Journal of Scientific Computing (2024) 100:46
https://doi.org/10.1007/s10915-024-02592-4

First-Order Greedy Invariant-Domain Preserving Approximation for Hyperbolic Problems: Scalar Conservation Laws, and p-System

Jean-Luc Guermond[1] · **Matthias Maier**[1] · **Bojan Popov**[1,4] · **Laura Saavedra**[2] · **Ignacio Tomas**[3]

Received: 4 October 2023 / Revised: 6 June 2024 / Accepted: 9 June 2024

Abstract
The paper focuses on first-order invariant-domain preserving approximations of hyperbolic systems. We propose a new way to estimate the artificial viscosity that has to be added to make explicit, conservative, consistent numerical methods invariant-domain preserving and entropy inequality compliant. Instead of computing an upper bound on the maximum wave speed in Riemann problems, we estimate a minimum wave speed in the said Riemann problems such that the approximation satisfies predefined invariant-domain properties and predefined entropy inequalities. This technique eliminates non-essential fast waves from the construction of the artificial viscosity, while preserving pre-assigned invariant-domain properties and entropy inequalities.

Keywords Conservation equations · Hyperbolic systems · Invariant domains · Convex limiting · Finite element method

Mathematics Subject Classification 35L65 · 65M60 · 65M12 · 65N30

1 Introduction

Let us consider the hyperbolic system of conservation equations $\partial_t \boldsymbol{u} + \nabla \cdot \mathbb{f}(\boldsymbol{u}) = \boldsymbol{0}$, where $\boldsymbol{u}$ denotes a conserved state taking values in $\mathbb{R}^m$ and $\mathbb{f}(\boldsymbol{u})$ an associated flux taking values in $\mathbb{R}^{m\times d}$, where d is the space dimension. Most explicit approximation methods for solving

✉ Jean-Luc Guermond
guermond@tamu.edu

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

2 Departamento de Matemática Aplicada a la Ingeniería Aeroespacial, E.T.S.I. Aeronáutica y del Espacio, Universidad Politécnica de Madrid, 28040 Madrid, Spain

3 Department of Mathematics and Statistics, Texas Tech University, 2500 Broadway, Lubbock, TX 79409, USA

4 Department of Mathematics and Informatics, University of Sofia, 5 James Boucher Blvd., 1164 Sofia, Bulgaria

Published online: 27 June 2024

this type of system are based on some notion of numerical flux and involve some numerical dissipation. For instance, all the first-order methods based on Lax's seminal paper [20, p. 163] involve numerical fluxes between pairs of degrees of freedom, say i, j, that take the following form $\frac{1}{2}(\mathbb{f}(\mathbf{U}_i)+\mathbb{f}(\mathbf{U}_j))\boldsymbol{n}_{ij}+\alpha_{ij}(\mathbf{U}_i-\mathbf{U}_j)$, where $\boldsymbol{n}_{ij}$ is some unit vector associated with the space discretization at hand and α_{ij} is an upper bound on the maximum wave speed in the Riemann problem using the flux $\mathbb{f}(\mathbf{U})\boldsymbol{n}_{ij}$ and the states $\mathbf{U}_i$ and $\mathbf{U}_j$ as left and right Riemann data. Denoting by $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j)$ the maximum wave speed in the Riemann problem in question, it is now well established that choosing α_{ij} such that $\alpha_{ij}\geq\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j)$ guarantees that some invariant-domain property can be extracted from the scheme; see e.g., Harten et al. [12], Tadmor [29, p. 375], Perthame and Shu [26, §5]. Using $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j)$ to construct invariant-domain preserving schemes dates back to the origins of computational fluid dynamics; we refer the reader for instance to [20, p. 163]. Recalling that the flux $\alpha_{ij}(\mathbf{U}_i-\mathbf{U}_j)$ is associated with numerical dissipation and therefore induces a loss of accuracy, a natural question to ask is whether it is possible to estimate a greedy value for α_{ij} in the open interval $(0,\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i,\mathbf{U}_j))$ guaranteeing that the scheme satisfies the desired invariant-domain properties and relevant entropy inequalities. It is the purpose of the present paper to give a positive answer to this question. The paper is the result of a research project that was initiated at the 9th International Conference on Numerical Methods for Multi-Material Fluid Flow, held in Trento, Italy, 9–13, September 2019. Some of the questions posed above and some answers thereto were outlined in [9].

To convince the reader that the program described above is feasible, let us consider the compressible Euler equations equipped with a γ-law, and let us consider the Riemann problem with the flux $\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n}$ and some left and right data $\boldsymbol{u}_L,\boldsymbol{u}_R$. Then denoting by λ_1^-,λ_1^+ the two wave speeds enclosing the 1-wave, λ_2 the speed of the 2-wave (i.e., the contact discontinuity), and λ_3^-,λ_3^+ the two wave speeds enclosing the 3-wave, we have $\lambda_1^-\leq\lambda_1^+\leq\lambda_2\leq\lambda_3^-\leq\lambda_3^+$, and the maximum wave speed in the Riemann problem is $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R):=\max(|\lambda_1^-|,|\lambda_3^+|)>|\lambda_2|$. If the Riemann data yields a solution that consists of just one contact discontinuity, one can establish that the amount of viscosity that is sufficient to satisfy all the invariant domain properties (in addition to local entropy inequalities) is just $|\lambda_2|$ (because the velocity and the pressure are constant in this case). Hence setting the graph viscosity wave speed α to be larger than or equal to $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ is needlessly over-diffusive since taking $\alpha=|\lambda_2|$ is sufficient in this case. Invoking the continuous dependence of the Riemann solution with respect to the data, one then realizes that a similar conclusion holds if the Riemann data is a small perturbation of a data set producing a contact discontinuity only. The situation described above is well illustrated by the multi-material Euler equations in Lagrangian coordinates. In this case the interface between two materials is a contact discontinuity that should keep its integrity over time. Let $v:=\boldsymbol{u}\cdot\boldsymbol{n}$ denote the component of the material velocity normal to the interface. The maximum wave speed in the Riemann problem using the two states on either sides of the interface gives $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)=\max(|\lambda_1^--v|,|\lambda_3^+-v|)$ in the Lagrangian reference frame, whereas the wave speed of the 2-wave is $\lambda_2-v=0$. In this case the amount of viscosity that is sufficient to satisfy all the invariant domain properties is $\alpha=\lambda_2-v=0$. Hence, if one instead uses $\alpha=\lambda_{\max}:=\max(|\lambda_1^--v|,|\lambda_3^+-v|)$ (as suggested e.g., in Guermond et al. [6] and most of the literature on the topic) one needlessly diffuses the contact discontinuity. The purpose of the present paper is to clarify the issues described above and derive a variation of the method presented in [4, 6] that is invariant-domain preserving, satisfies discrete entropy inequalities, and minimizes the amount of artificial viscosity used.

The first-order method presented in the paper can be made high-order and still be invariant-domain preserving by using one of many techniques developed to this effect and available in

the abundant literature dedicated to the topic. This can be done by adapting the flux transport corrected methodology from Zalesak [34, §II]. For instance, one can use methods inspired from [18, 19] when the functionals to limit are affine. When these functionals are nonlinear, one can use methods from Kuzmin and Turek [35], Kuzmin et al. [36] (for discontinuous finite elements) or from [7, 8] (for continuous finite elements). A complete list of all the excellent methods capable of achieving this goal cannot be cited here.

The paper is organized as follows. We formulate the problem and recall important concepts that are used in the paper in Sect. 2. We introduce in Sect. 3 the concept of greedy viscosity for any hyperbolic system. The key results of this section are the Definitions (3.7) and (3.8) and Theorem 3.6. The concept of greedy viscosity is then illustrated for scalar conservation equations in Sect. 4. The main result summarizing the content of this section are the definitions (4.2), (4.3) and Theorem 4.3. The concept is further illustrated for the p-system in Sect. 5. The ideas introduced in the paper are numerically illustrated in Sect. 6 for scalar conservation equations and for the p-system. Some of these tests are meant to illustrate that estimating a greedy wave speed in order to preserve the invariant-domain is not sufficient to converge to an entropy solution. Ensuring that entropy inequalities are satisfied is essential for this matter. We also show that using just one entropy is not sufficient for scalar conservation equations with a non-convex flux. Due to lack of space, the concept of greedy viscosity for systems like the compressible Euler equations equipped with a tabulated equation of state will be illustrated in a forthcoming second part of this work. A short outline of the performance of the method is given in the conclusions section, see Sect. 8.

2 Formulation of the Problem

In this section we formulate the question that is addressed in the paper and put it in context.

2.1 The Hyperbolic System

Our objective is to develop elementary and robust numerical tools to approximate hyperbolic systems in conservation form:

$$\begin{cases} \partial_t \boldsymbol{u} + \nabla\cdot\mathbb{f}(\boldsymbol{u}) = \boldsymbol{0}, \quad \text{for } (\boldsymbol{x},t) \in D\times\mathbb{R}_+, \\ \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}), \quad \text{for } \boldsymbol{x} \in \mathbb{R}^d. \end{cases} \tag{2.1}$$

Here d is the space dimension, D is a compact, connected, polygonal subset of $\mathbb{R}^d$. To avoid difficulties related to boundary conditions, we either solve the Cauchy problem or assume that the boundary conditions are periodic. The dependent variable (or state variable) $\boldsymbol{u}$ takes values in $\mathbb{R}^m$. The function $\mathbb{f} : \mathcal{A} \to (\mathbb{R}^m)^d$ is called flux. The domain of $\mathbb{f}$, i.e., $\mathcal{A} \subset \mathbb{R}^m$, is called *admissible set*. The state variable $\boldsymbol{u}$ is viewed as a column vector $\boldsymbol{u} = (u_1, \ldots, u_m)^\mathsf{T}$. The flux is a $m\times d$ matrix with entries $\mathbb{f}_{ik}(\boldsymbol{u}(\boldsymbol{x}))$, $i \in \{1{:}m\}$, $k \in \{1{:}d\}$ and $\nabla\cdot\mathbb{f}(\boldsymbol{u}(\boldsymbol{x}))$ is a column vector with entries $(\nabla\cdot\mathbb{f}(\boldsymbol{u}))_i = \sum_{k\in\{1:d\}} \partial_{x_k}\mathbb{f}_{ik}(\boldsymbol{u}(\boldsymbol{x}))$. For any $\boldsymbol{n} = (n_1 \ldots, n_d)^\mathsf{T} \in \mathbb{R}^d$, we denote $\mathbb{f}(\boldsymbol{u})\boldsymbol{n}$ the column vector with entries $\sum_{l\in\{1:d\}} \mathbb{f}_{il}(\boldsymbol{u})n_l$, where $i \in \{1{:}m\}$.

We assume in the entire paper that the admissible set $\mathcal{A} \subset \mathbb{R}^m$ is constructed such that for every pair of states $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A}\times\mathcal{A}$ and every unit vector $\boldsymbol{n}$ in $\mathbb{R}^d$, the following one-dimensional Riemann problem

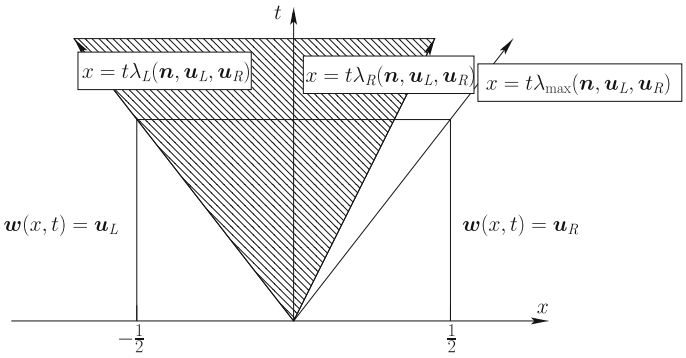


Fig. 1 Riemann problem and Riemann fan

$$\partial_t \boldsymbol{w} + \partial_x(\mathbb{f}(\boldsymbol{w})\boldsymbol{n}) = 0, \quad (x,t) \in \mathbb{R}\times\mathbb{R}_+, \qquad \boldsymbol{w}(x,0) = \begin{cases} \boldsymbol{u}_L, & \text{if } x < 0 \\ \boldsymbol{u}_R, & \text{if } x > 0, \end{cases} \tag{2.2}$$

has a unique solution satisfying adequate entropy inequalities. We assume that this solution is self-similar with self-similarity parameter $\xi := \frac{x}{t}$, and we set

$$\boldsymbol{v}\left(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R, \frac{x}{t}\right) := \boldsymbol{w}(x,t); \tag{2.3}$$

see for instance Lax [21], Toro [31]. Using Lax's notation, we denote by $\lambda_1^- \le \lambda_1^+$ the two wave speeds enclosing the 1-wave (i.e., the leftmost wave) and $\lambda_m^- \le \lambda_m^+$ the two wave speeds enclosing the m-wave (i.e., the rightmost wave). The key result that we are going to use in the paper is that $\boldsymbol{v}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R, \xi) = \boldsymbol{u}_L$ if $\xi \le \lambda_1^-$ and $\boldsymbol{v}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R, \xi) = \boldsymbol{u}_R$ if $\xi \ge \lambda_m^+$. We define a *left wave speed* $\lambda_L(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \lambda_1^-$ and a *right wave speed* $\lambda_R(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \lambda_m^+$. We also define *the maximum wave speed* of the Riemann problem to be

$$\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \max(|\lambda_L(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)|, |\lambda_R(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)|). \tag{2.4}$$

We will replace the notation $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ by $\lambda_{\max}$ when the context is unambiguous. For further reference, for every $t > 0$ we define

$$\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{v}\left(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R, \tfrac{x}{t}\right) \mathrm{d}x. \tag{2.5}$$

Notice that if $t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \frac{1}{2}$, then $\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ is the average of the entropy solution of the Riemann problem (2.2) over the Riemann fan. This property is illustrated in Fig. 1.

Definition 2.1 *(Invariant domain)* We say that $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ is invariant domain for (2.1) if the following holds true: (i) $\mathcal{B}$ is convex; (ii) for any pair $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{B}\times\mathcal{B}$, any unit vector $\boldsymbol{n} \in \mathbb{R}^d$, and all $t \in \left(0, \frac{1}{2\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)}\right)$, we have $\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{B}$, where $\overline{\boldsymbol{v}}$ is given by (2.5).

Lemma 2.2 (Invariance of the auxiliary states) *Let $\mathcal{B} \subset \mathcal{A}$ be any invariant domain for (2.1). Let $(\eta, \boldsymbol{q})$ be an entropy pair for (2.1). Let $\lambda > 0$, let $\boldsymbol{n} \in \mathbb{R}^d$ be a unit vector. For all $\boldsymbol{u}_L$, $\boldsymbol{u}_R$ in $\mathcal{A}$, consider the following auxiliary state:*

$$\overline{\boldsymbol{u}}_{LR}(\lambda) := \frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R) - \frac{1}{2\lambda}\left(\mathbb{f}(\boldsymbol{u}_R) - \mathbb{f}(\boldsymbol{u}_L)\right)\boldsymbol{n}. \tag{2.6}$$

Assume that $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathcal{B}$, *and* $\lambda \geq \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. *Then*

$$\overline{\boldsymbol{u}}_{LR}(\lambda) = \overline{\boldsymbol{v}}\left(\tfrac{1}{2\lambda}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R\right), \tag{2.7}$$

$$\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}, \tag{2.8}$$

$$\eta(\overline{\boldsymbol{u}}_{LR}(\lambda)) \leq \tfrac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) - \tfrac{1}{2\lambda}\left(\boldsymbol{q}(\boldsymbol{u}_R) - \boldsymbol{q}(\boldsymbol{u}_L)\right)\cdot\boldsymbol{n} \tag{2.9}$$

Proof See e.g., Lemma 2.1 and Lemma 2.2 in [4]. □

2.2 Agnostic Space Approximation

Without going into details, we now assume that we have at hand a fully discrete scheme where time is approximated by using the forward Euler time stepping and space is approximated by using some "centered" approximation of (2.1), i.e., without any artificial viscosity to stabilize the approximation. We denote by t^n the current time, $n \in \mathbb{N}$, and we denote by τ the current time step size; that is $t^{n+1} := t^n + \tau$ (we should write τ^n as the time step may vary at each time step, but we omit the super-index ^{n} to simplify the notation). Let us assume that the current approximation is a collection of states $\{\mathbf{U}_i^n\}_{i\in\mathcal{V}}$, where the index set $\mathcal{V}$ is used to enumerate all the degrees of freedom of the approximation. We assume that the "centered" update is given by $\mathbf{U}_i^{\mathrm{G},n+1}$ with

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{G},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} = \mathbf{0}. \tag{2.10}$$

The quantity m_i is called lumped mass and we assume that $m_i > 0$ for all $i \in \mathcal{V}$. The index set $\mathcal{I}(i)$ is called local stencil. This set collects only the degrees of freedom in $\mathcal{V}$ that interact with i. We set $\mathcal{I}(i)^* := \mathcal{I}(i)\backslash\{i\}$. The vector $\boldsymbol{c}_{ij} \in \mathbb{R}^d$ encodes the space discretization. We view $\frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}$ as a Galerkin (or centered or inviscid) approximation of $\nabla\cdot\mathbb{f}(\boldsymbol{u})$ at time t^n at some grid point (or cell) $i \in \mathcal{V}$. The superscript $^{\mathrm{G}}$ is meant to remind us that (2.10) is a Galerkin (or inviscid or centered) approximation of (2.1). That is, we assume that the consistency error in space in (2.10) scales optimally with respect to the mesh size for the considered approximation setting. We keep the discussion at this abstract level for the sake of generality; see Remark 2.3. The only requirement that we make on the coefficients $\boldsymbol{c}_{ij}$ is that the method is conservative; that is to say, we assume that

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji} \quad \text{and} \quad \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \mathbf{0}. \tag{2.11}$$

An immediate consequence of this assumption is that the total mass is conserved: $\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{G},n+1} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^n$.

Of course, Eq. (2.10) is in general not appropriate if the solution to (2.1) is not smooth. To recover some sort of stability (the exact notion of stability we have in mind is defined in Theorem 2.4) we modify the scheme by adding a graph viscosity based on the stencil $\mathcal{I}(i)$; that is, we compute the stabilized update $\mathbf{U}_i^{n+1}$ by setting:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{I}(i)^*} d_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n) = \mathbf{0}. \tag{2.12}$$

Here d_{ij}^n is the yet to be defined graph viscosity. We assume that

$$d_{ij}^n = d_{ji}^n > 0, \quad \text{if} \quad i \neq j. \tag{2.13}$$

The symmetry implies that the method remains conservative. The question addressed in the paper is the following: how large has d_{ij}^n to be chosen so that (2.12) preserves invariant domains and satisfies entropy inequalities (for some finite collection of entropies)?

Remark 2.3 *(Literature)* The algorithm (2.12) is a generalization of [20, p. 163]; see also Harten et al. [12], Tadmor [29, p. 375], Perthame and Shu [26, §5] and the literature cited in these references. The reader is referred to [4, 7] for realizations of the above algorithm with continuous finite elements. Realizations of the scheme with finite volumes and discontinuous elements are described in [8] and implemented in Kronbichler et al. [23], Maier et al. [16].

2.3 The Auxiliary Bar States

We now recall the main stability result established in [4]. The proof of this result is the source of inspiration for the rest of the paper. For all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$ we introduce the unit vector $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$. Given two states $\mathbf{U}_i^n$ and $\mathbf{U}_j^n$ in $\mathcal{A}$, we recall that $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is the maximum wave speed in the Riemann problem defined in Sect. 2.1 with left state $\mathbf{U}_i^n$, right state $\mathbf{U}_j^n$, and unit vector $\boldsymbol{n}_{ij}$. The guaranteed maximum speed (GMS) graph viscosity $d_{ij}^{\mathrm{GMS},n}$ is defined in [4] as follows:

$$d_{ij}^{\mathrm{GMS},n} := \max\left(\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}\right). \tag{2.14}$$

Theorem 2.4 (Local invariance) *Let $\mathcal{B} \subset \mathcal{A}$ be any invariant domain for (2.1). Let $(\eta, \boldsymbol{q})$ be any entropy pair for (2.1). Let $n \geq 0$ and $i \in \mathcal{V}$. Let d_{ij}^n be any graph viscosity such that $d_{ij}^n \geq d_{ij}^{\mathrm{GMS},n}$ and $d_{ij}^n > 0$. Assume that $0 < \tau \leq m_i / \sum_{j\in\mathcal{I}(i)^*} 2d_{ij}^n$. Assume that $\{\mathbf{U}_j^n\}_{j\in\mathcal{I}(i)} \subset \mathcal{B}$. Then the update $\{\mathbf{U}_i^{n+1}\}_{i\in\mathcal{V}}$ given by (2.12) satisfies the following properties:*

$$\mathbf{U}_i^{n+1} \in \mathcal{B}, \tag{2.15}$$

$$\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{n+1}) - \eta(\mathbf{U}_i^n)) + \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\boldsymbol{q}(\mathbf{U}_j^n) - \sum_{j\in\mathcal{I}(i)^*} d_{ij}^n(\eta(\mathbf{U}_j^n) - \eta(\mathbf{U}_j^n)) \leq 0. \tag{2.16}$$

Proof We refer to Theorem 4.1 and Theorem 4.5 in [4] for detailed proofs. But since these proofs contain ideas that are going to be used latter in the paper, we now reproduce the key arguments. Using the conservation property (2.11), i.e., $\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \mathbf{0}$, we rewrite (2.12) as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)^*}\left(2d_{ij}^n\mathbf{U}_i^n + (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^n(\mathbf{U}_j^n + \mathbf{U}_i^n)\right) = \mathbf{0}.$$

Then, recalling that $d_{ij}^n > 0$ by assumption, we introduce the auxiliary states

$$\overline{\mathbf{U}}_{ij}^n := \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{n}_{ij}\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}. \tag{2.17}$$

This allows us to rewrite (2.12) as follows:

$$\mathbf{U}_i^{n+1} = \left(1 - \sum_{j\in\mathcal{I}(i)^*}\frac{2\tau d_{ij}^n}{m_i}\right)\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)^*}\frac{2\tau d_{ij}^n}{m_i}\overline{\mathbf{U}}_{ij}^n. \tag{2.18}$$

Since we assumed that $1 - 2\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)^*} d_{ij}^n > 0$, the right-hand side in the above identity is a convex combination of the states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)}$ with the convention $\overline{\mathbf{U}}_{ii}^n := \mathbf{U}_i^n$. Setting

$\lambda_{ij} := \frac{d_{ij}^n}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}$ and recalling definition (2.6), we observe that $\overline{\mathbf{U}}_{ij}^n = \overline{\boldsymbol{u}}_{ij}(\lambda_{ij})$ (here, with slight abuse of notation, we write $\overline{\boldsymbol{u}}_{ij}(\lambda)$ instead of $\overline{\boldsymbol{u}}_{LR}(\lambda)$). Then the assumption $d_{ij}^n \geq d_{ij}^{\mathrm{GMS},n}$ implies that $\lambda_{ij} \geq \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$, and the rest of the proof readily follows by invoking Lemma 2.2 (in particular $\overline{\mathbf{U}}_{ij}^n = \overline{\boldsymbol{u}}_{ij}(\lambda_{ij}) = \overline{\boldsymbol{v}}(\frac{1}{2\lambda_{ij}}, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$). □

Remark 2.5 *(λ_ϵ and $\lambda_{\max}^{\mathcal{V}}$)* The expression (2.17) (and thereby the identity (2.18) as well) is ill-defined if $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) = 0$, (recall that Lemma 2.2 requires that one should take $\lambda \geq \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$). To avoid the division by zero issue, we introduce a small number $\epsilon \in (0, 1)$ and we define

$$\lambda_{\max}^{\mathcal{V}} := \max_{i\in\mathcal{V}, j\in\mathcal{I}(i)} \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n), \qquad \lambda_\epsilon := \epsilon\lambda_{\max}^{\mathcal{V}}, \tag{2.19a}$$

$$\lambda_{ij}^\sharp := \max(\lambda_\epsilon, \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)). \tag{2.19b}$$

Henceforth we assume that $\lambda_{\max}^{\mathcal{V}} > 0$, which implies $\lambda_\epsilon > 0$. Otherwise the wave speed is zero everywhere, the solution is constant in time, and there is nothing to update. We are now going to consider the auxiliary states $\overline{\boldsymbol{u}}_{ij}(\lambda)$ and (2.17) for $\lambda \in [\lambda_\epsilon, \lambda_{ij}^\sharp]$.

Remark 2.6 *(Key observation)*

The statements (2.15) and (2.16) in Theorem 2.4 are consequences of (2.8)–(2.9) in Lemma 2.2. And the assertions (2.8)–(2.9) hold true because $\lambda \geq \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ implies the identity (2.7), i.e., $\overline{\boldsymbol{u}}_{LR} = \overline{\boldsymbol{v}}(\frac{1}{2\lambda}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. We note, though, that $\lambda \geq \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ (and thus identity (2.7)) is just a *sufficient* condition for (2.8)–(2.9) to hold true. The remainder of the paper is dedicated to estimating a *greedy* wave speed $\lambda_{LR}^{\mathrm{grdy}} \in [\lambda_\epsilon, \lambda_{LR}^\sharp]$ (depending on $\boldsymbol{n}$, $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$) that is as small as possible so that (2.8)–(2.9) still holds, although (2.7) may no longer hold. For this wave speed $\lambda_{ij}^{\mathrm{grdy}}$ all the assertions in Theorem 2.4 still hold true after redefining the viscosity $d_{ij}^n := \max\left(\lambda_{ij}^{\mathrm{grdy}}\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{ji}^{\mathrm{grdy}}\|\boldsymbol{c}_{ji}\|_{\ell^2}\right)$.

This minimization program is reasonable since in the worst case scenario setting $\lambda = \lambda_{LR}^\sharp \geq \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ is always admissible, i.e., the minimizing set for λ is not empty.

Remark 2.7 *(Literature)* The importance of the auxiliary states $\overline{\boldsymbol{u}}_{LR}(\lambda)$, which are the backbone of Lax's scheme, has been recognized in Nessyahu and Tadmor [24, Eq. (2.6)]. That these states are averages of Riemann solutions provided λ is larger than $\lambda_{\max}$ is well documented in Harten et al. [12, §3.A] (see also the reference to a private communication with Harten at p. 375, line 12 in Tadmor [29]). A variant of Lemma 2.2 is invoked to prove Theorem 3.1 in [12]. This theorem is a somewhat simplified version of Theorem 2.4.

3 Greedy Wave Speed and Greedy Viscosity

The key idea of the paper is introduced in this section. Let $\mathcal{B}$ be a convex invariant domain for (2.1). In this entire section $\boldsymbol{n}$ is a unit vector and $\boldsymbol{u}_L, \boldsymbol{u}_R$ are two states in $\mathcal{B}$. The important results of this section are the definitions (3.7)–(3.8) and Theorem 3.6. Owing to Lemma 2.2, we know that the invariant-domain property (2.8) and the entropy inequality (2.9) hold for $\overline{\boldsymbol{u}}_{LR}(\lambda)$ if $\lambda \geq \lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. Our objective in this paper is to find a value of λ as small as possible in the interval $[\lambda_\epsilon, \lambda_{LR}^\sharp]$ so that (2.8) and (2.9) still hold (we no longer require that (2.7) be true). The actual estimation of this greedy wave speed in done Sect. 3.2.

3.1 Invariant Domain and Entropy: Structural Assumptions

As the notion of an invariant domain of the PDE system (2.1) is too general, we list in this section the properties that we want to preserve. We use the concept of quasiconcavity for his purpose. (The reader who is not familiar with this notion can replace the word *quasiconcavity* by *concavity* without losing the essence of what is said.)

Definition 3.1 *(Quasiconcavity)* Given a convex set $\mathcal{C} \subset \mathbb{R}^m$, we say that a function $\Psi : \mathcal{C} \to \mathbb{R}$ is quasiconcave if the set $L_\chi(\Psi) := \{\boldsymbol{u} \in \mathcal{C} \mid \Psi(\boldsymbol{u}) \geq \chi\}$ is convex for every $\chi \in \mathbb{R}$. The sets $\{L_\chi(\Psi)\}_{\chi\in\mathbb{R}}$ are called *upper level sets* or *upper contour sets*.

We now list the properties we are interested in and that we want to preserve. Let $L \in \mathbb{N}\backslash\{0\}$ and let us set $\mathcal{L}:=\{0:L\}$, $\mathcal{L}^*:=\{1:L\}$. We assume that there exists a collection of $L+1$ subsets $\{\mathcal{B}_l\}_{l\in\mathcal{L}}$ in $\mathbb{R}^m$, and a collection of L continuous quasiconcave functionals $\{\Psi_l : \mathcal{B}_{l-1} \to \mathbb{R}\}_{l\in\mathcal{L}^*}$ so that the following properties hold true:

$$\mathcal{B}_L \subset \mathcal{B}_{L-1} \subset \ldots \subset \mathcal{B}_0 := \mathbb{R}^m, \tag{3.1a}$$

$$\mathcal{B}_l = \{\boldsymbol{u} \in \mathcal{B}_{l-1} \mid \Psi_l(\boldsymbol{u}) \geq 0\}, \qquad \forall l \in \mathcal{L}^*, \tag{3.1b}$$

$$\mathcal{B}_L \subset \mathcal{B}, \tag{3.1c}$$

$$\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathcal{B}_l, \quad \text{and} \quad \overline{\boldsymbol{u}}_{LR}(\lambda^\sharp_{LR}) \in \mathcal{B}_l, \quad \forall l \in \mathcal{L}. \tag{3.1d}$$

Notice in passing that all the subsets $\{\mathcal{B}_l\}_{l\in\mathcal{L}}$ are convex since $\mathcal{B}_0 = \mathbb{R}^m$ and $\mathcal{B}_l = L_0(\Psi_l)$ for all $l \in \mathcal{L}^*$. These sets are also closed as the functional $\{\Psi_l\}_{l\in\mathcal{L}}$ are continuous. As $\mathcal{B}_l$ is convex for all $l \in \mathcal{L}$, the assumption (3.1d) then implies that $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_l$ for all $\lambda \in [\lambda^\sharp_{LR}, \infty)$ and all $l \in \mathcal{L}$. (The assumption (3.1d) is reasonable as we already know that $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}$ for all $\lambda \in [\lambda^\sharp_{LR}, \infty)$.)

As documented in Appendix A in Harten and Hyman [11] (and in Lemma 3.2 in [5]), computing a wave speed that guarantees a method to be invariant-domain preserving is not enough to ensure convergence to the entropy solution. Hence, in addition to invariant-domain properties, we also want to satisfy entropy inequalities. In order to clarify this objective, we assume to be given a finite set of entropy pairs for (2.1), say $\{(\eta_e, \boldsymbol{q}_e)\}_{e\in\mathcal{E}}$ with $\eta_e : \mathcal{B}_L \to \mathbb{R}$ and $\boldsymbol{q}_e : \mathcal{B}_L \to \mathbb{R}^d$ for all $e \in \mathcal{E}$. Let $\lambda^\flat_{LR}$ be the infimum of the set $\{\lambda \in [\lambda_\epsilon, \lambda^\sharp_{LR}] \mid \overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_L\}$; that is,

$$\lambda^\flat_{LR} := \inf\{\lambda \in [\lambda_\epsilon, \lambda^\sharp_{LR}] \mid \overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_L\}. \tag{3.2}$$

Note that $\lambda^\flat_{LR}$ is well defined because the minimizing set is not empty (it contains $\lambda^\sharp_{LR}$). This infimum is actually the minimum as $[\lambda_\epsilon, \lambda^\sharp_{LR}] \ni \lambda \mapsto \boldsymbol{u}_{LR}(\lambda)$ is continuous and $\mathcal{B}_L$ is closed. For every $e \in \mathcal{E}$, we introduce the function $\Phi^e : \left[\frac{1}{\lambda^\sharp_{LR}}, \frac{1}{\lambda^\flat_{LR}}\right] \to \mathbb{R}$ defined by

$$\Phi_e(t) := \eta_e\left(\overline{\boldsymbol{u}}_{LR}(\tfrac{1}{t})\right) - \frac{1}{2}(\eta_e(\boldsymbol{u}_L) + \eta_e(\boldsymbol{u}_R)) + \frac{t}{2}(\boldsymbol{q}_e(\boldsymbol{u}_R) - \boldsymbol{q}_e(\boldsymbol{u}_L))\cdot\boldsymbol{n}, \quad \forall e \in \mathcal{E}. \tag{3.3}$$

We have established in Lemma 2.2 that

$$\Phi_e(1/\lambda^\sharp_{LR}) \leq 0, \quad \forall e \in \mathcal{E}. \tag{3.4}$$

Our goal is to find a greedy wave speed $\lambda^{\text{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ as small as possible in $[\lambda^\flat_{LR}, \lambda^\sharp_{LR}]$ so that $\overline{\boldsymbol{u}}_{LR}(\lambda^{\text{grdy}}) \in \mathcal{B}_L$ and $\Phi_e(1/\lambda^{\text{grdy}}) \leq 0$, for all $e \in \mathcal{E}$.

Lemma 3.2 *The function* $\Phi_e : \left(\frac{1}{\lambda^{\sharp}_{LR}}, \frac{1}{\lambda^{\flat}_{LR}}\right) \to \mathbb{R}$ *is convex for all* $e \in \mathcal{E}$.

Proof Let $t_1, t_2 \in \left(\frac{1}{\lambda^{\sharp}_{LR}}, \frac{1}{\lambda^{\flat}_{LR}}\right)$ and $\theta \in [0, 1]$. Then using that

$$\begin{aligned}\overline{\boldsymbol{u}}_{LR}\left(\tfrac{1}{\theta t_1+(1-\theta)t_2}\right) &= \tfrac{\theta}{2}(\boldsymbol{u}_L+\boldsymbol{u}_R)+\tfrac{1-\theta}{2}(\boldsymbol{u}_L+\boldsymbol{u}_R)-(\tfrac{\theta}{2}t_1+\tfrac{1-\theta}{2}t_2)(\mathbb{f}(\boldsymbol{u}_R)-\mathbb{f}(\boldsymbol{u}_L))\boldsymbol{n}\\ &= \theta\overline{\boldsymbol{u}}_{LR}(\tfrac{1}{t_1})+(1-\theta)\overline{\boldsymbol{u}}_{LR}(\tfrac{1}{t_2}),\end{aligned}$$

the assertion follows from the convexity of η_e. □

Remark 3.3 *(Notation)* To be precise the entropy functional defined in (3.3) should be denoted by Φ^e_{LR} instead Φ_e as it depends on the pair $\boldsymbol{u}_L, \boldsymbol{u}_R$. Likewise, we should also use Ψ^l_{LR} instead of Ψ_l. In what follows the index $_{LR}$ reminds us of the dependence with respect the pair $\boldsymbol{u}_L$, $\boldsymbol{u}_R$ and the unit vector $\boldsymbol{n}$. We have chosen to use the symbols Φ_e and Ψ_l instead to simplify the notation.

Remark 3.4 *(Matryoshka doll structure)* The *Matryoshka doll structure* introduced in (3.1) is meant to reflect that the domain of definition of the functionals Ψ_l may become smaller and smaller as the index l increases. We illustrate this point with the compressible Euler equations with the equation of state $p(\boldsymbol{u}) := \frac{\gamma-1}{1-b\rho}\rho(e(\boldsymbol{u})-q)-\gamma p_\infty$, where $b \geq 0$, $\gamma > 1$, $q \in \mathbb{R}$, and $p_\infty \in \mathbb{R}$, and $e(\boldsymbol{u}) := E/\rho - \frac{1}{2}\|\boldsymbol{m}/\rho\|^2_\ell$. This equation of state is often called Nobel-Abel stiffened gas equation of state in the literature; see Le Métayer and Saurel [22]. In this case we have: $\Psi_1(\boldsymbol{u}) := \rho$, $\mathcal{B}_1 := \{\boldsymbol{u} = (\rho, \boldsymbol{m}, E)^{\mathsf{T}} \in \mathbb{R}^{d+2} \mid \rho > 0\}$; $\Psi_2(\boldsymbol{u}) := 1 - b\rho$, $\mathcal{B}_2 := \{\boldsymbol{u} \in \mathcal{B}_1 \mid 1 - b\rho > 0\}$; $\Psi_3(\boldsymbol{u}) := \rho(e(\boldsymbol{u}) - q) - p_\infty(1 - b\rho)$, $\mathcal{B}_3 := \{\boldsymbol{u} \in \mathcal{B}_2 \mid \rho(e(\boldsymbol{u}) - q) - p_\infty(1 - b\rho) > 0\}$. Notice that the constraint $\Psi_3(\boldsymbol{u}) > 0$ implies that $p(\boldsymbol{u}) + p_\infty > 0$ which is essential to be able to define the specific entropy $\eta(\boldsymbol{u}) = \log((1/\rho - b)^\gamma(p(\boldsymbol{u}) + p_\infty))$.

In practice, we are going to enforce sharper bounds than those shown above by making all the functionals $\{\Psi_l\}_{l\in\mathcal{L}}$ and all the sets $\{\mathcal{B}_l\}_{l\in\mathcal{L}}$ depend on the states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ (see Sects. 4, 5).

3.2 Algorithm for Estimating the Greedy Wave Speed

As mentioned above, the key idea of the paper is to define a greedy wave speed $\lambda^{\mathrm{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ in $[\lambda^{\flat}_{LR}, \lambda^{\sharp}_{LR}]$ so that $\overline{\boldsymbol{u}}_{LR}(\lambda^{\mathrm{grdy}}) \in \mathcal{B}_L$ and $\Phi_e(1/\lambda^{\mathrm{grdy}}) \leq 0$, for all $e \in \mathcal{E}$. We now present an algorithm that carries out this program (see Algorithm 1).

One starts by setting $\lambda_0(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \lambda_\epsilon$. Then one traverses the index set $\mathcal{L}^*$ in increasing order, and for each index l in $\mathcal{L}^*$ one computes the wave speed $\lambda_l(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ recursively defined by

$$\lambda_l := \min\{\lambda \in [\lambda_{l-1}, \lambda^{\sharp}_{LR}] \mid \Psi_l(\overline{\boldsymbol{u}}_{LR}(\lambda)) \geq 0\}. \tag{3.5}$$

Next, one (indiscriminately) traverses the index set $\mathcal{E}$ and computes the wave speed $\lambda_e(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ defined by

$$\lambda_e := \min\{\lambda \in [\lambda_L, \lambda^{\sharp}_{LR}] \mid \Phi_e(\lambda^{-1}) \leq 0\}. \tag{3.6}$$

One finally defines the greedy wave speed $\lambda^{\mathrm{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ as follows:

$$\lambda^{\mathrm{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \max_{e\in\mathcal{E}} \lambda_e. \tag{3.7}$$

Algorithm 1 Greedy wave speed

Input: $\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R$
Output: $\lambda^{\mathrm{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$
1: Compute $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$, $\lambda_0(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ and $\lambda^\sharp_{LR}$
2: **for** $l = 1$ **to** L **do**
3: Define Ψ_l^{LR} and compute $\lambda_l(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$; see (3.5)
4: **end for**
5: **for** $e \in \mathcal{E}$ **do**
6: Define $(\eta_e^{LR}, \boldsymbol{q}_e^{LR})$ and compute $\lambda_e(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$; see (3.6)
7: **end for**
8: Compute $\lambda^{\mathrm{grdy}}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$; see (3.7)

Techniques to compute the wave speed defined in (3.5) and (3.6) are explained in Sects. 4.1 and 4.2 for nonlinear scalar equations and in Sects. 5.3 and 5.4 for the p-system.

Lemma 3.5 *Assume that* (3.1) *hold true. Then,*

(i) λ_l *is well defined and* $\lambda_\epsilon \le \lambda_l \le \lambda^\sharp_{LR}$ *for all* $l \in \mathcal{L}$. *We have* $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_l$ *for all* $\lambda \in [\lambda_l, \lambda^\sharp_{LR}]$ *and all* $l \in \mathcal{L}$.
(ii) λ_e *is well defined. We have* $\Phi_e(\frac{1}{\lambda}) \le 0$ *for all* $\lambda \in [\lambda_e, \lambda^\sharp_{LR})]$ *and all* $e \in \mathcal{E}$.

Proof Recall that $\lambda^\sharp_{LR} := \max(\lambda_\epsilon, \lambda_{\max})$.
(i) We proceed by induction over $l \in \mathcal{L}$. The wave speed λ_ϵ is well defined (see (2.19a)) and $\lambda_\epsilon \in [\lambda_\epsilon, \lambda^\sharp_{LR}]$. Moreover, $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_0 := \mathbb{R}^m$ for all $\lambda \in [\lambda_\epsilon, \lambda^\sharp_{LR}]$. Hence, the induction assumption (i) holds for $l = 0$ since $\lambda_0 := \lambda_\epsilon$. Now let $l \in \mathcal{L}^*$ and let us prove that (i) holds. The induction assumption implies that the set $[\lambda_{l-1}, \lambda^\sharp_{LR}]$ is not empty (because $\lambda_{l-1} \le \lambda^\sharp_{LR}$), and $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_{l-1}$ for all $\lambda \in [\lambda_{l-1}, \lambda^\sharp_{LR}]$. This means in particular that $\Psi_l(\overline{\boldsymbol{u}}_{LR}(\lambda))$ is well defined for all $\lambda \in [\lambda_{l-1}, \lambda^\sharp_{LR}]$. Moreover, we have $\overline{\boldsymbol{u}}_{LR}(\lambda^\sharp_{LR}) \in \mathcal{B}_l$ owing to the assumption (3.1d). Hence the set $\{\lambda \in [\lambda_{l-1}, \lambda^\sharp_{LR}] \mid \Psi_l(\overline{\boldsymbol{u}}_{LR}(\lambda)) \ge 0\}$ is not empty. This set has a minimum since Ψ_l is continuous, the mapping $[\lambda_{l-1}, \lambda^\sharp_{LR}] \ni \lambda \mapsto \overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_{l-1}$ is continuous, and $[\lambda_{l-1}, \lambda^\sharp_{LR}]$ is compact. Hence λ_l is well defined and $\lambda_\epsilon \le \lambda_{l-1} \le \lambda_l \le \lambda^\sharp_{LR}$ (by definition). Let us now prove that $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_l$ for all $\lambda \in [\lambda_l, \lambda^\sharp_{LR}]$. We first observe that $\overline{\boldsymbol{u}}_{LR}(\lambda) = \theta\, \overline{\boldsymbol{u}}_{LR}(\lambda_l) + (1-\theta)\overline{\boldsymbol{u}}_{LR}(\lambda^\sharp_{LR})$ with $\theta := \frac{(\lambda^\sharp_{LR} - \lambda)\,\lambda_l}{(\lambda^\sharp_{LR} - \lambda_l)\,\lambda}$; hence, the set $\{\overline{\boldsymbol{u}}_{LR}(\lambda) \mid \lambda \in [\lambda_l, \lambda^\sharp_{LR}]\}$ is a line segment in $\mathbb{R}^m$. But both $\overline{\boldsymbol{u}}_{LR}(\lambda_l)$ and $\overline{\boldsymbol{u}}_{LR}(\lambda^\sharp_{LR})$ are members of $\{\boldsymbol{u} \in \mathcal{B}_{l-1} \mid \Psi_l(\boldsymbol{u}) \ge 0\} = \mathcal{B}_l$. Since $\mathcal{B}_l$ is convex, we conclude that the entire line segment $\{\overline{\boldsymbol{u}}_{LR}(\lambda) \mid \lambda \in [\lambda_l, \lambda^\sharp_{LR}]\}$ is in $\mathcal{B}_l$. This proves that the induction assumption holds true for l.

(ii) The argument in (i) proves that $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_L$ for all $\lambda \in [\lambda_L, \lambda^\sharp_{LR}]$. As the domain of η_e and $\boldsymbol{q}_e$ is $\mathcal{B}_L$, this argument proves that $\Phi_e(\frac{1}{\lambda})$ is well defined for all $\lambda \in [\lambda_L, \lambda^\sharp_{LR}]$ and all $e \in \mathcal{E}$. The continuity of Φ_e implies that λ_e is well defined as well. From the convexity of Φ_e established in Lemma 3.2 it follows that $\Phi_e(\frac{1}{\lambda}) \le 0$ for all $\lambda \in [\lambda_e, \lambda^\sharp_{LR}]$ since $\Phi_e(\frac{1}{\lambda_e}) \le 0$ and $\Phi_e(\frac{1}{\lambda^\sharp_{LR}}) \le 0$, see (3.4). □

3.3 Greedy Viscosity

We are now in a position to state the main results of Sect. 3. Using the same notation as in Sect. 2.3, let $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. With the greedy wave speed $\lambda^{\text{grdy}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ defined in (3.7), we define the greedy viscosity for the pair (i, j) at the time t^n as follows:

$$d_{ij}^{\text{grdy},n} = \max(\lambda^{\text{grdy}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda^{\text{grdy}}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}). \tag{3.8}$$

Note that if $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \geq \lambda_\epsilon$ (which is almost always the case), then

$$d_{ij}^{\text{GMS},n} \geq d_{ij}^{\text{grdy},n}. \tag{3.9}$$

The main result of the paper and the reason we have introduced the greedy wave speed is the following.

Theorem 3.6 (IDP Greedy viscosity) *Let $\mathcal{B}$ be an invariant domain for (2.1). Let $n \geq 0$, $i \in \mathcal{V}$. For all $j \in \mathcal{I}(i)^*$, let $\{\mathcal{B}_l^{ij}\}_{l\in\mathcal{L}}$ be a finite collection of convex sets, and let $\{\Psi_l^{ij} : \mathcal{B}_l^{ij} \to \mathbb{R}\}_{l\in\mathcal{L}^*}$ be a collection of continuous quasiconcave functionals. Let $\{(\eta_e^i, \boldsymbol{q}_e^i)\}_{e\in\mathcal{E}^i}$ be a finite set of entropy pairs for (2.1). (We use a superscript i on the entropy pairs to allow for the possibility to choose a different set of entropies for each index $i \in \mathcal{V}$.) Let $\{d_{ij}^{\text{grdy},n}\}_{j\in\mathcal{I}(i)^*}$ be the greedy graph viscosity defined by (3.8) and let $\{\mathbf{U}_i^{n+1}\}_{i\in\mathcal{V}}$ be the update defined in (2.12) with the choice $d_{ij}^n := d_{ij}^{\text{grdy},n}$. Assume the following:*

(i) $\{\mathcal{B}_l^{ij}\}_{l\in\mathcal{L}}$ and $\{\Psi_l^{ij}\}_{l\in\mathcal{L}^*}$ satisfy the assumptions in (3.1) for all $j \in \mathcal{I}(i)^*$;
(ii) τ is small enough so that $1 - 2\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)^*} d_{ij}^n > 0$.

Then the update $\{\mathbf{U}_i^{n+1}\}_{i\in\mathcal{V}}$ satisfies the following properties:

$$\mathbf{U}_i^{n+1} \in \text{conv}\Big(\bigcup_{j\in\mathcal{I}(i)^*} \mathcal{B}_L^{ij}\Big), \quad \textit{hence} \quad \mathbf{U}_i^{n+1} \in \mathcal{B}, \tag{3.10}$$

$$\frac{m_i}{\tau}(\eta_e^i(\mathbf{U}_i^{n+1}) - \eta_e^i(\mathbf{U}_i^n)) + \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\boldsymbol{q}_e^i(\mathbf{U}_j^n) - \sum_{j\in\mathcal{I}(i)^*} d_{ij}^n(\eta_e^i(\mathbf{U}_j^n) - \eta_e^i(\mathbf{U}_j^n)) \leq 0. \tag{3.11}$$

Proof We first recall that (2.12) can be rewritten as follows:

$$\mathbf{U}_i^{n+1} = \Big(1 - \sum_{j\in\mathcal{I}(i)^*} \frac{2\tau d_{ij}^n}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)^*} \frac{2\tau d_{ij}^n}{m_i}\overline{\mathbf{U}}_{ij}^n, \tag{3.12}$$

with the notation $\overline{\mathbf{U}}_{ij}^n := \overline{\boldsymbol{u}}_{ij}(\frac{d_{ij}^n}{\|\boldsymbol{c}_{ij}\|_{\ell^2}})$. Setting $\lambda_{ij} := \frac{d_{ij}^n}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}$ for all $j \in \mathcal{I}(i)^*$, we have $\overline{\mathbf{U}}_{ij}^n = \overline{\boldsymbol{u}}_{ij}(\lambda_{ij})$. As the assumptions in (3.1) hold and $\lambda^{\text{grdy}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is defined by (3.5)–(3.6)–(3.7) for all $j \in \mathcal{I}(i)^*$, we can apply Lemma 3.5. Then combining (3.7) with the identity $\lambda_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2} = d_{ij}^n = d_{ij}^{\text{grdy},n}$ implies $\lambda_{ij} \geq \lambda_{ij}^{\text{grdy}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$, and invoking Lemma 3.5(i), (3.1a) and (3.1c), we infer that

$$\overline{\mathbf{U}}_{ij}^n \in \bigcap_{l\in\mathcal{L}} \mathcal{B}_l^{ij} = \mathcal{B}_L^{ij} \subset \mathcal{B}.$$

Since we assumed that $1 - 2\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)^*} d_{ij}^n > 0$, the right-hand side in (3.12) is a convex combination of the states $\{\mathbf{U}_i^n\} \cup \{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)^*}$ which all lie in the convex hull

$\mathrm{conv}\Big(\bigcup_{j\in\mathcal{I}(i)^*}\mathcal{B}_L^{ij}\Big)$, and it follows that $\mathbf{U}_i^{n+1}\in\mathcal{B}$. Let us now establish the entropy inequality (3.11). From the convexity of η_e^i and (3.12) we obtain

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

Using Lemma 3.5(ii) and recalling that $\overline{\mathbf{U}}_{ij}^n=\overline{\boldsymbol{u}}_{ij}(\lambda_{ij})$, we infer that $\Phi_e(\frac{1}{\lambda_{ij}})\le 0$, i.e.,

$$2d_{ij}^n\eta_e^i(\overline{\mathbf{U}}_{ij}^n)\le d_{ij}^n(\eta_e^i(\mathbf{U}_i^n)+\eta_e^i(\mathbf{U}_j^n))-(\boldsymbol{q}_e^i(\mathbf{U}_j^n)-\boldsymbol{q}_e^i(\mathbf{U}_i^n))\cdot\boldsymbol{c}_{ij}.$$

Inserting this inequality in the previous inequality and using (2.11) gives (3.11). □

Remark 3.7 More generally, Theorem 3.6 holds true for any choice $\{d_{ij}^n\}_{j\in\mathcal{I}(i)^*}$ of graph viscosity provided that $d_{ij}^n\ge d_{ij}^{\mathrm{grdy},n}$ for all $j\in\mathcal{I}(i)^*$, $i\in\mathcal{V}$.

The result stated in Theorem 3.6 can be slightly refined by assuming a little more structure on the sets $\{\mathcal{B}_l^{ij}\}_{i\in\mathcal{I}(i)^*}$ for all $l\in\mathcal{L}$.

Corollary 3.8 (Localization) *Let the assumptions of Theorem 3.6 hold. Assume also that the following holds true for all $l\in\mathcal{L}^*$: There exists $i(l)\in\mathcal{I}(i)^*$ so that $\mathcal{B}_l^{ij}\subset\mathcal{B}_l^{ii(l)}$ for all $j\in\mathcal{I}(i)^*$. Then the update given by (2.12) satisfies the following local properties:*

$$\Psi_l^{ii(l)}(\boldsymbol{U}_i^{n+1})\ge 0,\quad\forall l\in\mathcal{L}^*. \tag{3.13}$$

Proof The assumption together with with property (3.10) from Theorem 3.6 implies that

$$\mathbf{U}_i^{n+1}\in\mathrm{conv}\Big(\bigcup_{j\in\mathcal{I}(i)^*}\mathcal{B}_L^{ij}\Big)\subset\Big(\bigcup_{j\in\mathcal{I}(i)^*}\mathcal{B}_l^{ij}\Big)=\mathcal{B}_l^{ii(l)}.$$

Hence, the statement is an immediate consequence of the definition of the set $\mathcal{B}_l^{ii(l)}$ given in (3.1b). □

Remark 3.9 Computing the maximal wave speed $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ for general hyperbolic systems typically requires solving a nonlinear scalar fixed point problem. Computing an upper bound on $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ is somewhat simpler as it requires to use iterative techniques that converge from above. Very accurate upper bounds are usually obtained in two to three iterations. The time spent to this task is in general negligible. For instance, the reader is referred to [3] where guaranteed upper bounds on $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ are given for the Euler equations with the co-volume equation of state (the source code for this method is available in the appendix of [3] and a source code computing $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ for a general equation of state is available at Clayton et al. [1]). Computing $\lambda^{\mathrm{grdy}}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ or an upper bound thereof is similar to estimating upper bounds for $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$. This can be easily done by using iterative techniques converging from above.

4 Scalar Conservation Equations

In this section we specialize the proposed definitions (3.7)–(3.8) on scalar conservation equations. Instead of using the notation $\mathbb{f}$ and $\boldsymbol{u}$, we now denote the flux by $\boldsymbol{f}$ and the dependent variable by u.

4.1 Maximum Principle

In the scalar case, the only invariant-domain property there is reduces to enforcing the maximum principle. We start by estimating a wave speed that does exactly that by following the algorithm (3.5) described in Sect. 3.2. We take care of the entropy inequalities (3.6) in Sect. 4.2

Let $u_L, u_R \in \mathcal{A}$ and let $\boldsymbol{n}$ be a unit vector in $\mathbb{R}^d$. (Computing $\lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ is a standard exercise; see e.g., Dafermos [2, Lem. 3.1], Holden and Risebro [15, § 2.2], Osher [25, Thm. 1].) We introduce two concave functionals to take care of the local minimum and maximum principle:

$$u_{LR}^{\min} := \min(u_L, u_R), \qquad u_{LR}^{\max} := \max(u_L, u_R), \tag{4.1a}$$

$$\Psi_1(u) := u - u_{LR}^{\min}, \qquad \Psi_2(u) := u_{LR}^{\max} - u. \tag{4.1b}$$

Accordingly, we set $\mathcal{B}_0 := \mathbb{R}$, $\mathcal{B}_1 := \{u \in \mathcal{B}_0 \mid \Psi_1(u) \geq 0\}$, $\mathcal{B}_2 := \{u \in \mathcal{B}_1 \mid \Psi_2(u) \geq 0\}$.

Lemma 4.1 *Let*

$$\lambda_{12}(\boldsymbol{n}, u_L, u_R) := \begin{cases} \frac{|(\boldsymbol{f}(u_R) - \boldsymbol{f}(u_L))\cdot\boldsymbol{n}|}{|u_R - u_L|} & \text{if } u_R \neq u_L \\ \max(|\boldsymbol{f}'(u_R)\cdot\boldsymbol{n}|, |\boldsymbol{f}'(u_L)\cdot\boldsymbol{n}|) & \text{if } u_R = u_L. \end{cases} \tag{4.2}$$

Then $\Psi_1(\overline{u}_{LR}(\lambda)) \geq 0$ *and* $\Psi_2(\overline{u}_{LR}(\lambda)) \geq 0$ *for all* $\lambda \geq \max(\lambda_{12}, \lambda_\epsilon)$. *(This also means* $\overline{u}_{LR}(\lambda) \in [u_{LR}^{\min}, u_{LR}^{\max}]$ *for all* $\lambda \geq \max(\lambda_{12}, \lambda_\epsilon)$.*)*

Proof Let $a := \frac{1}{2}(u_L + u_R)$, $b := (\boldsymbol{f}(u_R) - \boldsymbol{f}(u_L))\cdot\frac{\boldsymbol{n}}{2}$. Note that $a \in [u_{LR}^{\min}, u_{LR}^{\max}]$ and recall that $\overline{u}_{LR}(\lambda) := a - b\lambda^{-1}$, for all $\lambda \geq \lambda_\epsilon > 0$. We want to estimate the smallest value of λ in $[\lambda_\epsilon, \lambda_{LR}^\sharp]$ so that $\Psi_1(\overline{u}_{LR}(\lambda)) \geq 0$ and $\Psi_2(\overline{u}_{LR}(\lambda)) \geq 0$. That is, we want λ to be such that

$$-\frac{1}{2}(u_{LR}^{\max} - u_{LR}^{\min}) = a - u_{LR}^{\max} \leq b\lambda^{-1} \leq a - u_{LR}^{\min} = \frac{1}{2}(u_{LR}^{\max} - u_{LR}^{\min}).$$

This holds true if and only if $|b|\lambda^{-1} \leq \frac{1}{2}(u_{LR}^{\max} - u_{LR}^{\min})$. If $u_{LR}^{\max} - u_{LR}^{\min} \neq 0$, the smallest possible value of λ making this inequality to hold is $\lambda = \frac{|(\boldsymbol{f}(u_R) - \boldsymbol{f}(u_L))\cdot\boldsymbol{n}|}{u_{LR}^{\max} - u_{LR}^{\min}}$. If $u_{LR}^{\max} - u_{LR}^{\min} = 0$, every value of λ is admissible, but the only value of λ that is stable under perturbation of the two states is $\lambda = |\boldsymbol{f}'(u_R)\cdot\boldsymbol{n}|$ if $\boldsymbol{f}$ is of class C^1, and $\max(|\boldsymbol{f}'(u_R)\cdot\boldsymbol{n}|, |\boldsymbol{f}'(u_L)\cdot\boldsymbol{n}|)$ otherwise. □

We note that the wave speed identified in Lemma 4.1, $\frac{|(\boldsymbol{f}(u_R) - \boldsymbol{f}(u_L))\boldsymbol{n}|}{|u_R - u_L|}$, is the average speed, sometimes called Roe's average in the computational fluid dynamics literature. As the final wave speed defining the artificial viscosity is eventually larger than or equal to this quantity, Lemma 2 from Harten [10] implies that the scheme is *total variation non increasing* in one space dimension on the three point stencil (the wave speed λ_{12} also satisfies the necessary and sufficient condition formulated in Tadmor [30, Cor. 2.3]). It is well known that in the presence of sonic points this wave speed is not large enough to ensure that the approximation defined in (2.12) converges to the entropy solution (see, e.g., Harten and Hyman [11, App. A] or [5, Lem. 3.2] for a simple proof). This problem is addressed in the next section by augmenting the wave speed so as to make sure that some entropy inequalities are locally satisfied, i.e., (3.6) is satisfied.

4.2 Entropy Inequality

Now, following algorithm (3.6) described in Sect. 3.2, we further look for a wave speed, possibly larger than λ_{12}, so as to satisfy some entropy inequalities.

Lemma 4.2 *Let $k \in \mathbb{R}$. Let $\eta_k(u) := |u-k|$ be the Kružkov entropy associated with k and $\boldsymbol{q}_k(u) := \text{sign}(u-k)(\boldsymbol{f}(u)-\boldsymbol{f}(k))$ be the corresponding entropy flux. Let*

$$a_k := u_L + u_R - 2k, \qquad b := (\boldsymbol{f}(u_R)-\boldsymbol{f}(u_L))\cdot\boldsymbol{n},$$
$$c_k := \eta_k(u_L)+\eta_k(u_R), \qquad d_k := (\boldsymbol{q}_k(u_R)-\boldsymbol{q}_k(u_L))\cdot\boldsymbol{n}.$$

(Observe that $|a_k| = c_k$ if and only if $k \notin (u_{LR}^{\min}, u_{LR}^{\max})$.) Let $\lambda_{12}(\boldsymbol{n}, u_L, u_R)$ be defined as in Lemma 4.1, and let

$$\lambda(k,\boldsymbol{n},u_L,u_R) := \begin{cases} \lambda_{12}(\boldsymbol{n},u_L,u_R) & \text{if } k \notin (u_{LR}^{\min}, u_{LR}^{\max}) \\ \max\left(\frac{d_k+b}{c_k+a_k}, \frac{d_k-b}{c_k-a_k}, \lambda_{12}(\boldsymbol{n},u_L,u_R)\right) & \text{otherwise.} \end{cases} \tag{4.3}$$

Let $\Phi_k(\frac{1}{\lambda}) := \eta_k(\overline{u}_{LR}(\lambda)) - \frac{1}{2}(\eta_k(u_L)+\eta_k(u_R)) + \frac{1}{2\lambda}(\boldsymbol{q}_k(u_R)-\boldsymbol{q}_k(u_L))\cdot\boldsymbol{n}$. Then, for every $\lambda \geq \max(\lambda(k,\boldsymbol{n},u_L,u_R),\lambda_\epsilon)$ we have $\Phi_k(\frac{1}{\lambda}) \leq 0$.

Proof (1) Assume first that $k \notin (u_{LR}^{\min}, u_{LR}^{\max})$, i.e., $c_k = |a_k|$. The assumption $\lambda \geq \max(\lambda_{12},\lambda_\epsilon)$ implies that $\overline{u}_{LR}(\lambda) \in [u_{LR}^{\min}, u_{LR}^{\max}]$. Hence, $\text{sign}(\overline{u}_{LR}(\lambda)-k) = \text{sign}(\frac{1}{2}(u_R+u_L)-k)$. As a result, we have

$$\begin{aligned}\eta_k(\overline{u}_{LR}(\lambda)) &= \text{sign}(\overline{u}_{LR}(\lambda)-k)(\overline{u}_{LR}(\lambda)-k)\\ &= \text{sign}\left(\tfrac{1}{2}(u_R+u_L)-k\right)\left(\tfrac{1}{2}(u_R+u_L) - \tfrac{1}{2\lambda}(\boldsymbol{f}(u_R)-\boldsymbol{f}(u_L))\cdot\boldsymbol{n} - k\right).\end{aligned}$$

On the other hand, using that $\eta_k(u_R) = \text{sign}\left(\frac{1}{2}(u_R+u_L)-k\right)(u_R-k)$, $\boldsymbol{q}_k(u_R) = \text{sign}(\frac{1}{2}(u_R+u_L)-k)(\boldsymbol{f}(u_R)-\boldsymbol{f}(k))$, and the corresponding identities for $\eta_k(u_L)$ and $\boldsymbol{q}_k(u_L)$, we deduce that

$$\begin{aligned}&\tfrac{1}{2}\eta_k(u_L)+\tfrac{1}{2}\eta_k(u_R) - \tfrac{1}{2\lambda}(\boldsymbol{q}_k(u_R)-\boldsymbol{q}_k(u_L))\cdot\boldsymbol{n}\\ &\quad= \text{sign}(\tfrac{1}{2}(u_R+u_L)-k)\left(\tfrac{1}{2}(u_R+u_L) - \tfrac{1}{2\lambda}(\boldsymbol{f}(u_R)-\boldsymbol{f}(u_L))\cdot\boldsymbol{n} - k\right).\end{aligned}$$

Hence, we conclude that $\eta_k(\overline{u}_{LR}(\lambda)) = \eta_k(u_L)+\eta_k(u_R) - \lambda^{-1}(\boldsymbol{q}_k(u_R)-\boldsymbol{q}_k(u_L))\cdot\boldsymbol{n}$ for all $k \notin (u_{LR}^{\min}, u_{LR}^{\max})$.

(2) Let us now assume that $k \in (u_{LR}^{\min}, u_{LR}^{\max})$. Then we have that $c_k - |a_k| \geq 2\min(\eta_k(u_L),\eta_k(u_R)) > 0$. Hence definition (4.3) makes sense. Using the definitions for a_k, b, c_k, and d_k, we have $2\eta_k(\overline{u}_{LR}(\lambda)) = |a_k - \lambda^{-1}b|$. Then we want to find the smallest value of λ that guarantees that

$$|a_k - \lambda^{-1}b| \leq c_k - \lambda^{-1}d_k.$$

The above inequality is equivalent to

$$\lambda^{-1}(d_k-b) \leq c_k - a_k, \quad \text{and} \quad \lambda^{-1}(b+d_k) \leq c_k + a_k.$$

Using that $|a_k| < c_k$, we infer that

$$\lambda \geq \frac{d_k-b}{c_k-a_k}; \qquad \lambda \geq \frac{d_k+b}{c_k+a_k}.$$

The assertion follows readily. □

4.3 Summary

The following result summarizes what is proposed above. In particular, it shows how the Krůzkov entropies should be chosen.

Theorem 4.3 *Let $n \geq 0$, $i \in \mathcal{V}$, $\mathsf{U}_i^{\min,n} := \min_{j\in\mathcal{I}(i)} \mathsf{U}_j^n$, $\mathsf{U}_i^{\max,n} := \max_{j\in\mathcal{I}(i)} \mathsf{U}_j^n$. Let k_i be any real number in the range $(\mathsf{U}_i^{\min,n}, \mathsf{U}_i^{\max,n})$. Let $(\eta_{k_i}, \boldsymbol{q}_{k_i})$ be the associated Krůzkov entropy pair. For all $j \in \mathcal{I}(i)^*$, let $\lambda_{ij}^{\mathrm{grdy},n} := \max(\lambda_\epsilon, \lambda(k_i, \boldsymbol{n}_{ij}, \mathsf{U}_i^m, \mathsf{U}_j^n))$ and*

$$d_{ij}^{\mathrm{grdy},n} := \max(\lambda_{ij}^{\mathrm{grdy},n} \|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{ji}^{\mathrm{grdy},n} \|\boldsymbol{c}_{ji}\|_{\ell^2}). \tag{4.4}$$

Let U_i^{n+1} be given by (2.12) with the viscosity $d_{ij}^n = d_{ij}^{\mathrm{grdy},n}$ defined above. Assume that $1 - 2\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)^*} d_{ij}^n \geq 0$. Then

$$\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\min,n}, \mathsf{U}_i^{\max,n}] \tag{4.5}$$

$$\frac{m_i}{\tau}(\eta_{k_i}(\mathsf{U}_i^{n+1}) - \eta_{k_i}(\mathsf{U}_i^n)) + \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\boldsymbol{q}_{k_i}(\mathsf{U}_j^n) - \sum_{j\in\mathcal{I}(i)^*} d_{ij}^n(\eta_{k_i}(\mathsf{U}_j^n) - \eta_{k_i}(\mathsf{U}_j^n)) \leq 0. \tag{4.6}$$

Proof This is just a reformulation of Theorem 3.6. □

Remark 4.4 *(Entropy choice)* It is essential that k_i be chosen in $(\mathsf{U}_i^{\min,n}, \mathsf{U}_i^{\max,n})$; otherwise, we have $\lambda_{ij}^{\mathrm{grdy},n} := \max(\lambda_\epsilon, \lambda_{12}(\boldsymbol{n}_{ij}, \mathsf{U}_i^m, \mathsf{U}_j^n)$, and inequality (4.6) is just a restatement of the local maximum principle (i.e., $\mathsf{U}_i^{n+1} \in [\mathsf{U}^{\min,n}, \mathsf{U}^{\max,n}]$). It is also demonstrated in the numerical section that the choice of k_i in $(\mathsf{U}_i^{\min,n}, \mathsf{U}_i^{\max,n})$ should be random for the method to be robust when the flux $\boldsymbol{f}$ is not strictly convex or concave.

5 The *p*-System

In this section we illustrate the greedy viscosity idea on the one-dimensional p-system. The extension to the compressible Euler equations with arbitrary equation of state will be done in the forthcoming second part of this work.

5.1 The Model Problem

The p-system is a model for isentropic gas dynamics written in Lagrangian coordinates. The dependent variable has two components which are the specific volume, v, and the velocity, u. The system is written as follows:

$$\partial_t \begin{pmatrix} v \\ u \end{pmatrix} + \partial_x \begin{pmatrix} -u \\ p(v) \end{pmatrix} = 0, \qquad (x,t) \in \mathbb{R}\times\mathbb{R}_+. \tag{5.1}$$

The pressure $v \mapsto p(v)$ is assumed to be of class $C^2(\mathbb{R}_+;\mathbb{R})$ and be such that

$$p' < 0, \qquad 0 < p''. \tag{5.2}$$

As an illustration, we are going to restrict the discussion to the gamma-law, $p(v) = rv^{-\gamma}$, where $r > 0$ and $\gamma > 1$. We introduce the notation $\boldsymbol{u}{:=}(v, u)^\mathsf{T}$ and define the flux $\mathbb{f}(\boldsymbol{u}){:=}(-u, p(v))^\mathsf{T}$.

The admissible set for (5.1) is $\mathcal{A}{:=}(0, \infty){\times}\mathbb{R}$. The p-system ($\gamma > 1$) has two families of global Riemann invariants:

$$w_+(\boldsymbol{u}) = u + \int_v^\infty \sqrt{-p'(\xi)}\,\mathrm{d}\xi, \quad \text{and} \quad w_-(\boldsymbol{u}) = u - \int_v^\infty \sqrt{-p'(\xi)}\,\mathrm{d}\xi, \tag{5.3}$$

and it can be shown that

$$\mathcal{B}_{ab} := \{\boldsymbol{u} \in \mathcal{A} \mid a \le w_-(\boldsymbol{u}),\ w_+(\boldsymbol{u}) \le b\} \tag{5.4}$$

is an invariant domain for the system (5.1) for all $a < b \in \mathbb{R}$; see Hoff [14, Exp. 3.5, p. 597] for a proof in the context of parabolic regularization, or Young [33] for a direct proof. Note in passing that it is established in Hoff [13, Thm. 2.1] and [14, Thm. 4.1] that the Lax scheme is invariant-domain preserving for all $\mathcal{B}_{ab}$.

The p-system has many entropy pairs. We are going to use the following one:

$$\eta(\boldsymbol{u}) = \frac{1}{2}u^2 + \int_v^\infty p(\xi)\,\mathrm{d}\xi; \qquad q(\boldsymbol{u}) = up(v). \tag{5.5}$$

We now follow the principles explained in Algorithm 1 to estimate a greedy viscosity.

5.2 Maximum Wave Speed

Let us consider a left state $\boldsymbol{u}_i{:=}(v_i, u_i)^\mathsf{T}$, a right state $\boldsymbol{u}_j{:=}(v_j, u_j)^\mathsf{T}$, and a one-dimensional normal direction $n_{ij} \in \{-1, +1\}$ where $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. We now describe a procedure to compute (an upper bound of) the maximal wave speed $\lambda_{\max}(n_{ij}, \boldsymbol{u}_i, \boldsymbol{u}_j)$ that was introduced in (2.4) in Sect. 2.1. One first realizes that the Riemann problem with the flux $\mathbb{f}(\boldsymbol{u})n_{ij}$, left data data $(v_i, u_i)^\mathsf{T}$ and right data $(v_j, u_j)^\mathsf{T}$, is identical to the Riemann problem with the flux $\mathbb{f}(\boldsymbol{u})$ and data $\boldsymbol{u}_L{:=}(v_i, n_{ij}u_i)^\mathsf{T}$, $\boldsymbol{u}_R{:=}(v_j, n_{ij}u_j)^\mathsf{T}$. We now use the symbol n in lieu of n_{ij} and write $\lambda_{\max}(n, \boldsymbol{u}_L, \boldsymbol{u}_R)$ instead of $\lambda_{\max}(n_{ij}, \boldsymbol{u}_i, \boldsymbol{u}_j)$.

For the index $Z \in \{L, R\}$, we introduce

$$f_Z(v) := \begin{cases} -\sqrt{(p(v) - p(v_Z))(v_Z - v)}, & \text{if } v \le v_Z \\ \displaystyle\int_{v_Z}^{v} \sqrt{-p'(\xi)}\,\mathrm{d}\xi, & \text{if } v > v_Z. \end{cases} \tag{5.6}$$

and define $\phi(v) := f_L(v) + f_R(v) + u_L - u_R$. The function ϕ is increasing and concave with $\lim_{v\to+0}\phi(v) = -\infty$; see Young [33] for details. Notice that $\lim_{v\to+\infty}\phi(v) = w_+(\boldsymbol{u}_L) - w_-(\boldsymbol{u}_R)$. If $w_+(\boldsymbol{u}_L) - w_-(\boldsymbol{u}_R) \le 0$, then we set $v^*{:=}+\infty$ (vacuum appears in the Riemann solution in this case). If $w_+(\boldsymbol{u}_L) - w_-(\boldsymbol{u}_R) \ge 0$, the equation $\phi(v) = 0$ has a unique solution which we denote by v^*. Setting $v_{\min}{:=}\min(v_L, v_R)$, we have $\phi(v_{\min}) = u_L - u_R - \sqrt{(p(v_R) - p(v_L))(v_L - v_R)}$, and the following result is standard (see e.g., [33], [4, Lem. 2.5]):

$$\lambda_{\max}(n, \boldsymbol{u}_L, \boldsymbol{u}_R) = \begin{cases} \sqrt{\frac{p(v_{\min}) - p(v^*)}{v^* - v_{\min}}}, & \text{if } \phi(v_{\min}) > 0, \\ \sqrt{-p'(v_{\min})}, & \text{otherwise,} \end{cases} \tag{5.7}$$

Note that $\lambda_{\max}(n, \boldsymbol{u}_L, \boldsymbol{u}_R)$ is a decreasing function of v^*. The value of v^* can be found using Newton's method starting with a guess v^0 smaller than v^*. As ϕ is concave and increasing,

starting the Newton iterations on the left of v^* guarantees that at each step of Newton's method the new estimate is smaller than v^*, which in turn implies that the estimated maximum speed is an upper bound for the exact maximum speed. A starting guess v^0 with the above property can be computed as follows:

$$w_+^{\max} := \max(w_+(\boldsymbol{u}_L), w_+(\boldsymbol{u}_R)), \qquad w_-^{\min} := \min(w_-(\boldsymbol{u}_L), w_-(\boldsymbol{u}_R)) \tag{5.8a}$$

$$v^0 := (\gamma r)^{\frac{1}{\gamma-1}} \left(\frac{4}{(\gamma-1)(w_+^{\max} - w_-^{\min})} \right)^{\frac{2}{\gamma-1}}. \tag{5.8b}$$

Here, Eq. (5.8b) follows from finding the pair $u^0 := (v^0, u^0)^\mathsf{T}$ solving $w_+(\boldsymbol{u}^0) = w_+^{\max}$ and $w_-(\boldsymbol{u}^0) = w_-^{\min}$. This construction implies

$$\lambda_{\max}(n, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \widehat{\lambda}_{\max} := \sqrt{\frac{p(v_{\min}) - p(v^0)}{v^0 - v_{\min}}}. \tag{5.9}$$

5.3 Invariant-Domain Property

We first compute three wave speeds to guarantee a local invariant-domain property as in (3.5). Then we compute a fourth wave speed in Sect. 5.4 so as to ensure that a local entropy inequality holds for the above-defined entropy pair; see (3.6). Recall that

$$\overline{\boldsymbol{u}}_{LR}(\lambda) = \frac{1}{2} \begin{pmatrix} v_L + v_R + \frac{1}{\lambda}(u_R - u_L) \\ u_L + u_R - \frac{1}{\lambda}(p(v_R) - p(v_L)) \end{pmatrix}. \tag{5.10}$$

We introduce

$$\Psi_1(\boldsymbol{u}) := v, \qquad \Psi_2(\boldsymbol{u}) := w_+^{\max} - w_+(\boldsymbol{u}), \qquad \Psi_3(\boldsymbol{u}) := w_-(\boldsymbol{u}) - w_-^{\min}, \tag{5.11}$$

where $w_+^{\max}$ and $w_-^{\min}$ are defined in (5.8a). Observe that Ψ_1 is concave and Ψ_2 and Ψ_3 are both strictly concave due to (5.2). We define $\mathcal{B}_0 := \mathbb{R}^2$, $\mathcal{B}_1 := \{\boldsymbol{u} \in \mathbb{R}^2 \mid \Psi_1(v) > 0\} = \mathcal{A}$, $\mathcal{B}_2 := \{\boldsymbol{u} \in \mathcal{B}_1 \mid \Psi_2(\boldsymbol{u}) \ge 0\}$, and $\mathcal{B}_3 := \{\boldsymbol{u} \in \mathcal{B}_2 \mid \Psi_3(\boldsymbol{u}) \ge 0\}$. It is necessary to introduce Ψ_1 and $\mathcal{B}_1 = \mathcal{A}$ to make sure that the domain of definition of Ψ_2 and Ψ_3 is $\mathcal{A}$.

If $\boldsymbol{u}_L = \boldsymbol{u}_R$, then $\overline{\boldsymbol{u}}_{LR}(\lambda) = \boldsymbol{u}_L = \boldsymbol{u}_R$ for all $\lambda > 0$. In this case, we take $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_\epsilon$. Let us now assume that $\boldsymbol{u}_L \ne \boldsymbol{u}_R$. The smallest wave speed λ_1, greater than or equal to λ_ϵ, that ensures $\Psi_1(\overline{\boldsymbol{u}}_{LR}(\lambda)) > 0$ for all $\lambda > \lambda_1$ is given by

$$\lambda_1 = \max\left(\frac{u_L - u_R}{v_L + v_R}, \lambda_\epsilon \right). \tag{5.12}$$

Now we estimate λ_2. If $\Psi_2(\overline{\boldsymbol{u}}_{LR}(\lambda_1)) \ge 0$, then we set $\lambda_2 = \lambda_1$. If $\Psi_2(\overline{\boldsymbol{u}}_{LR}(\lambda_1)) < 0$ there are two cases. If $u_R - u_L \ge 0$ and $-(p(v_R) - p(v_L)) \le 0$, we have $\Psi_2(\overline{\boldsymbol{u}}_{LR}(\lambda)) \ge \Psi_2(\frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R)) \ge 0$ for all $\lambda > 0$ and we set $\lambda_2 := \lambda_1$. Otherwise, we observe that the curve $w_+(\boldsymbol{u}) = w_+^{\max}$ has a horizontal asymptote given by $\{u = w_+^{\max}\}$ and a vertical asymptote given by $\{v = 0\}$ and the condition ($u_R - u_L < 0$ or $-(p(v_R) - p(v_L)) > 0$) implies that the equation $\Psi_2(\overline{\boldsymbol{u}}_{LR}(\lambda)) = 0$ has a unique positive solution, λ_2^*, which can be computed using an iterative method, and we set $\lambda_2 = \lambda_2^*$; we omit the details for brevity. The argument to estimate λ_3 is analogous: If $\Psi_3(\overline{\boldsymbol{u}}_{LR}(\lambda_2)) \ge 0$, then we set $\lambda_3 = \lambda_2$. Otherwise, we observe that the curve $w_-(\boldsymbol{u}) = w_-^{\min}$ has a horizontal asymptote given by $\{u = w_-^{\min}\}$ and a vertical asymptote given by $\{v = 0\}$. Hence if $u_R - u_L \ge 0$ and $-(p(v_R) - p(v_L)) \ge 0$, we have $\Psi_3(\overline{\boldsymbol{u}}_{LR}(\lambda)) \ge \Psi_3(\frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R)) > 0$ for all $\lambda > 0$ and we set $\lambda_3 := \lambda_2$. Otherwise the

equation $\Psi_3(\overline{\boldsymbol{u}}_{LR}(\lambda)) = 0$ has a unique positive solution, λ_3^*, which can be computed using an iterative method, and we set $\lambda_3 = \lambda_3^*$. As asserted in Lemma 3.5, the process described above guarantees that $\overline{\boldsymbol{u}}_{LR}(\lambda) \in \mathcal{B}_3 := \{\boldsymbol{u} \in \mathcal{A} \mid \Psi_2(\boldsymbol{u}) \geq 0,\ \Psi_3(\boldsymbol{u}) \geq 0\}$ for all $\lambda \geq \lambda_3$.

5.4 Wave Speed Based on the Entropy Inequality

We now estimate a wave speed associated with one entropy inequality. The entropy functional in this case is

$$\Phi_e(t) := \eta\left(\overline{\boldsymbol{u}}_{LR}(\tfrac{1}{t})\right) - \frac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) + \frac{t}{2}(q(\boldsymbol{u}_R) - q(\boldsymbol{u}_L)), \tag{5.13}$$

where η and $\boldsymbol{q}$ are defined in (5.5). We have $\eta(\boldsymbol{u}) = \frac{1}{2}u^2 - \frac{1}{1-\gamma}vp(v)$ for the pressure gamma-law.

If $\boldsymbol{u}_L = \boldsymbol{u}_R$, then $\overline{\boldsymbol{u}}_{LR}(\lambda) = \boldsymbol{u}_L = \boldsymbol{u}_R$ for all $\lambda > 0$ and $\Phi_e(t) = 0$ for all $t \geq 0$. In this case, we take $\lambda_e = \lambda_3$. If $\boldsymbol{u}_L \neq \boldsymbol{u}_R$, we compute λ_e as defined in (3.6). More precisely, if $\Phi_e(\frac{1}{\lambda_3}) \leq 0$, then we set $\lambda_e = \lambda_3$. Otherwise, we observe that the equation $\Phi_e(\frac{1}{\lambda}) = 0$ has a unique solution in $[1/\lambda_{LR}^\sharp, 1/\lambda_3)$ because η defined in (5.5) is strictly convex and we also have established in (3.4) that $\Phi_e(\frac{1}{\lambda_{LR}^\sharp}) \leq 0$. Finally, we set $\lambda_e = \max(\lambda_e, \lambda_3)$. The greedy wave speed is obtained by setting $\lambda^{\text{grdy}}(n, \boldsymbol{u}_L, \boldsymbol{u}_R) := \lambda_e$. This algorithm is illustrated numerically in Sect. 7.

6 Numerical Illustrations with Scalar Conservation Equations

We start by illustrating the method for scalar conservation equations. To test the robustness of the method, we choose problems with fluxes that are not strictly convex and contain *sonic points*. Methods that underestimate the maximum wave speed (or just enforce the maximum principle) tend to fail when applied to this type of problems.

Here, we numerically show that computing the viscosity so as to enforce local entropy inequalities is sufficient to select the entropy solution provided that the family of entropies is rich enough. All the computations are done with continuous $\mathbb{P}_1$ finite elements and we take $\epsilon = 10^{-8}$ in (2.19a). The time stepping is done with the three stages, third-order, strong stability preserving Runge Kutta method [28]. The time step is computed by using the expression $\tau_n = \frac{\text{CFL}}{2} \max_{i\in\mathcal{V}} m_i / \sum_{j\in\mathcal{I}(i)^*} d_{ij}^{\text{grdy},n}$.

6.1 Piecewise Linear Flux

We consider a Riemann problem in one space dimension for the scalar conservation equation $\partial_t u + \partial_x f(u) = 0$ using the scalar flux $f(v) = 2 - v$ if $v \leq 2$ and $f(v) = 2v - 4$ otherwise. The initial data is $u_0(x) = 1$ if $x \leq 0$ and $u_0(x) = 3$ otherwise. This flux is convex and Lipschitz, but it is not strictly convex: the velocity is piecewise constant and discontinuous. This class of problems is thoroughly investigated in Petrova and Popov [27]. The solution is

$$u(x,t) = \begin{cases} 1 & \text{if } x \leq -t \\ 2 & \text{if } -t < x \leq 2t \\ 3 & \text{if } 2t < x. \end{cases} \tag{6.1}$$

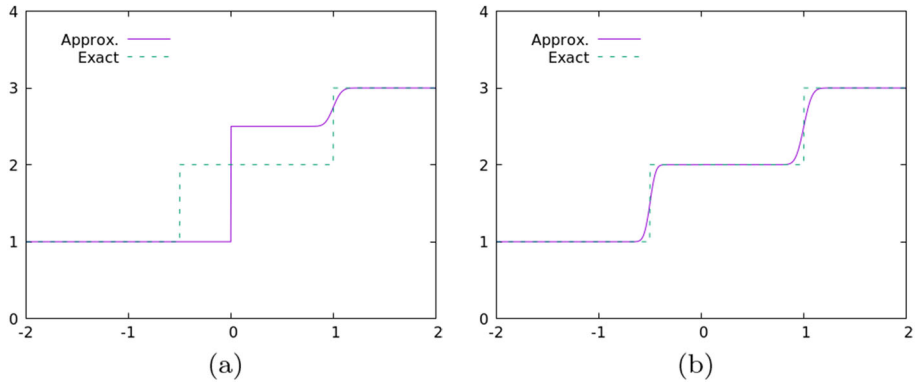


Fig. 2 Approximation of a scalar conservation equation with piecewise linear flux: **a** viscosity solely based on λ_{12} given in (4.2); **b** viscosity based on an entropy inequality, Eq. (4.3) with the choice $k_i = \frac{1}{2}(\mathsf{U}_i^{\min,n} + \mathsf{U}_i^{\max,n})$

The solution is composed of two contact waves (i.e., the characteristics do not cross) separated by an expansion wave. One contact wave moves to the left at speed -1, the other moves to the right at speed 2.

This example is meant to demonstrate that only using the wave speed λ_{12} defined in (4.2) to construct the graph viscosity (i.e., only using the Roe average) is not robust even in a case as simple as the one above. Using the wave speed λ_{12} guarantees that the maximum principle locally holds, but the approximation may converge to a nonentropic weak solution. We illustrate this phenomenon by applying the algorithm described in the paper over the domain $D = (-2, 2)$ using uniform meshes. The solution is computed to a final time $t = 0.5$ using CFL=0.75. We show in the left panel of Fig. 2 the solution obtained with the viscosity computed by using only λ_{12}. The graph of the exact solution is shown with a dashed line. We observe that the approximate solution does not converge to the exact solution. The leftmost discontinuity in the approximate solution is stationary instead of moving to the left at speed -1. The right panel shows the approximate solution using definition (4.3) for the wave speed with $k_i = \frac{1}{2}(\mathsf{U}_i^{\min,n} + \mathsf{U}_i^{\min,n})$ for every $i \in \mathcal{V}$. We have verified that the method using this definition for the wave speed converges with the expected rate (tables not shown here for brevity).

6.2 1D Non-convex Flux

We now consider a Riemann problem in one space dimension using the scalar flux $f(v) = \sin(v)$. The initial data is $u_0(x) = (2 + a)\pi$ if $x < 0$ and $u_0(x) = b\pi$ otherwise. Here, $a \in [\frac{1}{2}, 1]$ and $b \in [0, \frac{1}{2}]$ are two chosen parameters. Note that the flux is neither convex nor concave over the interval $[b\pi, (2 + a)\pi]$. Since $(2 + a)\pi > b\pi$, the solution is obtained by replacing the flux by its upper concave envelope which is $\widehat{f}(v) = \sin(v)$ for $v \in [b\pi, \frac{1}{2}\pi]$, $\widehat{f}(v) = 1$ for $v \in [\frac{1}{2}\pi, \frac{5}{2}\pi]$, and $\widehat{f}(v) = \sin(v)$ for $v \in [\frac{5}{2}\pi, (2 + a)\pi]$ (see, e.g., Dafermos [2, Lem. 3.1] and Holden and Risebro [15, §2.2]). We note that the entire the interval $v \in [\frac{1}{2}\pi, \frac{5}{2}\pi]$ is composed of sonic points. The exact solution is given by

$$u(x,t) = \begin{cases} (2+a)\pi & \text{if } x \le t\cos((2+a)\pi) \\ 3\pi - \arccos(|x/t|) & \text{if } t\cos((2+a)\pi) < x \le 0 \\ \arccos(x/t) & \text{if } 0 < x \le t\cos(b\pi) \\ b\pi & \text{if } t\cos(b\pi) < x. \end{cases} \tag{6.2}$$

It is a composite wave composed of an expansion followed by a stationary shock followed by a second expansion. The numerical tests reported below are done with $b = 0$ and $a = 1$ over the domain $D = (-1, 1)$.

Here again, tests done with the graph viscosity solely based on the Roe average λ_{12} yields a method that is not robust (figures and tables are not reported for brevity). We observe that the approximate solution is a stationary shock for every mesh refinement (i.e., the initial data does not evolve), which is clearly not the entropy solution. One can artificially try to avoid this problem by initializing the approximate solution at $t_0 > 0$ using the exact solution (6.2). If the mesh does not have a vertex located at $\{0\}$, then convergence starts only when the mesh size is less that t_0. On the other hand, we observe convergence with no pre-asymptotic range for every positive value of t_0 when the mesh has a vertex located at $\{0\}$. This behavior illustrates well the lack of robustness of methods that are solely based on the wave speed λ_{12}.

We now test the method based on the wave speed computed by using (4.3). The tests are done with CFL $= 0.5$. The relative errors in the L^1-norm and L^2-norm are computed at $t = 0.8$. We test two strategies to select the Krŭzkov entropy for each degree of freedom $i \in \mathcal{V}$. The first strategy consists of setting $k_i = \theta \mathsf{U}_i^{\min,n} + (1-\theta)\mathsf{U}_i^{\max,n}$ where $\theta = \frac{1}{2}$. The second strategy consists of setting $k_i = \theta_i \mathsf{U}_i^{\min,n} + (1-\theta_i)\mathsf{U}_i^{\max,n}$, where $\theta_i \in (0,1)$ is a uniformly distributed random number changing at every grid point $i \in \mathcal{V}$.

When using the first strategy with fixed $\theta = \frac{1}{2}$ we observe exactly the same problems as reported above when only using λ_{12}. Irrespective of the location of the grid points, the approximate solution is a stationary shock when one initializes the approximate solution with the exact solution at $t_0 = 0$. Initializing with the exact solution (6.2) at $t_0 = 10^{-8}$ still produces a stationary shock when the point $\{x = 0\}$ is not a vertex of the mesh, but a non trivial solution is obtained when the point $\{x = 0\}$ is a vertex of the mesh. We show in the right part of Table 1 convergence results using $t_0 = 10^{-8}$ and uniform meshes with odd numbers of grid points. We observe some kind of convergence on coarse meshes, but eventually the error stalls and stagnates as the mesh is further refined. We have observed this behavior for every constant value of θ. This is highly counter intuitive because the viscosity based on (4.3) is strictly larger than λ_{12}, and we have observed in the above paragraph that the approximate solution using λ_{12} converges to the entropy solution when the point $\{x = 0\}$ is a mesh vertex. Here again, we observe a clear lack of robustness even when the wave speed is augmented so as to guarantee one "entropy fix" per grid point.

We now discuss what happens when the Krŭzkov entropy is randomly chosen. All the problem mentioned above disappear when $\theta_i \in (0,1)$ is randomly chosen at every grid point. The method convergences whether there is a grid point at $\{0\}$ or not and whatever the initial time. In particular there is no problem setting $t_0 = 0$. We show convergence tests in the left panel of Table 1 with $t_0 = 0$. To be able to compare with the results displayed in the right part of the table, we have use the same meshes. The method is now clearly convergent and converges with the expected rates.

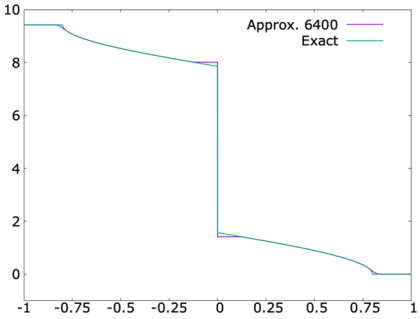
The conclusion of this section is that the method based on the greedy wave speed computed by using (4.3) with random Krŭzkov entropies is robust.

Remark 6.1 *(Robustness and "entropy stability")* The numerical tests performed in this section demonstrates that robustness comes from randomness of the Krŭzkov entropy. Note in

Table 1 1D two-sonic point problem

# dofs	Random entropy $\delta^1(t)$	Rate	$\delta^2(t)$	Rate	Average entropy $\delta^1(t)$	Rate	$\delta^2(t)$	Rate
51	1.96E−02	–	2.21E−02	–	2.41E−02	–	2.30E−02	–
101	1.39E−02	0.49	1.65E−02	0.42	1.81E−02	0.41	1.77E−02	0.38
201	9.17E−03	0.60	1.13E−02	0.55	1.33E−02	0.45	1.38E−02	0.35
401	5.87E−03	0.64	8.25E−03	0.45	9.81E−03	0.43	1.14E−02	0.28
801	3.66E−03	0.68	5.76E−03	0.52	7.54E−03	0.38	9.89E−03	0.20
1601	2.24E−03	0.71	4.15E−03	0.47	6.10E−03	0.31	9.06E−03	0.13
3201	1.38E−03	0.70	2.89E−03	0.52	5.20E−03	0.23	8.62E−03	0.07
6401	8.50E−04	0.70	2.06E−03	0.49	4.65E−03	0.16	8.39E−03	0.04

The second and fourth columns show relative errors in the L^1-norm and the L^2-norms using a random Krůzkov entropy with $k = \theta u_L + (1-\theta)u_R$, where $\theta \in (0,1)$ is a uniformly distributed random value. The sixth and eight columns report relative errors in the L^1-norm and the L^2-norms obtained for the average Krůzkov entropy with $k = \frac{1}{2}(u_L + u_R)$



(a) Solution at $t = 0.8$

# dofs	Square entropy $\eta(v) = \frac{1}{2}v^2$ $\delta^1(t)$	rate	$\delta^2(t)$	rate
50	2.22E-02	–	2.18E-02	–
100	1.63E-02	0.45	1.64E-02	0.41
200	1.15E-02	0.50	1.23E-02	0.42
400	8.08E-03	0.51	9.44E-03	0.38
800	5.82E-03	0.47	7.61E-03	0.31
1600	4.38E-03	0.41	6.50E-03	0.23
3200	3.48E-03	0.33	5.87E-03	0.15
6400	2.92E-03	0.25	5.52E-03	0.09

(b) Convergence table

Fig. 3 1D two-sonic point problem computed with the square entropy $\eta(v) = \frac{1}{2}v^2$. The "entropy stable" method does not converge to the entropy solution

passing that this series of tests casts doubt on the robustness of methods that are called *entropy stable* in the literature. Since these methods enforce only *one fixed* global entropy inequality (at the semi-discrete level), one may wonder whether they produce approximations that converge to the right solution for the above one-dimensional problem. In order to provide some numerical evidence in this matter, we adjust our method as introduced in Sect. 4.2 for the entropy $\eta(v) = \frac{1}{2}v^2$ which is usually invoked in the literature dedicated to entropy stable methods. Redoing the computations in the proof of Lemma 4.2 with the square entropy gives $\lambda = (2ab + d + \sqrt{\Delta})/(2(c - a^2))$ with $a{:=}\frac{1}{2}(u_L + u_R)$, $b{:=}\frac{1}{2}(\boldsymbol{f}(u_L) - \boldsymbol{f}(u_R))\cdot\boldsymbol{n}$, $c{:=}\eta(u_L)+\eta(u_R)$, $d = (\boldsymbol{q}(u_R) - \boldsymbol{q}(u_L))\cdot\boldsymbol{n}$, $\Delta{:=}(2ab+d)^2 - 4b^2(a^2-c)$. The method thus produced is locally and globally entropy stable with respect to $\eta(v) = \frac{1}{2}v^2$, i.e., Eq. (3.11) holds. Convergence tests with this method are reported in Fig. 3. These tests show that the approximation does not converge to the entropy solution (6.2). The convergence behavior is strange as the approximation seems to converge over a large pre-asymptotic range, but eventually, when the mesh is very fine, the approximation converges to a weak solution that is not the entropy solution. In conclusion, the method is definitely entropy stable for the square entropy but it is not convergent for non-convex fluxes; hence, it is not robust.

(a) λ_{12} with (4.2) (b) λ^{grdy} with (4.3), $\theta_i = \frac{1}{2}$ (c) λ^{grdy} with (4.3), θ_i random

Fig. 4 2D KPP problem with $\mathbb{P}_1$ elements on nonuniform Delaunay mesh (118850 grid points) at $t = 1$, CFL $= 0.5$, computed with three different strategies: **a** $\lambda_{\max} = \lambda_{12}$ using (4.2); **b** wave speed λ^{grdy} computed with (4.3) using $k_i = \theta \mathsf{U}_i^{\min,n} + (1-\theta)\mathsf{U}_i^{\max,n}$ with $\theta = \frac{1}{2}$; **c** wave speed λ^{grdy} computed with (4.3) using $k_i = \theta_i \mathsf{U}_i^{\min,n} + (1-\theta_i)\mathsf{U}_i^{\max,n}$ where $\theta_i \in (0, 1)$ is a uniformly random number changing for every $i \in \mathcal{V}$. Only the solution in the right panel is the correct entropy solution

6.3 The 2D KPP Problem

We finish our numerical examples by solving a two-dimensional scalar conservation equation with the non-convex flux $\boldsymbol{f}(u) := (\sin u, \cos u)^{\mathsf{T}}$

$$\partial_t u + \nabla \cdot \boldsymbol{f}(u) = 0, \quad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) = \begin{cases} \frac{14\pi}{4} & \text{if } \sqrt{x^2 + y^2} \le 1 \\ \frac{\pi}{4} & \text{otherwise,} \end{cases} \tag{6.3}$$

in the computational domain $D = [-2, 2]\times[-2.5, 1.5]$. The problem was originally proposed in Kurganov et al. [17]. The solution has a two-dimensional composite wave structure which high-order numerical schemes have difficulties to capture correctly. We approximate the solution with continuous $\mathbb{P}_1$ finite elements on nonuniform Delaunay triangulations up to a final time of $t = 1$. We show in Fig. 4 three results computed on a mesh with 118850 grid points with CFL = 0.5. The solution shown in the leftmost panel is obtained by only using the wave speed λ_{12} for computing the greedy viscosity. The solution in the midle panel is obtained with the the wave speed (4.3) and the Krůzkov entropy using $k_i = \theta \mathsf{U}_i^{\min,n} + (1-\theta)\mathsf{U}_i^{\max,n}$ with $\theta = \frac{1}{2}$. The solution in the rightmost panel is obtained with the the wave speed (4.3) and the Krůzkov entropy using $k_i = \theta_i \mathsf{U}_i^{\min,n} + (1-\theta_i)\mathsf{U}_i^{\max,n}$ where θ_i is a random number changing for every $i \in \mathcal{V}$. One may be mislead thinking that the solution in the middle panel is correct, but the only approximation that converges correctly is the one using the random entropy.

So, here again, our conclusion for scalar conservation equations is that robustness can be achieved for methods based on the greedy wave speed (4.3) provided the Krůzkov entropies are chosen randomly. Any other choice is not robust.

7 p-System

We test the method on the p-system using the equation of state $p(v) = \frac{1}{\gamma} v^{\gamma}$ with $\gamma = 3$. We consider a Riemann problem with left state $\boldsymbol{u}_L = (v_L, \sqrt{(1 - v_L)(p(v_L) - p(1))})$ and right state $\boldsymbol{u}_R = (v_R, -\sqrt{(1 - v_R)(p(v_R) - p(1))})$. The solution is composed of two shock

Table 2 Convergence tests for the p system for various choices of wave speed estimate

# dofs	$\widehat{\lambda}_{\max}$ $\delta^1(t)$	Rate	$\lambda_{\max}$ $\delta^1(t)$	Rate	λ^{grdy} $\delta^1(t)$	Rate
51	3.33E−01	–	1.93E−01	–	1.31E−01	–
101	2.41E−01	0.47	1.57E−01	0.30	1.18E−01	0.15
201	1.41E−01	0.78	6.58E−02	1.25	4.93E−02	1.25
401	7.59E−02	0.89	4.42E−02	0.58	3.65E−02	0.43
801	3.64E−02	1.06	2.09E−02	1.08	1.77E−02	1.05
1601	1.70E−02	1.10	9.07E−03	1.20	7.76E−03	1.19

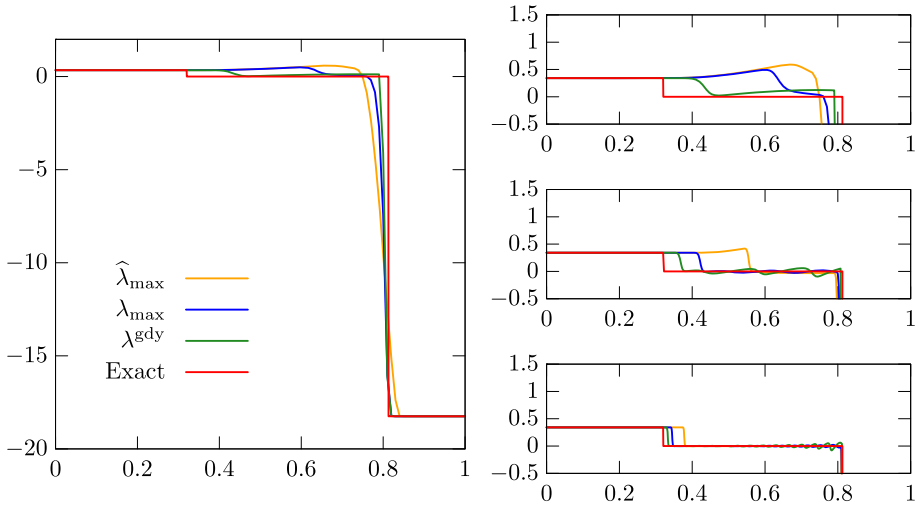


Fig. 5 Approximation of the u component in the p system, $t = 0.5$. Left: comparisons between the methods using $\widehat{\lambda}_{\max}$, $\lambda_{\max}$, and λ^{gdy} with 101 grid points. Right: Three refinements: 101 grid points (top), 401 grid points (middle), 1600 grid points (bottom)

waves when $v_L, v_R > 1$, and in this case $v^* = 1$, $u^* = 0$. For this test we set $v_L = 1.5$ and $v_R = 1000$. The left shock is weak and fast moving; the shock speed is close to -0.6849. The right shock is strong and slow; the shock speed is close to 1.827×10^{-2}.

The simulations are done in the computational domain $D{:=}(0, 1)$. The initial data is $\boldsymbol{u}_0(x) = \boldsymbol{u}_L$ if $x < 0.8$ and $\boldsymbol{u}_0(x) = \boldsymbol{u}_R$ otherwise. The relative error in the L^1-norm is computed at $t = 0.7$. The relative error is the sum of the relative error on v plus the relative error on u. Convergences test are done on a sequence of uniform meshes starting from 51 grid points to 1601 grid points. The results are shown in Table 2. The results in the first column are obtained by using the upper wave speed estimate $\widehat{\lambda}_{\max}$ given in (5.9). Those shown in the second column are obtained by using $\lambda_{\max}$ as defined in (5.7) where v^* is computed with a Newton method with 10^{-10} tolerance. Those shown in the right column are obtain with the greedy viscosity λ^{grdy} defined in Sects. 5.3–5.4.

We show in Fig. 5 the graph of the u component at the final time $t = 0.7$. In the left panel the approximation is done with 101 uniform grid points. We show a closer view of the plateau separating the two shocks in the right panel. The number of grid points used in each case is: 101 in the top right panel; 401 in middle right panel; and 1600 in the bottom right panel. This

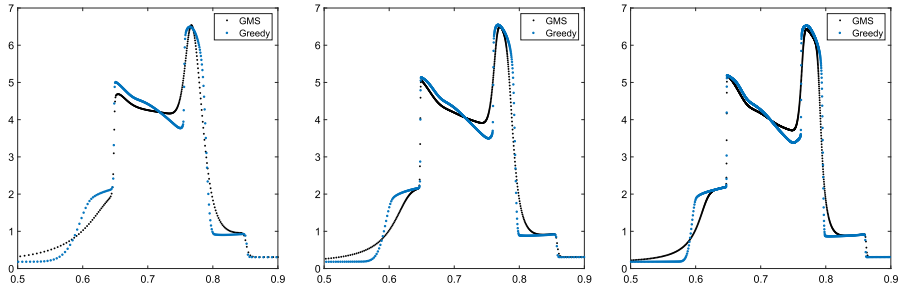


Fig. 6 Woodward–Colella blast wave. Density at $t = 0.038$. GMS versus Greedy viscosity, from left to right: $I = 400, 800, 1600$

series of simulations demonstrate well the gain in accuracy that can potentially be gained by using the greedy viscosity technique described in this paper.

8 Conclusions

We have presented a general strategy to compute the artificial viscosity in first-order approximation methods for hyperbolic systems. The technique is based on the estimation of a minimum wave speed guaranteeing that the approximation satisfies predefined invariant-domain properties and predefined entropy inequalities. This approach eliminates non-essential fast waves from the construction of the artificial viscosity, while preserving preassigned invariant-domain properties and entropy inequalities. One should however keep in mind that being invariant-domain preserving is in general not enough to have a method that is robust. Likewise ensuring only one entropy inequality is not a guarantee of robustness.

We finish by briefly demonstrating the performance of the proposed methodology when applied to the compressible Euler equations. For each pair $(i, j), i, \in \mathcal{V}, j \in \mathcal{I}(i)$, the greedy viscosity is computed by first computing a wave speed that guarantees that the density satisfies local lower and upper bounds extracted from the local Riemann problem. This wave speed is then augmented by making sure that the specific entropy satisfies a local bound. Finally, the wave speed is possibly again augmented to guarantee a local entropy inequality. The details are reported in a forthcoming second part of this work. As a preview, we consider the Woodward–Colella blast wave problem [32]. We show in Fig. 6 the density profile at $t = 0.038$. We compare for three different mesh sizes the results obtained with the wave speed $\lambda^{\max}$ (labelled with the acronym "GMS" for guaranteed maximum speed) with those obtained with the greedy wave speed λ^{greedy} (labelled with the acronym "Greedy"). The superiority of the greedy viscosity over the low-order standard method is evident, particularly in the region of the leftmost contact wave.

Acknowledgements The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. https://www.tacc.utexas.edu.

Funding DMS-1912847 (MM), DMS-2045636 (MM), by the Air Force Office of Scientific Research, under grant/contract number FA9550-23-1-0007 (JLG, MM, BP), the Army Research Office, under grant number W911NF-19-1-0431 (JLG, BP), the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contracts B640889, B641173 (JLG, BP), and by the European Union-NextGenerationEU, through the

National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008 (BP), and by the Spanish MCIN/AEI(10.13039/501100011033) under the grant PID2020-114173RB-I00 (LS).

Data availability Two codes have been written for this project: one in Fortran2018 and one in C++. The Fortran code is available upon request. The C++ code is part of the Ryujin library https://github.com/conservation-laws/ryujin.

References

1. Clayton, B., Guermond, J.-L., Popov, B.: Upper bound on the maximum wave speed in Riemann problems for the Euler equations with tabulated equation of state (2021). https://doi.org/10.5281/zenodo.4685868
2. Dafermos, C.M.: Polygonal approximations of solutions of the initial value problem for a conservation law. J. Math. Anal. Appl. **38**, 33–41 (1972)
3. Guermond, J.-L., Popov, B.: Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations. J. Comput. Phys. **321**, 908–926 (2016)
4. Guermond, J.-L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016)
5. Guermond, J.-L., Popov, B.: Invariant domains and second-order continuous finite element approximation for scalar conservation equations. SIAM J. Numer. Anal. **55**(6), 3120–3146 (2017)
6. Guermond, J.-L., Popov, B., Saavedra, L., Yang, Y.: Invariant domains preserving arbitrary Lagrangian Eulerian approximation of hyperbolic systems with continuous finite elements. SIAM J. Sci. Comput. **39**(2), A385–A414 (2017)
7. Guermond, J.-L., Nazarov, M., Popov, B., Tomas, I.: Second-order invariant domain preserving approximation of the Euler equations using convex limiting. SIAM J. Sci. Comput. **40**(5), A3211–A3239 (2018)
8. Guermond, J.-L., Popov, B., Tomas, I.: Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. Comput. Methods Appl. Mech. Eng. **347**, 143–175 (2019)
9. Guermond, J.-L., Popov, B.B., Tomas, I.: Invariant domain preserving schemes for hyperbolic systems of conservation laws. Multimat, August 2019, Trento, Italy. https://www.osti.gov/biblio/1641823-invariant-domain-preserving-schemes-hyperbolic-systems-conservation-laws (2019b)
10. Harten, A.: High resolution schemes for hyperbolic conservation laws. J. Comput. Phys. **49**(3), 357–393 (1983). (**ISSN 0021-9991**)
11. Harten, A., Hyman, J.M.: Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. J. Comput. Phys. **50**(2), 235–269 (1983)
12. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**(1), 35–61 (1983)
13. Hoff, D.: A finite difference scheme for a system of two conservation laws with artificial viscosity. Math. Comput. **33**(148), 1171–1193 (1979)
14. Hoff, D.: Invariant regions for systems of conservation laws. Trans. Am. Math. Soc. **289**(2), 591–610 (1985)
15. Holden, H., Risebro, N.H.: Front tracking for hyperbolic conservation laws, 2nd Edn., vol. 152 of Applied Mathematical Sciences. Springer, Heidelberg (2015)
16. Kronbichler, M., Maier, M., Tomas, I.: Graph-based methods for hyperbolic systems of conservation laws using discontinuous space discretizations, Part I: building blocks (2024)
17. Kurganov, A., Petrova, G., Popov, B.: Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws. SIAM J. Sci. Comput. **29**(6), 2381–2401 (2007)
18. Kuzmin, D., Turek, S.: Flux correction tools for finite elements. J. Comput. Phys. **175**(2), 525–558 (2002)
19. Kuzmin, D., Löhner, R., Turek, S.: Flux-corrected transport: principles, algorithms, and applications. In: Scientific Computation. Springer (2012). ISBN 9789400740372
20. Lax, P.D.: Weak solutions of nonlinear hyperbolic equations and their numerical computation. Commun. Pure Appl. Math. **7**, 159–193 (1954)
21. Lax, P.D.: Hyperbolic systems of conservation laws. II. Commun. Pure Appl. Math. **10**, 537–566 (1957)
22. Le Métayer, O., Saurel, R.: The Noble-Abel stiffened-gas equation of state. Phys. Fluids **28**(4), 046102 (2016)
23. Maier, M., Shadid, J., Tomas, I.: Structure-preserving finite-element schemes for the Euler–Poisson equations. Commun. Comput. Phys. **33**(3), 647–691 (2023). (**ISSN 1991-7120**)
24. Nessyahu, H., Tadmor, E.: Nonoscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys. **87**(2), 408–463 (1990)

25. Osher, S.: The Riemann problem for nonconvex scalar conservation laws and Hamilton–Jacobi equations. Proc. Am. Math. Soc. **89**(4), 641–646 (1983)
26. Perthame, B., Shu, C.-W.: On positivity preserving finite volume schemes for Euler equations. Numer. Math. **73**(1), 119–130 (1996)
27. Petrova, G., Popov, B.: Linear transport equations with discontinuous coefficients. Commun. Partial Differ. Equ. **24**(9–10), 1849–1873 (1999)
28. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**(2), 439–471 (1988)
29. Tadmor, E.: Numerical viscosity and the entropy condition for conservative difference schemes. Math. Comput. **43**(168), 369–381 (1984)
30. Tadmor, E.: The large-time behavior of the scalar, genuinely nonlinear Lax-Friedrichs scheme. Math. Comput. **43**(168), 353–368 (1984)
31. Toro, E.F.: Riemann solvers and numerical methods for fluid dynamics. In: A practical introduction, 3rd edn. Springer, Berlin (2009)
32. Woodward, P., Colella, P.: The numerical simulation of two-dimensional fluid flow with strong shocks. J. Comput. Phys. **54**(1), 115–173 (1984). (**ISSN 0021-9991**)
33. Young, R.: 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., pp. 219–234. American Mathematical Society, Providence (2002)
34. Zalesak, S.T.: Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys. **31**(3), 335–362 (1979)
35. Zhang, X., Shu, C.-W.: On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. J. Comput. Phys. **229**(23), 8918–8934 (2010)
36. Zhang, X., Shu, C.-W.: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **467**(2134), 2752–2776 (2011)