition of ϕ we derive that $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) = \sqrt{-p'(v^*)}$ if and only if $\phi(\min(v_L, v_R)) = u_L - u_R - \sqrt{(v_L - v_R)(p(v_R) - p(v_L))} \geq 0$. This finishes the proof of the first part of the lemma.

The exact value of v^* can be found using Newton's method starting with a guess $v^0 \leq v^*$. This guarantees that at each step of Newton's method the estimated maximum speed is an upper bound for the exact maximum speed. One can obtain such a guess v^0 by using the invariant domain property (2.13); i.e., we define the state $\boldsymbol{u}^0 := (v^0, u^0)$ by $w_1^{\max} = w_1(\boldsymbol{u}^0)$ and $w_2^{\min} = w_2(\boldsymbol{u}^0)$, thereby giving

$$v^0 = (\gamma r)^{\frac{1}{\gamma-1}} \left(\frac{4}{(\gamma-1)(w_1^{\max} - w_2^{\min})} \right)^{\frac{2}{(\gamma-1)}}.$$

The invariant domain property guarantees that $v^0 \leq v^*$. Hence, the result is established. ∎

Remark 2.6. Note that the estimate on $\lambda_{\max}(\boldsymbol{u}_L, \boldsymbol{u}_R)$ given in Lemma 2.5 is valid whether or not vacuum is created in the Riemann solution.

Remark 2.7. We only consider the case where both $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ are not vacuum states in Lemma 2.5, since the algorithm that we propose never produces vacuum states if vacuum is not present in the initial data.

2.5. Example 3: Euler. Consider the compressible Euler equations

$$\partial_t \boldsymbol{c} + \nabla\cdot(\boldsymbol{f}(\boldsymbol{c})) = 0, \quad \boldsymbol{c} = \begin{pmatrix} \rho \\ \boldsymbol{m} \\ E \end{pmatrix}, \quad \boldsymbol{f}(\boldsymbol{c}) = \begin{pmatrix} \boldsymbol{m} \\ \boldsymbol{m}\otimes\frac{\boldsymbol{m}}{\rho} + p\mathbb{I} \\ \frac{\boldsymbol{m}}{\rho}(E+p) \end{pmatrix}, \tag{2.14}$$

where the independent variables are the density ρ, the momentum vector field $\boldsymbol{m}$, and the total energy E. The velocity vector field $\boldsymbol{u}$ is defined by $\boldsymbol{u} := \boldsymbol{m}/\rho$ and the internal energy density e by $e := \frac{E}{\rho} - \frac{1}{2}|\boldsymbol{u}|^2$. The quantity p is the pressure. The symbol $\mathbb{I}$ denotes the identity matrix in $\mathbb{R}^d$. Let s be the specific entropy of the system, and assume that $-s(e,\rho^{-1})$ is strictly convex. It is known that

$$A_r := \{(\rho, \boldsymbol{m}, E) \mid \rho \geq 0, e \geq 0, s \geq r\} \tag{2.15}$$

is an invariant set for the Euler system for any $r \in \mathbb{R}$. It is shown in Frid [9, Thms. 7 and 8] that the set A_r is convex and is an invariant domain for the Lax–Friedrichs scheme.

Let $\boldsymbol{n} \in S^{d-1}(\boldsymbol{0},1)$, and let us formulate the Riemann problem (2.1) for the Euler equations. This problem was first described in the context of dimension splitting schemes with $d = 2$ in Chorin [4, p. 526]. The general case is treated in Colella [6, p. 188], see also Toro [29, Chap. 4.8]. We make a change of basis and introduce $\boldsymbol{t}_1, \dots, \boldsymbol{t}_{d-1}$ so that $\{\boldsymbol{n}, \boldsymbol{t}_1, \dots, \boldsymbol{t}_{d-1}\}$ forms an orthonormal basis of $\mathbb{R}^d$. With this new basis we have $\boldsymbol{m} = (m, \boldsymbol{m}^\perp)^\mathsf{T}$, where $m := \rho u$, $u := \boldsymbol{u}\cdot\boldsymbol{n}$, $\boldsymbol{m}^\perp := \rho(\boldsymbol{u}\cdot\boldsymbol{t}_1, \dots, \boldsymbol{u}\cdot\boldsymbol{t}_{d-1}) := \rho\boldsymbol{u}^\perp$. The projected equations are

$$\partial_t \boldsymbol{c} + \partial_x(\boldsymbol{n}\cdot\boldsymbol{f}(\boldsymbol{c})) = \boldsymbol{0}, \quad \boldsymbol{c} = \begin{pmatrix} \rho \\ m \\ \boldsymbol{m}^\perp \\ E \end{pmatrix}, \quad \boldsymbol{n}\cdot\boldsymbol{f}(\boldsymbol{c}) = \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + p \\ u\boldsymbol{m}^\perp \\ u(E+p) \end{pmatrix}. \tag{2.16}$$

Using the density ρ and the specific entropy s as dependent variables for the pressure, $p(\rho, s)$, the linearized Jacobian is

$$\begin{pmatrix} u & \rho & \boldsymbol{0}^\mathsf{T} & 0 \\ \rho^{-1}\partial_\rho p & u & \boldsymbol{0}^\mathsf{T} & \rho^{-1}\partial_s p \\ \boldsymbol{0} & \boldsymbol{0} & u\mathbb{I} & \boldsymbol{0} \\ 0 & 0 & \boldsymbol{0}^\mathsf{T} & u \end{pmatrix}.$$

The eigenvalues are u, with multiplicity d, $u + \sqrt{\partial_\rho p(\rho, s)}$, with multiplicity 1, and $u - \sqrt{\partial_\rho p(\rho, s)}$, with multiplicity 1. One key observation is that the Jacobian does not depend on $\boldsymbol{m}^\perp$; see Toro [29, p. 150]. As a consequence, the solution of the Riemann problem with data $(\boldsymbol{c}_L, \boldsymbol{c}_R)$ is such that (ρ, u, p) is obtained as the solution to the one-dimensional Riemann problem

$$\partial_t \begin{pmatrix} \rho \\ m \\ \mathcal{E} \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + p \\ u(\mathcal{E}+p) \end{pmatrix} = 0, \quad \text{with} \quad \rho e = \mathcal{E} - \frac{m^2}{2\rho}, \tag{2.17}$$

with data $\boldsymbol{c}_L^{\boldsymbol{n}} := (\rho_L, \boldsymbol{m}_L \cdot \boldsymbol{n}, \mathcal{E}_L)$, $\boldsymbol{c}_R^{\boldsymbol{n}} := (\rho_R, \boldsymbol{m}_R \cdot \boldsymbol{n}, \mathcal{E}_R)$, where $\mathcal{E}_Z = E_Z - \frac{1}{2}\frac{\|\boldsymbol{m}_Z^{\perp}\|_{\ell^2}^2}{\rho_Z}$, $Z \in \{L, R\}$. Moreover, for an ideal gas obeying the caloric equation of state $p = (\gamma - 1)\rho e$, it can be shown (see Toro [29, p. 150]) that $\boldsymbol{m}^{\perp}$ is the solution of the transport problem $\partial_t \boldsymbol{m}^{\perp} + \partial_x(u\boldsymbol{m}) = 0$. The bottom line of this argumentation is that the maximum wave speed in (2.16) is

$$\lambda_{\max}(\boldsymbol{c}_L, \boldsymbol{c}_R) = \max(|\lambda_1^-(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})|, |\lambda_3^+(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})|),$$

where $\lambda_1^-(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})$ and $\lambda_3^+(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})$ are the two extreme wave speeds in the Riemann problem (2.17) with data $(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})$.

We now determine the values of $\lambda_1^-(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})$ and $\lambda_3^+(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}})$. We only consider the case where both states, $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$, are not vacuum states, since the algorithm that we are proposing never produces vacuum states if vacuum is not present in the initial data. That is, we assume $\rho_L, \rho_R > 0$ and $p_L, p_R \geq 0$. Then the local sound speed is given by $a_Z = \sqrt{\frac{\gamma p_Z}{\rho_Z}}$, where Z is either L or R. We introduce the notation $A_Z := \frac{2}{(\gamma+1)\rho_Z}$, $B_Z := \frac{\gamma-1}{\gamma+1} p_Z$ and the functions

$$\phi(p) := f(p, L) + f(p, R) + u_R - u_L, \tag{2.18}$$

$$f(p, Z) := \begin{cases} (p - p_Z)\left(\frac{A_Z}{p + B_Z}\right)^{\frac{1}{2}} & \text{if } p \geq p_Z, \\ \frac{2a_Z}{\gamma - 1}\left(\left(\frac{p}{p_Z}\right)^{\frac{\gamma-1}{2\gamma}} - 1\right) & \text{if } p < p_Z, \end{cases} \tag{2.19}$$

where again Z is either L or R. It is shown in Toro [29, Chap. 4.3.1] that the function $\phi(p) \in C^1(\mathbb{R}_+; \mathbb{R})$ is monotone increasing and concave down. Observe that $\phi(0) = u_R - u_L - \frac{2a_L}{\gamma-1} - \frac{2a_R}{\gamma-1}$. Therefore, ϕ has a unique positive root if and only if the nonvacuum condition

$$u_R - u_L < \frac{2a_L}{\gamma - 1} + \frac{2a_R}{\gamma - 1} \tag{2.20}$$

holds; see Toro [29, (eq. 4.40), p. 127]; we denote this root by p^*, i.e., $\phi(p^*) = 0$, and p^* can be found via Newton's method. If (2.20) does not hold, we set $p^* = 0$. Then it can be shown that whether or not there is formation of vacuum, we have

$$\lambda_1^-(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}}) = u_L - a_L\left(1 + \frac{\gamma + 1}{2\gamma}\left(\frac{p^* - p_L}{p_L}\right)_+\right)^{\frac{1}{2}}, \tag{2.21}$$

$$\lambda_3^+(\boldsymbol{c}_L^{\boldsymbol{n}}, \boldsymbol{c}_R^{\boldsymbol{n}}) = u_R + a_R\left(1 + \frac{\gamma + 1}{2\gamma}\left(\frac{p^* - p_R}{p_R}\right)_+\right)^{\frac{1}{2}}, \tag{2.22}$$

where $z_+ := \max(0, z)$.

Remark 2.8 (fast algorithm). Note that if both $\phi(p_L) > 0$ and $\phi(p_R) > 0$, there is no need to compute p^*, since in this case $\lambda_1^-(u_L, u_R) = u_L - a_L$ and $\lambda_3^+(u_L, u_R) = u_R + a_R$; i.e., two rarefaction waves are present in the solution with a possible formation of vacuum. This observation is important since traditional techniques for computing p^* may require a large number of iterations in this situation; see Toro [29, p. 128]. Note finally that there is no need to compute p^* exactly since one needs only an upper bound on $\lambda_{\max}$. A very fast algorithm, with guaranteed upper bound on $\lambda_{\max}$ up to

any prescribed accuracy ϵ of the type $\lambda_{\max} \le \tilde{\lambda}_{\max} \le (1+\epsilon)\lambda_{\max}$, is described in Guermond and Popov [11]. It is also shown therein that

$$\tilde{p}^* = \left(\frac{a_L + a_R - \frac{\gamma-1}{2}(u_R - u_L)}{a_L\, p_L^{-\frac{\gamma-1}{2\gamma}} + a_R\, p_R^{-\frac{\gamma-1}{2\gamma}}} \right)^{\frac{2\gamma}{\gamma-1}} \tag{2.23}$$

is an upper bound for p^* for $\gamma \in (1, \frac{5}{3}]$. Therefore, denoting by $\tilde{\lambda}_1^-(u_L, u_R)$ and $\tilde{\lambda}_3^+(u_L, u_R)$ the wave speeds computed with $\tilde{p}^*$ instead of p^* in (2.21)–(2.22), we have $\tilde{\lambda}_1^-(u_L, u_R) \le \lambda_1^-(u_L, u_R) \le \lambda_3^+(u_L, u_R) \le \tilde{\lambda}_3^+(u_L, u_R)$, which in turn implies that $\tilde{\lambda}_{\max} = \max(|\tilde{\lambda}_1^-(u_L, u_R)|, |\tilde{\lambda}_3^+(u_L, u_R)|)$ is a guaranteed upper bound for $\lambda_{\max}$, i.e., $\lambda_{\max} \le \tilde{\lambda}_{\max}$.

3. First-order method. In this section we describe an explicit first-order finite element technique that, up to a CFL restriction, preserves all convex invariant sets of (1.1) that contain reasonable approximations of $\boldsymbol{u}_0$. Although most of the arguments invoked in this section are quite standard and mimic Lax's one-dimensional finite volume scheme, we are not aware of the existence of such a finite element–based scheme in the literature.

3.1. The finite element space. We want to approximate the solution of (1.1) with continuous finite elements. Let $(\mathcal{T}_h)_{h>0}$ be a shape-regular sequence of matching meshes. The elements in the mesh sequence are assumed to be generated from a finite number of reference elements denoted $\widehat{K}_1, \dots, \widehat{K}_\varpi$. For example, the mesh $\mathcal{T}_h$ could be composed of a combination of triangles and parallelograms in two space dimensions ($\varpi = 2$ in this case); it could also be composed of a combination of tetrahedra, parallelepipeds, and triangular prisms in three space dimensions ($\varpi = 3$ in this case). The diffeomorphism mapping $\widehat{K}_r$ to an arbitrary element $K \in \mathcal{T}_h$ is denoted $T_K : \widehat{K}_r \longrightarrow K$, and its Jacobian matrix is denoted $\mathbb{J}_K$, $1 \le r \le \varpi$. 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 notation).

Then we define the scalar-valued and vector-valued finite element spaces

$$P(\mathcal{T}_h) = \{v \in \mathcal{C}^0(D;\mathbb{R}) \mid v_{|K} \circ T_K \in \widehat{P}\ \forall K \in \mathcal{T}_h\}, \qquad \boldsymbol{P}(\mathcal{T}_h) = [P(\mathcal{T}_h)]^m, \tag{3.1}$$

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

$$\sum_{i\in\{1:n_{\rm sh}\}} \widehat{\theta}_i(\widehat{\boldsymbol{x}}) = 1 \quad \forall \widehat{\boldsymbol{x}} \in \widehat{K}. \tag{3.2}$$

The global shape functions in $P(\mathcal{T}_h)$ are denoted by $\{\varphi_i\}_{i\in\{1:I\}}$. Recall that these functions form a basis of $P(\mathcal{T}_h)$. Let $\mathsf{j} : \mathcal{T}_h \times \{1{:}n_{\rm sh}\} \longrightarrow \{1{:}I\}$ be the connectivity array. This array is defined such that

$$\varphi_{\mathsf{j}(i,K)}(\boldsymbol{x}) = \widehat{\theta}_i((T_K)^{-1}(\boldsymbol{x})) \quad \forall i \in \{1{:}n_{\rm sh}\},\ \forall K \in \mathcal{T}_h. \tag{3.3}$$

This definition, together with the partition of unity property, implies that

$$\sum_{i\in\{1:I\}} \varphi_i(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in D. \tag{3.4}$$

We denote by S_i the support of φ_i and by $|S_i|$ the measure of S_i, $i \in \{1{:}I\}$. We also define by $S_{ij} := S_i \cap S_j$ the intersection of the two supports S_i and S_j. Let E be a union of cells in $\mathcal{T}_h$; we define by $\mathcal{I}(E) := \{j \in \{1{:}I\} \mid |S_j \cap E| \neq 0\}$ the set that contains the indices of all the shape functions whose support on E is of nonzero measure. We are going to regularly invoke $\mathcal{I}(K)$ and $\mathcal{I}(S_i)$ and the partition of unity property $\sum_{i\in\mathcal{I}(K)} \varphi_i(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in K$.

Let $\mathcal{M} \in \mathbb{R}^{I\times I}$ be the consistent mass matrix with entries $\int_{S_{ij}} \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, and let $\mathcal{M}^L$ be the diagonal lumped mass matrix with entries

$$m_i := \int_{S_i} \varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{3.5}$$

The partition of unity property implies that $m_i = \sum_{j\in\mathcal{I}(S_i)} \int \varphi_j(\boldsymbol{x})\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$; i.e., the entries of $\mathcal{M}^L$ are obtained by summing the rows of $\mathcal{M}$. One key assumption that we use in the rest of the paper is that

$$m_i > 0 \quad \forall i \in \{1{:}I\}. \tag{3.6}$$

This assumption is satisfied by many finite element families.

3.2. The scheme. Let $\boldsymbol{u}_{h0} = \sum_{i=1}^I \mathbf{U}_i^0\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a reasonable approximation of $\boldsymbol{u}_0$ (we will be more precise in the following sections). Let $n \in \mathbb{N}$, τ be the time step, t^n be the current time, and let us set $t^{n+1} = t^n + \tau$. Let $\boldsymbol{u}_h^n = \sum_{i=1}^I \mathbf{U}_i^n\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be the space approximation of $\boldsymbol{u}$ at time t^n, and set $\boldsymbol{u}_h^{n+1} = \sum_{i=1}^I \mathbf{U}_i^{n+1}\varphi_i$, where $\mathbf{U}_i^{n+1}$ is yet to be determined. Using the observation that $\boldsymbol{f}(\boldsymbol{u}_h^n) = \sum_{j\in\{1:I\}} \boldsymbol{f}(\mathbf{U}_j^n)\varphi_j$ if $\boldsymbol{f}$ is linear, we adopt the following ansatz, which is formally second-order accurate in space:

$$\int_D \nabla\cdot(\boldsymbol{f}(\boldsymbol{u}_h^n))\varphi_i\,\mathrm{d}\boldsymbol{x} \approx \sum_{j\in\mathcal{I}(S_i)} \boldsymbol{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}, \tag{3.7}$$

where the coefficients $\boldsymbol{c}_{ij} \in \mathbb{R}^d$ are defined by

$$\boldsymbol{c}_{ij} = \int_D \varphi_i\nabla\varphi_j\,\mathrm{d}\boldsymbol{x}. \tag{3.8}$$

We propose to compute $\boldsymbol{u}_h^{n+1}$ by

$$m_i\frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(S_i)} \boldsymbol{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - \mathbf{U}_j^n d_{ij}^n = 0, \tag{3.9}$$

where the lumped mass matrix is used for the approximation of the time derivative. The coefficient d_{ij}^n is an artificial viscosity for the pair (i,j) that is yet to be clearly identified. For the time being we assume that

$$d_{ij}^n \geq 0, \quad \text{if} \quad i \neq j, \quad d_{ij}^n = d_{ji}^n, \quad \text{and} \quad d_{ii}^n := \sum_{i\neq j\in\mathcal{I}(S_i)} - d_{ji}^n. \tag{3.10}$$

Remark 3.1 (conservation). The definition $d_{ii}^n := \sum_{i\neq j\in\mathcal{I}(S_i)} - d_{ji}^n$ implies that $\sum_{j\in\mathcal{I}(S_i)} d_{ji}^n = 0$, which in turn implies that

$$\int_D \boldsymbol{u}_h^{n+1}\,\mathrm{d}\boldsymbol{x} = \int_D \boldsymbol{u}_h^n\,\mathrm{d}\boldsymbol{x} - \tau\int_D \nabla\cdot\left(\sum_{j=1}^I \boldsymbol{f}(\mathbf{U}_j^n)\varphi_j\right)\mathrm{d}\boldsymbol{x} \quad \forall n \geq 0, \tag{3.11}$$

i.e., the method is conservative. Note also that the symmetry assumption in (3.10) implies $d_{ii}^n := \sum_{i\neq j\in\mathcal{I}(S_i)} -d_{ij}^n$, which is often easier to compute.

3.3. The convex combination argument. We motivate the choice of the artificial viscosity coefficients d_{ij}^n in this section. Observing that the partition of unity property $\sum_{j\in\mathcal{I}(S_i)}\varphi_j = 1$ and (3.10) imply conservation, i.e.,

$$\sum_{j\in\mathcal{I}(S_i)} \boldsymbol{c}_{ij} = 0, \qquad \sum_{j\in\mathcal{I}(S_i)} d_{ij}^n = 0, \tag{3.12}$$

we rewrite (3.9) as follows:

$$m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau} = -\sum_{j\in\mathcal{I}(S_i)} (\boldsymbol{f}(\mathbf{U}_j^n)-\boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{c}_{ij} + d_{ij}^n(\mathbf{U}_j^n+\mathbf{U}_i^n). \tag{3.13}$$

Using again conservation, i.e., $d_{ii} = -\sum_{i\neq j\in\mathcal{I}(S_i)} d_{ij}^n$, we finally arrive at

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n\left(1-\sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^n}{m_i}\right) + \sum_{i\neq j\in\mathcal{I}(S_i)}\frac{2\tau d_{ij}^n}{m_i}\overline{\mathbf{U}}_{ij}^{n+1}, \tag{3.14}$$

where we have introduced the auxiliary quantities

$$\overline{\mathbf{U}}_{ij}^{n+1} := \frac{1}{2}(\mathbf{U}_j^n+\mathbf{U}_i^n) - (\boldsymbol{f}(\mathbf{U}_j^n)-\boldsymbol{f}(\mathbf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^n}. \tag{3.15}$$

A first key observation is that (3.14) is a convex combination provided τ is small enough and provided (3.6) holds. A second key observation at this point is that upon setting $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$, we realize that $\overline{\mathbf{U}}_{ij}^{n+1}$ is exactly of the form $\overline{\boldsymbol{u}}(t,\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)$ as defined in (2.6), with a fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^n$. The CFL condition $t\lambda_{\max}(\boldsymbol{n}_{ij},\boldsymbol{u}_L,\boldsymbol{u}_R)\leq\frac{1}{2}$ in Lemma 2.1 motivates the following definition for the viscosity coefficients d_{ij}^n:

$$d_{ij}^n := \max(\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{3.16}$$

where we recall that $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)$ is defined in the assumption (2.5).

Remark 3.2 (symmetry). If either φ_i or φ_j is zero at the boundary of the computational domain D, one integration by parts gives $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$, which in turn implies $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n) = \lambda_{\max}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)$. In conclusion $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2} = \lambda_{\max}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}$ if either φ_i or φ_j is an interior shape function.

Remark 3.3 (upwinding). Note that in the scalar one-dimensional case when the flux f is linear and the shape functions are piecewise linear, (3.9) gives the usual upwinding first-order method.

4. Stability analysis. We analyze the stability properties of the scheme (3.9) with the viscosity defined in (3.16).

4.1. Invariant domain property. We now prove the main result of the paper.

THEOREM 4.1 (local invariance). *Let $n\geq 0$ and let $i\in\{1{:}I\}$. Assume* (3.6). *Assume that τ is small enough so that $1+2\tau\frac{d_{ii}^n}{m_i}\geq 0$. Let $B\subset\mathcal{A}$ be a convex invariant set for* (1.1) *such that $\{\mathbf{U}_j^n \mid j\in\mathcal{I}(S_i)\}\subset B$; then $\mathbf{U}_i^{n+1}\in B$.*

Proof. Owing to the local CFL assumption $1 + 2\tau\frac{d_{ii}^n}{m_i} \geq 0$, (3.14) defines $\mathbf{U}_i^{n+1}$ as a convex combination of $\mathbf{U}_i^n$ and the collection of states $\{\overline{\mathbf{U}}_{ij}^{n+1}\}_{j\in\mathcal{I}(S_i)}$. Observe that upon defining $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$, the quantity $\overline{\mathbf{U}}_{ij}^{n+1}$ defined in (3.15) is exactly of the form $\overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ as defined in (2.6), with the flux $\boldsymbol{f}\cdot\boldsymbol{n}_{ij}$ and the fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij^n}$. The definition (3.16) implies that $d_{ij}^n \geq \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$, which is the CFL condition $t\lambda_{\max}(\boldsymbol{n}_{ij}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$ necessary for the conclusions of Lemma 2.1 to hold. This proves that $\overline{\mathbf{U}}_{ij}^{n+1} := \overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \in B$ for all $j \in \mathcal{I}(S_i)$ since B is a convex invariant set.

The convexity of B implies that $\mathbf{U}_i^{n+1} \in B$, since $\mathbf{U}_i^n \in B$ by assumption and we have established above that $\overline{\mathbf{U}}_{ij}^{n+1} \in B$ for all $j \in \mathcal{I}(S_i)$. ∎

COROLLARY 4.2 (global invariance). *Let $n \in \mathbb{N}$. Assume that τ is small enough so that the global CFL condition* $\min_{i\in\{1:I\}}\left(1 + 2\tau\frac{d_{ii}^n}{m_i}\right) \geq 0$ *holds. Let $B \subset \mathcal{A}$ be a convex invariant set. Assume that $\{\mathbf{U}_i^n \mid i \in \{1{:}I\}\} \subset B$. Then $\{\mathbf{U}_i^{n+1} \mid i \in \{1{:}I\}\} \subset B$.*

Proof. The statement is a direct consequence of Theorem 4.1. ∎

Note that the above results say that the invariant domain property holds for the coefficients $\mathbf{U}^{n+1}$, but they do not say whether this property holds for $\boldsymbol{u}_h^{n+1}$. In order to be able to extract some information on the approximate solution $\boldsymbol{u}_h^{n+1}$, we introduce an additional assumption on the reference shape functions. More specifically, we assume that the basis $\{\widehat{\theta}_i\}_{i\in\{1:n_{\mathrm{sh}}\}}$ has the following positivity property:

$$\widehat{\theta}_i(\boldsymbol{x}) \geq 0 \quad \forall \widehat{\boldsymbol{x}} \in \widehat{K}. \tag{4.1}$$

This property holds for linear Lagrange elements on simplices, quadrangular elements, and hexahedra. This assumption holds also for first-order prismatic elements in three space dimensions. It holds true also for Bernstein–Bezier finite elements of any polynomial degree; see e.g., Lai and Schumaker [20, Chap. 2] and Ainsworth [1]. This assumption implies that $\varphi_i(\boldsymbol{x}) \geq 0$ for all $i \in \{1{:}I\}$ and all $\boldsymbol{x} \in D$. Note that (4.1) implies (3.6).

COROLLARY 4.3 (global invariance on $\boldsymbol{u}_h^{n+1}$). *Assume that* (4.1) *holds. Let $B \subset \mathcal{A}$ be a convex invariant set containing the initial data $\boldsymbol{u}_0$. Assume that $\{\mathbf{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$. Let $N \in \mathbb{N}$. Assume that τ is small enough so that the CFL condition $1 + 2\tau\frac{d_{ii}^n}{m_i} \geq 0$ holds for all $i \in \{1{:}I\}$ and all $n \in \{0{:}N\}$. Then $\{\mathbf{U}_i^n \mid i \in \{1{:}I\}\} \subset B$ and $\boldsymbol{u}_h^n \in B$ for all $n \in \{0{:}N+1\}$.*

Proof. The CFL condition, together with (3.6), implies that $\{\mathbf{U}_i^n \mid i \in \{1{:}I\}\} \subset B$ for any $n \in \{0{:}N+1\}$ as established in Corollary 4.2. Let $\boldsymbol{x} \in D$; then the expansion $\boldsymbol{u}_h^n(\boldsymbol{x}) = \sum_{i\in\{1:I\}} \mathbf{U}_i^n\varphi_i(\boldsymbol{x})$ is in the convex hull of $\{\mathbf{U}_i^n\}_{i\in\{1:I\}}$ owing to the partition of unity property (3.4) and the positivity assumption (4.1). Hence, $\boldsymbol{u}_h^n(\boldsymbol{x}) \in B$ for any $\boldsymbol{x} \in D$ since B is convex. ∎

Remark 4.4 (construction of $\boldsymbol{u}_h^0$). Let $B \subset \mathcal{A}$ be a convex invariant set containing the initial data $\boldsymbol{u}_0$. If $\boldsymbol{P}(\mathcal{T}_h)$ is composed of piecewise linear Lagrange elements, then defining $\boldsymbol{u}_h^0$ to be the Lagrange interpolant of $\boldsymbol{u}_0$, we have $\{\mathbf{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$. Similarly, if $\boldsymbol{P}(\mathcal{T}_h)$ is composed of Bernstein finite elements of degree two and higher, then defining $\boldsymbol{u}_h^0$ to be the Bernstein quasi-interpolant of $\boldsymbol{u}_0$, we have $\{\mathbf{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$; see Lai and Schumaker [20, eq. (2.72)]. Note that the approximation of $\boldsymbol{u}_0$ is only second-order accurate in this case, independently of the polynomial degree

of the Bernstein polynomials; see [20, Thm. 2.45]. In both cases the assumptions of Corollary 4.3 hold.

Remark 4.5. The arguments invoking the convex combination (3.14) and the one-dimensional Riemann averages (3.15) are similar in spirit to those used in the proof of Theorem 3 in Perthame and Shu [27].

We now give an interpretation of the CFL condition $1+2\tau\frac{d_{ii}^n}{m_i}\geq 0$ in terms of the local mesh size. The local minimum mesh size, $\underline{h}_K$, for any $K\in\mathcal{T}_h$ is defined as follows:

$$\underline{h}_K := \frac{1}{\max_{i\neq j\in\mathcal{I}(K)}\|\nabla\varphi_i\|_{\boldsymbol{L}^\infty(S_{ij})}}, \tag{4.2}$$

and the global minimum mesh size is $\underline{h}:=\min_{K\in\mathcal{T}_h}\underline{h}_K$. Due to the shape regularity assumption, the quantities $\underline{h}_K$ and $h_K:=\operatorname{diam}(K)$ are uniformly equivalent, but it will turn out that using $\underline{h}_K$ instead of h_K gives a sharper estimate of the CFL number. Let us recall that $n_{\mathrm{sh}}:=\operatorname{card}(\mathcal{I}(K))$, and let us define $\vartheta_K:=\frac{1}{n_{\mathrm{sh}}-1}$. Note that

$$0<\vartheta_{\min}:=\min_{(\mathcal{T}_h)_{h>0}}\min_{K\in\mathcal{T}_h}\vartheta_K<+\infty, \tag{4.3}$$

since there are at most ϖ reference elements defining the mesh sequence. We also introduce the mesh-dependent quantities

$$\mu_{\min}:=\min_{K\in\mathcal{T}_h}\min_{i\in\mathcal{I}(K)}\frac{1}{|K|}\int_K\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\quad \mu_{\max}:=\max_{K\in\mathcal{T}_h}\max_{i\in\mathcal{I}(K)}\frac{1}{|K|}\int_K\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{4.4}$$

Note that $\mu_{\min}=\mu_{\max}=\frac{1}{n_{\mathrm{sh}}}=\frac{1}{d+1}$ for meshes uniquely composed of simplices and that $\mu_{\min}=\mu_{\max}=2^{-d}$ for meshes uniquely composed of parallelograms and cuboids.

LEMMA 4.6 (CFL). *Let* $\lambda_{\max}^n:=\max_{i\in\{1:I\}}\max_{j\in\mathcal{I}(S_i)}\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)$ *for* $n\geq 0$. *Assume that* $2\tau\frac{\lambda_{\max}^n}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}}\leq 1$; *then* $\min_{i\in\{1:I\}}\left(1+2\tau\frac{d_{ii}^n}{m_i}\right)\geq 0$.

Proof. Note first that

$$\|\boldsymbol{c}_{ij}\|_{\ell^2}\leq\int_{S_{ij}}\|\nabla\varphi_j\|_{\ell^2}\varphi_i\,\mathrm{d}\boldsymbol{x}\leq\underline{h}^{-1}\int_{S_{ij}}\varphi_i\,\mathrm{d}\boldsymbol{x}\leq\underline{h}^{-1}\mu_{\max}|S_{ij}|.$$

The definition of d_{ii} implies that

$$-d_{ii}\leq\frac{\lambda_{\max}^n}{\underline{h}}\mu_{\max}\sum_{i\neq j\in\mathcal{I}(S_i)}|S_{ij}|\leq\frac{\lambda_{\max}^n}{\underline{h}}\frac{\mu_{\max}}{\vartheta_{\min}}|S_i|.$$

Then using that $\mu_{\min}|S_i|\leq m_i$, we infer that $-2\tau\frac{d_{ii}}{m_i}\leq 2\tau\frac{\lambda_{\max}^n}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}}\leq 1$, which concludes the proof. ∎

4.2. Discrete entropy inequality. We now derive a local entropy inequality.

THEOREM 4.7. *Let* $B\subset\mathcal{A}$ *be a convex invariant set for* (1.1). *Let* $(\eta,\boldsymbol{q})$ *be an entropy pair for* (1.1). *Assume that* (2.7) *holds for any Riemann data* $(\boldsymbol{u}_L,\boldsymbol{u}_R)$ *in* B *and any* $\boldsymbol{n}\in S^{d-1}(\mathbf{0},1)$. *Let* $n\geq 0$ *and* $i\in\{1{:}I\}$. *Assume also that the local CFL condition* $1+2\tau\frac{d_{ii}^n}{m_i}\geq 0$ *holds: then we have the following local entropy inequality:*

$$\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{n+1})-\eta(\mathbf{U}_i^n))+\int_D\nabla\cdot\left(\sum_{j\in\mathcal{I}(S_i)}\boldsymbol{q}(\mathbf{U}_j^n)\varphi_j\right)\varphi_i\,\mathrm{d}\boldsymbol{x}\;-\sum_{j\in\mathcal{I}(S_i)}d_{ij}^n\eta(\mathbf{U}_j^n)\leq 0. \tag{4.5}$$

Proof. Let $(\eta, \boldsymbol{q})$ be an entropy pair for the system (1.1). Let $i \in \{1{:}I\}$; then recalling (3.14), the CFL condition and the convexity of η imply that

$$\eta(\mathbf{U}_i^{n+1}) \le \left(1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d_{ij}^n}{m_i}\right)\eta(\mathbf{U}_i^n) + \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d_{ij}^n}{m_i}\eta(\overline{\mathbf{U}}_{ij}^{n+1}).$$

Owing to Lemma 2.2 we have

$$\eta(\overline{\mathbf{U}}_{ij}^{n+1}) \le \frac{1}{2}(\eta(\mathbf{U}_i^n) + \eta(\mathbf{U}_j^n)) - t(\boldsymbol{q}(\mathbf{U}_j^n)\cdot\boldsymbol{n}_{ij} - \boldsymbol{q}(\mathbf{U}_i^n)\cdot\boldsymbol{n}_{ij}),$$

with $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^n$; hence,

$$\begin{aligned}\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{n+1}) - \eta(\mathbf{U}_i^n)) &\le \sum_{i\neq j\in\mathcal{I}(S_i)} 2d_{ij}^n(\eta(\overline{\mathbf{U}}_{ij}^{n+1}) - \eta(\mathbf{U}_i^n))\\ &\le \sum_{i\neq j\in\mathcal{I}(S_i)} d_{ij}^n(\eta(\mathbf{U}_j^n) - \eta(\mathbf{U}_i^n)) - \|\boldsymbol{c}_{ij}\|_{\ell^2}(\boldsymbol{q}(\mathbf{U}_j^n)\cdot\boldsymbol{n}_{ij} - \boldsymbol{q}(\mathbf{U}_i^n)\cdot\boldsymbol{n}_{ij}).\end{aligned}$$

The conclusion follows from the definitions of $\boldsymbol{n}_{ij}$, $\boldsymbol{c}_{ij}$, and d_{ij}^n. ∎

Remark 4.8. One recovers (3.9) from (4.5) with $\eta(\boldsymbol{v}) = \boldsymbol{v}$. Note also that (4.5) gives the global entropy inequality $\sum_{1\le i\le I} m_i\eta(\mathbf{U}_i^{n+1}) \le \sum_{1\le i\le I} m_i\eta(\mathbf{U}_i^n)$.

Remark 4.9. The meaning of the entropy inequality (2.7) might be somewhat ambiguous in some cases, especially when $\boldsymbol{u}$ is a measure. Since it is only the inequality (2.8) that is really needed in the proof of Theorem 4.7, we could replace the assumption (2.7) by (2.8). This would avoid having to invoke measure solutions, since $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ should always be finite for the Riemann problem (2.1) to have a reasonable (physical) meaning.

4.2.1. Cell-based versus edge-based viscosity. In the formulation (3.9) the term $\sum_{j\in\mathcal{I}(S_i)} d_{ij}^n\mathbf{U}_j$ models some edge-based dissipation; i.e., d_{ij}^n is a dissipation coefficient associated with the pair of degrees of freedom of indices (i, j). This formulation is related in spirit to that of local extremum diminishing (LED) schemes developed for scalar conservation equations in Kuzmin and Turek [18, eqs. (32)–(33)]; see also Jameson [15, section 2.1]. However, it is a bit difficult to understand that we are modeling some artificial dissipation by just staring at (3.9).

We now propose an alternative point of view using a cell-based viscosity. The traditional way to introduce dissipation in the finite element world consists of invoking the weak form of the Laplacian operator $-\nabla\cdot(\nu\nabla\psi)$. For instance, assuming that the viscosity field ν is piecewise constant over each mesh cell $K \in \mathcal{T}_h$, we write

$$\int_D -\nabla\cdot(\nu\nabla\psi)\varphi_i\,\mathrm{d}\boldsymbol{x} = \sum_{K\subset S_i} \nu_K \int_K \nabla\psi\cdot\nabla\varphi_i\,\mathrm{d}\boldsymbol{x}. \tag{4.6}$$

Unfortunately, it has been shown in Guermond and Nazarov [10] that the bilinear form $(\psi, \varphi) \longrightarrow \int_K \nabla\psi\cdot\nabla\varphi\,\mathrm{d}\boldsymbol{x}$ is not robust with respect to the shape of the cells. More specifically, the convex combination argument, which is essential to proving the maximum principle for scalar conservation equations in arbitrary space dimension with continuous finite elements, can be made to work only if $\int_{S_{ij}} \nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}\boldsymbol{x} < 0$ 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 a

lot of meshes—in particular in three space dimensions. To avoid this difficulty, it is proposed in [10] to replace (4.6) by $\sum_{K\subset S_i}\nu_K b_K(\psi,\varphi_i)$, where

$$b_K(\varphi_j,\varphi_i)=\begin{cases}-\vartheta_K|K| & \text{if } i\neq j, \quad i,j\in\mathcal{I}(K),\\ |K| & \text{if } i=j, \quad i,j\in\mathcal{I}(K),\\ 0 & \text{if } i\notin\mathcal{I}(K) \text{ or } j\notin\mathcal{I}(K).\end{cases} \tag{4.7}$$

The essential properties of b_K can be summarized as follows.

LEMMA 4.10. *There is $c>0$ depending only on the collection $\{(\widehat{K}_r,\widehat{P}_r,\widehat{\Sigma}_r)\}_{1\le r\le\varpi}$ and the shape-regularity such that the following identities hold for all $K\in\mathcal{T}_h$ and all $u_h,v_h\in P(\mathcal{T}_h)$:*

$$b_K(\varphi_i,\varphi_j)=b_K(\varphi_j,\varphi_i),\qquad b_K\left(\varphi_i,\sum_{j\in\mathcal{I}(K)}\varphi_j\right)=0, \tag{4.8}$$

$$b_K(u_h,v_h)=\vartheta_K|K|\sum_{i\in\mathcal{I}(K)}\sum_{\mathcal{I}(K)\ni j<i}(\mathsf{U}_i-\mathsf{U}_j)(\mathsf{V}_i-\mathsf{V}_j), \tag{4.9}$$

$$b_K(u_h,u_h)\ge ch_K^2\|\nabla u_h\|^2_{\boldsymbol{L}^2(K)}. \tag{4.10}$$

For instance, when K is a simplex and $\widehat{K}$ is the regular simplex, i.e., all the edges are of unit length, it can be shown that $b_K(\varphi_i,\varphi_i)=\kappa\int_K\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_j)\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_i)\,\mathrm{d}\boldsymbol{x}$ and $b_K(\varphi_j,\varphi_i)=-\frac{\kappa}{1-n_{\mathrm{sh}}}\int_K\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_j)\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_i)\,\mathrm{d}\boldsymbol{x}$ for $j\neq i$, with $\kappa=\frac12(1+\frac1d)$. Note also that $b_K(\varphi_j,\varphi_i)\sim h_K^2\int_K(\nabla\varphi_j)\cdot(\nabla\varphi_i)\,\mathrm{d}\boldsymbol{x}$ if K is a regular simplex, thereby showing the connection between b_K and the more familiar bilinear form associated with the Laplacian. One key argument from Guermond and Nazarov [10] is the recognition that the bilinear form defined in (4.7) has all the good characteristics of the Laplacian-based diffusion (see Lemma 4.10) and makes the convex combination argument work independently of the space dimension and the shape-regularity of the mesh family.

Hence, instead of (3.9), we could also compute $\boldsymbol{u}_h^{n+1}$ by

$$m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau}+\int_D\nabla\cdot\left(\sum_{j\in\mathcal{I}(S_i)}\boldsymbol{f}(\mathbf{U}_j^n)\varphi_j\right)\varphi_i\,\mathrm{d}\boldsymbol{x}+\sum_{K\in\mathcal{T}_h}\nu_K^n\sum_{j\in\mathcal{I}(K)}\mathbf{U}_j^n b_K(\varphi_j,\varphi_i)=0, \tag{4.11}$$

where $\{\nu_K^n\}_{K\in\mathcal{T}_h}$ is a piecewise constant artificial viscosity scalar field.

THEOREM 4.11. *Let $\{\nu_K^n\}_{K\in\mathcal{T}_h}$ be defined by*

$$\nu_K^n:=\max_{i\neq j\in\mathcal{I}(K)}\frac{\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}}{\sum_{T\subset S_{ij}}-b_T(\varphi_j,\varphi_i)}. \tag{4.12}$$

Then the conclusions of Theorems 4.1 *and* 4.7 *hold under the local CFL condition* $1+2\tau\frac{\tilde{d}_{ii}^n}{m_i}\ge0$.

Proof. Let us denote $\tilde{d}_{ij}^n:=-\sum_{K\in S_{ij}}\nu_K^n b_K(\varphi_j,\varphi_i)$; then (4.11) can be recast as

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

which in turn implies that

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n \left(1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau\tilde{d}_{ij}^n}{m_i}\right) + \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau\tilde{d}_{ij}^n}{m_i}\overline{\mathbf{U}}_{ij}^{n+1},$$

where we have introduced the auxiliary quantities

$$\overline{\mathbf{U}}_{ij}^{n+1} := \frac{1}{2}(\mathbf{U}_j^n + \mathbf{U}_i^n) - (\boldsymbol{n}_{ij}\cdot\boldsymbol{f}(\mathbf{U}_j^n) - \boldsymbol{n}_{ij}\cdot\boldsymbol{f}(\mathbf{U}_i^n))\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2\tilde{d}_{ij}^n}.$$

Here again $\overline{\mathbf{U}}_{ij}^{n+1}$ is of the form $\overline{\boldsymbol{u}}(t,\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)$ as defined in (2.6) with the fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2\tilde{d}_{ij}^n$, and hence we need to make sure that $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}/2\tilde{d}_{ij}^n \leq \frac{1}{2}$ to preserve the invariant domain property. Recalling that d_{ij}^n has been defined by $d_{ij}^n := \lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$ (see (3.10)), the above condition reduces to showing that $d_{ij}^n \leq \tilde{d}_{ij}^n$. The definitions of ν_K^n and $\tilde{d}_{ij}^n$ imply that

$$\tilde{d}_{ij}^n = -\sum_{K\in S_{ij}} \nu_K^n b_K(\varphi_j,\varphi_i) \geq -\sum_{K\in S_{ij}} \frac{d_{ij}^n}{\sum_{T\subset S_{ij}} -b_T(\varphi_j,\varphi_i)} b_K(\varphi_j,\varphi_i) = d_{ij}^n,$$

whence we get the desired result. ∎

LEMMA 4.12 (CFL). *Let* $\lambda_{\max}^n := \max_{i\in\{1:I\}}\max_{j\in\mathcal{I}(S_i)}\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)$ *for* $n \geq 0$. *Assume that* $2\tau\frac{\lambda_{\max}^n}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}} \leq 1$; *then* $\min_{i\in\{1:I\}}\left(1 + 2\tau\frac{\tilde{d}_{ii}^n}{m_i}\right) \geq 0$.

Proof. From the proof of Lemma 4.6 we have $\|\boldsymbol{c}_{ij}\|_{\ell^2} \leq \underline{h}^{-1}\mu_{\max}|S_{ij}|$, and hence

$$\nu_K^n \leq \frac{\lambda_{\max}^n\mu_{\max}}{\underline{h}} \max_{k\neq l\in I(K)} \frac{|S_{kl}|}{\sum_{T\subset S_{kl}} -b_T(\varphi_k,\varphi_l)},$$

which in turn implies that

$$\tilde{d}_{ij}^n \leq \frac{\lambda_{\max}^n\mu_{\max}}{\underline{h}} \sum_{K\subset S_{ij}} -b_K(\varphi_i,\varphi_j) \max_{k\neq l\in I(K)} \frac{|S_{kl}|}{\sum_{T\subset S_{kl}} -b_T(\varphi_k,\varphi_l)}.$$

Recalling the definition of $b_T(\varphi_k,\varphi_l)$, we have $\sum_{T\subset S_{kl}} -b_T(\varphi_k,\varphi_l) \geq \vartheta_{\min}\sum_{T\subset S_{kl}}|T| = \vartheta_{\min}|S_{kl}|$; hence,

$$\tilde{d}_{ij}^n \leq \frac{\lambda_{\max}^n\mu_{\max}}{\vartheta_{\min}\underline{h}} \sum_{K\subset S_{ij}} -b_K(\varphi_i,\varphi_j) = \frac{\lambda_{\max}^n\mu_{\max}}{\vartheta_{\min}\underline{h}} \sum_{K\subset S_{ij}} \vartheta_K|K|.$$

Finally, we have

$$-\tilde{d}_{ii} := \sum_{i\neq j\in\mathcal{I}(S_i)} \tilde{d}_{ij}^n \leq \frac{\lambda_{\max}^n\mu_{\max}}{\vartheta_{\min}\underline{h}} \sum_{i\neq j\in\mathcal{I}(S_i)} \sum_{K\subset S_{ij}} \vartheta_K|K| = \frac{\lambda_{\max}^n\mu_{\max}}{\vartheta_{\min}\underline{h}}|S_i|.$$

This means that the bound on $-\tilde{d}_{ii}$ is the same as that on $-d_{ii}$ in the proof of Lemma 4.12. Then using that $\mu_{\min}|S_i| \leq m_i$, we infer that $-2\tau\frac{\tilde{d}_{ii}}{m_i} \leq 2\tau\frac{\lambda_{\max}^n}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}} \leq 1$, which concludes the proof. ∎

5. Numerical illustrations. In this section we illustrate the method described in the paper, i.e., (3.9)–(3.16), and discuss possible variants.

5.1. Invariant domain property and convergence issues. In this section we give a counterexample showing that a method that is formally first-order consistent and satisfies the invariant domain property may not necessarily be convergent.

To illustrate our point, we focus our attention on scalar conservation equations and consider an algebraic approach that is sometimes used in the literature; see e.g., Kuzmin, Löhner, and Turek [19, p. 163] and Kuzmin and Turek [18, eqs. (32)–(33)]. Instead of constructing a convex combination involving (entropy satisfying) intermediate states as in (3.14), we rewrite (3.13) as

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau}=-\sum_{i\neq j\in\mathcal{I}(S_i)}(\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n))\cdot\boldsymbol{c}_{ij}+\sum_{j\in\mathcal{I}(S_i)}d_{ij}^n\mathsf{U}_j^n \tag{5.1}$$

or, equivalently,

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau}=-\sum_{i\neq j\in\mathcal{I}(S_i)}\frac{\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_j^n-\mathsf{U}_i^n}\cdot\boldsymbol{c}_{ij}(\mathsf{U}_j^n-\mathsf{U}_i^n)+\sum_{j\in\mathcal{I}(S_i)}d_{ij}^n\mathsf{U}_j^n. \tag{5.2}$$

Let us set $k_{ij}:=\frac{\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_j^n-\mathsf{U}_i^n}\cdot\boldsymbol{c}_{ij}$, (with $k_{ij}:=\boldsymbol{c}_{ij}\cdot D_{\boldsymbol{u}}\boldsymbol{f}(\mathsf{U}_i^n)$ if $\mathsf{U}_j^n=\mathsf{U}_i^n$); then

$$\mathsf{U}_i^{n+1}=\mathsf{U}_i^n\left(1-\frac{\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(S_i)}(-k_{ij}+d_{ij}^n)\right)+\sum_{i\neq j\in\mathcal{I}(S_i)}\frac{\tau}{m_i}(-k_{ij}+d_{ij}^n)\mathsf{U}_j^n. \tag{5.3}$$

Let us finally set

$$d_{ij}^n:=\max(0,k_{ij},k_{ji}),\ i\neq j,\quad\text{and}\quad d_{ii}:=-\sum_{i\neq j\in\mathcal{I}(S_i)}d_{ij}^n. \tag{5.4}$$

This choice implies that $-k_{ij}+d_{ij}^n\geq 0$ for all $i\in\{1{:}N\}$, $j\in\mathcal{I}(S_i)$. As a result, $\mathsf{U}_i^{n+1}\in\operatorname{conv}\{\mathsf{U}_j^n,\ j\in\mathcal{I}(S_i)\}$ under the appropriate CFL condition; hence, the solution process $u_h^n\longmapsto u_h^{n+1}$ described above in (5.3)–(5.4) satisfies the maximum principle. Although this technique looks reasonable a priori, it turns out that it is not diffusive enough to handle general fluxes as discussed in Guermond and Popov [12, section 3.3]. The convergence result established in [12] requires an estimation of the wave speed that is more accurate than just the average speed $\boldsymbol{n}_{ij}\cdot\frac{\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_j^n-\mathsf{U}_i^n}$, which is invoked in the above definition. This definition of the wave speed is correct in shocks, i.e., if the Riemann problem with data $(\mathsf{U}_i^n,\mathsf{U}_j^n)$ is a simple shock, but it may not be sufficient if the Riemann solution is an expansion or a composite wave, which is likely to be the case if $\boldsymbol{f}$ is not convex.

We now illustrate numerically the observation made above. We consider the so-called KPP problem proposed in Kurganov, Petrova, and Popov [17]. It is a two-dimensional scalar conservation equation with a nonconvex flux,

$$\partial_t u+\nabla\cdot\boldsymbol{f}(u)=0,\quad u(\boldsymbol{x},0)=u_0(\boldsymbol{x})=\begin{cases}\frac{14\pi}{4} & \text{if } \sqrt{x^2+y^2}\leq 1,\\ \frac{\pi}{4} & \text{otherwise,}\end{cases} \tag{5.5}$$

where $\boldsymbol{f}(u)=(\sin u,\cos u)$. This is a challenging test case for many high-order numerical schemes because the solution has a two-dimensional composite wave structure. For example, it has been shown in [17] that some central-upwind schemes based on WENO5, Minmod 2, and SuperBee reconstructions converge to nonentropic solutions.

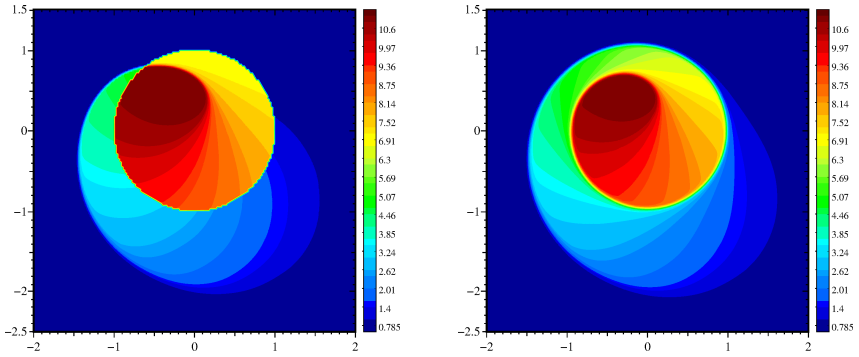

FIG. 1. *KPP solution with continuous* $\mathbb{P}_1$ *elements (*29871 *nodes,* 59100 *triangles). Left: entropy violating solution using* (3.9) *and* (5.4)*; Right: entropy satisfying solution using* (3.9) *and* (3.16).

The computational domain $[-2,2]\times[-2.5,1.5]$ is triangulated using nonuniform meshes, and the solution is approximated up to $t = 1$ using continuous $\mathbb{P}_1$ finite elements (29871 nodes, 59100 triangles). The time stepping is done with SSP RK3. The solution shown in the left panel of Figure 1 is obtained using (5.4) for the definition of d_{ij}^n. The numerical solution produces very sharp, nonoscillating, entropy violating shocks—the reason being that the artificial viscosity is not large enough. Note that the solution is maximum principle satisfying (the local maximum principle is satisfied at every grid point and every time step) and no spurious oscillations are visible. The numerical process converges to a nice-looking (wrong) piecewise smooth weak solution. The numerical solution shown in the right panel of Figure 1 is obtained by using our definition of d_{ij}^n, (3.16) (note that the results obtained with (4.11)–(4.12) together with (3.16) are indistinguishable from this solution). The expected helicoidal composite wave is clearly visible; this is the unique entropy satisfying solution.

In conclusion, the above counterexample shows that satisfying the invariant domain property/maximum principle does not imply convergence, even for a first-order method. It is also essential that the method satisfy local entropy inequalities to be convergent; this is the case of our method (3.9)–(3.16) (see Theorem 4.7), but it is not the case of the algebraic method (5.3)–(5.4).

Remark 5.1. The reader should be aware that we are citing Kuzmin, Löhner, and Turek [19, p. 163] and Kuzmin and Turek [18, eqs. (32)–(33)] a little bit out of context. The scheme as originally presented in these references was only meant to solve the linear transport equation, and as such it is a perfectly good method. Problems arise with (5.4) only when one extends the methodology to nonlinear fluxes, as we did in (5.2).

5.2. Special meshes. The construction of the intermediate states in (3.13) is not unique. For instance, we can extend a construction used by Hoff [13, Cor. 1] in one space dimension for the p-system. Let us assume that $i \in \{1,\dots,N\}$ is such that for every $j \in \mathcal{I}(S_i)\setminus\{i\}$, there is a unique $\sigma_i(j) \in \mathcal{I}(S_i)\setminus\{i,j\}$ such that $\boldsymbol{c}_{ij} := \int_{S_i} \varphi_i \nabla\varphi_j \,\mathrm{d}\boldsymbol{x} = -\int_{S_i} \varphi_i \nabla\varphi_{\sigma_i(j)} \,\mathrm{d}\boldsymbol{x} =: -\boldsymbol{c}_{i\sigma_i(j)}$. This property holds if the mesh and the reference finite elements have symmetry properties and if φ_i is an interior shape function. For instance, this property holds for $\mathbb{P}_1$ Lagrange elements if the mesh is centrosymmetric, i.e., the support of φ_i is symmetric with respect to the Lagrange node $\boldsymbol{a}_i$ associated with φ_i, for any $i \in \{1,\dots,N\}$. Then we can rewrite

(3.9) as

$$m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau}=d_{ii}\mathbf{U}_i^n+\sum_{j\in\mathcal{J}(S_i)}-(\boldsymbol{f}(\mathbf{U}_j^n)-\boldsymbol{f}(\mathbf{U}_{\sigma_i(j)}^n))\cdot\boldsymbol{c}_{ij}+d_{ij}^n\mathbf{U}_j^n+d_{i\sigma_i(j)}^n\mathbf{U}_{\sigma_i(j)}^n, \tag{5.6}$$

where the set $\mathcal{J}(S_i)\subset\mathcal{I}(S_i)$ is such that $\sigma_i:\mathcal{J}(S_i)\longrightarrow\sigma_i(\mathcal{J}(S_i))$ is bijective and $\mathcal{J}(S_i)\cup\sigma_i(\mathcal{J}(S_i))=\mathcal{I}(S_i)\setminus\{i\}$. Then upon recalling that $d_{ii}:=-\sum_{j\in\mathcal{J}(S_i)}(d_{ij}^n+d_{i\sigma_i(j)}^n)$, we have

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^n\left(1-\sum_{j\in\mathcal{J}(S_i)}\frac{\tau}{m_i}(d_{ij}^n+d_{i\sigma_i(j)}^n)\right)+\sum_{j\in\mathcal{J}(S_i)}\frac{\tau(d_{ij}^n+d_{i\sigma_i(j)}^n)}{m_i}\overline{\mathbf{U}}_{ij}^{n+1}, \tag{5.7}$$

where we have defined the intermediate state $\overline{\mathbf{U}}_{ij}^{n+1}$ by

$$\overline{\mathbf{U}}_{ij}^{n+1}=\frac{d_{i\sigma_i(j)}^n}{d_{ij}^n+d_{i\sigma_i(j)}}\mathbf{U}_{\sigma_i(j)}^n+\frac{d_{ij}^n}{d_{ij}^n+d_{i\sigma_i(j)}^n}\mathbf{U}_j^n-(\boldsymbol{f}(\mathbf{U}_j^n)-\boldsymbol{f}(\mathbf{U}_{\sigma_i(j)}^n))\cdot\frac{\boldsymbol{c}_{ij}}{d_{ij}^n+d_{i\sigma_i(j)}^n}. \tag{5.8}$$

The state $\overline{\mathbf{U}}_{ij}^{n+1}$ is of the form $\overline{\boldsymbol{u}}(t,\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n):=\int_{\alpha_L}^{\alpha_R}\boldsymbol{u}(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)(x,t)\,\mathrm{d}x$, where $\alpha_L=-\frac{d_{i\sigma_i(j)}^n}{d_{ij}^n+d_{i\sigma_i(j)}^n}$, $\alpha_R=\frac{d_{ij}^n}{d_{ij}^n+d_{i\sigma_i(j)}^n}$, and $t:=\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{d_{ij}^n+d_{i\sigma_i(j)}^n}$, provided

$$d_{i\sigma_i(j)}^n\geq(\lambda_1^-)^-(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}\quad\forall j\in\mathcal{J}(S_i), \tag{5.9}$$

$$d_{ij}^n\geq(\lambda_m^+)^+(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}\quad\forall j\in\mathcal{J}(S_i), \tag{5.10}$$

where we defined $x^+=\max(x,0)$ and $x^-=-\min(x,0)$. A sufficient condition that implies both of the above inequalities and is independent of the choice of the set $\mathcal{J}_i(S_i)$ is

$$\min(d_{ij}^n,d_{i\sigma_i(j)}^n)\geq\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2},\quad j\in\mathcal{J}(S_i). \tag{5.11}$$

Note that the above argument holds only if φ_i is an interior shape function satisfying the symmetry property $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{i\sigma_i(j)}$. If this is not the case, then we can always use the lower bound $d_{ij}^n\geq\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$.

In conclusion, the diffusion matrix $(d_{ij}^n)_{1\leq i,j\leq N}$ can be constructed as follows:

(1) For every shape function φ_i satisfying the symmetry property $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{i\sigma_i(j)}$ for every $j\in\mathcal{J}(S_i)$, we define $\widetilde{d}_{ij}^n=\widetilde{d}_{i\sigma_i(j)}^n=\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$.

(2) For every other shape function not satisfying the symmetry property mentioned above, we define $\widetilde{d}_{ij}=\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$.

(3) We construct the diffusion matrix by setting $d_{ij}^n:=\max(\widetilde{d}_{ij},\widetilde{d}_{ji})$ for $j\neq i$ and $d_{ii}:=-\sum_{i\neq j\in\mathcal{I}(S_i)}d_{ij}^n$. This construction guarantees conservation, i.e., $\sum_{i\in\mathcal{I}(S_j)}d_{ij}^n=0$, and first-order consistency, i.e., $\sum_{j\in\mathcal{I}(S_i)}d_{ij}^n=0$.

Remark 5.2. Quite surprisingly, in the case of scalar linear transport, the above construction and the construction done in section 3.3 (see definition (3.16)) give the same scheme (i.e., the same CFL).

5.3. Invariant domain property versus monotonicity. We show in this section that the invariance property and what is usually understood in the literature as monotonicity are two different concepts, and just looking at monotonicity may be misleading.

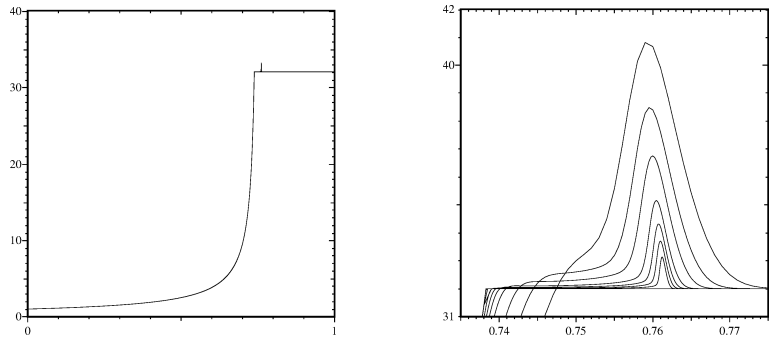

FIG. 2. *Left: v-profile for the p-system at $t = 0.75$, with 10^5 grid points. Right: close-up view of the v-profile for various grids composed of 10^3, 2×10^3, 4×10^3, 10^4, 2×10^4, 4×10^4, 10^5 grid points.*

5.3.1. p-system. We consider the p-system and solve the Riemann problem corresponding to the initial data $(v_L, u_L) = (1, 0)$, $(v_R, u_R) = (2^{\frac{2}{\gamma-1}}, \frac{1}{\gamma-1})$. The computational domain is the segment $[0, 1]$, and the separation between the left and right states is set at $x_0 = 0.75$. The solution is a single rarefaction wave from the first family (i.e., $w_1(v_L, u_L) = w_1(v_R, u_R)$):

$$v(x,t) = \begin{cases} 1 & \text{if } \frac{x-x_0}{t} \le -1, \\ (\frac{x_0-x}{t})^{\frac{-2}{\gamma+1}} & \text{if } -1 \le \frac{x-x_0}{t} \le -2^{-\frac{\gamma+1}{\gamma-1}}, \\ 2^{\frac{2}{\gamma-1}} & \text{otherwise,} \end{cases} \tag{5.12}$$

$$u(x,t) = \begin{cases} 0 & \text{if } \frac{x-x_0}{t} \le -1, \\ \frac{2}{\gamma-1}\left(1 - (\frac{x_0-x}{t})^{\frac{\gamma-1}{\gamma+1}}\right) & \text{if } -1 \le \frac{x-x_0}{t} \le -2^{-\frac{\gamma+1}{\gamma-1}}, \\ \frac{1}{\gamma-1} & \text{otherwise.} \end{cases} \tag{5.13}$$

This case is such that $(v^*, u^*) = (v_R, u_R)$, and hence the second wave corresponding to the eigenvalues $\lambda_2^{\pm}$ is not present. We use continuous piecewise linear finite elements with the algorithm (3.9)–(3.16). The time stepping is done with the SSP RK3 technique. We show the profile of v at $t = 0.75$ in Figure 2 for meshes composed of 10^3, 2×10^3, 4×10^3, 10^4, 2×10^4, 4×10^4, 10^5, and 2×10^5 cells. We observe that the profile is not monotone. There is an overshoot to the right of the foot of the (left-moving) wave. Actually, this overshoot does not violate the invariant domain property; we have verified numerically that at every time step and for every grid point in each mesh, the numerical solution is in the smallest invariant domain of type (2.13) that contains the piecewise linear approximation of the initial data. This result seems a bit surprising, but it is perfectly compatible with Theorem 4.1. Since the numerical solution cannot stay on the exact rarefaction wave (green line connecting $\mathbf{U}_L$ and $\mathbf{U}_L$ in Figure 3), the second wave reappears in the form of an overshoot at the end of the rarefaction wave (see right panel of Figure 2).

Let $(\mathbf{U}_L, \dots, \mathbf{U}_L, \mathbf{U}_R, \dots, \mathbf{U}_R)$ be the initial sequence of degrees of freedom. After one time step, two additional points appear in the phase space, denoted in Figure 3 by $\mathbf{U}_1^1$ and $\mathbf{U}_2^1$. Because of the invariant domain property, these points are under the rarefaction wave. Then the sequence of degrees of freedom at time $t = \tau$ is $(\mathbf{U}_L, \dots, \mathbf{U}_L, \mathbf{U}_1^1, \mathbf{U}_2^1, \mathbf{U}_R, \dots, \mathbf{U}_R)$. Four additional points $\mathbf{U}_1^2, \dots, \mathbf{U}_4^2$ appear after two time steps, and the sequence of degrees of freedom at time $t = 2\tau$ is

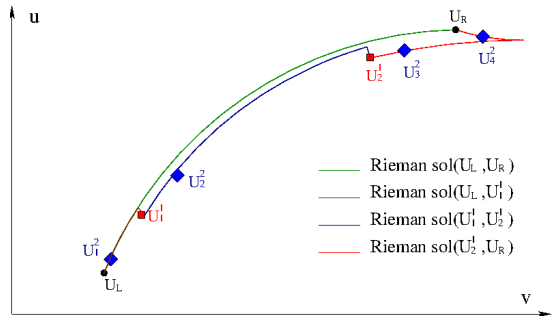

FIG. 3. *The overshooting mechanism for a single rarefaction wave in the phase space for the p-system. Initial data are shown in black; additional points after one time step are shown in red and after two time steps are shown in blue. Observe the position of* $\mathbf{U}_4^2$. *(See online version for color.)*

TABLE 1
Convergence rates on the L^1*-norm for the p-system.*

$1/h$	v	Rate	u	Rate
10^3	1.8632(–2)	-	7.2261(–3)	
2×10^3	1.0350(–2)	0.85	3.9239(–3)	0.88
4×10^3	5.6769(–3)	0.87	2.1173(–3)	0.89
10^4	2.5318(–3)	0.88	9.2888(–4)	0.90
2×10^4	1.3644(–3)	0.89	4.9541(–4)	0.91
4×10^4	7.3151(–4)	0.90	2.6319(–4)	0.91
1×10^5	2.9695(–4)	0.98	1.1352(–4)	0.92
2×10^5	1.5838(–4)	0.91	5.9869(–5)	0.92

$(\mathbf{U}_L,\dots,\mathbf{U}_L,\mathbf{U}_1^2,\dots,\mathbf{U}_4^2,\mathbf{U}_R,\dots,\mathbf{U}_R)$. The point $\mathbf{U}_4^2$ is the one whose v-component may overshoot because the exact solution of the Riemann problem with the left state $\mathbf{U}_2^1$ and the right state $\mathbf{U}_R$ is composed of two rarefaction waves, and the maximum value of v on these rarefactions is necessarily larger than v_R (see red line in Figure 3). Note that this is not a Gibbs phenomenon at all; in particular the amplitude of the overshoot decreases as the mesh is refined, as shown in the close-up view in the right panel of Figure 2. This phenomenon is actually very common in numerical simulations of hyperbolic systems but is rarely discussed; it is sometimes called "start up error" in the literature; see, for example, the comments on page 592 in Kurganov and Tadmor [16] and the comments at the bottom of page 1005 in Liska and Wendroff [23]. The (relative) L^1-norm of the error on both v and u at $t = 0.75$ is shown in Table 1. The method converges with an order close to 0.9.

5.3.2. Euler in one dimension (Leblanc shocktube). We consider now the compressible Euler equations. We solve the Riemann problem, also known in the literature as the Leblanc shocktube. The data are as follows: $\gamma = \frac{5}{3}$ and

$$\rho_L = 1.000, \quad u_L = 0.0, \quad p_L = 0.1/3,$$
$$\rho_R = 0.001, \quad u_R = 0.0, \quad p_R = 10^{-10}/3.$$

The structure of the solution is standard; it consists of a rarefaction wave moving to the left, a contact discontinuity in the middle, and a shock moving to the right. The density profile is monotone. We solve this problem with the algorithm (3.9)–(3.16) using piecewise linear finite elements. The density profile computed with 50,000,

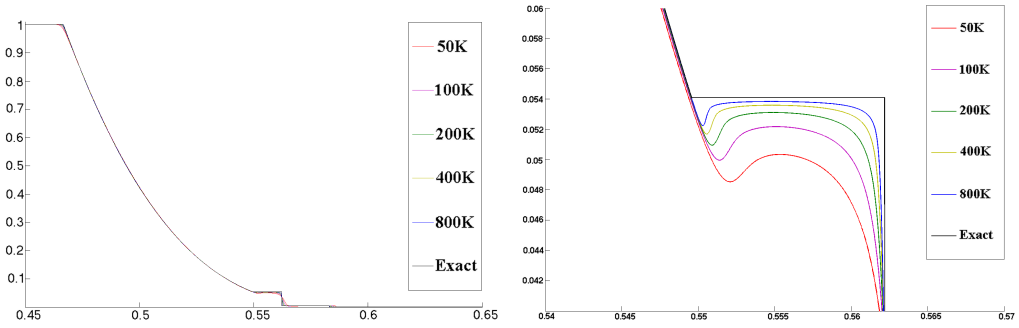

FIG. 4. *Left: density profile for the Leblanc shocktube at $t = 0.1$. Right: close-up of the density profile at the foot of the rarefaction wave.*

100,000, 200,000, 400,000 and 800,000 grid points is shown in the left panel of Figure 4. The right panel shows a close-up view of the region at the foot of the expansion wave. Of course, the scheme does not have any problem with the positivity of the density and the internal energy, but we observe that the numerical profile is not monotone; there is a small dip at the foot of the expansion. There is nothing wrong here, since, for each mesh, the numerical solution is guaranteed by Theorem 4.1 to be in the smallest convex invariant set that contains the Riemann data. This phenomenon is similar to what has been observed for the p-system in the previous section. This example shows again that the invariant domain property is a different concept from monotonicity, and just looking at monotonicity or local extrema diminishing properties is not enough to understand hyperbolic systems.

When taking a closer look at the shock region in Figure 4, one observes that the shock moves as the mesh is refined. The reader may then wonder whether the method is conservative, since it is known that nonconservative methods produce weak solutions that are nonentropic, and this usually shows up as bad shock location. This is not the case here since, as stated in Remark 3.1, the method is indeed conservative; actually, the method can be proved to converge to the entropy solution with a convergence rate at least better than $h^{\frac{1}{4}}$ in the $L^\infty_t(L^1_x)$-norm for scalar conservation equations and for initial data with bounded variations; see Guermond and Popov [12]. We illustrate the converge properties of the method in Table 2, where we show the convergence rate on the density in the L^1-norm in space at $t = 0.1$ for various meshes. The asymptotic rate is between 0.73 and 0.74.

TABLE 2
Convergence rates on the L^1-norm of the density for the Leblanc shocktube.

$1/h$	ρ	Rate
8,000	7.5213(–4)	-
16,000	4.779(–4)	0.65
32,000	2.9379(–4)	0.70
64,000	1.7709(–4)	0.73
128,000	1.0608(–4)	0.74
256,000	6.3500(–5)	0.74
512,000	3.8237(–5)	0.73

Sometimes in the literature, authors look at the velocity and the internal energy to demonstrate some of the phenomena described above. We refrain from doing this here, since these two quantities are not convex functions of the conservative variables; therefore, one cannot prove that they satisfy a local invariance property.

6. Concluding remarks. We have proposed a numerical method for solving hyperbolic systems using continuous finite elements and forward Euler time stepping. The properties of the method are based on the introduction of an artificial dissipation that is defined so that any convex invariant set is an invariant domain for the method. The main results of the paper are Theorems 4.1 and 4.7. The method is formally first-order accurate with respect to space and can be made higher-order with respect to the time step by using any explicit strong stability preserving time stepping technique. Although the argumentation of the proof of Theorem 4.1 relies on the notion of Riemann problems, the algorithm does not require solving any Riemann problem. The only information needed is an upper bound on the local maximum speed. Our next objective is to work on a generalization of the FCT technique (see Kuzmin, Löhner, and Turek [19]) to make the method at least formally second-order accurate in space while it is still domain invariant.

Finally, let us mention that all of the proofs in the paper are based on the discrete representation (3.9). Since this representation is not particular to continuous finite elements, our argumentation can be used to analyze other schemes that can be put into the form (3.9).

REFERENCES

[1] M. Ainsworth, *Pyramid algorithms for Bernstein–Bézier finite elements of high, nonuniform order in any dimension*, SIAM J. Sci. Comput., 36 (2014), pp. A543–A569, doi:10.1137/130914048.

[2] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math. (2), 161 (2005), pp. 223–342.

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

[4] A. J. Chorin, *Random choice solution of hyperbolic systems*, J. Comput. Phys., 22 (1976), pp. 517–533.

[5] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[6] P. Colella, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200, doi:10.1016/0021-9991(90)90233-Q.

[7] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21, doi:10.2307/2006218.

[8] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Sciences] 325, Springer-Verlag, Berlin, 2000, doi:10.1007/3-540-29089-3_14 .

[9] H. Frid, *Maps of convex sets and invariant regions for finite-difference systems of conservation laws*, Arch. Ration. Mech. Anal., 160 (2001), pp. 245–269, doi:10.1007/s00205010016.

[10] 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.

[11] J.-L. Guermond and B. Popov, *Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations*, J. Comput. Phys., 321 (2016), pp. 908–926.

[12] 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, doi:10.1137/140990863.

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

[14] D. Hoff, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610, doi:10.2307/2000254.

[15] A. Jameson, *Positive schemes and shock modelling for compressible flows*, Internat. J. Numer. Methods Fluids, 20 (1995), pp. 743–776, doi:10.1002/fld.1650200805.

[16] A. Kurganov and E. Tadmor, *Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 584–608, doi:10.1002/num.10025.

[17] 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, doi:10.1137/040614189.
[18] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558.
[19] D. Kuzmin, R. Löhner, and S. Turek, eds., *Flux-Corrected Transport. Principles, Algorithms, and Applications*, Sci. Comput., Springer, Berlin, 2005.
[20] M.-J. Lai and L. L. Schumaker, *Spline Functions on Triangulations*, Encyclopedia Math. Appl. 110, Cambridge University Press, Cambridge, UK, 2007, doi:10.1017/CBO9780511721588.
[21] P. D. Lax, *Hyperbolic systems of conservation laws.* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[22] P.-L. Lions, B. Perthame, and P. E. Souganidis, *Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates*, Comm. Pure Appl. Math., 49 (1996), pp. 599–638.
[23] R. Liska and B. Wendroff, *Comparison of several difference schemes on* $1D$ *and* $2D$ *test problems for the Euler equations*, SIAM J. Sci. Comput., 25 (2003), pp. 995–1017, doi:10.1137/S1064827502402120.
[24] T. P. Liu, *The Riemann problem for general systems of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.
[25] T. Nishida, *Global solution for an initial boundary value problem of a quasilinear hyperbolic system*, Proc. Japan Acad., 44 (1968), pp. 642–646.
[26] S. Osher, *The Riemann problem for nonconvex scalar conservation laws and Hamilton-Jacobi equations*, Proc. Amer. Math. Soc., 89 (1983), pp. 641–646, doi:10.2307/2044598.
[27] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130, doi:10.1007/s002110050187.
[28] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Science] 258, Springer-Verlag, New York, Berlin, 1983.
[29] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction*, 3rd ed., Springer-Verlag, Berlin, 2009.
[30] R. Young, *The p-system.* I. *The Riemann problem*, in The Legacy of the Inverse Scattering Transform in Applied Mathematics (South Hadley, MA, 2001), Contemp. Math. 301, Amer. Math. Soc., Providence, RI, 2002, pp. 219–234, doi:10.1090/conm/301/05166.