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

JEAN-LUC GUERMOND† AND BOJAN POPOV†

Abstract. We propose a numerical method to solve general hyperbolic systems in any space dimension using forward Euler time stepping and continuous finite elements on non-uniform grids. The properties of the method are based on the introduction of an artificial dissipation that is defined so that any convex invariant sets 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 Hoff [12, 13], and Frid [8].

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

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

1. Introduction.

The objective of this paper is to investigate a first-order approximation technique for nonlinear hyperbolic systems using continuous finite elements and explicit time stepping on non-uniform 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 $\boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n}$ 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, 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 non-uniform grids in any space dimension. The algorithm is described in §3.2, see (3.5) with definitions (3.4)-(3.8)-(3.13). It is a somewhat loose adaptation of the non-staggered Lax-Friedrichs scheme to continuous finite elements. The key results

*This material is based upon work supported in part by the National Science Foundation grants DMS-1217262, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA99550-12-0358, and by the Army Research Office under grant/contract number W911NF-15-1-0517. Draft version, February 24, 2016

†Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA.

of the paper are Theorem 4.1 and Theorem 4.2. 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 4.2 that the approximate solution satisfies a discrete entropy inequality for every entropy pair of the hyperbolic system. Similar results have been established for various finite volumes schemes by Hoff [12, 13], Perthame and Shu [25], Frid [8] for the compressible Euler equations and the p-system. Our scheme has no restriction on the nature of the hyperbolic system, besides the speed of propagation being finite. To the best of our knowledge, we are not aware of any similar scheme in the continuous finite element literature.

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 introduces 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 the terminology of Chueh et al. [4], Hoff [12, 13], Frid [8]. The reader who is familiar with the notions of invariant domains and Riemann problems may skip this section and go directly to §3, although the reader should be aware that our definitions of invariant sets and domains are slightly different from those of [4, 12, 13, 8].

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 to say there exists an (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 following 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 any 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 [19] and Bressan [2, Thm 5.3]. In particular there are $2m$ numbers

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

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 [2, 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 $t\geq 0$ we have

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

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

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

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 [22, Thm.1.2] and Dafermos [7, Thm. 9.5.1]. We refer to Osher [24, 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 to Bianchini and Bressan [1, Section 14] for characterizations of the Riemann solution using viscosity regularization. We also refer to Young [28, 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 $\boldsymbol{u}_R$ 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 $\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R) := \int_{-\frac12}^{\frac12}\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t)\,\mathrm{d}x$ and assume that $t\,\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)\leq\frac12$, then*

$$\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R) = \frac12(\boldsymbol{u}_L+\boldsymbol{u}_R) - t\big(\boldsymbol{f}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{f}(\boldsymbol{u}_L)\cdot\boldsymbol{n}\big). \tag{2.6}$$

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

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

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 $t\,\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)\leq\frac12$, Then*

$$\eta(\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)) \leq \tfrac12(\eta(\boldsymbol{u}_L)+\eta(\boldsymbol{u}_R)) - t(\boldsymbol{q}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{u}_L)\cdot\boldsymbol{n}). \tag{2.8}$$

Proof. Under the CFL assumption $t\,\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)\leq\frac12$, the inequality (2.7) implies that

$$\int_{-\frac12}^{\frac12}\eta(\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R))(x,t)\,\mathrm{d}x \leq \tfrac12(\eta(\boldsymbol{u}_L)+\eta(\boldsymbol{u}_R)) - t(\boldsymbol{q}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{u}_L)\cdot\boldsymbol{n}). \tag{2.9}$$

Jensen's inequality $\eta(\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)) \leq \int_{-\frac12}^{\frac12}\eta(\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t))\,\mathrm{d}x$ 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 et al. [4], Hoff [13], Smoller [26], Frid [8]. 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 $A \subset \mathcal{A} \subset \mathbb{R}^m$ is invariant for* (1.1) *if for any pair $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in A \times A$, any unit vector $\boldsymbol{n} \in \mathcal{S}^{d-1}(\mathbf{0}, 1)$, 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 A.*

Note that, 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 $\frac{1}{I} \int_I \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x \in A$. 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 the language by saying that a member of $\boldsymbol{X}_h$, say $\boldsymbol{u}_h$, is in the set $A \subset \mathbb{R}^m$ when actually we mean that $\{\boldsymbol{u}_h(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}\} \subset A$.

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

For scalar conservation equations the notions of invariant sets and invariant domains are closely related to the maximum principle, see Example 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 [23], and the general theory of positively invariant regions was developed in Chueh et al. [4]. Applications and extensions to numerical methods were developed in Hoff [12, 13] and Frid [8].

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 2.4 and Lions et al. [20]. For results on general hyperbolic systems, we refer to Frid [8], 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 [8, Thm. 7 and Thm. 8]. Similar results have been established for various finite volume schemes in two-space dimension for the Euler equations in Perthame and Shu [25, 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 scheme. 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 [6].

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 is 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 $\gamma \ge 1$, see Hoff [13, Exp. 3.5, p. 597] for a proof in the context of parabolic regularization, or use the results from Young [28] for a direct proof. Moreover, A_{ab} is an invariant domain for the Lax-Friedrichs scheme, see Hoff [12, Thm. 2.1] and Hoff [13, 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

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

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 have also have* $v^0 \le \min(v_L, v_R, v^*)$*, i.e.,* $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) \le \sqrt{-p'(v^0)}$*, 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 know 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 $\lim_{v\to+\infty} \phi(v) \ge 0$, see Young [28] 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 up 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 $v^* \le \min(v_L, v_R)$, i.e., the solution contains two shock waves which is equivalent to $\phi(\min(v_L, v_R)) \ge 0$. Using the definition of ϕ we derive that $\lambda_{\max}(\boldsymbol{u}_L,\boldsymbol{u}_R) = \sqrt{-p'(v^*)}$ if and only if $\phi(\min(v_L, v_R)) = u_L - u_R - \sqrt{(v_L - v_R)(p(v_R) - p(v_L))} \ge 0$. 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 $v^0 \le v^*$. 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 $v^0 \le v^*$. Hence, the result is established. ∎

Remark 2.1. Note that the estimate on $\lambda_{\max}(\boldsymbol{u}_L, \boldsymbol{u}_R)$ given in Lemma 2.5 is valid whether vacuum is created or not in the Riemann solution.

Remark 2.2. 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 in this paper never produces vacuum states if vacuum is not present in the initial data.

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

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

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 $e := E - \frac{1}{2}|\boldsymbol{u}|^2$. 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

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

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

Let $\boldsymbol{n} \in S^{d-1}(\mathbf{0}, 1)$ 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 [3, p. 526]. The general case is treated in Colella [5, p. 188], see also Toro [27, Chapter 4.8]. We make a change of basis and introduce $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{d-1}$ so that $\{\boldsymbol{n}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_{d-1}\}$ forms an orthonormal basis of $\mathbb{R}^d$. With this new basis we have $\boldsymbol{m} = (m, \boldsymbol{m}^\perp)^\mathsf{T}$, where $m := \rho u$, $u := \boldsymbol{u}\cdot\boldsymbol{n}$, $\boldsymbol{m}^\perp := \rho(\boldsymbol{u}\cdot\boldsymbol{t}_1, \ldots, \boldsymbol{u}\cdot\boldsymbol{t}_{d-1}) := \rho\boldsymbol{u}^\perp$. The projected equations are

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

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

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

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 [27, 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

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

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 [27, 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 in this paper 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

following notations $A_Z := \frac{2}{(\gamma+1)\rho_Z}$, $B_Z := \frac{\gamma-1}{\gamma+1}p_Z$ 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 [27, Chapter 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 non-vacuum condition

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

holds, see Toro [27, (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 there is formation of vacuum or not, 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 2.3. (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 to compute p^* may require a large number of iterations in this situation, see Toro [27, 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 $\lambda_{\max} \leq \tilde{\lambda}_{\max} \leq (1+\epsilon)\lambda_{\max}$, is described in Guermond and Popov [10].

3. First order method.

We describe in this section 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 affine matching meshes. The elements in the mesh sequence are assumed to be generated from a finite number of reference elements denoted $\widehat{K}_1, \ldots, \widehat{K}_\varpi$. 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 affine 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 Lagrange finite elements $\{(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)\}_{1\le r\le\varpi}$ (the index $r \in \{1{:}\varpi\}$ will be omitted in the rest of the paper to alleviate the notation). Then we define the scalar-valued and vector-valued Lagrange 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). Denoting $n_{\rm sh} := \dim \widehat{P}$ and denoting by $\{\widehat{\boldsymbol{a}}_i\}_{i\in\{1:n_{\rm sh}\}}$ the Lagrange nodes of $\widehat{K}$, we assume that the space $\widehat{P}$ is such that

$$\min_{1\le\ell\le n_{\rm sh}} \widehat{v}(\widehat{\boldsymbol{a}}_\ell) \le \widehat{v}(\widehat{\boldsymbol{x}}) \le \max_{1\le\ell\le n_{\rm sh}} \widehat{v}(\widehat{\boldsymbol{a}}_\ell), \quad \forall \widehat{v} \in \widehat{P}, \forall \widehat{\boldsymbol{x}} \in \widehat{K}. \tag{3.2}$$

Denoting by $\mathbb{P}_1$ and $\mathbb{Q}_1$ the set of multivariate polynomials of total and partial degree at most 1, respectively; the above assumption holds for $\widehat{P} = \mathbb{P}_1$ when K is a simplex and $\widehat{P} = \mathbb{Q}_1$ when K is a parallelogram or a cuboid. This assumption holds also for first-order prismatic elements in three space dimensions.

Let $\{\boldsymbol{a}_i\}_{i\in\{1:I\}}$ be the collection of all the Lagrange nodes in the mesh $\mathcal{T}_h$, and let $\{\varphi_i\}_{i\in\{1:I\}}$ be the corresponding global shape functions. Recall that $\{\varphi_i\}_{i\in\{1:I\}}$ forms a basis of $P(\mathcal{T}_h)$ and $\varphi_i(\boldsymbol{a}_j) = \delta_{ij}$. The Lagrange interpolation operator in $\mathcal{P}(\mathcal{T}_h)$ is denoted $\Pi_h : \mathcal{C}^0(\overline{D}) \longrightarrow \mathcal{P}(\mathcal{T}_h)$. Recall that $\Pi_h(v) = \sum_{1\le i<I} v(\boldsymbol{a}_i)\varphi_i$. 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 $S_{ij} := S_i \cap S_j$ the intersection of the two supports S_i and S_j. Let E be a union of cells in $\mathcal{T}_h$; we define $\mathcal{I}(E) := \{j \in \{1{:}I\} \mid |S_j \cap E| \neq 0\}$ 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 $\boldsymbol{x} \in K$.

We define the operator $\mathsf{C} : P(\mathcal{T}_h) \longrightarrow \mathbb{R}^I$ so that $\mathsf{C}(v_h)$ is the coordinate vector of v_h in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, i.e., $v_h = \sum_{i=1}^I \mathsf{C}(v_h)_i\varphi_i$. Note that $\mathsf{C}(v_h)_i = v_h(\boldsymbol{a}_i)$. We are also going to use capital letters for the coordinate vectors to alleviate the notation; for instance we shall write $\mathsf{V} = \mathsf{C}(v_h)$ when the context is unambiguous. Note finally that the above assumptions on the mesh and the reference elements imply the following property: for all $\boldsymbol{x} \in K$ and all $K \in \mathcal{T}_h$,

$$\min_{\ell\in\mathcal{I}(K)} \mathsf{C}(v_h)_\ell \le v_h(\boldsymbol{x}) \le \max_{\ell\in\mathcal{I}(K)} \mathsf{C}(v_h)_\ell, \quad \forall v_h \in P(\mathcal{T}_h). \tag{3.3}$$

We define similarly the $\boldsymbol{\mathsf{C}} : \boldsymbol{P}(\mathcal{T}_h) \longrightarrow (\mathbb{R}^m)^I$, i.e., $\boldsymbol{v}_h = \sum_{i=1}^I \boldsymbol{\mathsf{C}}(\boldsymbol{v}_h)_i\varphi_i$, or equivalently $\boldsymbol{\mathsf{C}}(\boldsymbol{v}_h)_i = \boldsymbol{v}_h(\boldsymbol{a}_i)$.

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

$$m_i := \int_{S_i} \varphi_i(\boldsymbol{x})\,\mathrm{d}x. \tag{3.4}$$

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

3.2. The scheme. Let $\boldsymbol{u}_{h0} \in \mathcal{P}(\mathcal{T}_h)$ be a reasonable approximation of $\boldsymbol{u}_0$ (we shall 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 $\boldsymbol{u}_h^n \in \mathcal{P}(\mathcal{T}_h)$ be the space approximation of $\boldsymbol{u}$ at time t^n and set $\mathbf{U}^n = \mathbf{C}(\boldsymbol{u}_h^n)$. We propose to compute $\boldsymbol{u}_h^{n+1}$ by

$$m_i \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} + \int_D \nabla\cdot(\Pi_h \boldsymbol{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}x - \sum_{j\in\mathcal{I}(S_i)} d_{ij}\mathbf{U}_j^n = 0, \tag{3.5}$$

where $\mathbf{U}^{n+1} = \mathbf{C}(\boldsymbol{u}_h^{n+1})$ and the lumped mass matrix is used for the approximation of the time derivative. The coefficient d_{ij} is an artificial viscosity for the pair (i,j) that has yet to be clearly identified. For the time being we assume that

$$d_{ij} \geq 0, \quad \text{if} \quad i \neq j, \quad d_{ij} = d_{ji}, \quad \text{and} \quad d_{ii} := \sum_{i\neq j\in\mathcal{I}(S_i)} -d_{ji}. \tag{3.6}$$

Using that $\nabla\cdot(\Pi_h \boldsymbol{f}(\boldsymbol{u}_h^n)) = \sum_j \boldsymbol{f}(\mathbf{U}_j^n)\cdot\nabla\varphi_j$, the above equation simplifies into

$$m_i \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(S_i)} \boldsymbol{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - \mathbf{U}_j^n d_{ij} = 0, \tag{3.7}$$

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

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

Remark 3.1. (Conservation) The definition $d_{ii} := \sum_{i\neq j\in\mathcal{I}(S_i)} -d_{ji}$ implies that $\sum_{j\in\mathcal{I}(S_i)} d_{ji} = 0$, which in turn implies conservation, i.e., $\int_D \boldsymbol{u}_h^{n+1}\,\mathrm{d}x = \int_D \boldsymbol{u}_h^n\,\mathrm{d}x + \int_D \nabla\cdot(\Pi_h \boldsymbol{f}(\boldsymbol{u}_h^n))\,\mathrm{d}x$. Note also that the symmetry assumption in (3.6) implies $d_{ii} := \sum_{i\neq j\in\mathcal{I}(S_i)} -d_{ij}$, which is often easier to compute.

3.3. The convex combination argument. We motivate the choice of the artificial viscosity coefficients d_{ij} in this section. Observing that the partition of unity property $\sum_{j\in\mathcal{I}(S_i)} \varphi_j = 1$ and (3.6) imply conservation, i.e.,

$$\sum_{j\in\mathcal{I}(S_i)} \boldsymbol{c}_{ij} = 0, \qquad \sum_{j\in\mathcal{I}(S_i)} d_{ij} = 0. \tag{3.9}$$

we re-write (3.7) as follows:

$$m_i \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} = -\sum_{j\in\mathcal{I}(S_i)} (\boldsymbol{f}(\mathbf{U}_j^n) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{c}_{ij} + d_{ij}(\mathbf{U}_j^n + \mathbf{U}_i^n). \tag{3.10}$$

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

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n\Big(1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d_{ij}}{m_i}\Big) + \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d_{ij}}{m_i}\overline{\mathbf{U}}_{ij}^{n+1}. \tag{3.11}$$

where we have introduced the auxiliary quantities

$$\overline{\mathbf{U}}_{ij}^{n+1} := \frac{1}{2}(\mathbf{U}_j^n + \mathbf{U}_i^n) - (\boldsymbol{f}(\mathbf{U}_j^n) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}}. \tag{3.12}$$

A first key observation is that (3.11) is a convex combination provided τ is small enough. A second key observation at this point is that upon setting $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$, $\overline{\mathbf{U}}_{ij}^{n+1}$ is exactly of the form $\overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)$ as defined in (2.6) with a fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}$. 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}

$$d_{ij} := \max(\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{3.13}$$

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

Remark 3.2. (Symmetry) If either $\boldsymbol{a}_i$ or $\boldsymbol{a}_j$ is an interior node in the mesh, one integration by parts implies that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$, which in turn implies $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j) = \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j, \mathbf{U}_i)$. In conclusion $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)\|\boldsymbol{c}_{ij}\|_{\ell^2} = \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j, \mathbf{U}_i)\|\boldsymbol{c}_{ji}\|_{\ell^2}$ if either $\boldsymbol{a}_i$ or $\boldsymbol{a}_j$ is an interior node.

Remark 3.3. (Upwinding) Note that in the scalar one-dimensional case when the flux f is linear, (3.5) gives the usual upwinding first-order method.

4. Stability analysis. We analyze the stability properties of the scheme (3.5) with the viscosity defined in (3.13).

4.1. Invariant domain property. Upon defining $h_K := \operatorname{diam}(K)$, the global maximum mesh size is denoted $h = \max_{K\in\mathcal{T}_h} h_K$. The local minimum mesh size, $\underline{h}_K$, for any $K \in \mathcal{T}_h$ is defined as follows:

$$\underline{h}_K := \frac{1}{\max_{i\neq j\in\mathcal{I}(K)} \|\nabla\varphi_i\|_{\boldsymbol{L}^\infty(S_{ij})}}, \tag{4.1}$$

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 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 $n_{\mathrm{sh}} := \operatorname{card}(\mathcal{I}(K))$ and let us define $\vartheta_K := \frac{1}{n_{\mathrm{sh}}-1}$. Note that

$$0 < \vartheta_{\min} := \min_{(\mathcal{T}_h)_{h>0}} \min_{K\in\mathcal{T}_h} \vartheta_K < +\infty, \tag{4.2}$$

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

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

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

THEOREM 4.1. *Let $A \subset \mathcal{A}$ be an invariant set for (1.1) in the sense of Definition 2.3. Assume that A is convex and*

$$\lambda_{\max}(A) := \max_{\boldsymbol{n}\in S^{d-1}(\mathbf{0},1)} \max_{\boldsymbol{u}_L,\boldsymbol{u}_R\in A} \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) < \infty, \tag{4.4}$$

where $S^{d-1}(\mathbf{0}, 1)$ is the unit sphere in $\mathbb{R}^d$. Assume that $\boldsymbol{u}_{h0} \in A$ and τ is such that

$$2\tau\frac{\lambda_{\max}(A)}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}} \leq 1. \tag{4.5}$$

Then

(i) A is an invariant domain for the solution process $\boldsymbol{u}_h^n \longmapsto \boldsymbol{u}_h^{n+1}$ for all $n \geq 0$.
(ii) Given $n \geq 0$ and $i \in \{1{:}I\}$, let $B \subset A$ be a convex invariant set such that $\mathbf{U}_l^n \in B$ for all $l \in \mathcal{I}(S_i)$, then $\mathbf{U}_i^{n+1} \in B$.

Proof. We prove the statement (i) by induction. Assume that $\boldsymbol{u}_h^n \in A$ for some $n \geq 0$; we are going to prove that $\boldsymbol{u}_h^{n+1} \in A$. Note that $\boldsymbol{u}_{h0} \in A$ by assumption. Let $i \in \{1{:}I\}$ and consider the update (3.5) rewritten in the form (3.11). 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 (3.12) 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 $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}$. The definition

$$d_{ij} \geq \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \tag{4.6}$$

is the CFL condition for the conclusions of Lemma 2.1 to hold with fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}$. Since A is a convex invariant set we have $\overline{\mathbf{U}}_{ij}^{n+1} := \overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \in A$ for all $j \in \mathcal{I}(S_i)$. Let us now prove that (3.11) is indeed a convex combination by proving that $1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau d_{ij}}{m_i} = 1 + \frac{2\tau d_{ii}}{m_i} \geq 0$. Note first that

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

The definition of d_{ii} implies that

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

The using that $\mu_{\min}|S_i| \leq m_i$, we infer that

$$-2\tau\frac{d_{ii}}{m_i} \leq 2\tau\frac{\lambda_{\max}(A)}{\underline{h}}\frac{\mu_{\max}}{\mu_{\min}\vartheta_{\min}} \leq 1,$$

which proves the result owing to the CFL assumption (4.5). Hence (3.11) defines $\mathbf{U}_i^{n+1}$ as a convex combination between $\mathbf{U}_i^n$ and the collection of states $\{\overline{\mathbf{U}}_{ij}^{n+1}\}_{j\in\mathcal{I}(S_i)}$. The convexity of A implies that $\mathbf{U}_i^{n+1} \in A$, since $\mathbf{U}_i^n \in A$ by assumption and we have established above that $\overline{\mathbf{U}}_{ij}^{n+1} \in A$ for all $j \in \mathcal{I}(S_i)$. The space approximation being piecewise linear, a convexity argument implies again that $\boldsymbol{u}_h^{n+1} \in A$, which proves the induction assumption.

Note in passing that we have also proved the following local invariance property: given any convex invariant set $B \subset A$ that contains $\{\mathbf{U}_l^n\}_{l\in\mathcal{I}(S_i)}$, $\mathbf{U}_i^{n+1}$ is also in B, i.e., the local statement (ii) holds. This completes the proof. □

Remark 4.1. The arguments invoking the convex combination (3.11) and the one-dimensional Riemann averages (3.12) are similar in spirit to those used in the proof of Theorem 3 in Perthame and Shu [25].

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

THEOREM 4.2. *Let $A \subset \mathcal{A}$ 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 A, and any $\boldsymbol{n} \in S^{d-1}(\mathbf{0}, 1)$. Assume also that (4.4) and (4.5) hold, then we have the following for any $n \geq 0$ and any $i \in \{1{:}I\}$:*

$$\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{n+1}) - \eta(\mathbf{U}_i^n)) + \int_D \nabla\cdot(\Pi_h\boldsymbol{q}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}x + \sum_{i\neq j\in\mathcal{I}(S_i)} d_{ij}\eta(\mathbf{U}_j^n) \leq 0. \tag{4.7}$$

Proof. Let $(\eta, \boldsymbol{q})$ be an entropy pair for the system (1.1). Let $i \in \{1{:}I\}$, then recalling (3.11), the CFL condition and the convexity of η imply that

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

Owing to Lemma 2.2 we have

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

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

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

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

Remark 4.2. One recovers the equation (3.5) from (4.7) with $\eta(\boldsymbol{v}) = \boldsymbol{v}$. Note also that (4.7) 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 4.3. 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 4.2, 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 vs. edge-based viscosity. In the formulation (3.5) the term $\sum_{j\in\mathcal{I}(S_i)} d_{ij}\mathbf{U}_j$ models some edge-based dissipation, i.e., d_{ij} 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 [17, Eq. (32)-(33)], see also Jameson [14, §2.1]. It is however a bit difficult to understand that we are modeling some artificial dissipation by just staring at (3.5).

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:

$$\int_D -\nabla\cdot(\nu\nabla\psi)\varphi_i\,\mathrm{d}x = \sum_{K\subset S_i} \nu_K \int_K \nabla\psi\cdot\nabla\varphi_i\,\mathrm{d}x. \tag{4.8}$$

Unfortunately, it has been shown in Guermond and Nazarov [9] that the bilinear form $(\psi, \varphi) \longrightarrow \int_K \nabla\psi\cdot\nabla\varphi\,\mathrm{d}x$ is not robust with respect to the shape of the cells. More specifically, the convex combination argument, which is essential to prove 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 [9] to replace (4.8) by $\sum_{K\subset S_i}\nu_K b_K(\psi,\varphi_i)$, where

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

The essential properties of b_K can be summarized as follows:

LEMMA 4.3. *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 $u_h,v_h\in P(\mathcal{T}_h)$:*

$$b_K(\varphi_i,\varphi_j)=b_K(\varphi_j,\varphi_i),\qquad b_K(\varphi_i,\sum_{j\in\mathcal{I}(K)}\varphi_j)=0.\tag{4.10}$$

$$b_K(u_h,v_h)=\vartheta_K|K|\sum_{i\in\mathcal{I}(K)}\sum_{\mathcal{I}(K)\ni j<i}(\mathsf{U}_i-\mathsf{U}_j)(\mathsf{V}_i-\mathsf{V}_j)\tag{4.11}$$

$$b_K(u_h,u_h)\ge ch_K^2\|\nabla u_h\|^2_{\boldsymbol{L}^2(K)}.\tag{4.12}$$

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 $b_K(\varphi_i,\varphi_i)=\kappa\int_K\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_j)\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_i)\,\mathrm{d}x$ and $b_K(\varphi_j,\varphi_i)=-\frac{\kappa}{1-n_{\mathrm{sh}}}\int_K\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_j)\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla\varphi_i)\,\mathrm{d}x$ for $j\neq i$, with $\kappa=\frac12(1+\frac1d)$. Note also that $b_K(\varphi_j,\varphi_i)\sim h_K^2\int_K(\nabla\varphi_j)\cdot(\nabla\varphi_i)\,\mathrm{d}x$ 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 [9] is the recognition that the bilinear form defined in (4.9) has all the good characteristics of the Laplacian-based diffusion (see Lemma 4.3) and makes the convex combination argument to work independently of the space dimension and the shape-regularity of the mesh family.

Hence, instead of (3.5), we could also compute $\boldsymbol{u}_h^{n+1}$ by

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau}+\int_D\nabla\cdot(\Pi_h\boldsymbol{f}(\boldsymbol{u}_h^n))\varphi_i\,\mathrm{d}x+\sum_{K\in\mathcal{T}_h}\nu_K^n\sum_{j\in\mathcal{I}(K)}\mathsf{U}_j^n b_K(\varphi_j,\varphi_i)=0,\tag{4.13}$$

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

THEOREM 4.4. *Let $\{\nu_K^n\}_{K\in\mathcal{T}_h}$ be defined by*

$$\nu_K:=\max_{i\neq j\in\mathcal{I}(K)}\frac{\lambda_{\max}(\boldsymbol{n}_{ij},\mathsf{U}_i,\mathsf{U}_j)\|\boldsymbol{c}_{ij}\|_{\ell^2}}{\sum_{T\subset S_{ij}}-b_T(\varphi_j,\varphi_i)}.\tag{4.14}$$

Then the conclusions of Theorem 4.1 and Theorem 4.2 hold under the assumptions (4.4) *and* (4.5) *and with the solution process $\boldsymbol{u}_h^n\longrightarrow\boldsymbol{u}_h^{n+1}$, $n\ge0$, defined by* (4.13).

Proof. Let us denote $\tilde{d}_{ij}:=-\sum_{K\in S_{ij}}\nu_K^n b_K(\varphi_j,\varphi_i)$, then (4.13) can be recast as follows:

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

which in turn implies that

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

where we have introduced the auxiliary quantities

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

Here again $\overline{\mathbf{U}}_{ij}^{n+1}$ is of the form $\overline{\boldsymbol{u}}(t,\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j)$ as defined in (2.6) with the fake time $t = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2\tilde{d}_{ij}$, hence we need to make sure that $\lambda_{\max}(\boldsymbol{n}_{ij},\boldsymbol{u}_L,\boldsymbol{u}_R)\|\boldsymbol{c}_{ij}\|_{\ell^2}/2\tilde{d}_{ij} \le \frac{1}{2}$ to preserve the invariant domain property. Recalling that d_{ij} has been defined by $d_{ij} := \lambda_{\max}(\boldsymbol{n}_{ij},\boldsymbol{u}_L,\boldsymbol{u}_R)\|\boldsymbol{c}_{ij}\|_{\ell^2}$ (see (3.6)), the above condition reduces to showing that $d_{ij} \le \tilde{d}_{ij}$. The definitions of ν_K and $\tilde{d}_{ij}$ implies that

$$\begin{aligned}\tilde{d}_{ij} &= -\sum_{K\in S_{ij}} \nu_K^n b_K(\varphi_j,\varphi_i) \ge -\sum_{K\in S_{ij}} \frac{\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j)\|\boldsymbol{c}_{ij}\|_{\ell^2}}{\sum_{T\subset S_{ij}} -b_T(\varphi_j,\varphi_i)} b_K(\varphi_j,\varphi_i)\\ &\ge -\sum_{K\in S_{ij}} \frac{d_{ij}}{\sum_{T\subset S_{ij}} -b_T(\varphi_j,\varphi_i)} b_K(\varphi_j,\varphi_i) = d_{ij},\end{aligned}$$

whence the desired result. We now prove that $1-\sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\tau\tilde{d}_{ij}}{m_i} \ge 0$ under the CFL condition (4.5). From the proof of Theorem 4.1 we have $d_{ij} \le \lambda_{\max}(A)\underline{h}^{-1}\mu_{\max}|S_{ij}|$, hence

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

which in turn implies that

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

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

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

Finally we have

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

This means that the bound on $-\tilde{d}_{ii}$ is the same as that on $-d_{ii}$ in the proof of Theorem 4.1. This concludes the proof. □

5. Numerical illustrations. We illustrate in this section the method described in the paper, i.e., (3.5)-(3.13), and discuss possible variants.

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

To illustrate or point, let us focus our attention on scalar conservation equations and let us consider an algebraic approach that is sometimes used in the literature,

see e.g., Kuzmin et al. [18, p. 163], Kuzmin and Turek [17, Eq. (32)-(33)]. Instead of constructing a convex combination involving (entropy satisfying) intermediate states like in (3.11), we re-write (3.10) as follows:

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

Or, equivalently

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

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 $k_{ij}:=0$ if $\mathsf{U}_j^n=\mathsf{U}_i^n$), then

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

Let us finally set

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

This choice implies that $-k_{ij}+d_{ij}\geq 0$ for all $i\in\{1\!:\!N\}$, $j\in\mathcal{I}(S_i)$. As a result, $\mathsf{U}_i^{n+1}\in\text{conv}\{\mathsf{U}_j^n,\ j\in\mathcal{I}(S_i)\}$ 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 [11, §3.3]. The convergence result established in [11] 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 $(\mathsf{U}_i,\mathsf{U}_j)$ 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 et al. [16]. It is a two-dimensional scalar conservation equation with a non-convex 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 [16] that some central-upwind schemes based on WENO5, Minmod 2 and SuperBee reconstructions converge to non-entropic solutions.

The computational domain $[-2,2]\times[-2.5,1.5]$ is triangulated using non-uniform 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 5.1 is obtained using (5.4) for the definition of d_{ij}. The numerical solution produces very sharp, non-oscillating, 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 5.1 is obtained by using our definition of d_{ij}, (3.13) (note in passing that the results obtained with (4.13)-(4.14) together with (3.13) are indistinguishable from this solution). The expected helicoidal composite wave is clearly visible; this is the unique entropy satisfying solution.

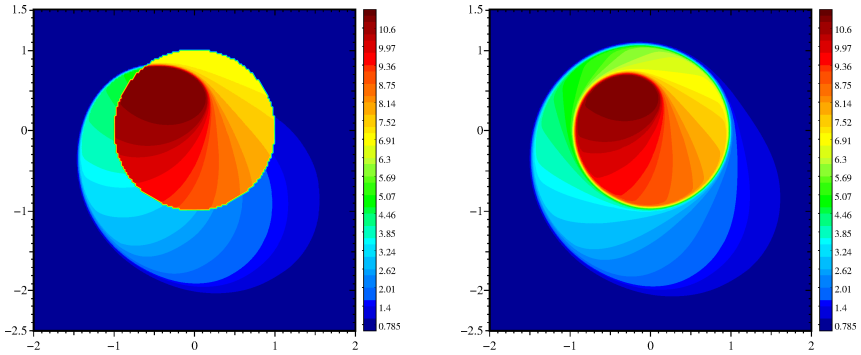

FIG. 5.1. *KPP solution with continuous $\mathbb{P}_1$ elements (29871 nodes, 59100 triangles). Left: entropy violating solution using* (3.5)-(5.4)*; Right: entropy satisfying solution using* (3.5)-(3.13).

In conclusion, the above counter-example 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 satisfies local entropy inequalities to be convergent; this is the case of our method (3.5)-(3.13) (see Theorem 4.2), 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 et al. [18, p. 163], Kuzmin and Turek [17, Eq. (32)-(33)] a little bit out of context. The scheme as originally presented in the above 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 nonconvex fluxes, as we did in (5.2).

5.2. Special meshes. The construction of the intermediate states in (3.10) is not unique. For instance we can extend a construction used by Hoff [12, Cor. 1] in one space dimension for the p-system. Let us assume that $i \in \{1,\dots,N\}$ is such that 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 $\boldsymbol{c}_{ij} := \int_{S_i} \phi_i \nabla\phi_j \,\mathrm{d}x = -\int_{S_i} \phi_i \nabla\phi_{\sigma_i(j)} \,\mathrm{d}x =: -\boldsymbol{c}_{i\sigma_i(j)}$. This property holds in one space dimension for any mesh if $\boldsymbol{a}_i$ is an interior node. It holds in higher space dimension provided the mesh has symmetry properties and $\boldsymbol{a}_i$ is an interior node; for instance it holds if the mesh is centrosymmetric, i.e., the support of ϕ_i is symmetric with respect to the node $\boldsymbol{a}_i$ for any $i \in \{1,\dots,N\}$. Then we can re-write (3.7) as follows:

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

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 $d_{ii} := -\sum_{j\in\mathcal{J}(S_i)}(d_{ij} +$

$d_{i\sigma_i(j)}$), we have

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

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

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

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 $\alpha_L = -\frac{d_{i\sigma_i(j)}}{d_{ij}+d_{i\sigma_i(j)}}$, $\alpha_R = \frac{d_{ij}}{d_{ij}+d_{i\sigma_i(j)}}$ and $t := \frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{d_{ij}+d_{i\sigma_i(j)}}$, provided

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

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

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

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

Note that the above argument holds only if $\boldsymbol{a}_i$ is an interior node 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 (4.6), i.e., $d_{ij} \geq \lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$.

In conclusion the diffusion matrix $(d_{ij})_{1\leq i,j\leq N}$ can be constructed as follows: (1) For every node i satisfying the symmetry property $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{i\sigma_i(j)}$ for every $j\in\mathcal{J}(S_i)$, we define $\widetilde{d}_{ij} = \widetilde{d}_{i\sigma_i(j)} = \lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_{\sigma_i(j)}^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$; (2) For every other index i 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 $d_{ij} := \max(\widetilde{d}_{ij},\widetilde{d}_{ji})$ for $j\neq i$ and $d_{ii} := -\sum_{i\neq j\in\mathcal{I}(S_i)}d_{ij}$. This construction guarantees conservation, i.e., $\sum_{i\in\mathcal{I}(S_j)}d_{ij} = 0$ and first-order consistency, i.e., $\sum_{j\in\mathcal{I}(S_i)}d_{ij} = 0$.

Remark 5.2. Quite surprisingly, in the case of scalar linear transport the above construction and the construction done in §3.3, (see definition (3.13)) give the same scheme (i.e., the same CFL).

5.3. Invariant domain property vs. 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.

5.4. 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., $w_1(v_L,u_L) = w_1(v_R,u_R)$):

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

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

This case is such that $(v^*, u^*) = (v_R, u_R)$, hence the second wave corresponding to the eigenvalues $\lambda_2^{\pm}$ is not present. We use continuous piecewise linear finite elements with the algorithm (3.5)-(3.13). The time stepping is done with the SSP RK3 technique. We show the profile of v at $t = 0.75$ in Figure 5.2 for meshes composed of $10^3, 2\times10^3, 4\times10^3, 10^4, 2\times10^4, 4\times10^4,\ 10^5,\ 2\times10^5$ cells. We observe that the profile

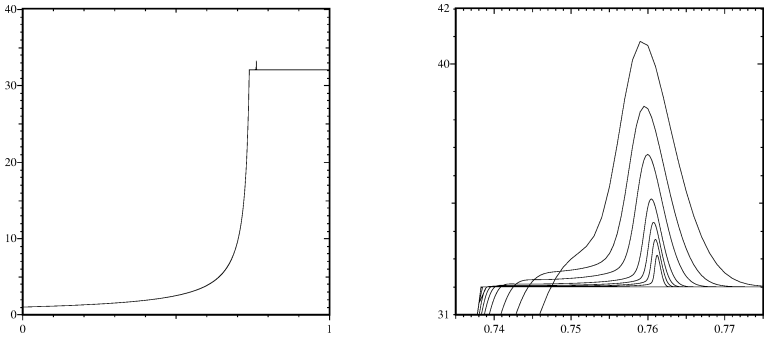

FIG. 5.2. *Left: v-profile for the p-system at $t = 0.75$, 10^5 grid points. Right: close up view of the v-profile for various grid sizes: $10^3, 2\times10^3, 4\times10^3, 10^4, 2\times10^4, 4\times10^4,\ 10^5$ grid points.*

is not monotone. There is an overshoot at the right of the foot of the (left-going) 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 5.3), the second wave reappears in the form of an overshoot at the end of the rarefaction wave (see right panel of the Figure 5.2).

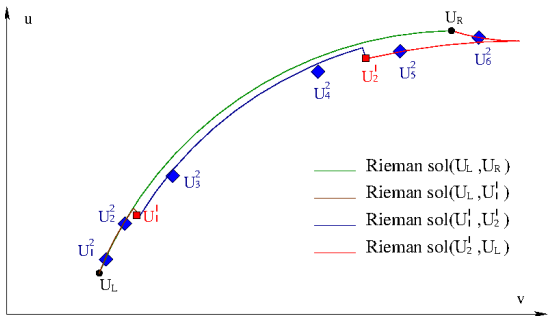

FIG. 5.3. *The overshooting mechanism for a single rarefaction wave in the phase space for the p-system. Initial data in black; additional points after one time step in red; after two time steps in blue. Observe the position of $\mathbf{U}_6^2$.*

Let $(\mathbf{U}_L \ldots, \mathbf{U}_L, \mathbf{U}_R \ldots, \mathbf{U}_R)$ be the initial sequence of degrees of freedom. After one time step two additional points appear in the phase space, denoted on Fig-

ure 5.3 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 $(\mathbf{U}_L \ldots, \mathbf{U}_L, \mathbf{U}_1^1, \mathbf{U}_2^2, \mathbf{U}_R \ldots, \mathbf{U}_R)$. Six additional points $\mathbf{U}_1^2, \ldots, \mathbf{U}_6^2$ appear after two time steps and the sequence of degrees of freedom at time $t = 2\tau$ is $(\mathbf{U}_L \ldots, \mathbf{U}_L, \mathbf{U}_1^2, \ldots, \mathbf{U}_6^2, \mathbf{U}_R \ldots, \mathbf{U}_R)$. The point $\mathbf{U}_6^2$ 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 5.3). 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 the Figure 5.2. 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 [15] and the comments at the bottom of page 1005 in Liska and Wendroff [21]. The (relative) L^1-norm of the error on both v and u at $t = 0.75$ is shown in Table 5.1. The method converges with an order close to 0.9.

$1/h$	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

TABLE 5.1
Convergence rates for the p-system

5.5. Euler in 1D (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

$$\rho_L = 1.000, \quad u_L = 0.0, \quad p_L = 0.1$$
$$\rho_R = 0.001, \quad u_R = 0.0, \quad p_R = 10^{-15}.$$

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 (3.5)-(3.13) using piecewise linear finite elements. The density profile computed with 50,000, 100,000, 200,000, 400,000 and 800,000 grid points is shown in the left panel of Figure 5.4. The right panel in the figure 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

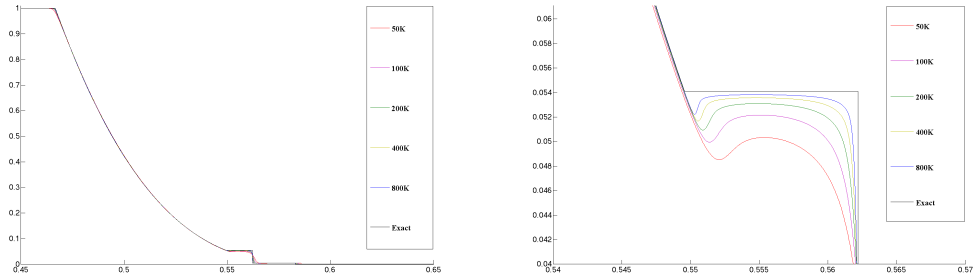

FIG. 5.4. *Left: Density profile for the Leblanc Shocktube at $t = 0.1$. Right: close up view of the density profile at the foot of the rarefaction wave.*

different concept than monotonicity, and just looking at monotonicity is not enough to understand hyperbolic systems.

6. Concluding remarks. We have proposed a numerical method to solve 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 sets is an invariant domain for the method. The main result of the paper are Theorem 4.1 and Theorem 4.2. 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 to solve 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 et al. [18]) to make the method at least formally second-order accurate in space and still be domain invariant.

References.

[1] S. Bianchini and A. Bressan. Vanishing viscosity solutions of nonlinear hyperbolic systems. *Ann. of Math. (2)*, 161(1):223–342, 2005.

[2] A. Bressan. *Hyperbolic systems of conservation laws*, volume 20 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2000. The one-dimensional Cauchy problem.

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

[4] K. N. Chueh, C. C. Conley, and J. A. Smoller. Positively invariant regions for systems of nonlinear diffusion equations. *Indiana Univ. Math. J.*, 26(2):373–392, 1977.

[5] P. Colella. Multidimensional upwind methods for hyperbolic conservation laws. *J. Comput. Phys.*, 87(1):171–200, 1990.

[6] M. G. Crandall and A. Majda. Monotone difference approximations for scalar conservation laws. *Math. Comp.*, 34(149):1–21, 1980.

[7] C. M. Dafermos. *Hyperbolic conservation laws in continuum physics*, volume 325 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2000.

[8] H. Frid. Maps of convex sets and invariant regions for finite-difference systems of conservation laws. *Arch. Ration. Mech. Anal.*, 160(3):245–269, 2001.
[9] J.-L. Guermond and M. Nazarov. A maximum-principle preserving C^0 finite element method for scalar conservation equations. *Comput. Methods Appl. Mech. Engrg.*, 272:198–213, 2013.
[10] J.-L. Guermond and B. Popov. Fast estimation of the maximum speed in a riemann problem. 2015. Submitted.
[11] 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(1):57–85, 2016.
[12] D. Hoff. A finite difference scheme for a system of two conservation laws with artificial viscosity. *Math. Comp.*, 33(148):1171–1193, 1979.
[13] D. Hoff. Invariant regions for systems of conservation laws. *Trans. Amer. Math. Soc.*, 289(2):591–610, 1985.
[14] A. Jameson. Positive schemes and shock modelling for compressible flows. *Internat. J. Numer. Methods Fluids*, 20(8-9):743–776, 1995. Finite elements in fluids—new trends and applications (Barcelona, 1993).
[15] A. Kurganov and E. Tadmor. Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers. *Numer. Methods Partial Differential Equations*, 18(5):584–608, 2002.
[16] A. Kurganov, G. Petrova, and B. Popov. Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws. *SIAM Journal on Scientific Computing*, 29(6):2381–2401, 2007.
[17] D. Kuzmin and S. Turek. Flux correction tools for finite elements. *Journal of Computational Physics*, 175(2):525–558, 2002.
[18] D. Kuzmin, R. Löhner, and S. Turek. *Flux–Corrected Transport*. Scientific Computation. Springer, 2005. 3-540-23730-5.
[19] P. D. Lax. Hyperbolic systems of conservation laws. II. *Comm. Pure Appl. Math.*, 10:537–566, 1957.
[20] 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(6):599–638, 1996.
[21] 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(3):995–1017 (electronic), 2003.
[22] T. P. Liu. The Riemann problem for general systems of conservation laws. *J. Differential Equations*, 18:218–234, 1975.
[23] T. Nishida. Global solution for an initial boundary value problem of a quasilinear hyperbolic system. *Proc. Japan Acad.*, 44:642–646, 1968.
[24] S. Osher. The Riemann problem for nonconvex scalar conservation laws and Hamilton-Jacobi equations. *Proc. Amer. Math. Soc.*, 89(4):641–646, 1983.
[25] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1):119–130, 1996.
[26] J. Smoller. *Shock waves and reaction-diffusion equations*, volume 258 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science]*. Springer-Verlag, New York-Berlin, 1983.
[27] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, third edition, 2009. A practical introduction.
[28] R. Young. The p-system. I. The Riemann problem. In *The legacy of the inverse*

scattering transform in applied mathematics (South Hadley, MA, 2001), volume 301 of *Contemp. Math.*, pages 219–234. Amer. Math. Soc., Providence, RI, 2002.