Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 347 (2019) 143–175
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems☆

Jean-Luc Guermond*, Bojan Popov, Ignacio Tomas

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

Received 6 July 2018; accepted 30 November 2018
Available online 14 December 2018

Abstract

We introduce an approximation technique for nonlinear hyperbolic systems with sources that is invariant domain preserving. The method is discretization-independent provided elementary symmetry and skew-symmetry properties are satisfied by the scheme. The method is formally first-order accurate in space. Then, we introduce a series of higher-order methods. When these methods violate the invariant domain properties, they are corrected by a limiting technique that we call convex limiting. After limiting, the resulting methods satisfy all the invariant domain properties that are imposed by the user (see Theorem 7.24) and is formally high-order accurate. The two key novelties are that (i) limiting is done by enforcing bounds on quasiconcave functionals; (ii) the bounds that are enforced on the solution at each time step are necessarily satisfied by the low-order approximation.

MSC: 65M60; 65M10; 65M15; 35L65

Keywords: Hyperbolic systems; Second-order accuracy; Convex invariant sets; Limiting; Graph viscosity; Finite volumes; Finite elements

1. Introduction

The present paper is concerned with the approximation of hyperbolic systems in conservation form with a source term:

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

The space dimension d is arbitrary. The dependent variable $\boldsymbol{u}$ takes values in $\mathbb{R}^m$ and the flux $\mathbb{f}$ takes values in $(\mathbb{R}^m)^d$. In this paper $\boldsymbol{u}$ is considered as a column vector $\boldsymbol{u} = (u_1,\dots,u_m)^\mathsf{T}$. The flux is a matrix with entries $\mathbb{f}_{ij}(\boldsymbol{u}(\boldsymbol{x}))$, $1\le i\le m$, $1\le j\le d$ and $\nabla\cdot\mathbb{f}(\boldsymbol{u}(\boldsymbol{x}))$ is a column vector with entries $(\nabla\cdot\mathbb{f}(\boldsymbol{u}))_i = \sum_{1\le j\le d}\partial_{x_j}\mathbb{f}_{ij}(\boldsymbol{u}(\boldsymbol{x}))$. For any $\boldsymbol{n} = (n_1\dots,n_d)^\mathsf{T}\in\mathbb{R}^d$, we denote $\mathbb{f}(\boldsymbol{u})\boldsymbol{n}$ the column vector with entries $\sum_{1\le l\le d}\mathbb{f}_{il}(\boldsymbol{u})n_l$, where $i\in\{1{:}m\}$. To simplify questions regarding boundary conditions, we assume that either periodic boundary conditions are enforced,

☆ This material is based upon work supported in part by the National Science Foundation grants DMS-1619892, DMS-1620058, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397, and by the Army Research Office under grant/contract number W911NF-15-1-0517.
* Corresponding author.
E-mail address: guermond@math.tamu.edu (J.-L. Guermond).

https://doi.org/10.1016/j.cma.2018.11.036
0045-7825/

or the initial data is compactly supported or constant outside a compact set. In both cases we denote by $D \subseteq \mathbb{R}^d$ the spatial domain where the approximation is constructed. The domain D is the d-torus in the case of periodic boundary conditions. In the case of the Cauchy problem, D is a compact, polygonal portion of $\mathbb{R}^d$ large enough so that the domain of influence of $\boldsymbol{u}_0$ is always included in D over the entire duration of the simulation.

The objective of the paper is to generalize the techniques that was introduced in Guermond et al. [1] for the approximation of the compressible Euler equations using continuous finite elements. We want to present an approximation technique that is almost discretization independent and works with any hyperbolic system with source term, under some mild assumptions on the source. The formalism encompasses finite volumes, continuous finite elements and discontinuous finite elements. The method is formally second-order or higher-order in space and can be made (at least) fourth-order accurate in time by using explicit Runge–Kutta SSP methods. The key ingredients of the method are as follows: (i) A low-order invariant domain preserving approximation technique using a graph viscosity. (The viscosity is based on the connectivity graph of the degrees of freedom of the method. One viscosity coefficient is computed on every edge of the graph.) (ii) A high-order approximation technique. (The method may not be fully entropy consistent and may step out of the local invariant domain); (iii) A *convex limiting* technique with guaranteed bounds. (The bounds in question are obtained by computing auxiliary states on every edge of the connectivity graph. The convex limiting method works for any quasiconcave functional, i.e., it is possible to limit any quasiconcave functional of the approximate solution.)

The paper is organized as follows. We recall elementary properties of the hyperbolic system (1.1) in Section 2. The theory for the low-order method is explained in Section 3. The main result of this section is Theorem 3.6. The auxiliary states, which play a key role in the convex limiting technique are defined in (3.8). The method is illustrated in the context of finite volumes, continuous finite elements, and discontinuous finite elements in Section 4. A brief overview of explicit Runge–Kutta Strong Stability Preserving methods is made in Section 5. The key result of this section is a reformulation of the Shu–Osher Theorem 5.4 which does not involve any norm. We show therein that only convexity matters. It seems that the result, as reformulated, is not well known in the literature. We show in Section 6 how higher-order schemes can be constructed. These methods are not necessarily invariant domain preserving. In passing we revisit an idea initially proposed by Jameson et al. [2, Eq. (12)] which consists of constructing a second-order graph viscosity by using a smoothness indicator. In Theorem 6.5 we prove that a high-order scheme based on the smoothness indicator of a conserved scalar component of the system does indeed preserve the bounds (for that component) that are naturally satisfied by the first-order method. In Theorem 6.8 we present another invariant domain preserving result for one scalar component of the conserved variables, but in this case the graph viscosity is computed by using a gap estimate (see Lemma 6.4) instead of a smoothness indicator. To the best of our knowledge, it seems that both results are original in the context of hyperbolic systems. The convex limiting technique is presented in Section 7, the key results of this section are Lemmas 7.15, 7.20 and Theorem 7.21. All these results are recapitulated into Theorem 7.24, which in some sense summarizes the content of the present paper. The idea of using the auxiliary states (3.8) and convex limiting has originally been proposed in Guermond et al. [1] for the Euler equations. The proposed generalization to general hyperbolic systems with source term for generic discretizations seems to be new.

Computations illustrating the performance of the abstract results stated in the paper can be found in Guermond and Popov [3], Guermond et al. [1] for the compressible Euler equations, and in Azerad et al. [4], Guermond et al. [5] for the shallow water equations.

2. Preliminaries

We recall in this section key properties about the system (1.1) that will be used repeatedly in the paper. The reader who is familiar with hyperbolic systems with source terms, Riemann problems, and invariant sets is invited to jump to Section 3.

2.1. Riemann problem space average and maximum wave speed

We consider (1.1) without source term in this subsection, i.e., $\boldsymbol{S}(\boldsymbol{u}) = 0$. Instead of trying to give a precise meaning to the solutions of (1.1), which is either a very technical task or a completely open problem, we instead assume that there is a clear notion of solution for the Riemann problem. That is to say we assume that there exists a nonempty

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

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

has a unique (entropy satisfying) self-similar solution denoted by $\boldsymbol{v}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R, \xi)$, where $\xi = \frac{x}{t}$ is the self-similarity parameter, see for instance Lax [6], Toro [7]. The key result that we are going to use in this paper is that there exists a maximum wave speed henceforth denoted $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$ such that $\boldsymbol{v}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R, \xi) = \boldsymbol{v}_L$ if $\xi \le -\lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$ and $\boldsymbol{v}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R, \xi) = \boldsymbol{v}_R$ if $\xi \ge \lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$. We assume that $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$ can be estimated from above efficiently; for instance, we refer the reader to Guermond and Popov [8] where guaranteed upper bounds on the maximum wave speed are given for the Euler equations with the co-volume equation of state. The following elementary result, which we are going to invoke repeatedly, is an important consequence of the finite speed of propagation assumption:

Lemma 2.1 (*Average Over the Riemann Fan*)**.** *Let $(\eta, \boldsymbol{q})$ be an entropy pair for the system (1.1). Let $\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{v}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R, \xi)\,\mathrm{d}x$ be the average of the Riemann solution over the Riemann fan at time t. Assume that $t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) \le \frac{1}{2}$, then the following holds true:*

$$\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) = \frac{1}{2}(\boldsymbol{v}_L + \boldsymbol{v}_R) - t\big(\mathbb{f}(\boldsymbol{v}_R)\boldsymbol{n} - \mathbb{f}(\boldsymbol{v}_L)\boldsymbol{n}\big). \tag{2.2}$$

$$\eta(\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)) \le \tfrac{1}{2}(\eta(\boldsymbol{v}_L) + \eta(\boldsymbol{v}_R)) - t(\boldsymbol{q}(\boldsymbol{v}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{v}_L)\cdot\boldsymbol{n}). \tag{2.3}$$

2.2. Invariant sets and invariant domains

We introduce in this section the notions of invariant sets and invariant domains. Our definitions are slightly different from those in Chueh et al. [9], Hoff [10], Smoller [11], Frid [12]. We associate invariant sets with solutions of Riemann problems and define invariant domains only for an approximation process; our definition has some similarities with Eq. (2.14) in Zhang and Shu [13].

Definition 2.2 (*Invariant Set*)**.** We say that a set $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ is invariant for (1.1) if $\mathcal{B}$ is convex and 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 any $t > 0$ such that $t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) \le \frac{1}{2}$, the average of the entropy solution of the Riemann problem (2.1) over the Riemann fan, say $\overline{\boldsymbol{v}}(t, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$, remains in $\mathcal{B}$, and if there exists $\tau_0 > 0$ such that for any $\mathbf{U} \in \mathcal{B}$ and any $\tau \le \tau_0$ the quantity $\mathbf{U} + \tau \boldsymbol{S}(\mathbf{U})$ is in $\mathcal{B}$.

We now introduce the notion of invariant domain for an approximation process. Let I be a positive natural number and let $\boldsymbol{R}_h : (\mathbb{R}^m)^I \to (\mathbb{R}^m)^I$ be a mapping over $(\mathbb{R}^m)^I$. Henceforth we abuse the language by saying that a member of $(\mathbb{R}^m)^I$, say $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_I)$, is in the set $\mathcal{B} \subset \mathbb{R}^m$ to actually mean that $\mathbf{U}_i \in \mathcal{B}$ for all $i \in \{1{:}I\}$.

Definition 2.3 (*Invariant Domain*)**.** A convex invariant set $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ is said to be an invariant domain for the mapping $\boldsymbol{R}_h : (\mathbb{R}^m)^I \to (\mathbb{R}^m)^I$ if and only if for any state $\mathbf{U}$ in $\mathcal{B}$, the state $\boldsymbol{R}_h(\mathbf{U})$ is also in $\mathcal{B}$.

For scalar conservation equations the notions of invariant sets and invariant domains are closely related to the notion of maximum principle. In the case of nonlinear hyperbolic systems, the maximum principle property does not apply and must be replaced by the notion of an invariant domain. To the best of our knowledge, the definition of invariant sets for the Riemann problem was introduced in Nishida [14], and the general theory of positively invariant regions was developed in Chueh et al. [9]. The analysis and development of numerical methods preserving invariant regions was considered in Hoff [10,15], Frid [12]. The objective of this paper is to generalize the invariant domain preserving method originally developed in Guermond and Popov [3] and the (invariant domain preserving) convex limiting technique introduced in Guermond et al. [1].

Remark 2.4 (*Siff Source Terms*)**.** The assumption that there exists a uniform τ_0 so that $\mathcal{B} + \tau \boldsymbol{S}(\mathcal{B}) \subset \mathcal{B}$ for all $\tau \in [0, \tau_0]$ is not reasonable for hyperbolic systems with stiff source terms since it imposes a very severe restriction on the time step. In this case other strategies must be adopted. We are going to restrict ourselves in the present paper to source terms that are moderately stiff in the sense of Definition 2.2, and we postpone the extension of the present work to systems with stiff source terms to a future publication. □

2.3. Examples

We briefly go over some examples of systems with source terms and show that the proposed definition for invariant sets is meaningful/useful.

2.3.1. Euler + co-volume EOS

For the compressible Euler equations with covolume of state the dependent variable is $\boldsymbol{u} = (\rho, \boldsymbol{m}, E)^\intercal$, where ρ is the density, $\boldsymbol{m}$ is the momentum, and E is the total energy. The flux is $\mathbb{f}(\boldsymbol{u}) = (\rho\boldsymbol{v}, \boldsymbol{m}\otimes\boldsymbol{v} + p\mathbb{I}, \boldsymbol{v}(E+p))^\intercal$ where $\boldsymbol{v} := \boldsymbol{m}/\rho$ and the pressure is given by the equation of state $p(1-b\rho) = (\gamma-1)e\rho$. The constant $b \geq 0$ is called the covolume and $\gamma > 1$ is the ratio of specific heats. We have $\mathcal{A} := \{\boldsymbol{u} \mid 1 \geq 1 - b\rho \geq 0, e(\boldsymbol{u}) \geq 0\}$ and it is shown in Guermond and Popov [8] that $\mathcal{B} := \{\boldsymbol{u} \mid 1 \geq 1 - b\rho \geq 0, e(\boldsymbol{u}) \geq 0, \Phi(\boldsymbol{u}) \geq \Phi_0\}$ is an invariant set for any $\Phi_0 \in \mathbb{R}$, where $e(\boldsymbol{u}) := E/\rho - \frac{1}{2}\boldsymbol{v}^2$ is the specific internal energy, and $\Phi(\boldsymbol{u})$ is the specific physical entropy. In this paper we call internal energy the quantity $\varepsilon(\boldsymbol{u}) := \rho e(\boldsymbol{u})$.

2.3.2. Shallow water

Saint–Venant's shallow water model describes the time and space evolution of a body of water evolving in time under the action of gravity assuming that the deformations of the free surface are small compared to the water elevation and the bottom topography z varies slowly. The dependent variable is $\boldsymbol{u} = (h, \boldsymbol{q})^\intercal$, where h is the water height and $\boldsymbol{q}$ is the flow rate in the direction parallel to the bottom. The flux is $\mathbb{f}(\boldsymbol{u}) = (\boldsymbol{q}, \boldsymbol{q}\otimes\boldsymbol{v} + \frac{1}{2}gh^2\mathbb{I})^\intercal$, where $\boldsymbol{v} := \boldsymbol{q}/h$ and g is the gravity constant. The source including the influence of the topography and Manning's friction law is $\boldsymbol{S}(\boldsymbol{u}) = (0, gh\nabla z - gn^2h^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2})$, where n is Manning's roughness coefficient, and γ is an experimental parameter often close to $\frac{4}{3}$.

It is well-known that $\mathcal{A} = \mathcal{B} := \{\boldsymbol{u} \mid h \geq 0\}$ is an invariant set for the system without source term. Let $\boldsymbol{u} \in \mathcal{B}$ and $\tau > 0$, then $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u}) = (h, \boldsymbol{q} + \tau(gh\nabla z - gn^2h^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2}))^\intercal$, and it is clear that $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u}) \in \mathcal{B}$ for any $\tau \geq 0$ because $h \geq 0$ by definition. Hence $\mathcal{B}$ is an invariant set according to Definition 2.2 with $\tau_0 = \infty$.

2.3.3. ZND model

We now consider the Zel'dovich–von Neumann–Döring model for compressible reacting flows. The dependent variable is $\boldsymbol{u} = (\rho_1, \rho_2, \boldsymbol{m}, E)^\intercal$, where ρ_1 is the density of the burned gas (fuel), ρ_2 is the density of the unburned gas, $\boldsymbol{m}$ is the momentum of the mixture, and E is the total energy. The flux is $\mathbb{f}(\boldsymbol{u}) = (\rho_1\boldsymbol{v}, \rho_2\boldsymbol{v}, \boldsymbol{m}\otimes\boldsymbol{v} + p\mathbb{I}, \boldsymbol{v}(E+p))^\intercal$ where $\boldsymbol{v} := \boldsymbol{m}/(\rho_1+\rho_2)$ and the pressure is given by an appropriate equation of state; for instance, for ideal polytropic gases it is common to adopt the so called γ-law, $p = (\gamma-1)(E - \frac{1}{2}\rho\boldsymbol{v}^2 - q_0\rho_2)$, where q_0 is the specific energy of the unburned gas. Denoting by $T := p/(\rho_1+\rho_2)$, the source term is $\boldsymbol{S}(\boldsymbol{u}) = (\kappa(T)\rho_2, -\kappa(T)\rho_2, \boldsymbol{0}, 0)^\intercal$, where $\kappa(T) = \kappa_0 \mathrm{e}^{-T_0/T}$, where $\kappa_0 \geq 0$ is the reaction rate constant and T_0 is the ignition temperature (up to multiplication by the gas constant R).

Denoting $\rho := \rho_1 + \rho_2$ and setting $e(\boldsymbol{u}) := (E - \frac{1}{2}\rho\boldsymbol{v}^2 - q_0\rho_2)/\rho$, it can be shown that $\mathcal{A} = \mathcal{B} := \{\boldsymbol{u} \mid \rho_1 \geq 0, \rho_2 \geq 0, e(\boldsymbol{u}) \geq 0\}$ is an invariant set for the homogeneous system, i.e., when $\boldsymbol{S} \equiv \boldsymbol{0}$. One can convince oneself that this is indeed true by realizing that when $\boldsymbol{S} \equiv \boldsymbol{0}$, upon denoting $E' := E - q_0\rho_2$, the dependent variable $(\rho, \boldsymbol{m}, E')$ solves the compressible Euler equations, and it is well-known that $\{\boldsymbol{u} \mid \rho \geq 0, E' - \frac{1}{2}\rho\boldsymbol{v}^2 \geq 0\}$ is an invariant set.

Now let us establish that for any $\boldsymbol{u} \in \mathcal{B}$ and any $\tau \leq \tau_0 := \kappa_0^{-1}$, the quantity $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u})$ is in $\mathcal{B}$. Let $\boldsymbol{u} \in \mathcal{B}$ and let $\tau \geq 0$, then $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u}) = (\rho_1 + \tau\kappa(T)\rho_2, \rho_2 - \tau\kappa(T)\rho_2, \boldsymbol{m}, E)^\intercal$. Since $T := (\gamma-1)e(\boldsymbol{u}) \geq 0$, $\rho_1 \geq 0$, $\rho_2 \geq 0$, and $\tau \geq 0$, it is clear that $\rho_1 + \tau\kappa(T)\rho_2 \geq 0$. Moreover, $\rho_2 - \tau\kappa(T)\rho_2 = \rho_2(1 - \tau\kappa(T)) \geq \rho_2(1 - \tau\kappa_0)$; hence $\rho_2 - \tau\kappa(T)\rho_2 \geq 0$ provided $\tau \leq \tau_0 := \kappa_0^{-1}$. Finally, observing that $\rho := \rho_1 + \tau\kappa(T)\rho_2 + \rho_2 - \tau\kappa(T)\rho_2 > 0$, we have $\rho e(\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u})) = E - \frac{1}{2}\rho\boldsymbol{v}^2 - q_0\rho_2(1 - \tau\kappa(T)) \geq E - \frac{1}{2}\rho\boldsymbol{v}^2 - q_0\rho_2 = e(\boldsymbol{u}) \geq 0$, thereby proving that $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u}) \in \mathcal{B}$.

2.3.4. Euler equations with sources

In some astrophysical applications one may want to solve the compressible Euler equations with Coriolis effects, gravitation effects and some heat transfer effects due to the emission and/or absorption of radiation. The dependent variables and the flux are the same as those of Euler's equations, but the source term is $(0, -2\boldsymbol{\Omega}\times\boldsymbol{m} - \rho\nabla\Phi, -\boldsymbol{m}\cdot\nabla\Phi + \rho H)^\intercal$, where $\boldsymbol{\Omega}$ is the angular velocity of the system, Φ some given gravitation potential, and ρH is a term that aggregates all the cooling and heating effects. One invariant domain for the homogeneous system is $\mathcal{A} = \mathcal{B} := \{\boldsymbol{u} \mid \rho \geq 0, e(\boldsymbol{u}) \geq 0\}$. Let $\boldsymbol{u} \in \mathcal{B}$ and $\tau \geq 0$. Then $\boldsymbol{u} + \tau\boldsymbol{S}(\boldsymbol{u}) = (\rho, \boldsymbol{m} - 2\tau\boldsymbol{\Omega}\times\boldsymbol{m} - \tau\rho\nabla\Phi, E - \tau\boldsymbol{m}\cdot\nabla\Phi + \tau\rho H)^\intercal$.

The density of the state $\boldsymbol{u}+\tau\boldsymbol{S}(\boldsymbol{u})$ is ρ, which is nonnegative by definition. The specific internal energy of the state $\boldsymbol{u}+\tau\boldsymbol{S}(\boldsymbol{u})$ is bounded from below as follows: $e(\boldsymbol{u}+\tau\boldsymbol{S}(\boldsymbol{u})) \geq e(\boldsymbol{u}) - \tau^2(4\boldsymbol{\Omega}^2\boldsymbol{v}^2 + (\nabla\Phi)^2) + \tau H$. For instance, for a γ-law equation of state, we have $e(\boldsymbol{u}) = c(\boldsymbol{u})^2/(\gamma(\gamma-1))$, where $c(\boldsymbol{u})$ is the speed of sound, and $e(\boldsymbol{u}+\tau\boldsymbol{S}(\boldsymbol{u})) \geq 0$, if $\tau^2 \leq \frac{c(\boldsymbol{u})^2}{2\gamma(\gamma-1)(4\boldsymbol{\Omega}^2\boldsymbol{v}^2+(\nabla\Phi)^2)}$ and $\tau H \geq -\frac{e(\boldsymbol{u})}{2}$. If $\nabla\Phi=\boldsymbol{g}$ is a constant, then the first condition is satisfied if $\tau^2 \leq \frac{c(\boldsymbol{u})^2}{2\gamma(\gamma-1)(4\boldsymbol{\Omega}^2M(\boldsymbol{u})^2c(\boldsymbol{u})^2+\|\boldsymbol{g}\|_{\ell^2}^2)}$ where $M(\boldsymbol{u})$ is the local Mach number; assuming that one can establish that $M(\boldsymbol{u}) \leq M_{\max}$ uniformly w.r.t. $\boldsymbol{u}$, and $\inf c(\boldsymbol{u}) \geq c_{\min} > 0$, which is required for hyperbolicity to hold, then the first condition holds if $\tau \leq \frac{c_{\min}}{(2\gamma(\gamma-1)(4\boldsymbol{\Omega}^2M_{\max}^2c_{\min}^2+\|\boldsymbol{g}\|_{\ell^2}^2))^{\frac{1}{2}}}$. One can also verify that many astrophysical models for the heat transfer effect lead to existence of $\tau_0' > 0$ such that $\tau H \geq -\frac{e(\boldsymbol{u})}{2}$ for all $\tau \leq \tau_0'$; the details are left to the reader.

3. Abstract low-order approximation

In this section we describe a generic invariant domain preserving technique for approximating solutions to (1.1). In order to stay general we present the method without referring to any particular discretization technique, we are going to use instead the graph theoretic language to describe the method. The method is illustrated with finite volumes, continuous elements, and discontinuous elements in Section 4.

3.1. The low-order scheme

To identify properly the time stepping technique, 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 now address the approximation in space by assuming that we have at hand some finite-dimensional vector space X_h with some basis $\{\varphi\}_{i\in\mathcal{V}}$, where $\varphi_i^n : D \to \mathbb{R}$, for all $i \in \mathcal{V}$. We introduce $\boldsymbol{X}_h^n := (X_h)^m$ and denote the approximation of $\boldsymbol{u}(\cdot,t^n)$ in $\boldsymbol{X}_h$ by $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{V}} \mathbf{U}_i^n\varphi_i$, with $\mathbf{U}_i^n \in \mathcal{A} \subset \mathbb{R}^m$ for all $i \in \mathbb{R}^m$. We do not need to know for the time being what the basis functions $\{\varphi_i\}_{i\in\mathcal{V}}$ are, but we assume that this setting allows us to construct an inviscid (very accurate) approximation of $\boldsymbol{u}(\cdot,t^{n+1})$ in $\boldsymbol{X}_h$, denoted $\boldsymbol{u}_h^{\mathrm{G},n+1} := \sum_{i\in\mathcal{V}} \mathbf{U}_i^{\mathrm{G},n+1}\varphi_i$, as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{G},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} = m_i\boldsymbol{S}(\mathbf{U}_i^n), \tag{3.1}$$

for any $i \in \mathcal{V}$, where the numbers $\{m_i\}_{i\in\mathcal{V}}$ are assumed to be positive. Note here that we use the forward Euler time stepping. Higher-order time stepping schemes will be considered in Section 5. For any $i \in \mathcal{V}$, the set $\mathcal{I}(i)$ is a (small) subset of $\mathcal{V}$, which we call stencil at i or adjacency list at i. We assume that the following property holds: $j \in \mathcal{I}(i)$ iff $i \in \mathcal{I}(j)$. We assume also that the $\mathbb{R}^d$-valued matrix $\{\boldsymbol{c}_{ij}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ has the following properties:

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji} \quad \text{and} \quad \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \mathbf{0}. \tag{3.2}$$

The quantities m_i, $\{\boldsymbol{c}_{ij}\}_{j\in\mathcal{I}(i)}$, and the set $\mathcal{I}(i)$ depend on the discretization that is chosen. We are going to be more specific in Section 4. We think of (3.1) as the "centered" consistent approximation of (1.1) that delivers optimal accuracy (for the considered setting) for smooth solutions.

Notice that the above construction allows us to introduce an undirected finite graph $(\mathcal{V},\mathcal{E})$, where for any pair $(i,j) \in \mathcal{V}\times\mathcal{V}$, we say that (i,j) is an edge of the graph, i.e., $(i,j) \in \mathcal{E}$, iff $i \in \mathcal{I}(j)$ and $j \in \mathcal{I}(i)$. We say that $(\mathcal{V},\mathcal{E})$ is the connectivity graph of the approximation.

Since (3.1) is "centered", it cannot handle properly shocks and discontinuous data. To address this issue we introduce some artificial dissipation. We do so by using the graph Laplacian associated with the connectivity graph $(\mathcal{V},\mathcal{E})$. We assume that the graph viscosity $\{d_{ij}^{\mathrm{L},n}\}_{(i,j)\in\mathcal{E}}$ is scalar and has the following properties:

$$d_{ij}^{\mathrm{L},n} = d_{ji}^{\mathrm{L},n} > 0, \quad \text{if} \quad i \neq j. \tag{3.3}$$

Although the diagonal value $d_{ii}^{\mathrm{L},n}$ is not needed, we adopt the convention $d_{ii}^{\mathrm{L},n} := -\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n}$. This convention will help us shorten some expressions later. We are now in position to define the first-order method on which the rest of the paper is built. We call low-order update $\mathbf{U}_i^{\mathrm{L},n+1}$ the quantity computed as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = m_i\boldsymbol{S}(\mathbf{U}_i^n), \tag{3.4}$$

for all $i \in \mathcal{V}$. Without further assumptions, the scheme has built-in conservation properties; more specifically, the following holds true.

Lemma 3.1 (*Conservation*)**.** *Assume that $\boldsymbol{S} \equiv \boldsymbol{0}$, then the scheme (3.2)–(3.4) is conservative in the sense that the following identity holds for any $n \in \mathbb{N}$:*

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^n. \tag{3.5}$$

Proof. Using that $\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \boldsymbol{0}$, we rewrite (3.4) in the form

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

Defining $\mathbf{F}_{ij}^{\mathrm{L},n} := (\mathbb{f}(\mathbf{U}_j^n)+\mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)$, the above identity implies that $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{L,n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^n + \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L},n}$. The assertion is a consequence of the skew-symmetry of $\boldsymbol{c}_{ij}$ and the symmetry of $d_{ij}^{\mathrm{L},n}$, i.e., $\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L},n} = \boldsymbol{0}$. □

Remark 3.2 (*Consistency*)**.** Although the consistency question will be addressed later, let us say at this point that consistency is not an immediate consequence of (3.2) and (3.3). Consistency will be achieved provided one can show that $\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n)$ is an approximation of $\partial_t \boldsymbol{u}$ (i.e., a moment with a shape function), $\sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}$ is an approximation of $\nabla\cdot\mathbb{f}(\boldsymbol{u})$ (i.e., a moment with a shape function), and $m_i \boldsymbol{S}(\mathbf{U}_i^n)$ is an approximation of $\boldsymbol{S}(\boldsymbol{u})$ (i.e., a moment with a shape function). Note that if all the values $\{\mathbf{U}_j\}_{j\in\mathcal{I}(i)}$ are constant, the graph viscosity term $\sum_{j\in\mathcal{I}(i)} d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)$ vanishes; which in some sense implies that (3.4) is a first-order consistent perturbation of (3.1). The scalars m_i and the vectors $\{\boldsymbol{c}_{ij}\}_{j\in\mathcal{I}(i)}$ are not uniquely defined and they may take different forms depending on the method of choice. In sections Sections 4.1–4.3 we will describe three methods based on finite volumes, continuous finite elements, and discontinuous finite elements, all of which can be written in the form (3.2)–(3.4). □

Remark 3.3 (*Algebraic-Fluxes*)**.** For further reference it will be useful to define the following quantity which we henceforth refer to as low-order algebraic flux:

$$\mathbf{F}_{ij}^{\mathrm{L},n} := (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n). \tag{3.6}$$

Algebraic fluxes will be instrumental for the development of limiting techniques in Section 7.3. In particular, the scheme (3.4) is conveniently rewritten as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L},n} = m_i \boldsymbol{S}(\mathbf{U}_i^n). \quad \square \tag{3.7}$$

Remark 3.4 (*Well-Balancing*)**.** In general, systems with a source term have time-independent solutions, i.e., fields solving $\nabla\cdot\mathbb{f}(\boldsymbol{u}) = \boldsymbol{S}(\boldsymbol{u})$, and it is often a desirable feature of numerical schemes that they preserve these steady states. This lead to the notion of well-balancing introduced in Bermudez and Vazquez [16], Greenberg and Leroux [17]; we also refer to Huang and Liu [18, §3] for early ideas on well-balancing. Although, well-balancing is a very important notion, it will not be addressed in this paper. □

3.2. Invariant domain preserving graph viscosity

Now we propose a definition of the graph viscosity that makes the algorithm (3.4) invariant domain preserving. Recall that the discretization setting is still unspecified. Most of the arguments presented in this subsection are generalizations of those in §3.2, §4.1 and §4.2 of Guermond and Popov [3].

Since $\sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_i^n)\boldsymbol{c}_{ij} = \boldsymbol{0}$ (see property (3.2)) we can rewrite the scheme (3.4) as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} 2d_{ij}^{\mathrm{L},n}\mathbf{U}_i^n + (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n + \mathbf{U}_i^n) = m_i \boldsymbol{S}(\mathbf{U}_i^n).$$

Then, upon introducing the auxiliary states (recalling that $d_{ij}^{\mathrm{L},n} > 0$ by assumption),

$$\overline{\mathbf{U}}_{ij}^n := \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}, \tag{3.8}$$

with the convention $\overline{\mathbf{U}}_{ii}^n := \mathbf{U}_i^n$, the low-order scheme (3.4) can be rewritten as follows:

$$\mathbf{U}_i^{\mathrm{L},n+1} = \Big(1 - \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathbf{U}}_{ij}^n + \tau \boldsymbol{S}(\mathbf{U}_i^n). \tag{3.9}$$

A first key observation we make at this point about (3.9) is that upon setting $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$, we realize that $\overline{\mathbf{U}}_{ij}^n$ is exactly of the form $\overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ as defined in (2.2) with the fake time $t_{ij} = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^{\mathrm{L},n}$. Then Lemma 2.1 motivates the following definition for the graph viscosity coefficients $d_{ij}^{\mathrm{L},n}$:

$$d_{ij}^{\mathrm{L},n} := \max(\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{3.10}$$

where recall that $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is the maximum wave speed defined in Section 2.1.

Lemma 3.5 (*Invariance of The Auxiliary States*)**.** *Let $\mathcal{B} \subset \mathcal{A}$ be a convex invariant set for (1.1) such that $\mathbf{U}_i^n, \mathbf{U}_j^n \in \mathcal{B}$. The state $\overline{\mathbf{U}}_{ij}^n$ defined in (3.8), with $d_{ij}^{\mathrm{L},n}$ as defined in (3.10), belongs to $\mathcal{B}$.*

Proof. Let us set $t_{ij} := \|\boldsymbol{c}_{ij}\|_{\ell^2}/(2d_{ij}^{\mathrm{L},n})$, then according to Lemma 2.1, we have $\overline{\mathbf{U}}_{ij}^n := \overline{\boldsymbol{u}}(t_{ij}, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \in \mathcal{B}$ if $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)t_{ij} \le \frac{1}{2}$. But the definition (3.10) implies that $d_{ij}^{\mathrm{L},n} \ge \lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}$, which is the CFL condition $t_{ij}\lambda_{\max}(\boldsymbol{n}_{ij}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \frac{1}{2}$ for the conclusions of Lemma 2.1 to hold. This proves that $\overline{\mathbf{U}}_{ij}^n := \overline{\boldsymbol{u}}(t, \boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \in \mathcal{B}$ for all $j \in \mathcal{I}(i)$ since $\mathcal{B}$ is a convex invariant set. □

A second important observation about (3.9) is that $\mathbf{U}_i^{\mathrm{L},n+1} - \tau \boldsymbol{S}(\mathbf{U}_i^n)$ is a convex combination of $\mathbf{U}_i^n$ and the states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ provided τ is small enough. This is the key to the following result.

Theorem 3.6 (*Local Invariance*)**.** *Let $n \ge 0$ and let $i \in \mathcal{V}$. Assume that τ is small enough so that $1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i} \ge 0$ and $2\tau \le \tau_0$. Let $\mathcal{B} \subset \mathcal{A}$ be a convex invariant set for (1.1) such that $\mathbf{U}_j^n \in \mathcal{B}$ for all $j \in \mathcal{I}(i)$, then $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{B}$.*

Proof. Using the definition $d_{ii}^{\mathrm{L},n} := \sum_{j\in\mathcal{I}(i)\backslash\{i\}} -d_{ij}^{\mathrm{L},n}$, we first notice that (3.9) can be rewritten as follows:

$$\mathbf{U}_i^{\mathrm{L},n+1} = \frac{1}{2}\Big(\Big(1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \frac{4\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathbf{U}}_{ij}^n\Big) + \frac{1}{2}\big(\mathbf{U}_i^n + 2\tau\boldsymbol{S}(\mathbf{U}_i^n)\big). \tag{3.11}$$

With obvious notation, let us rewrite the above equation as follows $\mathbf{U}_i^{\mathrm{L},n+1} = \frac{1}{2}\mathbf{W}_1 + \frac{1}{2}\mathbf{W}_2$. Owing to the local CFL assumption $1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i} \ge 0$, $\mathbf{W}_1$ is a convex combination of $\mathbf{U}_i^n$ and the collection of states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)\backslash\{i\}}$. But we have established in Lemma 3.5 that $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}$. Then, the convexity of $\mathcal{B}$ implies $\mathbf{W}_1$ is in $\mathcal{B}$. Since $\mathcal{B}$ is an invariant set according to Definition 2.2 and $\mathbf{U}_i^n \in \mathcal{B}$ by assumption, the condition $2\tau \le \tau_0$ implies that $\mathbf{W}_2 := \mathbf{U}_i^n + 2\tau\boldsymbol{S}(\mathbf{U}_i^n)$ is a member of $\mathcal{B}$. In conclusion, the convexity of $\mathcal{B}$ implies that $\mathbf{U}_i^{\mathrm{L},n+1} = \frac{1}{2}\mathbf{W}_1 + \frac{1}{2}\mathbf{W}_2$ is in $\mathcal{B}$. □

Corollary 3.7 (*Global Invariance*)**.** *Let $n \in \mathbb{N}$. Assume that the global CFL condition $\min_{i\in\mathcal{V}}\big(1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i}\big) \ge 0$ holds and $2\tau \le \tau_0$. Let $\mathcal{B} \subset \mathcal{A}$ be a convex invariant set. Assume that $\mathbf{U}_i^n \in \mathcal{B}$ for all $i \in \mathcal{V}$, then $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{B}$ for all $i \in \mathcal{V}$.*

Theorem 3.8 (*Entropy Inequality*)**.** *Let $n \ge 0$ and $i \in \mathcal{V}$. Assume also that the local CFL condition holds $1 + 2\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i} \ge 0$ and $2\tau \le \tau_0$, then the following local entropy inequality holds true for any entropy pair $(\eta, \boldsymbol{q})$ of the system (1.1):*

$$\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{\mathrm{L},n+1})-\eta(\mathbf{U}_i^n))+\sum_{j\in\mathcal{I}(i)}\boldsymbol{q}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}-d_{ij}^{\mathrm{L},n}(\eta(\mathbf{U}_j^n)-\eta(\mathbf{U}_i^n)) \tag{3.12}$$
$$\le m_i\boldsymbol{S}(\mathbf{U}_i^n)\cdot\nabla\eta(\mathbf{U}_i^{\mathrm{L},n+1}).$$

Proof. Let $i\in\mathcal{V}$ and let $(\eta,\boldsymbol{q})$ be an entropy pair for the system (1.1). Then recalling (3.9), the CFL condition and the convexity of η imply that

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

Lemma 2.1 implies that $\eta(\overline{\mathbf{U}}_{ij}^n)\le\frac{1}{2}(\eta(\mathbf{U}_i^n)+\eta(\mathbf{U}_j^n))-t_{ij}(\boldsymbol{q}(\mathbf{U}_j^n)\cdot\boldsymbol{n}_{ij}-\boldsymbol{q}(\mathbf{U}_i^n)\cdot\boldsymbol{n}_{ij})$, with $t_{ij}=\|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^{\mathrm{L},n}$; hence,

$$\frac{m_i}{\tau}(\eta(\mathbf{U}_i^{\mathrm{L},n+1}-\tau\boldsymbol{S}(\mathbf{U}_i^n))-\eta(\mathbf{U}_i^n))\le\sum_{j\in\mathcal{I}(i)\backslash\{i\}}2d_{ij}^{\mathrm{L},n}(\eta(\overline{\mathbf{U}}_{ij}^n)-\eta(\mathbf{U}_i^n))$$
$$\le\sum_{j\in\mathcal{I}(i)\backslash\{i\}}d_{ij}^{\mathrm{L},n}(\eta(\mathbf{U}_j^n)-\eta(\mathbf{U}_i^n))-\|\boldsymbol{c}_{ij}\|_{\ell^2}(\boldsymbol{q}(\mathbf{U}_j^n)\cdot\boldsymbol{n}_{ij}-\boldsymbol{q}(\mathbf{U}_i^n)\cdot\boldsymbol{n}_{ij}).$$

Moreover, the convexity of η implies that

$$\eta(\mathbf{U}_i^{\mathrm{L},n+1})-\tau\boldsymbol{S}(\mathbf{U}_i^n)\cdot\nabla\eta(\mathbf{U}_i^{\mathrm{L},n+1})\le\eta(\mathbf{U}_i^{\mathrm{L},n+1}-\tau\boldsymbol{S}(\mathbf{U}_i^n)).$$

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

Remark 3.9 (*Terminology*)**.** In order to refer to the scheme (3.4) with (3.10), following [1] we will use the acronym GMS-GV, standing for Guaranteed Maximum Speed Graph Viscosity. □

Remark 3.10 (*Symmetry*)**.** Since $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{ji}$ we note that $\overline{\mathbf{U}}_{ij}^n=\overline{\mathbf{U}}_{ji}^n$ (see definition (3.8)) which in turn implies that $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)=\lambda_{\max}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)$. In conclusion $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}=\lambda_{\max}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}$. Note that these properties may not hold at the boundary if nontrivial boundary conditions are applied. □

Remark 3.11 (*Positivity*)**.** It may happen that estimating a guaranteed upper bound $\lambda_{\max}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$ on the maximum wave speed in the Riemann problem is difficult. In this case one has to come up with some informed guess. We now give a lower bound on $\lambda_{\max}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$ that guaranties positivity if it happens that some components of $\mathbf{U}$, say U, has to be positive (think of the density and the total energy in the Euler equations or the water height in the shallow water equations). Let $\boldsymbol{f}_{\mathsf{U}}:\mathcal{A}\to\mathbb{R}^d$ be the component of $\mathbb{f}$ that corresponds to the component U of $\mathbf{U}$. Assume that $\mathcal{B}:=\{\mathbf{U}\in\mathcal{A}\mid\mathsf{U}>0\}$ is an invariant set for (2.1), assume also that the estimate on the maximum wave speed is such that the $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\ge\max\big(\frac{f_{\mathsf{U}}(\mathbf{U}_j^n)\cdot\boldsymbol{n}_{ij}}{\mathsf{U}_j^n},0\big)$, then under the same CFL condition as in Theorem 3.6 and Corollary 3.7, the set $\widetilde{\mathcal{B}}:=\{\mathbf{U}\in\mathbb{R}^m\mid\mathsf{U}>0\}\subseteq\mathcal{B}$ is such that $\big(\mathbf{U}_i^n\in\mathcal{B},\ \forall i\in\mathcal{V}\big)\Rightarrow\big(\mathbf{U}_i^{\mathrm{L},n+1}\in\widetilde{\mathcal{B}},\ \forall i\in\mathcal{V}\big)$. Let us finally illustrate the above result in one space dimension. For instance, for finite volumes and for piecewise linear continuous elements in one space dimension, one has $\boldsymbol{c}_{ij}=\frac{1}{2}\boldsymbol{n}_{ij}$ (see Section 4). Then, for the density in the Euler equations, or for the water height in the Saint–Venant equations, the above estimate becomes $\lambda_{\max}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\ge\max\big(\frac{1}{2}\boldsymbol{n}_{ij}\cdot\mathbf{V}(\mathbf{U}_j^n),0\big)$ where $\mathbf{V}(\mathbf{U})$ is the velocity. One recognizes here the standard upwind estimate. □

4. Examples of discretizations

In this section we illustrate the GMS-GV scheme described in Section 3 in the following three space discretization settings: finite volumes, continuous finite elements, and discontinuous elements.

4.1. Finite volumes

We now illustrate the construction of the abstract low-order scheme (3.2)–(3.4) in the context of finite volumes (FV).

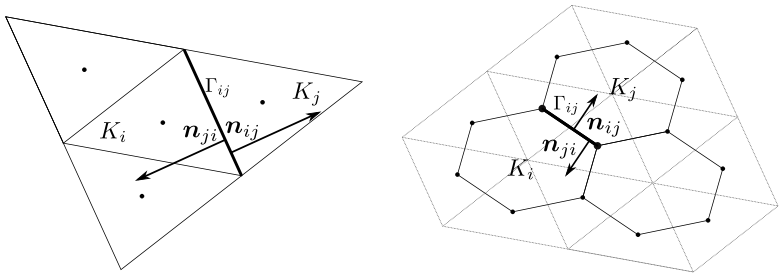

Fig. 1. Finite volume patch arising from a cell-centered discretization (left) and a vertex-centered discretization (right).

4.1.1. Technical preliminaries

We unify our presentation by putting into a single framework the so-called cell-centered and vertex-centered finite volume techniques, see Fig. 1. We refer the reader to Barth and Ohlberger [19], Eymard et al. [20] for comprehensive reviews on the finite volume techniques. For any manifold $E \subset \mathbb{R}^d$ of dimension l we denote by $|E|$ the l-Lebesgue measure of E. We assume that we have at hand a partition of the computational domain D into polygonal (polyhedral) cells $\{K_i\}_{i\in\mathcal{V}}$. We henceforth denote by $\mathcal{T}_h$ this collection of cells. For any pair of cells K_i, K_j having a common interface, we denote by $\Gamma_{ij} := \partial K_i \cap \partial K_j$ the interface in question. The unit vector on Γ_{ij} pointing from K_i to K_j is denoted $\boldsymbol{n}_{ij}$.

4.1.2. Definitions of $(\mathcal{V},\mathcal{E})$, m_i, and $\boldsymbol{c}_{ij}$

We define the connectivity graph $(\mathcal{V},\mathcal{E})$ by identifying the vertices of this graph with the cells in $\mathcal{T}_h$, and we say that a pair of cells K_i, K_j form an edge of the graph, i.e., $(i,j) \in \mathcal{E}$, iff the cells K_i and K_j share an interface, i.e., $\partial K_i \cap \partial K_j$ is a $(d-1)$-manifold of positive measure. For any $i \in \mathcal{V}$ we define the adjacency list $\mathcal{I}(i)$ to be the list of all the cells in $\mathcal{T}_h$ sharing an interface with K_i, i.e., $\mathcal{I}(i) := \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\}$, see Fig. 1. Denoting by $\mathbb{I}_{K_j}$ the indicator function of the cell K_j, we set $X_h := \operatorname{span}\{\mathbb{I}_{K_j}\}_{j\in\mathcal{V}}$ and then define the approximation space $\boldsymbol{X}_h := (X_h)^m = \{\sum_{j\in\mathcal{V}} \mathbf{V}_j \mathbb{I}_{K_j} \mid \mathbf{V}_j \in \mathbb{R}^m, \forall j \in \mathcal{V}\}$.

Let $\boldsymbol{u}_h^n = \sum_{j\in\mathcal{V}} \mathbf{U}_j^n \mathbb{I}_{K_j} \in \boldsymbol{X}_h$ be the approximation of $\boldsymbol{u}$ at time t^n, then most first-order finite volume schemes are written as follows

$$\frac{|K_i|}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \mathbf{F}_{ij}^{\mathrm{L},n} = |K_i| S(\mathbf{U}_i^n),$$

where $\mathbf{F}_{ij}^{\mathrm{L},n}$ is usually the Lax–Friedrichs/Rusanov flux (integrated over Γ_{ij}):

$$\mathbf{F}_{ij}^{\mathrm{L},n} := \frac{|\Gamma_{ij}|}{2}(\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{n}_{ij} - \alpha_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n), \tag{4.1}$$

where $\alpha_{ij}^{\mathrm{L},n}$ is some wave speed. Hence, we recover the generic expression (3.4) for the finite volume framework by setting

$$m_i := |K_i|, \quad \boldsymbol{c}_{ij} := \frac{|\Gamma_{ij}|}{2}\boldsymbol{n}_{ij}, \ \forall j \in \mathcal{I}(i)\backslash\{i\}, \quad \boldsymbol{c}_{ii} := \mathbf{0}, \quad d_{ij}^{\mathrm{L},n} := \alpha_{ij}^{\mathrm{L},n}. \tag{4.2}$$

The definition of $\boldsymbol{c}_{ij}$ immediately implies that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ij}$, and the Stokes theorem implies that $\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \frac{1}{2}\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \boldsymbol{n}_{ij}|\Gamma_{ij}| = \frac{1}{2}\int_{\partial K_i} \boldsymbol{n}\, \mathrm{d}s = \mathbf{0}$, which is the conservation property stated in (3.2). Note that $\mathbf{F}_{ij}^{\mathrm{L},n} = -\mathbf{F}_{ji}^{\mathrm{L},n}$ since $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ji}$. Let us mention in passing that while any family of vectors of the form $\boldsymbol{c}_{ij} = \alpha\boldsymbol{n}_{ij}|\Gamma_{ij}|$ satisfies the conservation constraint (3.2), only the factor $\alpha = \frac{1}{2}$ leads to a consistent discretization of the divergence operator.

4.2. Continuous finite elements

We describe in this section one possible implementation of the abstract low-order scheme (3.2)–(3.4) in the context of continuous finite elements (cG). The set of the d-variate polynomials of degree at most $k \in \mathbb{N}$ is denoted $\mathbb{P}_{k,d}$. The reader who is familiar with [1,3,21] is invited to move to Section 4.3.

4.2.1. Technical preliminaries

Let $(\mathcal{T}_h)_{h>0}$ be a shape-regular sequence of unstructured matching meshes. To keep some level of generality we assume that the elements in the mesh are generated from a finite number of reference elements denoted $\widehat{K}_1, \ldots, \widehat{K}_\varpi$. For example, the mesh $\mathcal{T}_h$ could be composed of a combination of triangles and parallelograms in dimension two (we would have $\varpi = 2$ in this case); it could also be composed of a combination of tetrahedra, parallelepipeds, and triangular prisms in dimension three (we would have $\varpi = 3$ in this case). The diffeomorphism mapping $\widehat{K}_r$ to an arbitrary element $K \in \mathcal{T}_h$ is denoted $T_K : \widehat{K}_r \longrightarrow K$. We now introduce a set of reference finite elements $\{(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)\}_{1\le r\le \varpi}$ (the index $r \in \{1{:}\varpi\}$ will be omitted in the rest of the paper to simplify the notation), and we define the following scalar-valued and vector-valued continuous finite element spaces:

$$X_h = \{v \in \mathcal{C}^0(D;\mathbb{R}) \mid v_{|K}\circ T_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}, \qquad \boldsymbol{X}_h = [X_h]^m. \tag{4.3}$$

The global shape functions are denoted by $\{\varphi_i\}_{i\in\mathcal{V}}$ and we assume that they satisfy the partition of unity property $\sum_{i\in\mathcal{V}} \varphi_i(\boldsymbol{x}) = 1$, for all $\boldsymbol{x} \in D$.

4.2.2. Definitions of $(\mathcal{V},\mathcal{E})$, m_i, and $\boldsymbol{c}_{ij}$

We define the connectivity graph $(\mathcal{V},\mathcal{E})$ by identifying the shape functions $\{\varphi_i\}_{i\in\mathcal{V}}$ with the vertices of the graph. The edges are defined as follows: we say that two shape functions (or two degrees of freedom) form an edge, i.e., $(i,j) \in \mathcal{E}$, iff $\varphi_i\varphi_j \not\equiv 0$. For any $i \in \mathcal{V}$, the adjacency list $\mathcal{I}(i)$ is defined by setting $\mathcal{I}(i) := \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\}$.

Let $\mathcal{M}$ be the consistent mass matrix with entries $\int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $i, j \in \mathcal{V}$, and let $\mathcal{M}^L$ be the diagonal lumped mass matrix with entries

$$m_i := \int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{4.4}$$

The partition of unity property implies that $m_i = \sum_{j\in\mathcal{I}(i)} \int_D \varphi_j(\boldsymbol{x})\varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, i.e., the entries of $\mathcal{M}^L$ are obtained by summing the rows of $\mathcal{M}$. In the rest of the paper we assume that $m_i > 0$, for all $i \in \mathcal{V}$. This assumption is satisfied by many families of finite elements.

Let $\boldsymbol{u}_h^n = \sum_{j\in\mathcal{V}} \mathbf{U}_j^n\varphi_j \in \boldsymbol{X}_h$ be the approximation of $\boldsymbol{u}$ at time t^n, where $\boldsymbol{X}_h$ is the continuous finite element space defined in (4.3). We approximate $\mathbb{f}(\boldsymbol{u}_h^n)$ by $\sum_{j\in\mathcal{V}} \mathbb{f}(\mathbf{U}_j^n)\varphi_j$. If $\widehat{P}$ is composed of Lagrange elements, then $\sum_{j\in\mathcal{V}} \mathbb{f}(\mathbf{U}_j^n)\varphi_j$ is the Lagrange interpolation of $\mathbb{f}(\boldsymbol{u}_h^n)$, and in this case the approximation is fully consistent with the polynomial degree of $\widehat{P}$; otherwise, the approximation is formally at least second-order accurate in space since it is exact if $\mathbb{f}$ is linear. As a result, we have

$$\int_D \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i\,\mathrm{d}\boldsymbol{x} \approx \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\int_D \varphi_i\nabla\varphi_j\,\mathrm{d}\boldsymbol{x} = \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}, \tag{4.5}$$

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

$$\boldsymbol{c}_{ij} = \int_D \varphi_i\nabla\varphi_j\,\mathrm{d}\boldsymbol{x}, \quad \forall j \in \mathcal{I}(i). \tag{4.6}$$

Here we observe that the partition of unity property and definition (4.6) imply that $\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \sum_{j\in\mathcal{I}(i)} \int_D \varphi_i\nabla\varphi_j\,\mathrm{d}\boldsymbol{x} = \int_D \varphi_i\nabla\big(\sum_{j\in\mathcal{I}(i)}\varphi_j\big)\,\mathrm{d}\boldsymbol{x} = \mathbf{0}$. On the other hand, the skew-symmetry property $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ follows using integration by parts if D is the d-torus (which is the case for periodic boundary conditions) or if either φ_i or φ_j vanish at the boundary of D (which is the case when we solve the Cauchy problem).

4.3. Discontinuous finite elements

We finally describe in this section one possible implementation of the abstract low-order scheme (3.2)–(3.4) in the context of discontinuous finite elements (dG). This section builds on top of the definitions and notation already introduced in Section 4.2.1.

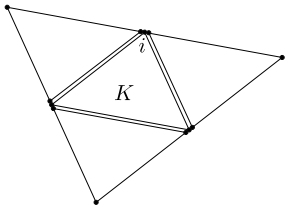

Fig. 2. Discontinuous $\mathbb{P}_{1,2}$ finite element patch (exploded view). Each black dot represents a scalar shape function. In this picture $i \in \mathcal{I}(K)$, $\mathrm{card}(\mathcal{I}(i)) = 7$, $\mathrm{card}(\mathcal{I}(K)) = 3$, $\mathrm{card}(\mathcal{I}(\partial K)) = 9$, $\mathrm{card}(\mathcal{I}(\partial K^{\mathsf{i}})) = 3$, $\mathrm{card}(\mathcal{I}(\partial K^{\mathsf{e}})) = 6$ and $\mathrm{card}(\mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}})) = 0$.

4.3.1. Technical preliminaries

Here we clarify/expand on the specific details related to discontinuous spaces. We define scalar-valued and vector-valued discontinuous finite element spaces as follows:

$$X_h = \{v \in L^1(D;\mathbb{R}) \mid v_{|K} \circ T_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}, \quad \boldsymbol{X}_h := [X_h]^m. \tag{4.7}$$

We denote by $\{\varphi_i\}_{i\in\mathcal{V}}$ the collection of global shape functions generated from the reference shape functions, i.e., $X_h = \mathrm{span}\{\varphi_i\}_{i\in\mathcal{V}}$. Each shape function has support on one cell only. We denote by $\mathcal{I}(K)$ the set of indices of the shape functions with support in K. Similarly, letting ∂K to be the boundary of the cell K, we denote by $\mathcal{I}(\partial K)$ the set of indices of the shape functions with non-vanishing trace on ∂K:

$$\mathcal{I}(K) := \big\{i \in \mathcal{V} \mid \varphi_{i|K} \not\equiv 0\big\}, \qquad \mathcal{I}(\partial K) := \big\{i \in \mathcal{V} \mid \varphi_{i|\partial K} \not\equiv 0\big\}. \tag{4.8}$$

Note that $\mathcal{I}(\partial K)$ not only includes indices of shape functions with support in $\mathcal{I}(K)$ but this set also includes indices of shape functions that do not have support in K (see Fig. 2 for additional geometrical insight). More precisely $\mathcal{I}(\partial K)$ is the union of two disjoint sets $\mathcal{I}(\partial K^{\mathsf{i}})$ and $\mathcal{I}(\partial K^{\mathsf{e}})$ defined as

$$\mathcal{I}(\partial K^{\mathsf{i}}) := \big\{i \in \mathcal{I}(K) \,\big|\, \varphi_{i|\partial K} \not\equiv 0\big\}, \qquad \mathcal{I}(\partial K^{\mathsf{e}}) := \mathcal{I}(\partial K)\backslash\mathcal{I}(\partial K^{\mathsf{i}}). \tag{4.9}$$

Finally, we assume that the finite element spaces are always constructed so that the sets of shape functions $\{\varphi_j\}_{j\in\mathcal{I}(K)}$ form a partition of unity over K and the shape functions $\{\varphi_j\}_{j\in\mathcal{I}(\partial K^{\mathsf{i}})}$, $\{\varphi_j\}_{j\in\mathcal{I}(\partial K^{\mathsf{e}})}$ form partitions of unity over ∂K, i.e.,

$$\sum_{j\in\mathcal{I}(K)} \varphi_{j|K} = 1, \qquad \sum_{j\in\mathcal{I}(\partial K^{\mathsf{i}})} \varphi_{j|\partial K} = 1, \quad \text{and} \quad \sum_{j\in\mathcal{I}(\partial K^{\mathsf{e}})} \varphi_{j|\partial K} = 1. \tag{4.10}$$

4.3.2. Definitions of $(\mathcal{V},\mathcal{E})$, m_i, and $\boldsymbol{c}_{ij}$

We start by defining the undirected graph $(\mathcal{V},\mathcal{E})$. The vertices are identified with the shape functions $\{\varphi_i\}_{i\in\mathcal{V}}$. Let $i \in \mathcal{V}$ and let K be the unique cell containing the support of φ_i. For any $i, j \in \mathcal{V}$, we say that the pair (i, j) is an edge of the connectivity graph, i.e., $(i, j) \in \mathcal{E}$, iff either $j \in \mathcal{I}(K)$ or $j \in \mathcal{I}(\partial K^{\mathsf{e}})$ and $\varphi_i\varphi_j|_{\partial K} \not\equiv 0$.

The consistent mass matrix and the lumped mass matrix are defined as in Section 4.2; in particular we set

$$m_i := \int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{4.11}$$

Let $\boldsymbol{u}_h^n = \sum_{j\in\mathcal{V}} \mathbf{U}_j^n\varphi_j \in \boldsymbol{X}_h$ be the approximation of $\boldsymbol{u}$ at time t^n, where $\boldsymbol{X}_h$ is a discontinuous finite element space defined in (4.7). Let $K \in \mathcal{T}_h$ and $i \in \mathcal{I}(K)$. The traditional heuristics for the derivation of dG schemes consists of integrating by parts on each cell K and introducing a numerical flux $\widehat{\mathbb{f}}$ on the boundary ∂K as follows:

$$\int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i\,\mathrm{d}\boldsymbol{x} \approx \int_K -\mathbb{f}(\boldsymbol{u}_h^n)\cdot\nabla\varphi_i\,\mathrm{d}\boldsymbol{x} + \int_{\partial K} \widehat{\mathbb{f}}\boldsymbol{n}_K\varphi_i\,\mathrm{d}s. \tag{4.12}$$

Upon denoting by $\boldsymbol{u}_h^{n,\mathsf{i}}$ the interior trace of $\boldsymbol{u}_h^n$ on ∂K and $\boldsymbol{u}_h^{n,\mathsf{e}}$ the exterior trace on ∂K, it is common to define the numerical flux as follows:

$$\widehat{\mathbb{f}}\boldsymbol{n}_K = \frac{1}{2}(\mathbb{f}(\boldsymbol{u}_h^{n,\mathsf{i}}) + \mathbb{f}(\boldsymbol{u}_h^{n,\mathsf{e}}))\boldsymbol{n}_K + \alpha_{\partial K}^n(\boldsymbol{u}_h^{n,\mathsf{i}} - \boldsymbol{u}_h^{n,\mathsf{e}}), \tag{4.13}$$

where $\alpha^n_{\partial K} > 0$ is usually some ad-hoc wave speed. The exact form of $\alpha^n_{\partial K}$ is unimportant for the time being; the sole purpose of the term $\alpha^n_{\partial K}(\boldsymbol{u}_h^{n,\mathsf{i}} - \boldsymbol{u}_h^{n,\mathsf{e}})$ is to stabilize the algorithm. We are just going to assume that this term introduces a first-order consistency error and that we are perfectly allowed to introduce further modifications to the discrete divergence operator (4.12) consistent with this assumption. Inserting (4.13) into (4.12) and integrating by parts, we obtain

$$\begin{aligned}\int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}\boldsymbol{x} &\approx \int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}\boldsymbol{x} \\ &+ \int_{\partial K} \tfrac{1}{2}(\mathbb{f}(\boldsymbol{u}_h^{n,\mathsf{e}}) - \mathbb{f}(\boldsymbol{u}_h^{n,\mathsf{i}}))\cdot\boldsymbol{n}_K\varphi_i \,\mathrm{d}s + \int_{\partial K} \alpha^n_{\partial K}(\boldsymbol{u}_h^{n,\mathsf{i}} - \boldsymbol{u}_h^{n,\mathsf{e}})\varphi_i \,\mathrm{d}s.\end{aligned} \tag{4.14}$$

We now consider an idea analogous to (4.5) and we replace $\mathbb{f}(\boldsymbol{u}_h^n)$ on the right-hand side of (4.14) by $\sum_{j\in\mathcal{V}} \mathbb{f}(\mathbf{U}_j^n)\varphi_j$ (where $\{\varphi_j\}_{j\in\mathcal{V}}$ are the shape functions of our discontinuous finite element space) to get:

$$\begin{aligned}\int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}\boldsymbol{x} &\approx \sum_{j\in\mathcal{I}(K)} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}^K \\ &+ \sum_{j\in\mathcal{I}(\partial K^{\mathsf{e}})} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}^\partial - \sum_{j\in\mathcal{I}(\partial K^{\mathsf{i}})} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}^\partial + \int_{\partial K} \alpha^n_{\partial K}(\boldsymbol{u}_h^{n,\mathsf{i}} - \boldsymbol{u}_h^{n,\mathsf{e}})\varphi_i \,\mathrm{d}s,\end{aligned} \tag{4.15}$$

with the notation

$$\boldsymbol{c}_{ij}^K := \int_K \varphi_i\nabla\varphi_j \,\mathrm{d}\boldsymbol{x}, \qquad \boldsymbol{c}_{ij}^\partial := \tfrac{1}{2}\int_{\partial K} \varphi_j\varphi_i\boldsymbol{n}_K \,\mathrm{d}s, \tag{4.16}$$

The three summations in (4.15) represent a consistent discretization of the divergence operator. In order to condense these three summations into a single one, and after noticing that j can belong to only one of three possible (disjoint) subsets: $\mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}})$, $\mathcal{I}(\partial K^{\mathsf{i}})$ or $\mathcal{I}(\partial K^{\mathsf{e}})$, we define the vector $\boldsymbol{c}_{ij}$ by setting:

$$\boldsymbol{c}_{ij} := \begin{cases} \boldsymbol{c}_{ij}^K & \text{if } j \in \mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}}), \\ (\boldsymbol{c}_{ij}^K - \boldsymbol{c}_{ij}^\partial) & \text{if } j \in \mathcal{I}(\partial K^{\mathsf{i}}), \\ \boldsymbol{c}_{ij}^\partial & \text{if } j \in \mathcal{I}(\partial K^{\mathsf{e}}). \end{cases} \tag{4.17}$$

Therefore, (4.15) can be rewritten as follows:

$$\int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n))\varphi_i \,\mathrm{d}\boldsymbol{x} \approx \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} + \int_{\partial K} \alpha^n_{\partial K}(\boldsymbol{u}_h^{n,\mathsf{i}} - \boldsymbol{u}_h^{n,\mathsf{e}})\varphi_i \,\mathrm{d}s. \tag{4.18}$$

Lemma 4.1. *The set of coefficients $\{\boldsymbol{c}_{ij}\}_{j\in\mathcal{I}(i)}$ defined in (4.17) satisfy the conservation properties (3.2).*

Proof. Let us start by proving the skew-symmetry property. Notice that (4.16) is equivalent to

$$\boldsymbol{c}_{ij} := \begin{cases} \boldsymbol{c}_{ij}^K - \boldsymbol{c}_{ij}^\partial & \text{if } j \in \mathcal{I}(K), \\ \boldsymbol{c}_{ij}^\partial & \text{if } j \in \mathcal{I}(\partial K^{\mathsf{e}}). \end{cases}$$

Let $j \in \mathcal{I}(K')$. Assume first that $K = K'$, then $\boldsymbol{c}_{ij} = \boldsymbol{c}_{ij}^K - \boldsymbol{c}_{ij}^\partial$. An integration by parts gives $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}^K + \boldsymbol{c}_{ij}^\partial$, which implies that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ because $i \in \mathcal{I}(K')$. Assume now that $K \neq K'$ but $i \in \mathcal{I}(\partial K'^{\mathsf{e}})$, then $\boldsymbol{c}_{ij} = \boldsymbol{c}_{ij}^\partial$. But $\boldsymbol{n}_K = -\boldsymbol{n}_{K'}$, hence $\boldsymbol{c}_{ij}^\partial = -\boldsymbol{c}_{ji}^\partial$, which means that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ because $i \in \mathcal{I}(\partial K'^{\mathsf{e}})$.

Let us now prove that $\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \mathbf{0}$. Using that $\mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}})$, $\mathcal{I}(\partial K^{\mathsf{i}})$, $\mathcal{I}(\partial K^{\mathsf{e}})$ is a partition of $\mathcal{I}(i)$ and definition (4.17), we have that

$$\begin{aligned}\sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} &= \sum_{j\in\mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}})} \boldsymbol{c}_{ij} + \sum_{j\in\mathcal{I}(\partial K^{\mathsf{i}})} \boldsymbol{c}_{ij} + \sum_{j\in\mathcal{I}(\partial K^{\mathsf{e}})} \boldsymbol{c}_{ij} \\ &= \sum_{j\in\mathcal{I}(K)\backslash\mathcal{I}(\partial K^{\mathsf{i}})} \boldsymbol{c}_{ij}^K + \sum_{j\in\mathcal{I}(\partial K^{\mathsf{i}})} (\boldsymbol{c}_{ij}^K - \boldsymbol{c}_{ij}^\partial) + \sum_{j\in\mathcal{I}(\partial K^{\mathsf{e}})} \boldsymbol{c}_{ij}^\partial \\ &= \sum_{j\in\mathcal{I}(K)} \boldsymbol{c}_{ij}^K - \sum_{j\in\mathcal{I}(\partial K^{\mathsf{i}})} \boldsymbol{c}_{ij}^\partial + \sum_{j\in\mathcal{I}(\partial K^{\mathsf{e}})} \boldsymbol{c}_{ij}^\partial.\end{aligned}$$

The partition of unity property on K (see (4.10)) implies that $\sum_{j\in\mathcal{I}(K)} \boldsymbol{c}_{ij}^K = \mathbf{0}$. The partition of unity property on ∂K (see (4.10)) implies that $\sum_{j\in\mathcal{I}(\partial K^{\mathrm{i}})} \boldsymbol{c}_{ij}^{\partial} = \int_{\partial K} \varphi_i \boldsymbol{n}_K \,\mathrm{d}s$ and $\sum_{j\in\mathcal{I}(\partial K^{\mathrm{e}})} \boldsymbol{c}_{ij}^{\partial} = \int_{\partial K} \varphi_i \boldsymbol{n}_K \,\mathrm{d}s$; hence, the last two summations cancel each other. This completes the proof. □

4.4. Graph viscosity for dG

It is important to notice at this stage, that the formulation of the viscous fluxes $\int_{\partial K} \alpha_{\partial K}^n (\boldsymbol{u}_h^{n,\mathrm{i}} - \boldsymbol{u}_h^{n,\mathrm{e}})\varphi_i \,\mathrm{d}s$ in (4.18) is not compatible with our pursuit of a purely algebraic formulation. Note that the dissipation in (4.18) is active only on ∂K and there is no dissipation in the bulk of K, at least when the polynomial degree of the approximation is larger than or equal to 1. More precisely, assume for the sake of simplicity that $\alpha_{\partial K}^n$ is constant over ∂K. Let us assume also that the shape functions are Lagrange-based and let $\{\boldsymbol{x}_i\}_{i\in\mathcal{V}}$ be the Lagrange nodes associated with $\{\varphi_i\}_{i\in\mathcal{V}}$. Then using the quadrature generated by the Lagrange nodes, one can legitimately approximate the integral $\int_{\partial K} \alpha_{\partial K}^n (\boldsymbol{u}_h^{n,\mathrm{i}} - \boldsymbol{u}_h^{n,\mathrm{e}})\varphi_i \,\mathrm{d}s$ by $m_i^{\partial} \alpha_{\partial K}^n (\boldsymbol{u}_h^{n,\mathrm{i}}(\boldsymbol{x}_i) - \boldsymbol{u}_h^{n,\mathrm{e}}(\boldsymbol{x}_i))$, where $m_i^{\partial} = \int_{\partial K} \varphi_i \,\mathrm{d}s$. This means that if i and j are in $\mathcal{I}(K)$ and $i \neq j$ (which we can assume since the polynomial degree is at least 1), then the "stabilizing" term $\int_{\partial K} \alpha_{\partial K}^n (\boldsymbol{u}_h^{n,\mathrm{i}} - \boldsymbol{u}_h^{n,\mathrm{e}})\varphi_i \,\mathrm{d}s$ does not contain any term proportional to $\boldsymbol{u}_h^{n,\mathrm{i}}(\boldsymbol{x}_i) - \boldsymbol{u}_h^{n,\mathrm{i}}(\boldsymbol{x}_j)$. That is to say, the traditional dG stabilization does not contain any stabilizing mechanism between the degrees of freedom that are internal to K. It is at this very point that we depart from the traditional dG formulation: we replace $\int_{\partial K} \alpha_{\partial K}^n (\boldsymbol{u}_h^i - \boldsymbol{u}_h^e)\varphi_i \,\mathrm{d}s$ by the graph Laplacian $-\sum_{j\in\mathcal{I}(i)} d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)$ which accounts for any possible interactions inside K and with the exterior traces on ∂K. Therefore, we finally replace (4.18) by

$$\int_K \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h))\varphi_i \,\mathrm{d}\boldsymbol{x} \approx \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{I}(i)} d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n), \tag{4.19}$$

and, thus modified, the final dG scheme exactly matches the generic form of the abstract scheme (3.4).

5. Runge Kutta SSP time integration

Increasing the time accuracy while keeping the invariant domain property can be done by using so-called Strong Stability Preserving (SSP) time discretization methods. The key idea is to achieve higher-order accuracy in time by making convex combinations of forward Euler steps. More precisely each time step of a SSP method is decomposed into substeps that are all forward Euler steps; the final update is constructed as a convex combination of the intermediate solutions. This section is meant to be a brief overview of SSP methods; we refer the reader to Ferracina and Spijker [22], Higueras [23], Gottlieb et al. [24] for more detailed reviews. The main result of this section is the Shu–Osher Theorem 5.4. Our formulation of the result is slightly different from the original statement to emphasize that this result is only about convexity (i.e., it does not involve any norm, seminorm, or convex functional). The reader familiar with this material is invited to move to Section 6.

5.1. SSPRK methods

We are going to illustrate the SSP concept with explicit Runge–Kutta methods. Let us consider a finite-dimensional vector space E, a subset $A \subset E$ and a (nonlinear) operator $L : [0,T]\times A \longrightarrow E$. We are interested in approximating in time the following problem $\partial_t u + L(t,u) = 0$ with appropriate initial condition. We assume that this system of ordinary differential equations makes sense (for instance L is continuous w.r.t. t and Lipschitz w.r.t. u). We further assume that there exists a convex subset $B \subset A$ and $\tau_{\max} > 0$ such that

$$v + \tau L(t,v) \subset B, \qquad \forall v \in B,\ \forall t \in [0,T],\ \forall \tau \in [0,\tau_{\max}]. \tag{5.1}$$

Consider a general s stages, explicit Runge–Kutta method identified by its Butcher tableau composed of a matrix $(a_{ij})_{\{1\le i,j\le s\}} \in \mathbb{R}^{s\times s}$ and a vector $(b_j)_{\{1\le j\le s\}} \in \mathbb{R}^s$

$$\begin{array}{c|cccccc} 0 & & & & & & \\ c_2 & a_{21} & & & & & \\ c_3 & a_{31} & a_{32} & & & & \\ \vdots & \vdots & & \ddots & & & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{5.2}$$

where $c_i := \sum_{j=1}^{i-1} a_{ij}$ for $i \in \{2{:}s\}$. Let us also set $c_1 := 0$. Let us assume now that the above s stages, explicit Runge–Kutta method has the following $(\alpha - \beta)$ representation: There are real coefficients α_{ki}, β_{ki} with $k \in \{0{:}i-1\}$ and $i \in \{1{:}s\}$ such that u^{n+1} is obtained by first setting $w^{(0)} = u^n$, then computing

$$w^{(i)} = \sum_{k=0}^{i-1} \alpha_{ik} w^{(k)} + \beta_{ik} \tau L(t^n + \gamma_k \tau, w^{(k)}), \quad i \in \{1{:}s\}, \tag{5.3}$$

and finally setting $u^{n+1} = w^{(s)}$, where $\sum_{0\le k\le i-1} \alpha_{ik} = 1$, $\gamma_k := c_{k+1}$, $\alpha_{ik} \ge 0$, and $\beta_{ik} \ge 0$, for all $k \in \{0{:}i-1\}$ and all $i \in \{1{:}s\}$. We further assume that $\beta_{ik} = 0$ if $\alpha_{ik} = 0$, $k \in \{0{:}i-1\}$, $i \in \{1{:}s\}$. Not every s stages, explicit Runge–Kutta method admits an $(\alpha-\beta)$ representation. Any Runge–Kutta method that admits an $(\alpha-\beta)$ representation as defined above is said to be SSP for a reason that will be stated in Theorem 5.4.

Example 5.1 (*Midpoint Rule*)**.** The midpoint rule, defined by the Butcher tableau

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \tag{5.4}$$

does not have a legitimate $(\alpha - \beta)$ representation, since it would require that $\beta_{20} + \alpha_{21} = 0$, which in turn would imply that either $\beta_{20} < 0$ or $\alpha_{21} < 0$. □

Example 5.2 (*SSPRK(2,2)*)**.** Heun's method, which is a second-order Runge–Kutta technique composed of two stages, is SSP. It has the following $(\alpha - \beta)$ tableau and can be implemented as follows:

α		β		γ	c_{os}
1		1		0	
$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	1	1

$$\begin{aligned} w^{(1)} &= u^n + \tau L(t^n, u^n), \\ w^{(2)} &= w^{(1)} + \tau L(t^{n+1}, w^{(1)}), \\ u^{n+1} &= \tfrac{1}{2} u^n + \tfrac{1}{2} w^{(2)}. \quad \square \end{aligned}$$

Example 5.3 (*SSPRK(3,3), SSPRK(4,3)*)**.** The following Runge–Kutta methods, which are third-order and composed of three substeps and four substeps, respectively, are SSP:

α			β			γ	c_{os}
1			1			0	
$\frac{3}{4}$	$\frac{1}{4}$		0	$\frac{1}{4}$		1	1
$\frac{1}{3}$	0	$\frac{2}{3}$	0	0	$\frac{2}{3}$	$\frac{1}{2}$	

α				β				γ	c_{os}
1				$\frac{1}{2}$				0	
0	1			0	$\frac{1}{2}$			$\frac{1}{2}$	2
$\frac{2}{3}$	0	$\frac{1}{3}$		0	0	$\frac{1}{6}$		1	
0	0	0	1	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	

For instance the SSPRK(3, 3) method can be implemented as follows:

$$\begin{aligned} w^{(1)} &= u^n + \tau L(t^n, u^n), & z^{(1)} &= w^{(1)} + \tau L(t^n + \tau, w^{(1)}), \\ w^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} z^{(1)}, & z^{(2)} &= w^{(2)} + \tau L(t^n + \tfrac{1}{2}\tau, w^{(2)}), \\ u^{n+1} &= \frac{1}{3} u^n + \frac{2}{3} z^{(2)}. \quad \square \end{aligned}$$

5.2. The key result

We henceforth denote

$$c_{\text{os}} := \inf_{\{\alpha_{ik}\neq 0,\ 1\le k+1\le i\le s\}} \alpha_{ik} \beta_{ik}^{-1}. \tag{5.5}$$

The following theorem is the main result of this section.

Theorem 5.4. *[Shu–Osher] Assume that the Runge–Kutta method with the Butcher tableau (5.2) is SSP. Let $B \subset A$ be convex. Let $u^n \in B$ and assume that $\tau \le c_{\text{os}} \tau_{\max}$, then $u^{n+1} \in B$.*

Proof. Let $n \geq 0$ and assume that $u^n \in B$. Let $i \in \{1{:}s\}$ and assume that $w^{(k)} \in B$ for all $k \in \{0{:}i-1\}$. Note that this assumption is satisfied for $i = 1$ since $w^{(0)} = u^n \in B$. Consider the kth term in (5.3), $0 \leq k \leq i-1$. If $\alpha_{ik} = 0$ then $\beta_{ik} = 0$ by construction, and there is nothing to sum. Assume now that $\alpha_{ik} > 0$. Let us denote $r_{ik} := \beta_{ik}/\alpha_{ik}$ and $z^{(i,k)} := w^{(k)} + r_{ik}\tau L(t^n + \gamma_k\tau, w^{(k)})$, then the condition $\tau \leq c_{\mathrm{os}}\tau_{\max}$ implies that $r_{ik}\tau \leq (\beta_{ik}/\alpha_{ik})c_{\mathrm{os}}\tau_{\max} \leq \tau_{\max}$, which, owing to (5.1), is sufficient to ascertain that $z^{(i,k)} \in B$ for all $k \in \{0{:}i-1\}$. Observing that $w^{(i)} = \sum_{k=1}^{i-1} \alpha_{ki} z^{(i,k)}$, the condition $\sum_{0\leq k\leq i-1} \alpha_{ik} = 1$ together with $0 \leq \alpha_{ik}$, $0 \leq k \leq i-1$, implies that $w^{(i)}$ is a convex combination of $z^{(i,0)}, \ldots, z^{(i,i-1)}$; hence $w^{(i)}$ is in B since B is convex. In conclusion $w^{(k)} \in B$ for all $k \in \{0{:}i\}$ and all $i \in \{1{:}s\}$, thereby proving that $u^{n+1} = w^{(s)} \in B$. □

Remark 5.5 (*Literature*)**.** Theorem 5.4 has been established in a slightly different form in Shu and Osher [25, Prop. 2.1] not explicitly invoking convexity. Although our proof is very similar to that in [25], the statement of Theorem 5.4 is slightly different since it only involves convexity; no norm or seminorm (as in Gottlieb et al. [26, p. 92]), or convex functional (as in [24, Eq. (1.3)]) is involved. This variant of the theorem does not seem to be very well known. □

Remark 5.6 (*Structure of B*)**.** In the original paper [25] and in [26], E is a normed vector space equipped with some norm $\|\cdot\|_E$. The assumption (5.1) then consists of stating that $\mathbb{I} + \tau L(t,\cdot)$ maps any ball B centered at 0 into B for any $s \in [0, \tau_{\max}]$ and any $t \in [0, T]$. In particular taking any $v \in E$ and defining B to be the ball of radius $\|v\|_B$ centered at 0, the assumption (5.1) amounts to saying that $\|v + \tau L(t,v)\|_B \leq \|v\|_B$, which is Eq. (1.3) in [26]. The norm that is used in [25] is the total variation. In the present paper the assumption (5.1) is more general. We are going to use it with the following structure: we are going to assume that there are two positive integers $I, m \in \mathbb{N}\backslash\{0\}$ such that $E = (\mathbb{R}^m)^I$. Here $\mathbb{R}^m$ is called the phase space. Then we assume that there is convex subset of the phase space $\mathcal{B} \subset \mathbb{R}^m$ such that the assumption (5.1) holds with $B := (\mathcal{B})^I$. All the convex arguments invoked in the rest of the paper extends to SSP RK techniques with this particular structure. □

6. High-order method

The algorithm that we are going to develop in Section 7 relies on the construction of the low-order invariant domain preserving solution $\mathbf{U}_i^{\mathrm{L},n+1}$ described in Sections 3.1–3.2 and a high-order solution $\mathbf{U}_i^{\mathrm{H},n+1}$ that possibly wanders outside the invariant domain. We are then going to limit the high-order solution by pushing it back into the invariant domain in the direction of the low-order solution. This limiting technique, which we call *convex limiting*, will be explained in Section 7. The purpose of the present section is to present various ways to construct $\mathbf{U}_i^{\mathrm{H},n+1}$.

6.1. Achieving high-order consistency

In this section we describe in broad terms how high-order consistency can be achieved.

6.1.1. Discretization-independent setting

Independently of the space discretization that is used, we henceforth assume that the high-order update $\mathbf{U}_i^{\mathrm{H},n+1}$ is computed as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{H},n} = m_i \boldsymbol{S}(\mathbf{U}_i^n), \tag{6.1}$$

where the high-order flux $\mathbf{F}_{ij}^{\mathrm{H},n}$ is assumed to be skew-symmetric; i.e., $\mathbf{F}_{ij}^{\mathrm{H},n} = -\mathbf{F}_{ji}^{\mathrm{H},n}$ for all $i \in \mathcal{V}$, $j \in \mathcal{I}(i)$ (under appropriate boundary conditions). The skew-symmetry implies that the high-order update is conservative; i.e., $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{H},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^n$ if $\boldsymbol{S} \equiv \mathbf{0}$. The expression (6.1) is the only information regarding the high-order update that will be necessary for the convex limiting technique to be presented in section Section 7.

There are many different techniques to compute high-order consistent fluxes $\mathbf{F}_{ij}^{\mathrm{H},n}$ which depend on the space discretization of choice. For the sake of completeness, we list some of those in Sections 6.1.2, 6.1.3, and 6.1.4. None of this material is essential to understand the convex limiting technique explained in Section 7.

6.1.2. High-order algebraic fluxes: Finite volumes

In the context of finite volume schemes, high-order algebraic fluxes $\mathbf{F}_{ij}^{\mathrm{H},n}$ are obtained as integrals of high-order numerical fluxes over the interfaces between volumes, i.e., $\mathbf{F}_{ij}^{\mathrm{H},n} := \int_{\Gamma_{ij}} \widehat{\mathbb{f}}\boldsymbol{n}_{ij}\,\mathrm{d}s$ where $\widehat{\mathbb{f}}\boldsymbol{n}_{ij}$ is some numerical flux. For instance, a widely popular choice of algebraic flux consists of setting:

$$\mathbf{F}_{ij}^{\mathrm{H},n} := \int_{\Gamma_{ij}} \left(\tfrac{1}{2}(\mathbb{f}(\boldsymbol{u}_h^{\mathrm{H},n,\mathsf{i}}) + \mathbb{f}(\boldsymbol{u}_h^{\mathrm{H},n,\mathsf{e}}))\cdot\boldsymbol{n}_{ij} + d_{ij}^{\mathrm{L},n}(\boldsymbol{u}_h^{\mathrm{H},n,\mathsf{i}} - \boldsymbol{u}_h^{\mathrm{H},n,\mathsf{e}}) \right) \mathrm{d}s, \tag{6.2}$$

where the superscripts e and i denote the exterior and interior traces respectively, and $\boldsymbol{u}_h^{\mathrm{H},n}$ is a discontinuous piecewise polynomial reconstruction (of degree at most k) recovered from the piecewise constant solution $\boldsymbol{u}_h^n = \sum_{j\in\mathcal{V}} \mathbf{U}_j^n \mathbb{I}_{K_j}$ satisfying the conservation constraint $\frac{1}{|K_i|}\int_{K_i}(\boldsymbol{u}_h^{\mathrm{H},n} - \boldsymbol{u}_h^n)\,\mathrm{d}\boldsymbol{x} = 0$. More precisely $\boldsymbol{u}_h^{\mathrm{H},n,\mathsf{i}}(\boldsymbol{x}) = \lim_{K_i\ni\boldsymbol{y}\to\boldsymbol{x}} \boldsymbol{u}_h^{\mathrm{H},n}(\boldsymbol{y})$ and $\boldsymbol{u}_h^{\mathrm{H},n,\mathsf{e}}(\boldsymbol{x}) = \lim_{K_j\ni\boldsymbol{y}\to\boldsymbol{x}} \boldsymbol{u}_h^{\mathrm{H},n}(\boldsymbol{y})$. In practice, (6.2) has to be computed using quadrature on the faces of the element. The choices of numerical flux $\widehat{\mathbb{f}}\boldsymbol{n}_{ij}$ and reconstruction $\boldsymbol{u}_h^{\mathrm{H},n}$ that could be used in (6.2) are not unique. There is a massive body of literature on this topic and it is well beyond the scope of the current paper to elaborate further in this direction; we refer the reader to Barth and Ohlberger [19], Kröner [27], Morton and Sonar [28] for additional background. For the purpose of the present paper, we are only going to assume that (6.1) holds with skew-symmetric algebraic fluxes $\mathbf{F}_{ij}^{\mathrm{H},n}$.

6.1.3. High-order algebraic flux: Continuous finite elements

We now turn our attention to continuous finite elements. In this case high-order consistency can be achieved by using a degenerate graph viscosity $d_{ij}^{\mathrm{H},n}$ such that $d_{ij}^{\mathrm{H},n} \ll d_{ij}^{\mathrm{L},n}$ in smooth regions while $d_{ij}^{\mathrm{H},n} \approx d_{ij}^{\mathrm{L},n}$ near shocks. Of course $d_{ij}^{\mathrm{H},n}$ must also satisfy the conservation constraints

$$d_{ij}^{\mathrm{H},n} = d_{ji}^{\mathrm{H},n} \geq 0 \quad \text{if } i \neq j, \quad \text{and} \quad \sum_{j\in\mathcal{I}(i)} d_{ij}^{\mathrm{H},n} = 0. \tag{6.3}$$

The algebraic flux looks as the one defined in (3.6) for the low-order method; the only difference here is that we use the high-order viscosities $\{d_{ij}^{\mathrm{H},n}\}_{j\in\mathcal{I}(i)}$:

$$\mathbf{F}_{ij}^{\mathrm{H},n} := (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n). \tag{6.4}$$

Higher-order accuracy in space can also be obtained by using the consistent mass matrix instead of the lumped mass matrix for the discretization of the time derivative. By reducing dispersive errors, this technique is known to yield superconvergence at the grid points; see Christon et al. [29], Guermond and Pasquetti [30]. In this case the high-order update is computed by solving the following mass matrix problem:

$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n) + (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = m_i \boldsymbol{S}(\mathbf{U}_i^n). \tag{6.5}$$

Noticing that $m_{ij} = \delta_{ij}m_i + m_{ij} - \delta_{ij}m_i$, we can rewrite (6.5) as

$$\begin{aligned} &\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \frac{(m_{ij} - \delta_{ij}m_i)}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n) \\ &\qquad + (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = m_i \boldsymbol{S}(\mathbf{U}_i^n). \end{aligned} \tag{6.6}$$

Since $\sum_{j\in\mathcal{I}(i)}(m_{ij} - \delta_{ij}m_i) = 0$, we add $-\sum_{j\in\mathcal{I}(i)} \frac{(m_{ij}-\delta_{ij}m_i)}{\tau}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n) = 0$ to the identity (6.6) to get

$$\begin{aligned} &\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{(m_{ij} - \delta_{ij}m_i)}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n - \mathbf{U}_i^{\mathrm{H},n+1} + \mathbf{U}_i^n) \\ &\qquad + (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = m_i \boldsymbol{S}(\mathbf{U}_i^n). \end{aligned} \tag{6.7}$$

Then (6.1) holds with the following definition for the high-order algebraic flux:

$$\begin{aligned} \mathbf{F}_{ij}^{\mathrm{H},n} :=& \frac{m_{ij}}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n - \mathbf{U}_i^{\mathrm{H},n+1} + \mathbf{U}_i^n) \\ &+ (\mathbb{f}(\mathbf{U}_j^n) + \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n). \end{aligned} \tag{6.8}$$

In the context of finite difference methods, a scheme with the above structure is said to be linearly implicit as the numerical fluxes depend linearly on the state $\mathbf{U}_j^{\mathrm{H},n+1}$.

We finally mention a third approach which has antidispersive properties that are similar to (6.5) but does not require solving a mass matrix problem a each time step. This method consists of approximating the inverse of $\mathcal{M}$ by $(\mathcal{M}^L)^{-1}(\mathcal{I}+(\mathcal{M}^L-\mathcal{M})(\mathcal{M}^L)^{-1})$, where $\mathcal{I}$ is the identity matrix. We refer the reader to Guermond et al. [31, §3.3] for the details.

6.1.4. High-order algebraic flux: Discontinuous finite elements

Just like for continuous finite elements, high-order consistency is space is obtained for discontinuous finite elements by replacing the low-order graph viscosity $d_{ij}^{\mathrm{L},n}$ by a high-order graph viscosity $d_{ij}^{\mathrm{H},n}$ satisfying the symmetry and positivity properties stated in (6.3). The corresponding flux in (6.1) is

$$\mathbf{F}_{ij}^{\mathrm{H},n} := (\mathbb{f}(\mathbf{U}_j^n)+\mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n), \qquad \forall \mathcal{I}(K)\cup\mathcal{I}(K^{\mathsf{e}}). \tag{6.9}$$

Like for continuous elements, superconvergence can be obtained by using the consistent mass matrix. A high-order discontinuous finite element scheme using the consistent mass matrix can be written as follows:

$$\begin{aligned}&\sum_{j\in\mathcal{I}(K)}\frac{m_{ij}}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1}-\mathbf{U}_j^n)+\sum_{j\in\mathcal{I}(i)}(\mathbb{f}(\mathbf{U}_j^n)+\mathbb{f}(\mathbf{U}_j^n))\boldsymbol{c}_{ij}\\ &-d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)=m_i\boldsymbol{S}(\mathbf{U}_i^n).\end{aligned} \tag{6.10}$$

Notice that the mass matrix only involves the dofs in $\mathcal{I}(K)$. As in the continuous case, noting that $m_{ij}=\delta_{ij}m_i+m_{ij}-\delta_{ij}m_i$, using the partition of unity properties, and proceeding as in (6.6)–(6.7), we obtain the following definition for the high-order flux $\mathbf{F}_{ij}^{\mathrm{H},n}$ that is used in (6.1):

$$\mathbf{F}_{ij}^{\mathrm{H},n} := \begin{cases}\dfrac{m_{ij}}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1}-\mathbf{U}_j^n-\mathbf{U}_i^{\mathrm{H},n+1}+\mathbf{U}_i^n)\\ \qquad+(\mathbb{f}(\mathbf{U}_j^n)+\mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij}-d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n) & \text{if } j\in\mathcal{I}(K),\\ (\mathbb{f}(\mathbf{U}_j^n)+\mathbb{f}(\mathbf{U}_i^n))\boldsymbol{c}_{ij}-d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n) & \text{if } j\in\mathcal{I}(K^{\mathsf{e}}).\end{cases} \tag{6.11}$$

6.2. Smoothness-based graph viscosity

The objective of this section is to present a method where the high-order graph viscosity in (6.4), (6.8), (6.9), or (6.11) is obtained by estimating the smoothness of some functional (e.g., an entropy) of the current solution.

6.2.1. Principles of the method

Let $\boldsymbol{u}_h^n=\sum_{i\in\mathcal{V}}\mathbf{U}_i^n\varphi_i$ be the current approximation and let $g:\mathcal{A}\to\mathbb{R}$ be some functional (examples will be given below). We define the smoothness indicator associated to g as follows:

$$\alpha_i^n := \frac{\left|\sum_{j\in\mathcal{I}(i)}\beta_{ij}(g(\mathbf{U}_j^n)-g(\mathbf{U}_i^n))\right|}{\max(\sum_{j\in\mathcal{I}(i)}|\beta_{ij}|\,|g(\mathbf{U}_j^n)-g(\mathbf{U}_i^n)|,\epsilon_i)}, \tag{6.12}$$

with $\epsilon_i=\epsilon\max_{j\in\mathcal{I}(i)}|g(\mathbf{U}_j^n)|$, where ϵ is very small number. This term avoids degeneracy when $g(\mathbf{U}_j^n)$ is constant for all $j\in\mathcal{I}(i)$; see Remark 6.1. The real numbers β_{ij} are selected to make the method linearity-preserving (see Berger et al. [32] for a review on linearity-preserving limiters in the finite volume literature). The reader is referred to Remark 6.2 for the details. Notice that $\alpha_i^n\in[0,1]$ for all $i\in\mathcal{V}$ and $\alpha_i^n=1$ if $g(\mathbf{U}_i)$ is a local extremum. This property will play an important role in the proof of Theorem 6.5 which is the main result of Section 6.2.

We now define the high-order graph viscosity by setting

$$d_{ij}^{\mathrm{H},n} := d_{ij}^{\mathrm{L},n}\max(\psi(\alpha_i^n),\psi(\alpha_j^n)), \tag{6.13}$$

where $\psi\in\mathrm{Lip}([0,1];[0,1])$ is any Lipschitz function from $[0,1]$ to $[0,1]$ such that $\psi(1)=1$. One typical example is $\psi(\alpha)=\left(\max\left(0,\frac{\alpha-\alpha_0}{1-\alpha_0}\right)\right)^q$ with $q\geq 2$ and $\alpha_0\in[0,1)$. For instance one can take $\alpha_0=\frac{1}{2}$ and $q=4$. One need to be careful though not to take α_0 too close to 1 and q not too large since we will see in Theorem 6.5 that the Lipschitz constant of ψ plays an important role in the properties of the method.

Remark 6.1 (*Choices for ϵ*)**.** Using double precision arithmetic, the regularization in (6.12) can be done with $\epsilon = 10^{-\frac{16}{2}}$. We have also observed that using $\epsilon = (m_i/|D|)^{\frac{3}{d}}$ maintains the second-order accuracy properties of the method in any L^q-norm, $q \in [1, \infty]$. □

Remark 6.2 (*Linearity-Preserving β_{ij}*)**.** To be linearity-preserving with continuous finite elements one should obtain $\alpha_i^n = 0$ if $g(\boldsymbol{u}_h^n)$ is linear on the support of the shape function φ_i. One simple choice for continuous finite elements consists of setting $\beta_{ij} = \int_D \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x$ (for the time being we do not require $\beta_{ij} > 0$ in (6.12)). For discontinuous elements, one could take $\beta_{ij} = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x - \int_{\partial K} \frac{1}{2} \nabla\varphi_j \cdot \boldsymbol{n}_K \varphi_i \,\mathrm{d}x$, where K is the unique cell such that $i \in \mathcal{I}(K)$ and $\boldsymbol{n}_K$ is the unit normal vector on ∂K pointing outward K. For finite volumes, one should get $\alpha_i^n = 0$ if a linear reconstruction fits all the data $\{g(\mathbf{U}_j^n)\}_{j\in\mathcal{I}(i)}$. For instance, one can use the mean-value coordinates; see Floater [33, Eq. 5.1] for the details. Let us finally remark that although using $\beta_{ij} = 1$ is not a priori linearity preserving, we have numerically verified that this choice works reasonably well on quasi-uniform meshes. □

If the coefficients β_{ij} are defined so the linearity-preserving property holds, then the numerator of (6.12) behaves like $h^2 \|D^2 g(\boldsymbol{u}(\boldsymbol{\xi}, t^n))\|_{\ell^2(\mathbb{R}^{d\times d})}$ at some point $\boldsymbol{\xi}$, whereas the denominator behaves like $h\|\nabla g(\boldsymbol{\zeta})\|_{\ell^2(\mathbb{R}^d)}$ at some point $\boldsymbol{\zeta}$. Therefore, we have $\alpha_i^n \approx h\|D^2 g(\boldsymbol{\xi})\|_{\ell^2(\mathbb{R}^{d\times d})}/\|\nabla g(\boldsymbol{\zeta})\|_{\ell^2(\mathbb{R}^d)}$, that is to say α_i^n is of order h in the regions where g is smooth and does not have a local extremum. This argument shows that $d_{ij}^{\mathrm{H},n}$ is one order smaller than $d_{ij}^{\mathrm{L},n}$ (in terms of mesh size). Hence it is reasonable to expect that the method using $d_{ij}^{\mathrm{H},n}$ is formally second-order accurate in space.

Example 6.3 (*Choosing $g(\mathbf{U})$*)**.** In the context of the shallow water equations one can use the water height as smoothness indicator. For the compressible Euler equations one can use the density. We are going to prove stability properties for these two choices in Theorem 6.5, (see also Example 6.6). In general it is a good idea to choose $g(\mathbf{U})$ to be entropy associated with (1.1) (with or without the source term). We refer the reader to Guermond et al. [1], where a full set of tests is reported for the compressible Euler equations with the γ-law. The computations therein are done with $g(\mathbf{U}) = \frac{\rho}{\gamma-1}\log(e(\mathbf{U})\rho^{1-\gamma})$, where $e(\mathbf{U})$ is the specific internal energy □

6.2.2. Stability for scalar components

We now establish some invariant domain preserving properties associated with the smoothness-based graph viscosity (6.12) when the coefficients β_{ij} are positive. We further specialize the setting by assuming that $g : \mathcal{A} \to \mathbb{R}$ is a projection onto one of the scalar components of $\mathbf{U}$. Without loss of generality we set $g(\mathbf{U}) = \mathsf{U}_1$ with the convention $\mathbf{U} := (\mathsf{U}_1, \ldots, \mathsf{U}_m)^{\mathsf{T}}$. From now on, we drop the index $_1$ to simplify the notation; that is, we set $g(\mathbf{U}) = \mathsf{U}$. We denote by $S : \mathcal{A} \to \mathbb{R}$ the corresponding scalar component of the source $\boldsymbol{S}$. One important assumption in this section is that $S \equiv 0$, i.e., the scalar component of the source acting on U is zero. The reader is referred to Example 6.6 for illustrations of the technique under consideration for the shallow water equations and the compressible Euler equations.

We have seen in Theorem 3.6 that the auxiliary states $\overline{\mathbf{U}}_{ij}^n$ defined in (3.8) play an important role in the stability analysis. These states are such that if $\mathbf{U}_i^n, \mathbf{U}_j^n \in \mathcal{B}$, where $\mathcal{B} \subset \mathcal{A}$ is some convex invariant set, then $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}$, provided that $1 + \frac{2\tau d_{ii}^{\mathrm{L},n}}{m_i} \geq 0$, and the low-order graph viscosity $d_{ij}^{\mathrm{L},n}$ is defined as in (3.10). We denote by $\overline{U}_{ij}^n$ the scalar component of $\overline{\mathbf{U}}_{ij}^n$ that is of interest to us. Then we set

$$\mathsf{U}_i^{\mathrm{M},n} := \max_{j\in\mathcal{I}(i)} \overline{U}_{ij}^n, \qquad \mathsf{U}_i^{\mathrm{m},n} := \min_{j\in\mathcal{I}(i)} \overline{U}_{ij}^n. \tag{6.14}$$

We set $\mathcal{I}(i^+) := \{j \in \mathcal{I}(i) \mid \mathsf{U}_i^n < \mathsf{U}_j^n\}$ and $\mathcal{I}(i^-) := \{j \in \mathcal{I}(i) \mid \mathsf{U}_j^n < \mathsf{U}_i^n\}$. To simplify the notation we set

$$\gamma_i^n := -\frac{2\tau d_{ii}^{\mathrm{L},n}}{m_i}, \qquad \gamma_i^{+,n} := \frac{2\tau}{m_i} \sum_{j\in\mathcal{I}(i^+)} d_{ij}^{\mathrm{L},n}, \qquad \gamma_i^{-,n} := \frac{2\tau}{m_i} \sum_{j\in\mathcal{I}(i^-)} d_{ij}^{\mathrm{L},n}. \tag{6.15}$$

The following key "gap lemma" will be invoked later.

Lemma 6.4 (*Gap Estimates*)**.** *Let $n \geq 0$, and $i \in \mathcal{V}$. We define the gap parameter*

$$\theta_i^n := \frac{\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}}, \ \textit{if } \mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0; \qquad \theta_i^n := \frac{1}{2}, \ \textit{otherwise}. \tag{6.16}$$

Assume that $\gamma_i^n < 1$. *Let* $\mathbf{U}_i^{n+1}$ *be the high-order update given by (6.1) using either the high-order cG flux (6.4) or the high-order dG flux (6.9) with any graph viscosity* $\{d_{ij}^{\mathrm{H},n}\}_{j\in\mathcal{I}(i)\setminus\{i\}}$ *defined by* $d_{ij}^{\mathrm{H},n} := d_{ij}^{\mathrm{L},n}\max(\psi_i^n,\psi_j^n)$ *with* $\psi_i^n, \psi_j^n \in [0,1]$. *Then,*

$$\mathsf{U}_i^{n+1} \le \mathsf{U}_i^{\mathrm{M},n} - (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{-,n}\right), \tag{6.17}$$

$$\mathsf{U}_i^{n+1} \ge \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\left(\theta_i^n(1-\gamma_i^n) - (1-\theta_i^n)(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{+,n}\right). \tag{6.18}$$

Proof. There is nothing to prove if $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} = 0$. Let us now assume that $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$. Subtracting (3.4) from (6.1) we obtain

$$\mathbf{U}_i^{\mathrm{H},n+1} = \mathbf{U}_i^{\mathrm{L},n+1} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n).$$

Let us focus on the scalar component U_i^n. Recalling the auxiliary states $\overline{\mathsf{U}}_{ij}^n$ defined in (3.8) and recalling that we have assumed $S \equiv 0$, the identity (3.11) gives $\mathsf{U}_i^{\mathrm{L},n+1} = (1-\gamma_i)\mathsf{U}_i^n + \sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$. Then setting $\mathsf{U}_i^{*,n} := \frac{1}{\gamma_i^n}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$, we have $\mathsf{U}_i^{\mathrm{L},n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n}$, and this in turn implies that

$$\mathsf{U}_i^{\mathrm{H},n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

(ii) Using that $\mathsf{U}_i^{*,n} \in \mathrm{conv}\{\overline{\mathsf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)\setminus\{i\}}$, we have $\mathsf{U}_i^{*,n} \le \mathsf{U}_i^{\mathrm{M},n}$, and we infer that

$$\mathsf{U}_i^{\mathrm{H},n+1} \le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

Then using that $d_{ij}^{\mathrm{H},n} \le d_{ij}^{\mathrm{L},n}$, since $\max(\psi_i^n,\psi_j^n) \le 1$, the above inequality gives

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

Now using that $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$ and that U_i^n is in the convex hull of $\mathsf{U}_i^{\mathrm{M},n}$ and $\mathsf{U}_i^{\mathrm{m},n}$, we have $\mathsf{U}_i^n = \theta_i^n\mathsf{U}_i^{\mathrm{M},n} + (1-\theta_i^n)\mathsf{U}_i^{\mathrm{m},n}$ where $\theta_i^n \in [0,1]$ has been defined in (6.16). Hence, $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n} = -\theta_i^n(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$ and $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n} = (1-\theta_i^n)(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$. With these definitions, the above inequality is rewritten as follows:

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

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

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

(iv) The other estimate is obtained similarly. More precisely, using that $\mathsf{U}_{ij}^{*,n} \ge \mathsf{U}_i^{\mathrm{m},n}$, we infer that

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

which completes the proof. □

We now formulate the main result of this section.

Theorem 6.5. *Let $\psi \in \mathrm{Lip}([0,1];[0,1])$ be such that $\psi(1)=1$ and with Lipschitz constant k_ψ. Consider the scheme (6.1) using either the high-order cG flux (6.4) or the high-order dG flux (6.9) with the graph viscosity defined in (6.13). Assume that $g(\mathbf{U})=\mathsf{U}$ in (6.12). Assume that all the coefficients β_{ij} in (6.12) are positive and there exists $\varpi^\sharp \in (0,\infty)$ uniform with respect to the mesh sequence $(\mathcal{T}_h)_{h>0}$, such that $\max_{i\in\mathcal{V}}(\max_{j\in\mathcal{I}(i)}\beta_{ij}/\min_{j\in\mathcal{I}(i)}\beta_{ij}) \le \varpi^\sharp$. Let $i\in\mathcal{V}$ and $n\ge 0$. Then, under the local CFL condition $\gamma_i^n \le \frac{1}{1+k_\psi c_\sharp}$, where $c_\sharp = \varpi^\sharp \max_{i\in\mathcal{V}}\mathrm{card}(\mathcal{I}(i))$ (this number is uniformly bounded with respect to the mesh sequence), the scheme is locally invariant domain preserving for the scalar component U: i.e., $\mathsf{U}_i^{\mathrm{H},n+1}\in[\mathsf{U}_i^{\mathrm{m},n},\mathsf{U}_i^{\mathrm{M},n}]$.*

Proof. Note first that if $\mathsf{U}_i^{\mathrm{M},n}=\mathsf{U}_i^{\mathrm{m},n}$, then $\mathsf{U}_i^{\mathrm{H},n+1}=\mathsf{U}_i^n\in[\mathsf{U}_i^{\mathrm{m},n},\mathsf{U}_i^{\mathrm{M},n}]$ irrespective of the value of $d_{ij}^{\mathrm{H},n}$, which proves the statement. Let us assume now that $\mathsf{U}_i^{\mathrm{M},n}\neq\mathsf{U}_i^{\mathrm{m},n}$. If $\theta_i^n=\frac{\mathsf{U}_i^n-\mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n}-\mathsf{U}_i^{\mathrm{m},n}}\in\{0,1\}$, then either $\mathsf{U}_i^n=\mathsf{U}_i^{\mathrm{m},n}$ or $\mathsf{U}_i^n=\mathsf{U}_i^{\mathrm{M},n}$. In this case, $\alpha_i^n=1$ and $\psi(\alpha_i^n)=1$; as a result, $d_{ij}^{\mathrm{H},n}=d_{ij}^{\mathrm{L},n}\max(1,\psi(\alpha_j))=d_{ij}^{\mathrm{L},n}$ for all $j\in\mathcal{I}(i)$, which implies that $\mathsf{U}_i^{\mathrm{H},n+1}=\mathsf{U}_i^{\mathrm{L},n+1}\in[\mathsf{U}_i^{\mathrm{m},n},\mathsf{U}_i^{\mathrm{M},n}]$. Finally, let us assume that $0<\theta_i^n<1$. Observing that $\|y|-|x\|=\max(-|x|+|y|,|x|-|y|)$, we infer that $-\|y|-|x\|\le|y|-|x|$ for all $x,y\in\mathbb{R}$. This inequality in turn implies that

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

where $c_\sharp=\varpi^\sharp\max_{i\in\mathcal{V}}\mathrm{card}(\mathcal{I}(i))$ is a number uniformly bounded with respect to the mesh sequence. Likewise we have

$$1-\alpha_i^n\le 2c_\sharp\theta_i^n.$$

Let k_ψ be the Lipschitz constant of ψ. Then $1-\psi(\alpha_i^n)=\psi(1)-\psi(\alpha_i^n)\le k_\psi(1-\alpha_i^n)$. This in turn implies that

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

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

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

The conclusion follows from Lemma 6.4. □

Example 6.6 (*Shallow Water/Euler Equations*)**.** The above technique can be used to solve the Saint–Venant equations. In this case one can use the water height as smoothness indicator. This technique can also be used to solve the compressible Euler equations. In this case one can use the density as smoothness indicator. Let us denote by U the scalar component that is chosen for the smoothness indicator. Then the scheme (6.1) using the high-order flux (6.4) or (6.9) with the graph viscosity defined in (6.13) with $g(\mathbf{U})=\mathsf{U}$ satisfies the local maximum/minimum principle $\mathsf{U}_i^{\mathrm{H},n+1}\in[\mathsf{U}_i^{\mathrm{m},n},\mathsf{U}_i^{\mathrm{M},n}]$ for all $i\in\mathcal{V}$ under the appropriate CFL condition. This means in particular that the water height (or the density) stays positive. □

Remark 6.7 (*Literature*)**.** The origins of the smoothness-based viscosity can be found in e.g., Jameson et al. [2, Eq. (12)], see also the second formula in the right column of page 1490 in Jameson [34]. A version of Theorem 6.5 for scalar conservation equations is proved in Guermond and Popov [21]. To the best of our knowledge, it seems that

Theorem 6.5 as stated here for hyperbolic systems and generic discretizations is original. The technique presented here shows similarities with that proposed in Burman [35, Thm. 4.1] and Barrenechea et al. [36, Eq. (2.4)–(2.5)]. The quantity $(\alpha_i^n)^p$, $p \geq 2$, is used in [35] to construct a nonlinear viscosity that yields the maximum principle and convergence to the entropy solution for Burgers' equation in one dimension. It is used in [36] for solving linear scalar advection–diffusion equations. □

6.3. Greedy graph viscosity

We continue with a technique entirely based on the observations made in Lemma 6.4, irrespective of any smoothness considerations. As in Section 6.2.2, we specialize the setting by assuming that there is one scalar component of $\mathbf{U}$, say U, for which the source term is zero, i.e., $S \equiv 0$.

Let $i \in \mathcal{V}$ and $n \geq 0$. Let θ_n^n, $\gamma_i^{-,n}$, and $\gamma_i^{+,n}$ be the quantities defined in (6.15)–(6.16) for all $i \in \mathcal{V}$. We recall that Lemma 6.4 is quite general and just requires that $S(\mathbf{U}) \equiv 0$ and $\psi_i^n, \psi_j^n \in [0,1]$. Let us set

$$\psi_i^n := \max\left(1 - 2(1-\gamma_i^n)\min\left(\frac{1}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{(1-\theta_i^n)}\right), 0\right), \tag{6.19}$$

if $\theta_i^n \notin \{0,1\}$ and $\psi_i^n = 1$ otherwise. Then we set

$$d_{ij}^{\mathrm{H},n} := d_{ij}^n \max(\psi_i^n, \psi_j^n), \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i)\backslash\{i\}. \tag{6.20}$$

We now formulate the main result of this section.

Theorem 6.8 (*Greedy Graph Viscosity*)**.** *Consider the scheme (6.1) using either the high-order cG flux (6.4) or the high-order dG flux (6.9) with the graph viscosity defined in (6.20) using the definitions (6.15)–(6.16) with* $\mathsf{U}_i^{\mathrm{m},n}$, $\mathsf{U}_i^{\mathrm{M},n}$ *defined in (6.14). Assume that* $\gamma_i^n \leq 1$*, then the scheme is locally invariant domain preserving for the scalar component* U*: i.e.,* $\mathsf{U}_i^{\mathrm{H},n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$.

Proof. Note first that if $\mathsf{U}_i^{\mathrm{M},n} = \mathsf{U}_i^{\mathrm{m},n}$, then $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ irrespective of the value of d_{ij}^n, which proves the statement. If $\theta_i^n \in \{0,1\}$, then $\psi_i^n = 1$ implies that $d_{ij}^n = d_{ij}^{\mathrm{L},n}\max(1,\psi_j^n) = d_{ij}^{\mathrm{L},n}$ for all $j \in \mathcal{I}(i)\backslash\{i\}$, which again implies that $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. Finally, let us assume that $0 < \theta_i^n < 1$. The definition of ψ_i^n in (6.19) implies that $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}$, which in turn gives $\theta_i^n(\psi_i^n - 1)\frac{1}{2}\gamma_i^{-,n} + (1-\gamma_i^n)(1-\theta_i^n) \geq 0$. This is the condition in Lemma 6.4 that shows that $\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\mathrm{M},n}$, see (6.17). Similarly, we have $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{+,n}}\frac{\theta_i^n}{1-\theta_i^n}$, which gives $(\psi_i^n - 1)(1-\theta_i^n)\frac{1}{2}\gamma_i^{+,n} + (1-\gamma_i^n)\theta_i^n \geq 0$. This is the condition in Lemma 6.4 that shows that $\mathsf{U}_i^{\mathrm{m},n} \leq \mathsf{U}_i^{n+1}$, see (6.18). □

Remark 6.9 (*Small CFL Number*)**.** Note in (6.19) that the quantity ψ_i^n is almost equal to 1 when U_i^n is not a local extremum and the local CFL number γ_i^n is small. This shows that the method becomes greedier as the CFL number decreases; thereby the name of the method. □

Remark 6.10. [Min–Max] The greedy graph viscosity based on (6.19) explicitly involves the bounds $\mathsf{U}_i^{\mathrm{m},n}$ and $\mathsf{U}_i^{\mathrm{M},n}$, whereas the smoothness-based graph viscosity using (6.12) does not. □

6.4. Commutator-based graph viscosity

The objective of this section is to construct the high-order graph viscosity so that the method is entropy consistent and close to be invariant domain preserving. In other words, we do not want to rely on the (yet to be explained) limiting process to enforce entropy consistency. For instance one naive choice consists of using $d_{ij}^{\mathrm{H},n} = 0$, which gives the maximum accuracy for smooth solutions, but as shown in Lemma 4.6 in Guermond and Popov [21] one can construct simple counterexamples with Burgers' equation such that the resulting method is maximum principle preserving, after limiting, but does not converge to the entropy solution. A better option consists of estimating an entropy residual/commutator as suggested in [21, §5.1], [1, §3.4], [5, §6.1].

The key idea consists of measuring the smoothness of an entropy by measuring how well the chain rule, $\nabla\cdot(\boldsymbol{F}(\boldsymbol{u})) = (\nabla\eta(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot(\mathbb{f}(\boldsymbol{u}))$, is satisfied by the discretization at hand. Given an entropy pair $(\eta(\boldsymbol{v}), \boldsymbol{F}(\boldsymbol{v}))$ for (1.1) we set $\eta_i^{\max,n} := \max_{j\in\mathcal{I}(i)}\eta(\mathbf{U}_j^n)$, $\eta_i^{\min,n} := \min_{j\in\mathcal{I}(i)}\eta(\mathbf{U}_j^n)$, $\epsilon_i = \epsilon\max_{j\in\mathcal{I}(i)}|\eta(\mathbf{U}_j^n)|$ and $\Delta\eta_i^n = \max(\frac{1}{2}(\eta_i^{\max,n}-\eta_i^{\min,n}), \epsilon_i)$, then the so-called entropy viscosity, or commutator-based graph viscosity, is defined by setting

$$N_i^n := \sum_{j\in\mathcal{I}(i)}(\boldsymbol{F}(\mathbf{U}_j^n) - (\nabla\eta(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}, \tag{6.21}$$

$$d_{ij}^{\mathrm{H},n} := \min(d_{ij}^{\mathrm{L},n}, \max(\frac{|N_i^n|}{\Delta\eta_i^n}, \frac{|N_j^n|}{\Delta\eta_j^n})). \tag{6.22}$$

The normalization in (6.22) and the choice of entropy are not unique; we refer the reader to [1] where relative entropies are used.

7. Convex limiting

In this section we develop a general limiting framework to preserve convex invariant sets and (more generally) quasiconcave constraints. This work is aligned with the ideas presented in Khobalatte and Perthame [37], Perthame and Qiu [38], Perthame and Shu [39] in the context of finite volume methods. We also refer the reader to Zhang and Shu [40,41], Jiang and Liu [42] for recent/related developments in the context of dG methods. The ideas presented in this section are slightly more general as they naturally extend beyond the Finite Volume/dG methods. The approach that we propose is related to flux-limiting techniques like the flux-corrected transport method by Boris and Book [43], Zalesak [44].

7.1. Quasiconcavity

We have seen in Section 3 that the low-order solution $\mathbf{U}_i^{\mathrm{L},n+1}$ satisfies some "convex bounds" and, in principle, we would like the high-order solution to satisfy these "convex bounds" as well. But, before proceeding any further, we need to define clearly what we mean by convex bounds. We also need to give a precise statement about the bounds that are naturally satisfied by the first-order method. These are the two objectives of the present section and the next one Section 7.2.

In general, the convex bounds mentioned above can be described in terms of upper contour sets of quasiconcave functions and lower contour sets of quasiconvex functions. For the sake of completeness we recall the definitions of quasiconcavity and quasiconvexity.

Definition 7.1 (*Quasiconcavity*)**.** Given a convex set $\mathcal{B}\subset\mathbb{R}^m$, we say that a function $\Psi : \mathcal{B}\to\mathbb{R}$ is quasiconcave if the set $L_\chi(\Psi) := \{\mathbf{U}\in\mathcal{B} \mid \Psi(\mathbf{U})\geq\chi\}$ is convex for any $\chi\in\mathbb{R}$. The sets $L_\chi(\Psi)$ are called upper contour sets.

We are going to make use of the following equivalent definition.

Lemma 7.2 (*Quasiconcavity*)**.** *Let $\mathcal{B}\subset\mathbb{R}^m$ be convex set. A function $\Psi : \mathcal{B}\to\mathbb{R}$ is quasiconcave iff for every finite set $\mathcal{S}\subset\mathbb{N}$, every corresponding set of convex coefficients $\{\lambda_j\}_{j\in\mathcal{S}}$ (i.e., $\sum_{j\in\mathcal{S}}\lambda_j = 1$ and $\lambda_j\geq 0$ for all $j\in\mathcal{S}$), and every corresponding collection of vectors $\{\mathbf{U}_j\}_{j\in\mathcal{S}}$ in $\mathcal{B}$, the following holds true:*

$$\Psi\Big(\sum_{j\in\mathcal{S}}\lambda_j\mathbf{U}_j\Big)\geq\min_{j\in\mathcal{S}}\Psi(\mathbf{U}_j). \tag{7.1}$$

Definition 7.3 (*Quasiconvexity*)**.** A function $\Psi : \mathcal{B}\to\mathbb{R}$ is quasiconvex if $-\Psi$ is quasiconcave.

Note that Jensen's inequality implies that concave/convex functions are quasiconcave/quasiconvex (respectively). The reader is referred to Avriel et al. [45] for further properties of quasiconcave/convex functions. We now give a result that is useful to prove that a function is quasiconcave.

Lemma 7.4. *Let $\mathcal{B}\subset\mathbb{R}^m$ be a convex set. Let $R : \mathcal{B}\to(0,\mathbb{R})$ be a positive function. Let $\Psi : \mathcal{B}\to\mathbb{R}$ and assume that the product $R\Psi$ is concave. Then Ψ is quasiconcave if one of the following two assumptions is satisfied: (i) R is affine or (ii) R is convex and Ψ is nonnegative.*

Proof. Let $\{\lambda_j\}_{j\in\mathcal{S}}$ be a set of convex coefficients. Let $\{\mathbf{U}_j^n\}_{j\in\mathcal{S}}$ be members of $\mathcal{B}$. Let us set $\chi := \min_{j\in\mathcal{S}} \Psi(\mathbf{U}_j)$. Let $\Phi(\mathbf{U}) := R(\mathbf{U})(\Psi(\mathbf{U})-\chi)$. Notice that if R is affine, or if R is convex and Ψ is nonnegative, then $-\chi R(\mathbf{U})$ is concave. As a result, Φ is concave since $R(\mathbf{U})\Psi(\mathbf{U})$ and $-\chi R(\mathbf{U})$ are both concave and the sum of two concave functions is concave (this may not be the case for the sum of quasiconcave functions). Notice also that $\min_{j\in\mathcal{S}} \Phi(\mathbf{U}_j) \geq 0$ because $R \geq 0$ and $\min_{j\in\mathcal{S}} \Psi(\mathbf{U}_j) - \chi \geq 0$. Hence

$$\Phi\Big(\sum_{j\in\mathcal{S}}\lambda_j\mathbf{U}_j\Big) = R\Big(\sum_{j\in\mathcal{S}}\lambda_j\mathbf{U}_j\Big)\Big(\Psi\Big(\sum_{j\in\mathcal{S}}\lambda_j\mathbf{U}_j\Big)-\chi\Big) \geq \sum_{j\in\mathcal{S}}\lambda_j\,\Phi(\mathbf{U}_j) \geq 0.$$

This in turn implies that $\Psi(\sum_{j\in\mathcal{S}}\lambda_j\mathbf{U}_j) \geq \chi = \min_{j\in\mathcal{S}}\Psi(\mathbf{U}_j)$, which proves the assertion owing to Lemma 7.2. □

Example 7.5 (*Entropy*)**.** Let $\eta : \mathcal{B} \to \mathbb{R}$ be any entropy for (1.1) (recall that entropies are convex by definition), then $\Psi(\mathbf{U}) = -\eta(\mathbf{U})$ is quasiconcave. □

Example 7.6 (*Specific Entropy*)**.** Let $\eta : \mathcal{B} \to \mathbb{R}$ be any entropy for (1.1). Let $R : \mathcal{B} \to (0,\infty)$ be a positive linear function, then Lemma 7.4 implies that $\Psi(\mathbf{U}) = -\eta(\mathbf{U})/R(\mathbf{U})$ is quasiconcave. One can think of this function as a specific entropy in the case of the shallow water equations ($R(\mathbf{U})$ is the water height), or the case of the Euler equations ($R(\mathbf{U})$ is the density). □

Let us now give examples of quasiconcave functionals in the context of the compressible Euler equations with an arbitrary equation of state. The conserved variables in this case are $\mathbf{U} := (\rho, \boldsymbol{m}, E)^\intercal$.

Example 7.7 (*Density*)**.** We set $\mathcal{B} := \mathbb{R}^{d+2}$, $\Psi(\mathbf{U}) := \rho$. The functional $\Psi : \mathcal{B} \to \mathbb{R}$ is linear, hence it is quasiconcave. Note the following functional $\Psi(\mathbf{U}) = -\rho$ is also quasiconcave. □

Example 7.8 (*Total Energy*)**.** We set $\mathcal{B} := \mathbb{R}^{d+2}$, $\Psi(\mathbf{U}) := E$. The functional $\Psi : \mathcal{B} \to \mathbb{R}$ is linear, hence it is quasiconcave. Note the following functional $\Psi(\mathbf{U}) = -E$ is also quasiconcave. □

Example 7.9 (*Internal Energy*)**.** We set $\mathcal{B} := \{\mathbf{U} = (\rho, \boldsymbol{m}, E)^\top \in \mathbb{R}^m \mid \rho > 0\}$ and introduce the internal energy $\varepsilon(\mathbf{U}) := E - \frac{\boldsymbol{m}^2}{2\rho}$. A direct computation shows that the functional $\varepsilon : \mathcal{B} \to \mathbb{R}$ has a negative semi-definite Hessian for every equation of state, thereby proving that ε is concave, hence quasiconcave. □

Let us now illustrate the use of Lemma 7.4 with $R(\mathbf{U}) = \rho$.

Example 7.10 (*Specific Internal Energy*)**.** Let $\mathcal{B} := \{\mathbf{U} = (\rho, \boldsymbol{m}, E)^\top \in \mathbb{R}^m \mid \rho > 0\}$, and introduce the specific internal energy $e(\mathbf{U}) := \frac{\varepsilon(\mathbf{U})}{\rho} = \frac{E}{\rho} - \frac{\boldsymbol{m}^2}{2\rho^2}$. Clearly $R(\mathbf{U}) := \rho$ is convex; moreover, $\Phi(\mathbf{U}) := R(\mathbf{U})e(\mathbf{U}) = E - \frac{\boldsymbol{m}^2}{2\rho} = \varepsilon(\mathbf{U})$ is the internal energy, which we know is a concave function for any equation of state. Hence we conclude from Lemma 7.4 that the specific internal energy is quasiconcave for any equation of state. Notice in passing that this argument proves that the set $\{\mathbf{U} := (\rho, \boldsymbol{m}, E)^\top \mid \rho \geq \rho_0, e(\mathbf{U}) \geq e_0\}$ is convex for any $\rho_0, e_0 \in (0,\infty)$. □

Example 7.11 (*Generalized Specific Entropies*)**.** We set $\mathcal{B} := \{\mathbf{U} \in \mathbb{R}^m \mid \rho > 0, e(\mathbf{U}) > 0\}$. Let $\eta : \mathcal{B} \to \mathbb{R}$ be a generalized entropy as defined in Harten [46, Eq. (2.10a)], Harten et al. [47, Thm. 2.1]. Then using Lemma 7.4 with $R(\mathbf{U}) = \rho$ and $\Psi(\mathbf{U}) = \eta(\mathbf{U})/R(\mathbf{U})$, we conclude that the specific entropy $s(\mathbf{U}) := \rho^{-1}\eta(\mathbf{U})$ is quasiconcave. Note in passing that we have proved that the set $\{\mathbf{U} := (\rho, \boldsymbol{m}, E)^\top \mid \rho > \rho_0, e(\mathbf{U}) > \rho_0, s(\mathbf{U}) > s_0\}$ is convex for any $\rho_0, e_0 > 0$ and any $s_0 \in \mathbb{R}$. We refer the reader to Theorem 8.2.2 from Serre [48] for other properties of this set. □

Example 7.12 (*Kinetic Energy*)**.** We set $\mathcal{B} := \{\mathbf{U} = (\rho, \boldsymbol{m}, E)^\top \in \mathbb{R}^m \mid \rho > 0\}$. Let $\Psi(\mathbf{U}) = -\frac{1}{2}\rho^{-1}\boldsymbol{m}^2$ be the (negative) kinetic energy. It is clear that $\Phi(\mathbf{U}) = -\frac{1}{2}\boldsymbol{m}^2$ is concave, then using Lemma 7.4 with $R(\mathbf{U}) = \rho$, we conclude that the (negative) kinetic energy is quasiconcave. □

We finish with a result that is useful to transform quasiconcave functionals.

Lemma 7.13. *Let $\Psi : \mathcal{B} \to \mathbb{R}$ be a quasiconcave function. Let $L : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function, then $L \circ \Psi$ is quasiconcave.*

Proof. Let us use the characterization (7.1). Since L is nondecreasing, we have $L \circ \Psi(\sum_{j\in\mathcal{S}} \lambda_j \mathbf{U}_j) \geq L(\min_{j\in\mathcal{S}} \Psi(\mathbf{U}_j))$. Let $k \in \mathcal{S}$ be such that $\Psi(\mathbf{U}_k) := \min_{j\in\mathcal{S}} \Psi(\mathbf{U}_j)$. Then, for any $j \in \mathcal{S}$, we have $\Psi(\mathbf{U}_k) \leq \Psi(\mathbf{U}_j)$, which implies that $L\circ\Psi(\mathbf{U}_k) \leq L\circ\Psi(\mathbf{U}_j)$. Hence $L(\min_{j\in\mathcal{S}} \Psi(\mathbf{U}_j)) = L(\Psi(\mathbf{U}_k)) = \min_{j\in\mathcal{S}} L(\Psi(\mathbf{U}_j))$. In conclusion $L \circ \Psi(\sum_{j\in\mathcal{S}} \lambda_j \mathbf{U}_j) \geq \min_{j\in\mathcal{S}} L \circ \Psi(\mathbf{U}_j)$, which proves the assertion. □

Example 7.14 (*Specific Entropy*)**.** Let us illustrate the use of Lemma 7.13 with the compressible Euler equations, and, to simplify the argument, let assume that the equation of state is the γ-law. Consider the physical specific entropy $\Psi(\mathbf{U}) = \frac{1}{\gamma-1}\log(\varepsilon(\mathbf{U})\rho^{-\gamma})$, where $\varepsilon(\mathbf{U})$ is the internal energy. This function is quasiconcave owing to Lemma 7.4 with $R(\mathbf{U}) = \rho$, since $\rho\,\Psi(\mathbf{U})$ is known to be concave. Then using Lemma 7.13 we conclude that $\tilde{\Psi}(\mathbf{U}) = \varepsilon(\mathbf{U})\rho^{-\gamma}$ is quasiconcave. □

7.2. Bounds

In this section we define the bounds that we are going to use to limit the high-order solution. The following result will play a key role in the rest of the paper, since it tells us precisely what are the "convex bounds" that the low-order solution produced by the GMS-GV scheme satisfies.

Lemma 7.15 (*Natural Bounds on the GMS-GV Scheme*)**.** *Let $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ be a convex set and $\Psi : \mathcal{B} \to \mathbb{R}$ be a quasiconcave functional. Let $n \geq 0$, $i \in \mathcal{V}$, and assume that $1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i} \geq 0$ and $2\tau \leq \tau_0$. Assume that $\mathbf{U}_j^n \in \mathcal{B}$ for all $j \in \mathcal{I}(i)$. Let $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)}$ be the auxiliary states defined in (3.8). Consider the following quantity:*

$$\Psi_i^{\min} := \min(\Psi(\mathbf{U}_i^n + 2\tau\, \boldsymbol{S}(\mathbf{U}_i^n)),\ \min_{j\in\mathcal{I}(i)} \Psi(\overline{\mathbf{U}}_{ij}^n)). \tag{7.2}$$

Then, the first-order update $\mathbf{U}_i^{\mathrm{L},n+1}$ computed with the GMS-GV scheme (see (3.4) plus (3.10)) is in $\mathcal{B}$ and satisfies the following inequality:

$$\Psi(\mathbf{U}_i^{\mathrm{L},n+1}) \geq \Psi_i^{\min}. \tag{7.3}$$

Proof. Using the assumptions, $1 + 4\tau\frac{d_{ii}^{\mathrm{L},n}}{m_i} \geq 0$ and $2\tau \leq \tau_0$, we first observe that (3.11) shows that $\mathbf{U}_i^{\mathrm{L},n+1}$ is a convex combination of the states $\mathbf{U}_i^n + 2\tau\,\boldsymbol{S}(\mathbf{U}_i^n)$ and $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{I}(i)\setminus\{i\}}$ which are all in $\mathcal{B}$; hence $\mathbf{U}_i^{\mathrm{L},n+1}$ is in $\mathcal{B}$. Then the conclusion follows readily by using the quasiconcavity property (7.1). □

Remark 7.16 (*Quasiconcavity vs. Quasiconvexity*)**.** Since any quasiconvex function can be transformed into a quasiconcave function by a sign change, the above lemma gives $\Psi(\mathbf{U}_i^{\mathrm{L},n+1}) \leq \Psi_i^{\max} := \max(\Psi(\mathbf{U}_i^n + 2\tau\,\boldsymbol{S}(\mathbf{U}_i^n)), \max_{j\in\mathcal{I}(i)} \Psi(\overline{\mathbf{U}}_{ij}^n))$ for any quasiconvex function $\Psi : \mathcal{B} \subset \mathcal{A} \to \mathbb{R}$. Therefore, in order to alleviate the language, we will henceforth refrain from mentioning quasiconvexity and will formulate every "convex bounds" in terms of quasiconcave functionals only. □

Remark 7.17 (*Invariant Set vs. Local Bound*)**.** Notice that Lemma 7.15 contains two statements that are of different nature. The first one is an invariant domain property: $(\mathbf{U}_j^n \in \mathcal{B},\ \forall j \in \mathcal{I}(i)) \Rightarrow (\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{B})$. Since $\mathcal{B}$ does not depend on $i \in \mathcal{V}$, this local assertion can be reformulated into a global statement $(\mathbf{U}_i^n \in \mathcal{B},\ \forall i \in \mathcal{V}) \Rightarrow (\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{B},\ \forall i \in \mathcal{V})$. The second statement $\Psi(\mathbf{U}_i^{\mathrm{L},n+1}) \geq \Psi_i^{\min}$ is a local bound that can be viewed as a local "generalized minimum principle". This bound cannot be made uniform; it is local in time and space, since $\Psi_i^{\min}$ depends on i and n. □

Remark 7.18 (*Relaxation*)**.** The reader must be aware that in general the bound $\Psi_i^{\min}$ defined in (7.2) must be slightly relaxed in order to go beyond second-order accuracy in space in the L^1-norm. We refer the reader to Section 7.6 for implementation details on relaxation techniques. □

7.3. Abstract framework

In the sections Sections 6.1.2–6.1.4 we have seen that most high-order methods can be written in the algebraic form

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{H},n+1}-\mathbf{U}_i^n)+\sum_{j\in\mathcal{I}(i)}\mathbf{F}_{ij}^{\mathrm{H},n}=m_i\boldsymbol{S}(\mathbf{U}_i^n), \tag{7.4}$$

with $\mathbf{F}_{ij}^{\mathrm{H},n}\in\mathbb{R}^m$ satisfying the skew-symmetry constraint $\mathbf{F}_{ij}^{\mathrm{H},n}=-\mathbf{F}_{ij}^{\mathrm{H},n}$ for all $j\in\mathcal{I}(i)$ (whether we use the consistent mass matrix for the discretization for the time derivative or not), where the superscript $^{\mathrm{H}}$ denotes high-order. Subtracting (3.7) from (7.4) and reorganizing we get $m_i\mathbf{U}_i^{\mathrm{H},n+1}=m_i\mathbf{U}_i^{\mathrm{L},n+1}+\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\tau(\mathbf{F}_{ij}^{\mathrm{L},n}-\mathbf{F}_{ij}^{\mathrm{H},n})$. This expression can be rewritten into the following important identity:

$$m_i\mathbf{U}_i^{\mathrm{H},n+1}=m_i\mathbf{U}_i^{\mathrm{L},n+1}+\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\mathbf{A}_{ij}^n, \tag{7.5}$$

where $\mathbf{A}_{ij}^n:=\tau(\mathbf{F}_{ij}^{\mathrm{L},n}-\mathbf{F}_{ij}^{\mathrm{H},n})\in\mathbb{R}^m$. The *convex limiting* technique to be explained in the next section relies heavily on (7.5). Note that (by construction) we have that $\mathbf{A}_{ij}^n=-\mathbf{A}_{ji}^n$, which means that $\sum_{i\in\mathcal{V}}m_i\mathbf{U}_i^{\mathrm{H},n+1}=\sum_{i\in\mathcal{V}}m_i\mathbf{U}_i^{\mathrm{L},n+1}$; that is to say, the high-order and the low-order solution have the same mass whether the source term $\boldsymbol{S}$ is present or not.

7.4. Convex limiting

Without loss of generality, we consider a family of quasiconcave functionals $\{\Psi_i\}_{i\in\mathcal{V}}$, $\Psi_i:\mathcal{B}\to\mathbb{R}$ where $\mathcal{B}\subset\mathbb{R}^m$ is a convex set and $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1})\geq 0$ for each $i\in\mathcal{V}$. Our goal is to modify the high-order update so that the modified high-order update satisfies the same quasiconcave constraints as the low-order solution and has the same mass as the high-order update.

Taking inspiration from the flux-corrected transport methodology, we introduce symmetric limiting parameters $\ell_{ij}=\ell_{ji}\in[0,1]$, $i,j\in\mathcal{V}$, and we define the limited solution $\mathbf{U}_i^{n+1}$ as follows:

$$m_i\mathbf{U}_i^{n+1}:=m_i\mathbf{U}_i^{\mathrm{L},n+1}+\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\ell_{ij}\mathbf{A}_{ij}^n. \tag{7.6}$$

Notice that $\mathbf{U}_i^{n+1}=\mathbf{U}_i^{\mathrm{L},n+1}$ if $\ell_{ij}=0$ for all $j\in\mathcal{I}(i)\backslash\{i\}$ and $\mathbf{U}_i^{n+1}=\mathbf{U}_i^{\mathrm{H},n+1}$ if $\ell_{ij}=1$ for all $j\in\mathcal{I}(i)\backslash\{i\}$; hence, $\Psi_i(\mathbf{U}_i^{n+1})\geq 0$ when $\ell_{ij}=0$. Our goal is to find a set of coefficients ℓ_{ij} as close to 1 as possible so that $\Psi_i(\mathbf{U}_i^{n+1})\geq 0$.

Lemma 7.19 (*Conservation*)**.** *The limiting process is conservative for any choice of coefficients ℓ_{ij} if $\ell_{ij}=\ell_{ji}$ for any $j\in\mathcal{I}(i)\backslash\{i\}$.*

Proof. the skew-symmetry of $\mathbf{A}_{ij}^n$ together with the symmetry of the limiter ℓ_{ij} implies that $\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\ell_{ij}\mathbf{A}_{ij}^n=\mathbf{0}$; therefore $\sum_{i\in\mathcal{V}}m_i\mathbf{U}_i^{n+1}=\sum_{i\in\mathcal{V}}m_i\mathbf{U}_i^{\mathrm{L},n+1}$. □

The expression (7.6) goes back to the flux-corrected transport framework pioneered by Boris and Book [43], Zalesak [44]. The reader can further explore some current developments for flux-corrected transport methods in the books Kuzmin et al. [49,50]. At this point we depart from the existing flux-corrected transport literature and follow [1] instead. We rewrite (7.6) as follows:

$$\mathbf{U}_i^{n+1}=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\lambda_j(\mathbf{U}_i^{\mathrm{L},n+1}+\ell_{ij}\mathbf{P}_{ij}^n),\qquad\text{with}\qquad\mathbf{P}_{ij}^n:=\frac{1}{m_i\lambda_j}\mathbf{A}_{ij}^n, \tag{7.7}$$

where $\{\lambda_j\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ is any set of strictly positive convex coefficients (see Remark 7.22), i.e., $\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\lambda_j=1$, $\lambda_j>0$ for all $j\in\mathcal{I}(i)\backslash\{i\}$. The following two lemmas should convince the reader that it is possible to estimate ℓ_{ij} efficiently by doing one-dimensional line-searches only.

Lemma 7.20. *Let $\Psi_i(\boldsymbol{u}) : \mathcal{B} \to \mathbb{R}$ be a quasiconcave function. Assume that the limiting parameters $\ell_{ij} \in [0,1]$ are such that $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n) \geq 0$, for all $j \in \mathcal{I}(i)\backslash\{i\}$, then the following inequality holds true:*

$$\Psi_i\Big(\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \lambda_j(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n)\Big) \geq 0.$$

Proof. Let $L_0(\Psi_i) := \{\mathbf{U} \in \mathcal{B} \mid \Psi_i(\mathbf{U}) \geq 0\}$. By definition all the limited states $\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n$ are in $L_0(\Psi_i)$ for all $j \in \mathcal{I}(i)\backslash\{i\}$. Since Ψ_i is quasiconcave, the upper contour set $L_0(\Psi_i)$ is convex. Hence, the convex combination $\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \lambda_j(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n)$ is in $L_0(\Psi_i)$, i.e., $\Psi_i\big(\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \lambda_j(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n)\big) \geq 0$, which concludes the proof. □

Theorem 7.21. *For every $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$, let ℓ_j^i be defined by*

$$\ell_j^i = \begin{cases} 1 & \text{if } \Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \mathbf{P}_{ij}^n) \geq 0, \\ \max\{\ell \in [0,1] \mid \Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n) \geq 0\} & \text{otherwise.} \end{cases} \tag{7.8}$$

The following two statements hold true: (i) $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n) \geq 0$ *for every* $\ell \in [0, \ell_j^i]$; (ii) *Setting* $\ell_{ij} = \min(\ell_j^i, \ell_i^j)$, *we have* $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n) \geq 0$ *and* $\ell_{ij} = \ell_{ji}$.

Proof. (i) First, if $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \mathbf{P}_{ij}^n) \geq 0$ we observe that $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n) \geq 0$ for any $\ell \in [0,1]$ because $\mathbf{U}_i^{\mathrm{L},n+1} \in L_0(\Psi_i)$, $\mathbf{U}_i^{\mathrm{L},n+1} + \mathbf{P}_{ij}^n \in L_0(\Psi_i)$ and $L_0(\Psi_i)$ is convex. Second, if $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \mathbf{P}_{ij}^n) < 0$, we observe that the segment $\{\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n \mid \ell \in [0,1]\}$ crosses the level set $\partial\{\Psi_i(\mathbf{U}) \geq 0\}$ because $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1}) \geq 0$. Notice also that the quasiconcavity of Ψ_i implies that ℓ_j^i is uniquely defined since the segment $\{\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n \mid \ell \in [0,1]\}$ can cross the level set $\partial\{\Psi_i(\mathbf{U}) \geq 0\}$ only once; moreover, for any $\ell \in [0, \ell_j^i]$ we have $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n) \geq 0$ because $\mathbf{U}_i^{\mathrm{L},n+1} \in L_0(\Psi_i)$, $\mathbf{U}_i^{\mathrm{L},n+1} + \ell_j^i\mathbf{P}_{ij}^n \in L_0(\Psi_i)$ and $L_0(\Psi_i)$ is convex. (ii) Since $\ell_{ij} = \min(\ell_j^i, \ell_i^j) \leq \ell_j^i$, the above construction implies that $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n) \geq 0$. Note finally that $\ell_{ij} = \min(\ell_j^i, \ell_i^j) = \ell_{ji}$. □

Remark 7.22 (*Choice of Convex Coefficients*)**.** There are infinitely many possible choices for the strictly positive convex coefficients $\{\lambda_j\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ in (7.7). Note that it is even possible to choose a different set $\{\lambda_j\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ for each $i \in \mathcal{V}$ without affecting the results presented in this paper. We have not made any theoretical attempt to exploit these additional degrees of freedom in order to optimize the convex limiting technique. All the computations reported in Guermond et al. [1] have been done with the simplest choice $\lambda_j := \frac{1}{\mathrm{card}(\mathcal{I}(i))-1}$ for all $j \in \mathcal{I}(i)\backslash\{i\}$ for all $i \in \mathcal{V}$. Other choices have been explored computationally but none turned out to be more efficient than the others. It might be interesting though to explore this question further; for instance, other choices of convex coefficients could help preserve some symmetries. □

Remark 7.23 (*Multiple Limiting*)**.** In general we have to consider families of quasiconcave functionals $\{\{\Psi_i\}_{i\in\mathcal{V}}\}_{l\in\mathcal{L}}$, $\Psi_i^l : \mathcal{B}^l \to \mathbb{R}$, where $\mathcal{B}^l \subset \mathbb{R}^m$ is the convex admissible set of the functional Ψ_i^l. The list $\mathcal{L}$ describes the nature of the functionals; this list could encompass any of the functionals shown in Examples 7.5 to 7.12. The list $\mathcal{L}$ is sometimes ordered in the sense that $\mathcal{B}^{l'} \subset \mathcal{B}^l$ if $l' \geq l$. Let us illustrate this concept with the compressible Euler equations. Usually one starts with $B^1 = \mathbb{R}^m$ to enforce a local minimum principle on the density (which implies positivity of the density). We can also take $\mathcal{B}^2 = \mathbb{R}$ to enforce a local maximum principle on the density by using $\Psi(\mathbf{U}) = -\rho$. Then we can consider $B^3 = \{\mathbf{U} \in \mathcal{B}^1 \mid \rho > 0\}$ to enforce a local minimum principle on the (specific) internal energy (which implies positivity of the (specific) internal energy). We finally set $B^4 = \{\mathbf{U} \in \mathcal{B}^2 \mid e(\mathbf{U}) > 0\}$ to enforce a local minimum principle on the specific entropy. □

The following result is the main conclusion of the paper.

Theorem 7.24. *Let $\{\Psi^l : \mathcal{B}^l \to \mathbb{R}\}_{l\in\mathcal{L}}$, be a family of quasiconcave functionals, where the sets $\mathcal{B}^l \subset \mathbb{R}^m$ are convex for all $l \in \mathcal{L}$. Let $\mathcal{B} : \{\mathbf{U} \in \mathbb{R}^m \mid \Psi^l(\mathbf{U}) \geq 0,\ \forall l \in \mathcal{L}\}$. Let $n \geq 0$. Assume that $\min_{i\in\mathcal{V}}(1 + 4\frac{d_{ii}^{\mathrm{L},n}}{m_i}) \geq 0$ and $\tau \leq 2\tau_0$. Consider the quasiconcave functionals $\{\Psi_i^l\}_{i\in\mathcal{V},l\in\mathcal{L}}$ defined by $\Psi_i^l(\mathbf{U}) = \Psi^l(\mathbf{U}) - \Psi_i^{l,\min}$ with $\Psi_i^{l,\min}$ defined in (7.2). Let*

$\ell_j^{i,l}$ be the limiter computed by using (7.8) for any $i \in \mathcal{V}$, $j \in \mathcal{I}(i)\backslash\{i\}$, $l \in \mathcal{L}$. Let $\ell_{ij} = \min(\min_{l\in\mathcal{L}} \ell_j^{i,l}, \min_{l\in\mathcal{L}} \ell_i^{j,l})$. Let $\mathbf{U}_i^{n+1}$ be defined in (7.7). Assume that $\mathcal{B}$ is an invariant set for (1.1), then $\mathcal{B}$ is an invariant domain, i.e., $(\mathbf{U}_i^n \in \mathcal{B},\ \forall i \in \mathcal{V}) \Rightarrow (\mathbf{U}_i^{n+1} \in \mathcal{B},\ \forall u \in \mathcal{V})$.

Proof. Notice first that $\mathcal{B}$ is convex since it is the intersection of convex sets $\mathcal{B} = \bigcap_{l\in\mathcal{L}}\{\mathbf{U} \in \mathbb{R}^m \mid \psi^l(\mathbf{U}) \geq 0\}$. Since $\mathcal{B}$ is a convex invariant set for (1.1), the CFL assumption together with Theorem 3.6 implies that $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{B}$ for all $i \in \mathcal{V}$. Then Theorem 7.21 can be applied because $\Psi_i^l(\mathbf{U}^{\mathrm{L},n+1}) \geq 0$. This theorem then implies that $\Psi^l(\mathbf{U}_i^{n+1}) \geq \Psi_i^{l,\min}$ for all $l \in \mathcal{L}$. Moreover, $\mathbf{U}_i^n + 2\tau \boldsymbol{S}(\mathbf{U}_i^n) \in \mathcal{B}$ and $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}$, then owing to the CFL assumption and definition (7.2), this implies that $\Psi_i^{l,\min} \geq 0$. In conclusion $\Psi^l(\mathbf{U}_i^{n+1}) \geq 0$ for all $l \in \mathcal{L}$, which implies that $\mathbf{U}_i^{n+1} \in \mathcal{B}$. □

Remark 7.25 (*SSP Extension*)**.** Owing to Remark 5.6, Theorem 7.24 extends to any SSP RK time stepping provided the limiting is done at the end of each elementary forward Euler substep. □

7.5. Implementation details

The objective of this section is to give further details on the convex limiting technique introduced above in order to help the reader to implement it.

7.5.1. Pseudocode of the limiting algorithm

Given a set of quasi-convex functionals $\{\Psi_i\}_{i\in\mathcal{V}}$, $\Psi_i : \mathcal{B} \to \mathbb{R}$, such that $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1}) \geq 0$ with convex set $\mathcal{B}$, Algorithm 1 enforces the quasi-concave constraints $\Psi_i(\mathbf{U}_i^{n+1}) \geq 0$ for each $i \in \mathcal{V}$. This pseudocode attempts to reflect as accurately as possible the way convex limiting is coded in practice. Basically, convex limiting is done in two loops over the set of the global degrees of freedom $\mathcal{V}$: the first loop (lines 1 to 14) computes the matrix ℓ_j^i in general non-symmetric form; the second loop (lines 15 to 19) computes the final symmetric limiter ℓ_{ij}. Lemma 7.20 explains why the limiters ℓ_j^i estimated in the first loop are large enough to enforce the constraint $\Psi_i(\mathbf{U}_i^{n+1}) \geq 0$ for each $i \in \mathcal{V}$. Theorem 7.21 explains why the symmetrization (shrinkage) of the limiters done in the second loop still produces limiters compatible with these constraints. We have found that initializing ℓ_i with the lines 2–6 instead of setting $\ell_i = 1$ reduces the number of times the line-search in line 11 is executed.

Algorithm 1 Convex Limiting

1: **for** $i \in \mathcal{V}$ **do**
2: **if** $\Psi_i(\mathbf{U}_i^{\mathrm{H},n+1}) \geq 0$ **then**
3: $\ell_i := 1$
4: **else**
5: $\ell_i := \max\{\ell \in [0,1] \mid \Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^{\mathrm{L},n+1})) \geq 0\}$
6: **end if**
7: **for** $j \in \mathcal{I}(i)\backslash\{i\}$ **do**
8: **if** $\Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell_i \mathbf{P}_{ij}^n) \geq 0$ **then**
9: $\ell_j^i := \ell_i$
10: **else**
11: $\ell_j^i := \max\{\ell \in [0,\ell_i] \mid \Psi_i(\mathbf{U}_i^{\mathrm{L},n+1} + \ell\, \mathbf{P}_{ij}^n) \geq 0\}$
12: **end if**
13: **end for**
14: **end for**
15: **for** $i \in \mathcal{V}$ **do**
16: **for** $j \in \mathcal{I}(i)\backslash\{i\}$ **do**
17: $\ell_{ij} := \min\{\ell_j^i, \ell_i^j\}$
18: **end for**
19: **end for**

7.5.2. Transforming $\Psi_i(\mathbf{U}) \geq 0$ into a quadratic constraint

As mentioned in the previous subsection, the line-search invoked in line 5 and line 11 of Algorithm 1 could be computationally expensive. However, it happens sometimes that the constraint of interest $\Psi_i(\mathbf{U}) \geq 0$ can be transformed into $\tilde{\Psi}_i(\mathbf{U}) \geq 0$ where $\tilde{\Psi}_i$ is a quadratic function, not necessarily quasi-concave. In this case it is possible to design a very efficient algorithm for the line-search.

Example 7.26 (*Internal Energy*)**.** To illustrate the above statement, let us consider the compressible Euler equations with some arbitrary equation of state. Let us set $\mathcal{B} = \{\mathbf{U} := (\rho, \boldsymbol{m}, E)^\top \mid \rho > 0\}$, $\varepsilon(\mathbf{U}) := E - \frac{|\boldsymbol{m}|^2_{\ell^2}}{2\rho}$ (internal energy), and $\Psi_i(\mathbf{U}) := \varepsilon(\mathbf{U}) - \varepsilon_i^{\min}$. We have seen in Example 7.9 that $\Psi_i : \mathcal{B} \to \mathbb{R}$ is quasiconcave (actually $\Psi_i : \mathcal{B} \to \mathbb{R}$ is concave). It is clear that one has $\Psi_i(\mathbf{U}) \geq 0$ iff $\tilde{\Psi}_i(\mathbf{U}) := \rho\varepsilon(\mathbf{U}) - \rho\varepsilon_i^{\min} \geq 0$ for all $\mathbf{U} \in \mathcal{B}$. Notice that $\rho\varepsilon(\mathbf{U}) = E\rho - \frac{1}{2}\boldsymbol{m}^2$ and $\rho\varepsilon_i^{\min}$ are quadratic polynomials of the conserved variables; hence, $\tilde{\Psi}_i(\mathbf{U})$ is quadratic (but a simple computation shows also that $\tilde{\Psi}_i$ is not quasiconcave). In conclusion, instead of doing the line-search with $\Psi_i(\mathbf{U}) := \varepsilon(\mathbf{U}) - \varepsilon_i^{\min}$, one can do the line-search with the quadratic functional $\tilde{\Psi}_i(\mathbf{U}) = \rho\varepsilon(\mathbf{U}) - \rho\varepsilon_i^{\min}$. □

We now state an abstract result that formalizes the above observation.

Lemma 7.27. *Let $\Psi : \mathcal{B} \subset \mathbb{R}^m \to \mathbb{R}$. Let $\mathbf{U}^{\mathrm{L}} \in \mathcal{B}$ and assume that $\Psi(\mathbf{U}^{\mathrm{L}}) \geq 0$. Let $\tilde{\Psi} : \mathcal{B} \to \mathbb{R}$, let $\mathbf{P} \in \mathbb{R}^m$, and assume that there is $\ell^{\max} \in [0, 1]$ such that $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ iff $\tilde{\Psi}(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^{\max}]$. Assume that $\tilde{\Psi}$ is quadratic and let $a := \frac{1}{2}\mathbf{P}^\top \mathrm{D}^2\tilde{\Psi}\mathbf{P}$, $b := \mathrm{D}\tilde{\Psi}(\mathbf{U}^{\mathrm{L}})\cdot\mathbf{P}$ and $c := \tilde{\Psi}(\mathbf{U}^{\mathrm{L}})$. Let $\ell^{\min}$ be the smallest positive root of the equation $a\ell^2 + b\ell + c = 0$, with the convention that $\ell^{\min} := 1$ if the equation has no positive root. Let $\ell^i_j := \min(\ell^{\min}, \ell^{\max})$, then $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^i_j]$.*

Proof. Let us first observe that $\tilde{\Psi}(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) = a\ell^2 + b\ell + c\ell =: g(\ell)$ for all $\ell \in [0, \ell^{\max}]$; hence, $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ iff $g(\ell) \geq 0$ for all $\ell \in [0, \ell^{\max}]$. If there is no positive root to the equation $a\ell^2 + b\ell + c = 0$, then the sign of $g(\ell)$ over $[0, \infty)$ is constant. The assumption $g(0) = c := \Psi(\mathbf{U}^{\mathrm{L}}) \geq 0$, implies that $g(\ell) \geq 0$ for all $\ell \in [0, \infty)$. That is, $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^{\max}]$, and in particular this is true for all $\ell \in [0, \ell^i_j]$ since in this case $\ell^i_j := \min(\ell^{\min}, \ell^{\max}) \leq \ell^{\max}$. Otherwise, if there is at least one positive root to the equation $g(\ell) = 0$, then denoting by $\ell^{\min}$ the smallest positive root, we have $g(\ell) \geq 0$ for all $\ell \in [0, \ell^{\min}]$ (if not, there would exist $\ell_1 \in (0, \ell^{\min})$ s.t. $g(\ell_1) < 0$ and the intermediate value theorem would imply the existence a root $\ell^* \in (0, \ell_1)$ which contradicts that $\ell^{\min}$ is the smallest positive root). This argument implies again that $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^i_j]$. □

Example 7.28 (*Kinetic Energy*)**.** Coming back to the compressible Euler equations or the shallow water equations, the above technique can be applied to enforce the local maximum principle on the kinetic energy $\Psi_i(\mathbf{U}) \geq 0$, with $\Psi_i(\mathbf{U}) = \Psi(\mathbf{U}) - \Psi_i^{\min}$ and $\Psi(\mathbf{U}) = -\frac{1}{2}\rho^{-1}\boldsymbol{m}^2$ with $\mathcal{B} = \{\mathbf{U} := (\rho, \boldsymbol{m}, E)^\mathsf{T} \mid \rho > 0\}$. (Notice that because of the sign convention $\Psi_i^{\min}$ is the *maximum* of the kinetic energy over the states $\{\overline{\mathbf{U}}^n_{ij}\}_{j\in\mathcal{I}(i)}$ and the state $\mathbf{U}^n_i + 2\tau\boldsymbol{S}(\mathbf{U}^n_i)$. Hence the constraint $\Psi_i(\mathbf{U}) \geq 0$ amounts to enforcing a local maximum principle on the kinetic energy.) We have shown in Example 7.12 that Ψ_i is quasiconcave. In this case Lemma 7.27 can be applied with the functional $\tilde{\Psi}_i(\mathbf{U}) = \rho\,\Psi_i(\mathbf{U}) = -\frac{1}{2}\boldsymbol{m}^2 - \rho\,\Psi_i^{\min}$ which is clearly quadratic. Note that $\tilde{\Psi}_i(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ iff $\Psi_i(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ provided $\rho(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$. Hence before applying Lemma 7.27, one must compute the limiter $\ell^{\max}$, which depends on $\mathbf{U}^{\mathrm{L}}$ and $\mathbf{P}$, such that $\rho(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^{\max}]$. This technique has been introduced in [5, §6.4] in the context of the shallow water equations. □

Remark 7.29 (*Parameter $\ell^{\max}$*)**.** The purpose of the parameter $\ell^{\max}$ appearing in the statement of Lemma 7.27 is to ascertain that stating that $\Psi(\mathbf{U} + \ell\mathbf{P}) \geq 0$ is equivalent to stating that $\tilde{\Psi}(\mathbf{U} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^{\max}]$. The limiter $\ell^{\max}$ depends on $\mathbf{U}^{\mathrm{L}}$ and $\mathbf{P}$ and must be computed before applying Lemma 7.27; see Example 7.28. □

7.5.3. Transforming $\Psi_i(\mathbf{U}) \geq 0$ into a concave constraint

It is sometimes possible to transform a quasiconcave constraint into a concave constraint. This type of transformation is useful, since designing efficient and robust line-search procedures for general quasiconcave functionals is not a trivial task, whereas it is always possible to use the Newton–secant algorithm presented in Section 7.5.4 for concave functionals.

For instance, let $\Psi : \mathcal{B} \to \mathbb{R}$ be a quasiconcave function, then referring to Lemma 7.4, it is sometimes possible to find $R : \mathcal{B} \to (0, \infty)$, positive and convex, such that $R\,\Psi$ is concave. This is indeed the case for any "specific" entropy as described in Example 7.6. The following lemma formalizes this observation.

Lemma 7.30. *Let $\mathcal{B} \subset \mathbb{R}^m$ be a convex set. Let $\Psi : \mathcal{B} \to \mathbb{R}$ and $R : \mathcal{B} \to (0, \infty)$. Assume that $\Phi := R\,\Psi : \mathcal{B} \to \mathbb{R}$ is concave. Let $\mathbf{U}^{\mathrm{L}} \in \mathcal{B}$ and assume that $\Psi(\mathbf{U}^{\mathrm{L}}) \geq 0$. Let $\mathbf{P} \in \mathbb{R}^m$ and let $\ell_{\max} \in [0, 1]$ be such that $\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P} \in \mathcal{B}$ for all $\ell \in [0, \ell^{\max}]$. Let $\Psi^{\min} \in \mathbb{R}$. Assume that either (i) R is affine or (ii) $\Psi^{\min} \geq 0$ and R is convex. Then the following statements hold true:*

(i) $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) - \Psi^{\min} \geq 0$ iff $\Phi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) - \Psi^{\min} R(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \geq 0$ for all $\ell \in [0, \ell^{\max}]$;
(ii) the map $[0, \ell^{\max}] \ni \ell \mapsto \Phi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) - \Psi^{\min} R(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) \in \mathbb{R}$ is concave.

Proof. (i) Since $\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P} \in \mathcal{B}$ for all $\ell \in [0, \ell^{\max}]$, we infer that $R(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) > 0$ for all $\ell \in [0, \ell^{\max}]$. Hence, the first assertion is a consequence of the assumption $R(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) > 0$ for all $\ell \in [0, \ell^{\max}]$. (ii) Observe that $-\Psi^{\min} R : \mathcal{B} \to \mathbb{R}$ is concave if $R : \mathcal{B} \to \mathbb{R}$ is affine. Observe also that $-\Psi^{\min} R : \mathcal{B} \to \mathbb{R}$ is concave if $R : \mathcal{B} \to \mathbb{R}$ is convex and $\Psi^{\min} \geq 0$. Hence the second assertion is just a consequence of the concavity of $\Phi : \mathcal{B} \to \mathbb{R}$. □

Example 7.31 (*Specific Entropy*)**.** Let us illustrate the use of Lemma 7.30 with the compressible Euler equations. Assume to simplify the argument that the equation of state is the γ-law. Consider the physical specific entropy $\Psi(\mathbf{U}) = \frac{1}{\gamma - 1} \log(\varepsilon(\mathbf{U})\rho^{-\gamma})$ and the quasiconcave constraint $\Psi(\mathbf{U}) - \Psi_i^{\min} \geq 0$. Line-searches for this quasiconcave functional may be delicate (lines 5 and 11 in Algorithm 1), not only because it is not strictly concave, but also because of the presence of the logarithm. We have seen in Example 7.14 that this constraint can be transformed into another quasiconcave constraint $\tilde{\Psi}(\mathbf{U}) - \tilde{\Psi}_i^{\min} \geq 0$ with $\tilde{\Psi}(\mathbf{U}) := \varepsilon(\mathbf{U})\rho^{-\gamma} = \exp((\gamma - 1)\,\Psi(\mathbf{U}))$. Let us assume that the solution at the previous time step $\mathbf{U}^n$ is such that $\tilde{\Psi}_i^{\min} \geq 0$ for all $i \in \mathcal{V}$, which is reasonable since it requires the internal energy and the density to be nonnegative at t^n. Then using $R(\mathbf{U}) = \rho^{\gamma}$, which is convex over $\mathcal{B} = \{\mathbf{U} \mid \rho > 0\}$, using that $R(\mathbf{U})\tilde{\Psi}(\mathbf{U}) = \varepsilon(\mathbf{U})$ is concave, and $\tilde{\Psi}_i^{\min} \geq 0$, and invoking Lemma 7.30, we finally transform (again) the above quasiconcave constraint into the concave constraint $\varepsilon(\mathbf{U}) - \rho^{\gamma}\,\tilde{\Psi}_i^{\min} \geq 0$. Notice in passing that, for the γ-law, enforcing positivity of the density and the above local minimum principle on the specific entropy ($\varepsilon(\mathbf{U}) - \rho^{\gamma}\,\tilde{\Psi}_i^{\min} \geq 0$) guarantees positivity of the internal energy. □

The parameter $\ell^{\max}$ appearing in the statement of Lemma 7.30 arises naturally when one performs convex limiting for more than one functional. More precisely, before applying 7.30 one must sure that $\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P} \in \mathcal{B}$ for all $\ell \in [0, \ell^{\max}]$ by convex limiting so that $R(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}) > 0$. For instance, in the setting of Example 7.31, the parameter $\ell^{\max}$ is the limiter that must be computed to ascertain that the density of the state $\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}$ is positive over the interval $[0, \ell^{\max}]$.

7.5.4. Line-search: The Newton–secant solver

Unless the function $g(\ell) := \Psi_i(\mathbf{U}^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}^n)$ has a special structure (say, linear or quadratic), the line-searches invoked at lines 5 and 11 in Algorithm 1 require the use of an iterative procedure. Without claiming originality, we now show how the line-searches can be done by using the Newton–secant algorithm to guarantee that $\Psi_i(\mathbf{U}^{\mathrm{L}} + \ell_j^i\mathbf{P}_{ij}^n) \geq 0$ independently of the tolerance that is given to the algorithm to estimate ℓ_j^i.

Let us assume that $g(\ell) \in \mathcal{C}^2([0, 1]; \mathbb{R})$ is strictly concave and $g(0) > 0$. Let us set $\ell_l^0 = 0$. Let us assume also that there exists $\ell_r^0 \in (0, 1]$ such that $g(\ell_r^0) < 0$. Hence there exists a unique number $\ell^* \in (\ell_l^0, \ell_r^0)$ such $g(\ell_l^0) > g(\ell^*) = 0 > g(\ell_r^0)$. Our goal is now to estimate iteratively ℓ^* from below, up to some fixed tolerance. Notice that in this particular setting Newton's algorithm converges from above; that is, Newton's algorithm will always return an approximate value of ℓ^* that is larger than ℓ^*, (unless g is quadratic). The following lemma describes an iterative process $(\ell_l^k, \ell_r^k) \to (\ell_l^{k+1}, \ell_r^{k+1})$, $k \geq 0$, such that

$$\ell_l^0 < \cdots < \ell_l^k < \ell_l^{k+1} < \cdots \leq \ell^* \leq \ldots < \ell_r^{k+1} < \ell_r^k < \cdots < \ell_r^0$$

Lemma 7.32 (*One Iteration Update*)**.** *Let $\ell_l^k < \ell_r^k$. Let $g \in C^2([\ell_l^k, \ell_r^k]; \mathbb{R})$. Assume that $g''(\ell) < 0$ for all $\ell \in [\ell_l^k, \ell_r^k]$. Assume that $g(\ell_l^k) > 0$ and $g(\ell_r^k) < 0$.*

(i) Let $s_l^k := \frac{g(\ell^{k,r}) - g(\ell^{k,l})}{\ell^{k,r} - \ell^{k,l}}$ and $s_r^k := g'(\ell^{k,r})$. Then $s_l^k < 0$ and $s_r^k < 0$.

(ii) Let ℓ_l^{k+1} and ℓ_r^{k+1} be defined by

$$\ell_l^{k+1} := \ell_l^k - \frac{g(\ell_l^k)}{s_l^k}, \qquad \ell_r^{k+1} := \ell_r^k - \frac{g(\ell_r^k)}{s_r^k}.$$

Then $\ell_l^k < \ell_l^{k+1} < \ell^* < \ell_r^{k+1} < \ell_r^k$.

Proof. The inequalities $\ell_l^k < \ell_l^{k+1} < \ell^*$ are standard properties of the secant algorithm. The inequalities $\ell^* < \ell_r^{k+1} < \ell_r^k$ are standard properties of Newton's algorithm. The details are left to the reader. □

Algorithm 2 Newton–Secant solver

Require: $k = 0, k_{\max} \geq 1, \ell_l < \ell_r, g(\ell_l) > 0, g(\ell_r) < 0, \text{tol} > 0$

1: **while** $k \leq k_{\max}$ **and** $\ell_r - \ell_l > \text{tol}$ **do**
2: $k := k + 1$
3: $\ell_l^{\text{aux}} := \ell_l$
4: **if** $g(\ell_l) > g(\ell_r)$ **then**
5: $s_l := \frac{g(\ell_r) - g(\ell_l)}{\ell_r - \ell_l}$ ▷ Condition $\ell_r - \ell_l > 0$ checked in line 1
6: $\ell_l := \ell_l - \frac{g(\ell_l)}{s_l}$
7: **else**
8: **break**
9: **end if**
10: **if** $\ell_l > \ell_r$ **or** $g(\ell_l) < 0$ **then** ▷ Assumes $g(\ell_r) < 0$
11: $\ell_l := \ell^{l,\text{aux}}$
12: **break**
13: **end if**
14: **if** $g'(\ell_r) <$ **then**
15: $\ell_r := \ell_r - \frac{g(\ell_r)}{g'(\ell_r)}$
16: **else**
17: **break**
18: **end if**
19: **if** $g(\ell_r) > 0$ **then** ▷ Condition $\ell_r - \ell_l > 0$ will be checked in line 1
20: **break**
21: **end if**
22: **end while**
23: **return** $\ell_i^j := \ell_l$

In Algorithm 2, line 1 checks the stopping criteria. The "break" statements (or "exit" statements, depending on the programming language) force the code out of the while loop, redirecting the control to Line 23. One may reach break statements due to roundoff errors. Lines 4–9 is the secant update (approximation from the left), while Lines 14–18 define the Newton update (approximation from the right). Lines 10–13 and 19–21 are sanity checks. The Newton–secant update preserves the order $\ell_l^k < \ell_l^{k+1} < \ell^* < \ell_r^{k+1} < \ell_r^k$ (see Lemma 7.32), however some crossover may occur after some iterations because of round-off errors (due to the nature of floating-point arithmetic). Notice that the output of interest is the one produced by the secant update (see line 23), since the output produced by Newton's method violates the inequality that we want to satisfy.

Remark 7.33 (*Deficiencies of Newton's Method*)**.** If we assume that $g(\ell)$ is strictly concave over $[0, 1]$, which is the case of interest here, one can construct counterexamples illustrating that Newton's method can either not converge or produce an output that violates the bound that we want to enforce. For instance, if the initial guess $\ell^0 \in [0, 1]$ for Newton's method is such that $\ell^0 > \ell^*$ (i.e., $g(\ell^0) < 0$), then Newton's method produces a sequence $\{\ell^k\}_{k\in\mathbb{N}}$ satisfying $\ell^* < \ell^k$ for all $k \in \mathbb{N}$. This implies that $g(\ell^k) < 0$ for all $k \in \mathbb{N}$, which is incompatible with the constraint that we want to satisfy. On the other hand, if g reaches a maximum at $\ell_c \in (0, \ell^*)$ and the initial guess is such that $\ell^0 \in (0, \ell_c)$,

then the sequence $\{\ell^k\}_{k\in\mathbb{N}}$ wanders outside the internal $[0, 1]$. Assuming that $g(\ell)$ is well defined outside $[0, 1]$, the sequence $\{\ell^k\}_{k\in\mathbb{N}}$ may converge to a negative solution. □

Remark 7.34 (*Actual Performance*)**.** The convergence rate of Algorithm 2 is at least 1.618 because it combines the second-order Newton method with the $\frac{\sqrt{5}+1}{2}$-order secant method. In practice, we have verified that Algorithm 2 rarely ever requires more than three iterations to reach tolerances such as $\text{tol} = 10^{-10}$ (see Guermond et al. [1]). Most frequently one exits the loop after reaching machine accuracy error. □

7.6. Relaxing the bounds

In general the quantity $\Psi_i^{\min}$ defined in (7.2) is accurate enough to make the limited high-order solution second-order in the L^1-norm in space. But it is too tight to make the method higher-order or even second-order in the L^∞-norm in the presence of smooth extrema. The situation is even worse when using the specific physical entropy to limit the high-order solution. For instance, it is observed in Khobalatte and Perthame [37, §3.3] that strictly enforcing the minimum principle on the specific (physical) entropy for the compressible Euler equations degrades the converge rate to first-order; it is said therein that "It seems impossible to perform second-order reconstruction satisfying the conservativity requirements ... and the maximum principle on $\varepsilon(\boldsymbol{u})$". We confirm this observation. To recover full accuracy in the L^∞-norm for smooth solutions, one must relax the bound $\Psi_i^{\min}$.

To avoid repeating ourselves, we refer the reader to Guermond et al. [1, §4.7] where we explain how the bound $\Psi_i^{\min}$ should be relaxed. In a nutshell, one proceeds as follows: For each $i \in \mathcal{V}$, we set

$$\Delta^2\Psi_i = \frac{1}{\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\beta_{ij}}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\beta_{ij}(\Psi(\mathbf{U}_i^n) - \Psi(\mathbf{U}_j^n)),$$

where the coefficients β_{ij} are meant to make the computation linearity-preserving (see Remark 6.2). Then we compute the average

$$\overline{\Delta^2\Psi_i} := \frac{1}{2\text{card}(\mathcal{I}(i))}\sum_{i\neq j\in\mathcal{I}(i)}(\frac{1}{2}\Delta^2\Psi_i + \frac{1}{2}\Delta^2\Psi_j),$$

and finally the relaxation is done by redefining $\Psi_i^{\min}$ as follows:

$$\overline{\Psi_i^{\min}} \leftarrow \max((1-\text{sign}(\Psi_i^{\min})r_i)\,\Psi_i^{\min},\ \Psi_i^{\min} - |\overline{\Delta^2\Psi_i}|),$$

where $r_i = (\frac{m_i}{|D|})^{\frac{1.5}{d}}$. Notice that $r_i \in (0, 1)$. The somewhat ad hoc threshold $(1-\text{sign}(\Psi_i^{\min})r_h)$ is never active when the mesh size is fine enough. This term is just meant to be a safeguard on coarse meshes. For instance, for the compressible Euler equations, when $\Psi(\mathbf{U})$ is either the density (or the internal energy), this threshold guarantees positivity of the density (or the internal energy) because in this case $(1 - \text{sign}(\Psi_i^{\min})r_i) \geq 0$. The exponent 1.5 is somewhat ad hoc; in principle one could take $r_i = (\frac{m_i}{|D|})^{\frac{\delta}{d}}$ with $\delta < 2$.

References

[1] J.-L. Guermond, M. Nazarov, B. Popov, I. Tomas, Second-order invariant domain preserving approximation of the Euler equations using convex limiting, SIAM J. Sci. Comput. 40 (5) (2018) A3211–A3239.

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

[3] J.-L. Guermond, B. Popov, Invariant domains and first-order continuous finite element approximation for hyperbolic systems, SIAM J. Numer. Anal. 54 (4) (2016) 2466–2489.

[4] P. Azerad, J.-L. Guermond, B. Popov, Well-balanced second-order approximation of the shallow water equation with continuous finite elements, SIAM J. Numer. Anal. 55 (6) (2017) 3203–3224.

[5] J.-L. Guermond, M. Quezada de Luna, B. Popov, C. Kees, M. Farthing, Well-balanced second-order finite element approximation of the shallow water equations with friction, SIAM J. Sci. Comput. 40 (6) (2018) A3873A3901.

[6] P.D. Lax, Hyperbolic systems of conservation laws. II, Comm. Pure Appl. Math. 10 (1957) 537–566.

[7] E.F. Toro, Riemann Solvers and Numerical Methods for fluid Dynamics, third ed., Springer-Verlag, Berlin, 2009, p. xxiv+724, A practical introduction.

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

[9] K.N. Chueh, C.C. Conley, J.A. Smoller, Positively invariant regions for systems of nonlinear diffusion equations, Indiana Univ. Math. J. 26 (2) (1977) 373–392.
[10] D. Hoff, Invariant regions for systems of conservation laws, Trans. Amer. Math. Soc. 289 (2) (1985) 591–610.
[11] J. Smoller, Shock Waves and Reaction-diffusion Equations, second ed., in: Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 258, Springer-Verlag, New York, ISBN: 0-387-94259-9, 1994, p. xxiv+632.
[12] H. Frid, Maps of convex sets and invariant regions for finite-difference systems of conservation laws, Arch. Ration. Mech. Anal. 160 (3) (2001) 245–269.
[13] X. Zhang, C.-W. Shu, Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms, J. Comput. Phys. 230 (4) (2011) 1238–1248.
[14] T. Nishida, Global solution for an initial boundary value problem of a quasilinear hyperbolic system, Proc. Japan Acad. 44 (1968) 642–646.
[15] D. Hoff, A finite difference scheme for a system of two conservation laws with artificial viscosity, Math. Comp. 33 (148) (1979) 1171–1193.
[16] A. Bermudez, M.E. Vazquez, Upwind methods for hyperbolic conservation laws with source terms, Comput. Fluids 23 (8) (1994) 1049–1071.
[17] J.M. Greenberg, A.Y. Leroux, A well-balanced scheme for the numerical processing of source terms in hyperbolic equations, SIAM J. Numer. Anal. 33 (1) (1996) 1–16.
[18] L.P. Huang, T.-P. Liu, A conservative, piecewise-steady difference scheme for transonic nozzle flow, Comput. Math. Appl. Part A 12 (4–5) (1986) 377–388, Hyperbolic partial differential equations, III.
[19] T. Barth, M. Ohlberger, Finite volume methods: foundation and analysis, in: Encyclopedia of Computational Mechanics, John Wiley & Sons, Ltd., 2004.
[20] R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in: Handbook of numerical analysis, Vol. VII, in: Handb. Numer. Anal., VII, North-Holland, Amsterdam, 2000, pp. 713–1020.
[21] J.-L. Guermond, B. Popov, Invariant domains and second-order continuous finite element approximation for scalar conservation equations, SIAM J. Numer. Anal. 55 (6) (2017) 3120–3146.
[22] L. Ferracina, M.N. Spijker, An extension and analysis of the Shu-Osher representation of Runge-Kutta methods, Math. Comp. 74 (249) (2005) 201–219.
[23] I. Higueras, Representations of Runge-Kutta methods and strong stability preserving methods, SIAM J. Numer. Anal. 43 (3) (2005) 924–948.
[24] S. Gottlieb, D.I. Ketcheson, C.-W. Shu, High order strong stability preserving time discretizations, J. Sci. Comput. 38 (3) (2009) 251–289.
[25] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (2) (1988) 439–471.
[26] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM Rev. 43 (1) (2001) 89–112 (electronic).
[27] D. Kröner, Numerical Schemes for Conservation Laws, in: Wiley-Teubner Series Advances in Numerical Mathematics, John Wiley & Sons, Ltd., Chichester; B. G. Teubner, Stuttgart, ISBN: 0-471-96793-9, 1997, p. viii+508.
[28] K.W. Morton, T. Sonar, Finite volume methods for hyperbolic conservation laws, Acta Numer. 16 (2007) 155–238.
[29] M.A. Christon, M.J. Martinez, T.E. Voth, Generalized Fourier analyses of the advection-diffusion equation-Part I: one-dimensional domains, Internat. J. Numer. Methods Fluids 45 (8) (2004) 839–887.
[30] J.-L. Guermond, R. Pasquetti, A correction technique for the dispersive effects of mass lumping for transport problems, Comput. Methods Appl. Mech. Engrg. 253 (2013) 186–198.
[31] J.-L. Guermond, M. Nazarov, B. Popov, Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations, SIAM J. Numer. Anal. 52 (4) (2014) 2163–2182.
[32] M. Berger, M.J. Aftosmis, S.M. Murman, Analysis of slope limiters on irregular grids, AIAA Paper 2005–0490, American Institute for Aeronautics and Astronautics, Reno, NV, USA, Also NASA TM NAS-05-007, 2005.
[33] M.S. Floater, Generalized barycentric coordinates and applications, Acta Numer. 24 (2015) 161–214.
[34] A. Jameson, Origins and further development of the Jameson-Schmidt-Turkel scheme, AIAA J. 55 (5) (2017).
[35] E. Burman, On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws, BIT 47 (4) (2007) 715–733.
[36] G.R. Barrenechea, E. Burman, F. Karakatsani, Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes, Numer. Math. 135 (2) (2017) 521–545.
[37] B. Khobalatte, B. Perthame, Maximum principle on the entropy and second-order kinetic schemes, Math. Comp. 62 (205) (1994) 119–131.
[38] B. Perthame, Y. Qiu, A variant of van Leer's method for multidimensional systems of conservation laws, J. Comput. Phys. 112 (2) (1994) 370–381.
[39] B. Perthame, C.-W. Shu, On positivity preserving finite volume schemes for euler equations, Numer. Math. 73 (1) (1996) 119–130.
[40] X. Zhang, C.-W. Shu, On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, J. Comput. Phys. 229 (23) (2010) 8918–8934.
[41] X. Zhang, C.-W. Shu, A minimum entropy principle of high order schemes for gas dynamics equations, Numer. Math. 121 (3) (2012) 545–563.
[42] Y. Jiang, H. Liu, An Invariant-region-preserving (IRP) limiter for compressible Euler equations, in: Proceedings of the Hyp2016 Conference, 2017.
[43] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11** (1973), no. 1, 38–69], J. Comput. Phys. 135 (2) (1997) 170–186.
[44] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, J. Comput. Phys. 31 (3) (1979) 335–362.
[45] M. Avriel, W.E. Diewert, S. Schaible, I. Zang, Generalized Concavity, in: Mathematical Concepts and Methods in Science and Engineering, vol. 36, Plenum Press, New York, ISBN: 0-306-42656-0, 1988, p. x+332.
[46] A. Harten, On the symmetric form of systems of conservation laws with entropy, J. Comput. Phys. 49 (1) (1983) 151–164.

[47] A. Harten, P.D. Lax, C.D. Levermore, W.J. Morokoff, Convex entropies and hyperbolicity for general euler equations, SIAM J. Numer. Anal. 35 (6) (1998) 2117–2127 (electronic).
[48] D. Serre, Systems of Conservation Laws, Vol. 2, Cambridge University Press, Cambridge, 2000, p. xii+269, Geometric structures, oscillations, and initial-boundary value problems, Translated from the 1996 French original by I. N. Sneddon.
[49] D. Kuzmin, R. Löhner, S. Turek, Flux–Corrected Transport, in: Scientific Computation, Springer, 2005, 3-540-23730-5.
[50] D. Kuzmin, R. Löhner, S. Turek, Flux-Corrected Transport: Principles, Algorithms, and Applications, in: Scientific Computation, Springer, ISBN: 9789400740372, 2012.