SIAM J. SCI. COMPUT.
Vol. 40, No. 5, pp. A3211–A3239

SECOND-ORDER INVARIANT DOMAIN PRESERVING APPROXIMATION OF THE EULER EQUATIONS USING CONVEX LIMITING*

JEAN-LUC GUERMOND†, MURTAZO NAZAROV‡, BOJAN POPOV†, AND IGNACIO TOMAS†

Abstract. A new second-order method for approximating the compressible Euler equations is introduced. The method preserves all the known invariant domains of the Euler system: positivity of the density, positivity of the internal energy, and the local minimum principle on the specific entropy. The technique combines a first-order, invariant domain preserving, guaranteed maximum speed method using a graph viscosity (GMS-GV1) with an invariant domain violating, but entropy consistent, high-order method. Invariant domain preserving auxiliary states, naturally produced by the GMS-GV1 method, are used to define local bounds for the high-order method, which is then made invariant domain preserving via a *convex limiting* process. Numerical tests confirm the second-order accuracy of the new GMS-GV2 method in the maximum norm, where the 2 stands for second-order. The proposed convex limiting is generic and can be applied to other approximation techniques and other hyperbolic systems.

Key words. hyperbolic systems, Riemann problem, invariant domain, entropy inequality, high-order method, exact rarefaction, quasi-convexity, limiting, finite element method

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

DOI. 10.1137/17M1149961

1. Introduction.

It is well known that solutions of the compressible Euler equations satisfy physical constraints like positivity of the density, positivity of the internal energy, and minimum principle on the specific entropy. Any numerical method that satisfies these constraints or similar constraints is usually called invariant domain preserving in the literature, e.g., Chueh, Conley, and Smoller [7]. The objective of the present work is to present an approximation technique for the compressible Euler equations that is explicit in time, second-order accurate in space and time, and invariant domain preserving. The method is presented in the context of continuous finite elements, but it is quite general and can be applied to other discretization settings like discontinuous Galerkin and finite volume techniques. Like in many other high-order approximation methods, the proposed technique consists of combining a first-order, invariant domain preserving, and entropy satisfying approximation with a high-order entropy consistent approximation. The high-order method is made invariant domain preserving and formally entropy compliant by adapting the artificial viscosity through a limiting process. One key novelty is that the density, the internal energy, and the specific entropy are limited by using bounds that are in the domain of dependence

*Submitted to the journal's Methods and Algorithms for Scientific Computing section October 2, 2017; accepted for publication (in revised form) June 29, 2018; published electronically October 2, 2018.

http://www.siam.org/journals/sisc/40-5/M114996.html

Funding: This work was partially supported by National Science Foundation grants DMS-1619892 and DMS-1620058, by the Air Force Office of Scientific Research, USAF, under grant/contract FA9550-15-1-0257, and by the Army Research Office under grant/contract W911NF-15-1-0517.

†Department of Mathematics, Texas A & M University, College Station, TX 77843 (guermond@math.tamu.edu, popov@math.tamu.edu, itomas@tamu.edu).

‡Division of Scientific Computing, Department of Information Technology, Uppsala University, 752 36 Uppsala, Sweden (murtazo@math.tamu.edu).

of the local data. Another novelty is that the limiting bounds are local for all the quantities, and these bounds are naturally satisfied by the low-order solution.

The so-called Flux Corrected Transport (FCT) method, introduced in Boris and Book [5] for approximating the one-dimensional mass conservation equation and later generalized to multiple dimensions in Zalesak [49], are among the first successful techniques to produce second-accuracy while imposing pointwise bounds such as positivity of the density. This methodology can be used to preserve the maximum principle for any scalar conservation equation. We refer the reader to Kuzmin and Turek [31], Kuzmin, Löhner, and Turek [32] for reviews on FCT. Unfortunately, the FCT algorithm, as proposed in the above references, is not well suited to enforce constraints that are not affine. For instance, it cannot be (easily) applied to guarantee the positivity of the specific internal energy and the minimum principle on the specific entropy since these are quasi-concave constraints. (Said constraints can be made concave by multiplication by the density as is shown in section 4.) In the context of finite volumes, efficient second-order limiting techniques for the specific internal energy and the specific entropy were first proposed in Khobalatte and Perthame [27], Perthame and Qiu [39], and Perthame and Shu [40]. These ideas have been extended to discontinuous Galerkin framework in a series of papers by Zhang and Shu [51, 52] and Jiang and Liu [26]. The key argument common to [27, 40, 39] is to rely upon convex combinations and concavity. In the present paper we are going to build on these ideas and propose a general postprocessing methodology to enforce general concave constraints for the Euler equations, which we call *convex limiting*. Instead of limiting slopes or reconstructed approximations, we adopt an algebraic point of view similar to FCT. The method is presented in the context of continuous finite elements, but since it is algebraic, it can be applied to finite volumes and discontinuous Galerkin approximation techniques as well.

This paper is organized as follows. The problem is formulated in section 2. The finite element setting and the notation are also introduced in this section. The time and space approximation using continuous finite elements is described in section 3. Both the first-order, invariant domain preserving, entropy satisfying method and the high-order invariant domain violating method are detailed in this section. The bulk of the novel material is reported in section 4. The main results of this section are the local bounds given in (4.1)–(4.4) together with Lemmas 4.3 and 4.4. The performance of the proposed method is illustrated in section 5.

2. Preliminaries. In this section, we introduce the Euler equations and the finite element setting. Some important properties of the Euler equations that are used later in this paper are also recalled. The reader who is familiar with the Euler equations, invariant domains, and the finite element theory is invited to jump to section 3.

2.1. The Euler equations. Let d be the space dimension, and let D be an open polyhedral domain in $\mathbb{R}^d$. We consider the compressible Euler equations in conservative form in $\mathbb{R}^d$:

$$\partial_t \rho + \nabla\cdot\boldsymbol{m} = 0, \tag{2.1a}$$

$$\partial_t \boldsymbol{m} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{m}) + \nabla p = 0, \tag{2.1b}$$

$$\partial_t E + \nabla\cdot(\boldsymbol{v}(E+p)) = 0, \tag{2.1c}$$

$$\rho(\boldsymbol{x},0) = \rho_0, \qquad \boldsymbol{m}(\boldsymbol{x},0) = \boldsymbol{m}_0, \qquad E(\boldsymbol{x},0) = E_0. \tag{2.1d}$$

The independent variables are $(\boldsymbol{x},t)\in D\times\mathbb{R}_+$. The dependent variables, henceforth called conserved variables, are the density, ρ, the momentum, $\boldsymbol{m}$, and the total energy, E. The quantity $\boldsymbol{v}:=\rho^{-1}\boldsymbol{m}$ is the velocity of the fluid particles. The pressure, p, is given by the equation of state, which we assume to be derived from a specific entropy, $s:\mathbb{R}_+\times\mathbb{R}_+\to\mathbb{R}$, defined through the thermodynamics identity: $T\,\mathrm{d}s:=\mathrm{d}e-p\rho^{-2}\,\mathrm{d}\rho$, where $e:=\rho^{-1}E-\frac{1}{2}\boldsymbol{v}^2$ is the specific internal energy. For instance, it is common to take $s(\rho,e)-s_0=\log(e^{\frac{1}{\gamma-1}}\rho^{-1})$ for a polytropic ideal gas. Using the notation $s_e(\rho,e):=\frac{\partial s}{\partial e}(\rho,e)$ and $s_\rho(\rho,e):=\frac{\partial s}{\partial\rho}(\rho,e)$, the equation of state then takes the form $p:=-\rho^2 s_\rho s_e^{-1}$. To simplify the notation we introduce the conserved variable $\boldsymbol{u}:=(\rho,\boldsymbol{m},E)^\mathsf{T}\in\mathbb{R}^{d+2}$ and the flux $\mathbb{f}(\boldsymbol{u}):=(\boldsymbol{m},\boldsymbol{v}\otimes\boldsymbol{m}+p\mathbb{I}_d,\boldsymbol{v}(E+p))^\mathsf{T}\in\mathbb{R}^{(d+2)\times d}$, where $\mathbb{I}_d$ is the $d\times d$ identity matrix. To avoid abuse of notation and ambiguities, we introduce

$$\Phi(\boldsymbol{u}):=s(\rho,e(\boldsymbol{u})), \tag{2.2}$$

where $e(\boldsymbol{u}):=\rho^{-1}E-\frac{\boldsymbol{m}^2}{2\rho^2}$, $\boldsymbol{m}^2:=\|\boldsymbol{m}\|_{\ell^2}^2$, and $\|\cdot\|_{\ell^2}$ is the Euclidean norm, i.e., Φ is the specific entropy expressed as a function of the conserved variables. Finally, we indicate internal energy by the quantity $\varepsilon:=\rho e=E-\frac{1}{2}\rho\boldsymbol{v}^2$.

The convention adopted in this paper is that for any vectors $\boldsymbol{a}$, $\boldsymbol{b}$, with entries $\{a_k\}_{k=1,\dots,d}$, $\{b_k\}_{k=1,\dots,d}$, the following holds: $(\boldsymbol{a}\otimes\boldsymbol{b})_{kl}=a_kb_l$ and $\nabla\cdot\boldsymbol{a}=\sum_{k=1,\dots,d}\partial_{x_k}a_k$, $(\nabla\boldsymbol{a})_{kl}=\partial_{x_l}a_k$. $\boldsymbol{a}\cdot\nabla=\sum_{k=1,\dots,d}a_k\partial_{x_k}$. Moreover, for any second-order tensor $\mathbb{g}$, with entries $\{\mathbb{g}_{kl}\}_{k=1,\dots,d+2}^{l=1,\dots,d}$, we define $(\nabla\cdot\mathbb{g})_k=\sum_{l=1,\dots,d}\partial_{x_l}\mathbb{g}_{kl}$, $(\mathbb{g}\boldsymbol{a})_k=\sum_{l=1,\dots,d}\mathbb{g}_{kl}a_l$, $(\boldsymbol{a}^\mathsf{T}\mathbb{g})_l=\sum_{l=1,\dots,d}a_k\mathbb{g}_{kl}$.

To avoid technicalities regarding boundary conditions, we assume that either periodic boundary conditions are enforced, or the initial data is compactly supported, in which case we are interested in the solution before the domain of influence of $(\rho_0,\boldsymbol{m}_0,E_0)$ reaches the boundary of D, i.e., homogeneous Dirichlet boundary conditions are enforced. (Some details are given in section 3.5 on how to deal with the slip boundary condition.)

2.2. Intrinsic properties. The well-posedness of (2.1) is an extremely difficult question that is far beyond the scope of the present paper. But to make sense of the approximation techniques to be presented in the rest of the paper we are going to rely on the notion of solution of a one-dimensional Riemann problem which is more tractable. For any unit vector $\boldsymbol{n}\in\mathbb{R}^d$, we consider the following Riemann problem:

$$\partial_t\boldsymbol{u}+\partial_x(\mathbb{f}(\boldsymbol{u})\boldsymbol{n})=0,\quad (x,t)\in\mathbb{R}\times\mathbb{R}_+,\qquad \boldsymbol{u}(x,0)=\begin{cases}\boldsymbol{u}_L & \text{if } x<0,\\ \boldsymbol{u}_R & \text{if } x>0,\end{cases} \tag{2.3}$$

and assume that there exists a so-called admissible set $\mathcal{A}$ such that for any pair of states $(\boldsymbol{u}_L,\boldsymbol{u}_R)\in\mathcal{A}\times\mathcal{A}$, this problem has a unique physical (entropy) solution henceforth denoted by $\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t)$. This assumption holds true for small data (see Bianchini and Bressan [4]), and we refer the reader to Godlewski and Raviart [10, Thm. II.3.1] for a similar statement for large data. One key property of the physical solution is that there exists a quantity $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$, called maximum wave speed, such that $\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t)=u_L$ if $x/t\le-\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ and $\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t)=u_R$ if $x/t\ge\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$.

We now introduce notions of invariant sets and invariant domains. (Our definition is slightly different from those in Chueh, Conley, and Smoller [7], Hoff [23], Frid [9].)

DEFINITION 2.1 (invariant set). *We say that a set $B \subset \mathcal{A} \subset \mathbb{R}^m$ is invariant for* (2.1a)–(2.1c) *if for any pair $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in B \times B$, any unit vector $\boldsymbol{n} \in \mathbb{R}^d$, and any $t > 0$, the average of the entropy solution of the Riemann problem* (2.3) *over the Riemann fan, $\frac{1}{2t\lambda_{\max}} \int_{-\lambda_{\max} t}^{\lambda_{\max} t} \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x$, remains in B.*

We are also going to make use of the notion of invariant domain for an approximation process. Let $\boldsymbol{X}_h \subset L^1(\mathbb{R}^d; \mathbb{R}^m)$ be a finite-dimensional approximation space, and let $S_h : \boldsymbol{X}_h \ni \boldsymbol{u}_h \longmapsto S_h(\boldsymbol{u}_h) \in \boldsymbol{X}_h$ be a mapping over $\boldsymbol{X}_h$. (Think of S_h as being a one-time-step approximation of (2.1).) Henceforth, we abuse the notation by saying that a member of $\boldsymbol{X}_h$, say $\boldsymbol{u}_h$, is in the set $B \subset \mathbb{R}^m$ when actually we mean that $\{\boldsymbol{u}_h(\boldsymbol{x}) \mid \boldsymbol{x} \in D\} \subset B$.

DEFINITION 2.2 (invariant domain). *A convex invariant set $B \subset \mathcal{A} \subset \mathbb{R}^m$ is said to be an invariant domain for the mapping S_h if and only if for any state $\boldsymbol{u}_h$ in B, the state $S_h(\boldsymbol{u}_h)$ is also in B.*

It is known that the set

$$A_{s^{\min}} := \{(\rho, \boldsymbol{m}, E) \mid \rho > 0, e > 0, s \geq s^{\min}\} \tag{2.4}$$

is invariant for the Euler system for any $s^{\min} \in \mathbb{R}$. It is also established in Serre [42, Thm. 8.2.2] that the set $A_{s^{\min}}$ is convex, and it is shown in Frid [9, Thms. 7 and 8] that it is an invariant domain for the Lax–Friedrichs scheme. The finite element method introduced in Guermond and Popov [13] also satisfies this invariant domain property; this finite element construction is recalled in section 3.1. It is generally acknowledged in the literature that physical solutions to (2.1) should satisfy entropy inequalities. More specifically, let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable and be such that

$$f'(s) > 0, \qquad f'(s)c_p^{-1} - f''(s) > 0 \qquad \forall (\rho, e) \in \mathbb{R}_+^2, \tag{2.5}$$

where $c_p(\rho, e) = T\partial_T s(p, T)$ is the specific heat at constant pressure. It is shown in Harten et al. [21, Thm. 2.1] that $\rho f(s)$ is strictly concave with respect to the conserved variables if and only if (2.5) holds, i.e., $\mathcal{A} \ni \boldsymbol{u} \mapsto -\rho f(\Phi(\boldsymbol{u}))$ is convex. Note that the so-called physical entropy $S(\boldsymbol{u}) := \rho\Phi(\boldsymbol{u})$ is obtained by setting $f(x) = x$. We say that a weak solution to (2.1) is an entropy solution if it satisfies the following inequality in the weak sense for every generalized entropy:

$$\partial_t(\rho f(s)) + \nabla\cdot(\boldsymbol{m} f(s)) \geq 0. \tag{2.6}$$

In particular, it is known (at least for γ-law equations of state) that the entropy solution to the Riemann problem (2.3) satisfies (2.6). Both the Lax–Friedrichs scheme and the finite element method introduced in [13] satisfy a discrete version of (2.6) for every generalized entropy.

The objective of the present work is to construct an explicit, second-order continuous finite element method that is consistent with (2.6) and for which $A_{s^{\min}}$ is an invariant domain.

2.3. Finite element setting. We are going to approximate the solution of (2.1) with continuous Lagrange finite elements. For this purpose, we introduce a shape-regular sequence of matching meshes $(\mathcal{T}_h)_{h>0}$ and assume that the elements in each mesh are generated from a small collection of reference elements denoted $\widehat{K}_1, \ldots, \widehat{K}_\varpi$. In two space dimensions for instance, the mesh $\mathcal{T}_h$ could be composed of a combination of parallelograms and triangles (i.e., $\varpi = 2$). In three space dimensions, $\mathcal{T}_h$ could also

be composed of a combination of triangular prisms, parallelepipeds, and tetrahedra ($\varpi = 3$). Given $K \in \mathcal{T}_h$, the geometric transformation mapping $\widehat{K}_r$ to $K \in \mathcal{T}_h$ is denoted $T_K : \widehat{K}_r \longrightarrow K$. We are going to construct the approximation space by using some reference Lagrange finite elements $\{(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)\}_{1\le r\le \varpi}$, where the objects $(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)$ are Ciarlet triples (we omit the index $r \in \{1\!:\!\varpi\}$ in the rest of the paper to simplify the notation). Given a reference Lagrange element $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$, we denote by $\{\widehat{\boldsymbol{x}}_l\}_{l\in\mathcal{L}}$ the reference Lagrange nodes and by $\{\widehat{\theta}_l\}_{l\in\mathcal{L}}$ the reference shape functions, i.e., $\mathrm{card}(\mathcal{L}) = \dim(\widehat{P}) =: n_{\mathrm{sh}}$ (note that the index r has been omitted). $\widehat{P}$ is the reference approximation space (usually a scalar-valued polynomial space) and $\widehat{\Sigma}$ is the set of the Lagrange degrees of freedom. Let $\mathbb{P}_{l,d}$ be the vector space composed of the d-variate polynomials of degree at most l. We henceforth assume that there is $k \ge 1$ such that $\mathbb{P}_{k,d} \subset \widehat{P}$. The reference degrees of freedom $\{\widehat{\sigma}_l\}_{l\in\mathcal{L}}$ are such that $\widehat{\sigma}_l(\widehat{p}) = \widehat{p}(\widehat{\boldsymbol{x}}_l)$ for all $l \in \mathcal{L}$ and all $\widehat{p} \in \widehat{P}$. By definition $\widehat{\theta}_l(\widehat{\boldsymbol{x}}_{l'}) = \delta_{ll'}$, for all $l.l' \in \mathcal{L}$, which in turn implies the partition of unity property: $\sum_{l\in\mathcal{L}} \widehat{\theta}_l(\widehat{\boldsymbol{x}}) = 1$ for all $\widehat{\boldsymbol{x}} \in \widehat{K}$. Setting $\{\widehat{w}_l := \int_{\widehat{K}} \theta_l \,\mathrm{d}x\}_{l\in\mathcal{L}}$, the following quadrature is $(k+1)$th order accurate since it is exact for all $\widehat{p} \in \widehat{P}$: $\int_{\widehat{K}} \widehat{p}(\widehat{\boldsymbol{x}}) \,\mathrm{d}\widehat{\boldsymbol{x}} = \sum_{l\in\mathcal{L}} \widehat{p}(\widehat{\boldsymbol{x}}_l)\widehat{w}_l$. We henceforth assume that

$$\int_{\widehat{K}} \widehat{\theta}_l \,\mathrm{d}\widehat{\boldsymbol{x}} =: \widehat{w}_l > 0, \qquad l \in \mathcal{L}. \tag{2.7}$$

There are numerous reference finite elements satisfying (2.7). For instance, all of the elements based on $\mathbb{Q}_{k,d}$ polynomials using tensor product of Gauss–Lobatto points on quadrangles or hexahedra satisfy (2.7). The question is slightly more nuanced for $\mathbb{P}_{k,d}$ polynomials on simplices, but one can use Fekete points (see Taylor, Wingate, and Vincent [44]) or various variations thereof for $k \ge 3$ (see Hesthaven [22], Warburton [47]). Note that the standard $\mathbb{P}_{1,d}$ Lagrange element satisfies (2.7) but $\mathbb{P}_{2,d}$ does not for $d \ge 2$.

We now introduce approximation spaces constructed as usual by using the pullback by the geometric transformation. More precisely, we define the following scalar-valued and vector-valued finite element spaces:

$$P(\mathcal{T}_h) = \{v \in \mathcal{C}^0(D;\mathbb{R}) \mid v_{|K}\circ T_K \in \widehat{P}\ \forall K \in \mathcal{T}_h\}, \qquad \boldsymbol{P}(\mathcal{T}_h) = [P(\mathcal{T}_h)]^{d+2}. \tag{2.8}$$

The global shape functions in $P(\mathcal{T}_h)$, which we recall form the basis of $P(\mathcal{T}_h)$, are denoted by $\{\varphi_i\}_{i\in\mathcal{I}}$, i.e., $\mathrm{card}(\mathcal{I}) = \dim(P(\mathcal{T}_h))$. For any $i \in \mathcal{I}$, we denote $\mathcal{I}(i) := \{j \in \mathcal{I} \mid \varphi_i\varphi_j \not\equiv 0\}$. The global Lagrange nodes are denoted $\{\boldsymbol{x}_i\}_{i\in\mathcal{I}}$. Upon introducing the connectivity array $\mathsf{j} : \mathcal{T}_h\times\mathcal{L} \longrightarrow \mathcal{I}$, we have $\varphi_{\mathsf{j}(l,K)}(\boldsymbol{x}) = \widehat{\theta}_l((T_K)^{-1}(\boldsymbol{x}))$, and $\boldsymbol{x}_{\mathsf{j}(K,l)} = T_K(\widehat{\boldsymbol{x}}_l)$ for all $l \in \mathcal{L}$ and all $K \in \mathcal{T}_h$. This implies that $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$. The local partition of unity property implies that

$$\sum_{i\in\mathcal{I}} \varphi_i(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in D. \tag{2.9}$$

Upon defining $m_i := \int_D \varphi_i(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$, the above definitions imply that the following quadrature is $(k+1)$th order accurate $\int_D v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \sum_{i\in\mathcal{I}} m_i v(\boldsymbol{x}_i)$ since it is exact for all $v \in P(\mathcal{T}_h)$. The matrix with entries $\int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ is called the consistent mass matrix and denoted by $\mathcal{M} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$. Using the above quadrature and the property $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$, the integral $\int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ can be approximated by m_i. Note that (2.9) implies that $\sum_{j\in\mathcal{I}} m_{ij} = m_i$. We henceforth denote by $\mathcal{M}^L$ the diagonal matrix with entries $(m_i)_{i\in\mathcal{I}}$ and refer to $\mathcal{M}^L$ as the lumped mass matrix. Note that (2.7) implies that $m_i > 0$ for all $i \in \mathcal{I}$.

Denoting by $\|\cdot\|_{\ell^2}$ the Euclidean norm in $\mathbb{R}^d$, we introduce the following two quantities which will play an important role in the rest of this paper:

$$\boldsymbol{c}_{ij} := \int_D \varphi_i \nabla\varphi_j \,\mathrm{d}x, \qquad \boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}, \qquad i,j \in \mathcal{I}. \tag{2.10}$$

Note that (2.9) implies $\sum_{j\in\mathcal{I}} \boldsymbol{c}_{ij} = \mathbf{0}$. Furthermore, if either φ_i or φ_j is zero on ∂D, then $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$. In particular, we have $\sum_{i\in\mathcal{I}} \boldsymbol{c}_{ij} = \mathbf{0}$ if φ_j is zero on ∂D. This property will be used to establish conservation. (The definition of $\boldsymbol{c}_{ij}$ is revisited in section 3.5 to account for the slip boundary condition.)

3. Time and space approximation. We describe in this section the first-order technique and the higher-order technique that will be used to construct the second-order invariant domain preserving and entropy compliant method which is the object of the present paper. The discussion is restricted to the forward Euler time stepping since higher-order accuracy in time is trivially achieved by using Strong Stability Preserving Runge–Kutta (SSP RK) time stepping. All of the numerical tests reported in section 5 are done with the SSP RK(3,3) method (three stages, third-order); see Shu and Osher [43, eq. (2.18)] and Kraaijevanger [28, Thm. 9.4].

3.1. Low-order approximation (GMS-GV1). The low-order method that we are going to use is fully described in [13] and is henceforth referred to as GMS-GV1 for Guaranteed Maximum Speed method with first-order Graph Viscosity. Let $\boldsymbol{u}_h^0 = \sum_{i\in\mathcal{I}} \mathbf{U}_i^0 \varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a reasonable approximation of the initial data $\boldsymbol{u}_0$. Let t_n be the current time, let τ be the current time step, and let us set $t_{n+1} = t_n + \tau$ for some $n \in \mathbb{N}$. Letting $\boldsymbol{u}_h^n = \sum_{i\in\mathcal{I}} \mathbf{U}_i^n \varphi_i$ be the approximation of $\boldsymbol{u}$ at time t_n, we estimate the low-order approximation $\boldsymbol{u}_h^{L,n+1} = \sum_{i\in\mathcal{I}} \mathbf{U}_i^{L,n+1} \varphi_i$ by setting

$$\frac{m_i}{\tau}(\mathbf{U}_i^{L,n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - d_{ij}^{L,n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = 0, \tag{3.1}$$

where $d_{ij}^{L,n}$ is called graph viscosity and is defined below. There is no need to define $d_{ii}^{L,n}$ for (3.1) to make sense, but to simplify the notation used later in the paper we denote $d_{ii}^{L,n} := -\sum_{i\neq j\in\mathcal{I}(i)} d_{ij}^{L,n}$. Notice that if all of the states $\boldsymbol{U}_j^n$ are equal to the same constant for all $j \in \mathcal{I}(i)$, then the perturbation $\sum_{j\in\mathcal{I}(i)} d_{ij}^{L,n}(\mathbf{U}_j^n - \mathbf{U}_i^n)$ is zero. This means that this term introduces a first-order consistency error. This type of perturbation in known as graph Laplacian in the literature, and we henceforth refer to it as graph viscosity or artificial viscosity. In the rest of this paper the graph viscosity coefficients $d_{ij}^{L,n}$ are defined for all $i \neq j \in \mathcal{I}$ by setting

$$d_{ij}^{L,n} := \max(\widehat{\lambda}_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \widehat{\lambda}_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{3.2}$$

where $\widehat{\lambda}_{\max}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$ is any upper bound on the maximum wave speed in the Riemann problem (2.3), $\lambda_{\max}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$. Of course the sharper the upper bound, the better. It is proved in Toro [46, p. 150] that the maximum wave speed in (2.3) is the same as in the following reduced Riemann problem:

$$\partial_t \boldsymbol{w} + \partial_x \boldsymbol{f}^{1\mathrm{D}}(\boldsymbol{w}) = 0, \tag{3.3}$$

with data $\boldsymbol{w}_L := (\rho_L, m_L, \mathcal{E}_L)^\mathsf{T}$, $\boldsymbol{w}_R := (\rho_R, m_R, \mathcal{E}_R)^\mathsf{T}$, where $m := \boldsymbol{m}\cdot\boldsymbol{n}$, $v := m/\rho$, $\boldsymbol{m}^\perp := \boldsymbol{m} - (\boldsymbol{m}\cdot\boldsymbol{n})\boldsymbol{n}$, $\mathcal{E} := E - \frac{1}{2}\frac{\|\boldsymbol{m}^\perp\|_{\ell^2}^2}{\rho}$, and flux $\boldsymbol{f}^{1\mathrm{D}}(\boldsymbol{w}) := (m, vm + p, v(\mathcal{E}+p))^\mathsf{T}$.

We also refer the reader to Guermond and Popov [14, sect. 2] for more details on this question. Let $v_L := \boldsymbol{m}_L\cdot\boldsymbol{n}/\rho_L$, $v_R := \boldsymbol{m}_R\cdot\boldsymbol{n}/\rho_R$ be the left and right velocities, and let c_L, c_R be the left and right sound speeds, respectively. The definition of the maximum speed $\lambda_{\max}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R) = \lambda_{\max}(\boldsymbol{w}_L,\boldsymbol{w}_R)$ is as follows:

$$\lambda_{\max}(\boldsymbol{w}_L,\boldsymbol{w}_R) = \max((\lambda_1^-)_-,(\lambda_3^+)_+), \tag{3.4}$$

where λ_1^- and λ_3^+ are the two extreme wave speeds enclosing the Riemann fan of the one-dimensional problem (3.3); for the co-volume equation of state, $p(1-b\rho) = (\gamma-1)e\rho$, $b\geq 0$, these two extreme wave speeds are given by

$$\lambda_1^-(p^*) = v_L - c_L\left(1+\frac{\gamma+1}{2\gamma}\left(\frac{p^*-p_L}{p_L}\right)_+\right)^{\frac{1}{2}}, \tag{3.5}$$

$$\lambda_3^+(p^*) = v_R + c_R\left(1+\frac{\gamma+1}{2\gamma}\left(\frac{p^*-p_R}{p_R}\right)_+\right)^{\frac{1}{2}}, \tag{3.6}$$

where the intermediate pressure p^* is obtained by solving a nonlinear problem; see Toro [46, Chap. 4] and [14, sect. 3]. Here we use the notation $z_+ := \max(0,z)$, and we recall that the local sound speed for the co-volume gas is $c = \sqrt{\frac{\gamma p}{\rho(1-b\rho)}}$. We recall in passing that the widely used estimate $\max(|v_L|+c_L,|v_R|+c_R)$ is *not* a guaranteed upper bound of $\lambda_{\max}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$.

A guaranteed upper bound on $\lambda_{\max}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$ for (3.3) with the co-volume equation of state is given in [13, Rem. 2.8] and [14, Lem. 4.3]. For the reader's convenience, we now recall how this upper bound is estimated. We start by observing that $\lambda_{\max}(\boldsymbol{w}_L,\boldsymbol{w}_R)(p^*)$ is a monotone increasing function of p^*. Therefore, instead of computing the exact pressure p^* which requires an iterative process, one can use an explicit upper bound on p^*, say $\widehat{p}\geq p^*$. Then, a guaranteed upper bound on $\lambda_{\max}(\boldsymbol{w}_L,\boldsymbol{w}_R)(p^*)$ is $\widehat{\lambda}_{\max}(\boldsymbol{w}_L,\boldsymbol{w}_R) := \max((\lambda_1^-(\widehat{p}))_-,(\lambda_3^+(\widehat{p}))_+)$. One such upper bound valid for $1<\gamma\leq\frac{5}{3}$ is established in [14, Lem. 4.3], and the value of $\widehat{p}$ in question is given by the so-called two-rarefaction approximation:

$$\widehat{p} := \left(\frac{c_L(1-b\rho_L)+c_R(1-b\rho_R)-\frac{\gamma-1}{2}(v_R-v_L)}{c_L(1-b\rho_L)\,p_L^{-\frac{\gamma-1}{2\gamma}}+c_R(1-b\rho_R)\,p_R^{-\frac{\gamma-1}{2\gamma}}}\right)^{\frac{2\gamma}{\gamma-1}}. \tag{3.7}$$

Remark 3.1 (other discretizations). Note that the expression (3.1) that is used to compute the update $\mathbf{U}_i^{n+1}$ is quite generic; many other discretizations of the Euler equations can be put in this abstract form. The notion of a continuous finite element only intervenes in the definition of the vectors $\boldsymbol{c}_{ij}$, the index set $\mathcal{I}(i)$, and the lumped mass matrix coefficients m_i. Other discretizations lead to other forms for $\boldsymbol{c}_{ij}$, $\mathcal{I}(i)$, and m_i. Almost everything that is said in the rest of the paper can be applied to these discretizations as well. ∎

3.2. The intermediate limiting states. We now deduce from (3.1) intermediate local states that will be useful to limit the yet-to-be-defined high-order solution. Using that $\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij}=0$, we rewrite (3.1) as follows:

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

Then, upon introducing the quantities

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

with the convention $\overline{\mathbf{U}}_{ii}^{n+1} = \mathbf{U}_i^n$, and the notation $\overline{\mathbf{U}}_{ij}^{n+1} = (\overline{\rho}_{ij}^{n+1}, \overline{\boldsymbol{m}}_{ij}^{n+1}, \overline{E}_{ij}^{n+1})^\intercal$, the low-order update $\mathbf{U}_i^{L,n+1}$ can be represented as a convex combination as follows:

$$\mathbf{U}_i^{L,n+1} = \left(1 - \sum_{i\neq j\in\mathcal{I}(i)} \frac{2\tau d_{ij}^{L,n}}{m_i}\right)\overline{\mathbf{U}}_{ii}^n + \sum_{i\neq j\in\mathcal{I}(i)} \left(\frac{2\tau d_{ij}^{L,n}}{m_i}\right)\overline{\mathbf{U}}_{ij}^{n+1}, \tag{3.9}$$

under the appropriate CFL condition. Lemma 2.1 from [14], inspired by Perthame and Shu [40, sect. 5] and Nessyahu and Tadmor [38, eq. (2.7)], implies that the intermediate state $\overline{\mathbf{U}}_{ij}^{n+1}$ is a space average of the Riemann solution $\boldsymbol{u}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$; that is to say, $\overline{\mathbf{U}}_{ij}^{n+1} = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{u}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)(x, t_{ij})\,\mathrm{d}x$ with $t_{ij} := \|\boldsymbol{c}_{ij}\|_{\ell^2}/(2d_{ij}^{L,n})$ provided $t_{ij}\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \leq \frac{1}{2}$. (Let us emphasize that though the time t_{ij} is related to the Riemann problem (3.3), this time has nothing to do with that of the PDE (2.1).) Notice that the definition of the low-order graph viscosity (3.2) does imply that $t_{ij}\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \leq \frac{1}{2}$. An immediate consequence of this structure is that $\overline{\mathbf{U}}_{ij}^{n+1}$ has positive density, positive internal energy, and satisfies the following minimum principle on the specific entropy: $\Phi(\overline{\mathbf{U}}_{ij}^{n+1}) \geq \min(\Phi(\mathbf{U}_i^n), \Phi(\mathbf{U}_j^n))$. Another consequence of the above observation is the following result.

THEOREM 3.2 (local invariance/entropy inequality). *Let $i \in \mathcal{I}$. Assume* (2.7) *and* $1 + 2\tau\frac{d_{ii}^{L,n}}{m_i} \geq 0$. (i) *Let* $s_i^{\min} = \min_{j\in\mathcal{I}(i)} \Phi(\mathbf{U}_j^n)$*; then* $\mathbf{U}_i^{L,n+1} \in A_{s_i^{\min}}$. (ii) *Let* $(\eta := \rho f(\Phi), \boldsymbol{q} := \boldsymbol{m} f(\Phi))$ *be a generalized entropy pair for* (2.1). *Then the following local entropy inequality holds:*

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

A practical interpretation of item (i) is that the low-order solution $\mathbf{U}_i^{L,n+1}$ has positive density, positive internal energy, and satisfies the local minimum principle on the specific entropy. Item (ii) shows that this solution is also entropy satisfying in some discrete sense. This result is proved in [13, Thm. 4.7] in a more general setting for any hyperbolic system with a convex entropy. Note that in mathematical papers the entropies are generally assumed to be convex, whereas the physical generalized entropies $\rho f(\Phi)$ are concave (this is just a matter of sign convention).

3.3. Smoothness-based approximation. In this and the following subsection we introduce high-order approximation techniques that will provide us with a provisional high-order solution $\mathbf{U}_j^{H,n+1}$. The method presented in this section is easy to implement but is inherently only second-order accurate in space.

We introduce a technique to reduce the graph viscosity that is based on a measure of the local smoothness of the solution in the spirit of the finite volume literature (see

e.g., Jameson, Schmidt, and Turkel [25, eq. (12)] and see the second formula in the right column of page 1490 in Jameson [24]). Given a scalar-valued function g and its finite element interpolant $g_h = \sum \mathsf{G}_i \varphi_i$, and denoting $\epsilon_i = \epsilon \max_{j\in\mathcal{I}(i)} |\mathsf{G}_j|$ where $\epsilon = 10^{-\frac{16}{2}}$ in double precision arithmetic, we define the smoothness indicator of g as follows:

$$\alpha_i(g_h) := \frac{\left|\sum_{j\in\mathcal{I}(i)} \beta_{ij}(\mathsf{G}_j - \mathsf{G}_i)\right|}{\max(\sum_{j\in\mathcal{I}(i)} |\beta_{ij}||\mathsf{G}_j - \mathsf{G}_i|, \epsilon_i)}, \tag{3.10}$$

where the real numbers β_{ij} are selected to obtain $\alpha_i(g_h) = 0$ if g_h is linear on the support of the shape function φ_i; this property is called linearity-preserving (see Berger, Aftosmis, and Murman [3] for a review on linearity-preserving limiters in the finite volume literature). One simple choice for the coefficients β_{ij} consists of setting $\beta_{ij} = \int_D \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x$ (note that we do not require $\beta_{ij} > 0$ in (3.10)). One can also use the mean-value coordinates, e.g., Floater [8, eq. 5.1]. 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.

Notice that $\alpha_i \in [0,1]$ for all $i \in \mathcal{I}$ and $\alpha_i = 1$ if G_i is a local extremum of g_h. Moreover, if the coefficients β_{ij} are defined so that the linearity-preserving property holds, then the numerator of (3.10) behaves like $h^2\|D^2 g(\boldsymbol{\xi})\|_{\ell^2(\mathbb{R}^{d\times d})}$ 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 $\alpha_i \approx h\|D^2 g(\boldsymbol{\xi})\|_{\ell^2(\mathbb{R}^{d\times d})}/\|\nabla g(\boldsymbol{\zeta})\|_{\ell^2(\mathbb{R}^d)}$, that is to say α_i is of order h in the regions where g is smooth and does not have a local extremum.

Let $\psi \in C^{0,1}([0,1];[0,1])$ be any positive function such that $\psi(1) = 1$. The high-order smoothness-based graph viscosity is defined by setting

$$d_{ij}^{H,n} := d_{ij}^{L,n} \max(\psi(\alpha_i^n(g_h)), \psi(\alpha_j^n(g_h))), \quad d_{ii}^{H,n} := -\sum_{i\neq j\in\mathcal{I}(i)} d_{ij}^{H,n}. \tag{3.11}$$

A typical choice for ψ consists of setting $\psi(\alpha) = \alpha^2$. Then the provisional high-order approximation is computed as follows:

$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{\tau}(\mathbf{U}_j^{H,n+1} - \mathbf{U}_j^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - d_{ij}^{H,n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = 0. \tag{3.12}$$

One choice for g that we consider in some numerical tests reported at the end of the paper consists of using the mathematical entropy; that is, $g(\boldsymbol{u}) = \rho\Phi(\boldsymbol{u}) = S(\boldsymbol{u})$. Other possible options consist of using generalized entropies of the Euler equations, $g(\boldsymbol{u}) = \rho f(\Phi(\boldsymbol{u}))$. In particular, taking $f(s) = 1$ gives $g(\boldsymbol{u}) = \rho$, which is an extreme case of generalized entropy; it is extreme in the sense that $-g(\boldsymbol{u})$ is convex but not strictly convex. Note in passing that it is shown in Guermond, Popov, and Tomas [18] that using $g(\boldsymbol{u}) = \rho$ guarantees positivity of the density provided the coefficients β_{ij} are defined to be positive, and mass lumping is used in (3.12). Another option, which is somewhat similar to that of Jameson, Schmidt, and Turkel [25, eq. (12)] and Jameson [24, p. 1490], consists of taking $g(\boldsymbol{u}) = p$. Note, however, that it might be better to take $p^{\frac{1}{\gamma}}$ to be entropy consistent, since $p^{\frac{1}{\gamma}}$ is an extreme generalized entropy for polytropic gas as shown in Harten [19, eq. (2.10a)]. Let us emphasize that strict convexity of the entropy is not needed here for the purpose of the present paper.

Notice that we use the consistent mass matrix in (3.12) to reduce dispersion error since it is known that the use of the consistent mass matrix corrects the dominating

dispersion error (at least for piecewise linear approximation); see Christon, Martinez, and Voth [6], Gresho, Sani, and Engelman [11], Guermond and Pasquetti [12], Thompson [45]. The beneficial effects of the consistent mass matrix are particularly visible when solving problems with nonsmooth solutions; see, e.g., [12, Fig. 5.5].

Although the above smoothness-based method has been used in the literature for many years (e.g., [25]) and has been shown to work properly for scalar conservation equations with convex flux (i.e., genuinely nonlinear flux), it is not robust. More precisely, it may fail to converge properly in the presence of composite waves; that is, the method may converge to a weak nonentropic solution. We refer the reader to Guermond and Popov [15, sect. 6.5] and Kurganov, Petrova, and Popov [29, sect. 4] where this effect is documented. For this reason, in the next section we describe a method that is based on a measure of the entropy production and, from our experience, is more robust than the smoothness-based technique.

3.4. Entropy viscosity commutator. We now introduce a method that is formally high-order for any polynomial degree, contrary to the one introduced in section 3.3. Our objective is to construct a high-order method that is entropy consistent and close to being invariant domain preserving. We do not want to rely on the yet to be explained limiting process to enforce entropy consistency. We refer the reader to Lemmas 3.2 and 4.6 and section 6.1 in Guermond and Popov [15] and Guermond and Popov [13, sect. 5.1] for counter-examples of methods that are invariant domain preserving but entropy violating. The heuristics we have in mind is that limiting should be understood as a light polishing applied to a method that is already entropy consistent and almost invariant domain preserving. Following an idea introduced in Guermond et al. [16, 17], we construct a high-order graph viscosity that is entropy consistent by estimating a nondimensional entropy residual. However, contrary to the techniques introduced in [16, 17], we do not want the time discretization to interfere with the estimation of the residual, so we now propose a slightly different approach. Given the current approximation $\boldsymbol{u}_h^n$, we estimate the next inviscid approximation by setting $\mathbf{U}_i^{G,n+1} := \mathbf{U}_i^n - \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}$. Essentially, $\mathbf{U}_i^{G,n+1}$ is the Galerkin approximation of $\boldsymbol{u}(t^{n+1})$. Let $(\eta(\boldsymbol{v}), \boldsymbol{F}(\boldsymbol{v}))$ be an entropy pair for (2.1). We estimate the entropy residual for the degree of freedom i by computing $\frac{m_i}{\tau}(\mathbf{U}_i^{G,n+1} - \mathbf{U}_i^n)\cdot\eta'(\mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \boldsymbol{F}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}$. By using the definition of $\mathbf{U}_i^{G,n+1}$, this is equivalent to computing $\sum_{j\in\mathcal{I}(i)}(\boldsymbol{F}(\mathbf{U}_j^n) - \eta'(\mathbf{U}_i^n)^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij} := N_i^n$. At this point we need to realize that making the following change on the entropy $\rho f(s) \to \rho(f(s) - \beta)$, where β is an arbitrary constant, does not change the value of N_i^n. To account for this invariance, we define the following relative entropy $\eta_i^n(\mathbf{U}) = \rho(f(\Phi(\mathbf{U})) - f(\Phi(\mathbf{U}_i^n)))$ and the corresponding entropy flux, $\boldsymbol{F}_i^n(\mathbf{U}) = \mathbf{M}(f(\Phi(\mathbf{U})) - f(\Phi(\mathbf{U}_i^n)))$, with the convention $\mathbf{U} := (\rho, \mathbf{M}, E)^{\mathsf{T}}$. These definitions are equivalent to $\eta_i^n(\mathbf{U}) = \eta(\mathbf{U}) - \frac{\rho}{\rho_i^n}\eta(\mathbf{U}_i^n)$ and $\boldsymbol{F}_i^n(\mathbf{U}) = \mathbf{M}(\frac{\eta(\mathbf{U})}{\rho} - \frac{\eta(\mathbf{U}_i^n)}{\rho_i^n})$. We set

$$N_i^n := \sum_{j\in\mathcal{I}(i)} (\boldsymbol{F}_i^n(\mathbf{U}_j^n) - ((\eta_i^n)'(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}, \tag{3.13}$$

$$D_i^n := \left|\sum_{j\in\mathcal{I}(i)} \boldsymbol{F}_i^n(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\right| + \sum_{k=1}^{d+2}\left|\partial_{u_k}\eta_i^n(\mathbf{U}_i^n)\right| \times \left|\sum_{j\in\mathcal{I}(i)} \mathbb{f}_{u_k}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\right|, \tag{3.14}$$

where $\mathbb{f}_{u_1}, \ldots, \mathbb{f}_{u_{d+2}}$ are the $\mathbb{R}^d$-valued components of the flux $\mathbb{f}$. We then construct a

normalized entropy viscosity ratio:

$$R_i^n = \frac{|N_i^n|}{D_i^n}. \tag{3.15}$$

Notice that $R_i^n \in [0,1]$ and the definition of R_i^n is invariant by the following scaling of the entropy: $\rho f(s) \to \rho(\alpha f(s) - \beta)$ for any $\alpha, \beta \in \mathbb{R}$. The quantity N_i^n appearing at the numerator in the definition of the entropy residual R_i^n can be interpreted as a commutator. More specifically, N_i^n can be rewritten as follows: $\int_D (\nabla\cdot(\Pi_h \boldsymbol{F}(\boldsymbol{u}_h^n)) - (\eta_i^n)'(\mathbf{U}_i^n)^\mathsf{T}\nabla\cdot(\Pi_h(\mathbb{f}(\boldsymbol{u}_h^n))))\varphi_i \,\mathrm{d}\boldsymbol{x}$, where $\Pi_h : C^0(D;\mathbb{R}^{d+2}) \to \boldsymbol{P}(\mathcal{T}_h)$ is the Lagrange interpolation operator. Notice in passing that $N_i^n = 0$ in the hypothetical case that $\eta : \mathbb{R}^{d+2} \to \mathbb{R}$ is linear. Finally, the high-order graph viscosity (or entropy viscosity (EV)) is defined by setting

$$d_{ij}^{H,n} = d_{ij}^{L,n} \max(|R_i^n|, |R_j^n|), \quad d_{ii}^{H,n} := -\sum_{i \neq j \in \mathcal{I}(i)} d_{ij}^{H,n}, \tag{3.16}$$

and the provisional high-order approximation is computed as follows:

$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{\tau}(\mathbf{U}_j^{H,n+1} - \mathbf{U}_j^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - d_{ij}^{H,n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = 0. \tag{3.17}$$

Note again that we use the consistent mass matrix to reduce dispersion errors, as explained in section 3.3.

Remark 3.3 (decay rate on R_i^n). Let us now convincingly show that R_i^n is at least one order smaller (in term of mesh size) than $d_{ij}^{L,n}$. Let use denote $F''_{\max}$ and $f''_{\max}$ the maximum over the convex hull of the states $\{\mathbf{U}_j^n\}_{j\in\mathcal{I}(i)}$ of the matrix norm (say the norm induced by the Euclidean norm in $\mathbb{R}^{d+2}$) of the Hessians $D^2\boldsymbol{F}$ and $D^2\mathbb{f}$. Then, denoting by N_i^n the numerator in (3.15), and recalling that $D\boldsymbol{F}(\mathbf{U}) = \eta'(\mathbf{U})^\mathsf{T} D\mathbb{f}(\mathbf{U})$, we have

$$\begin{aligned}\|N_i^n\|_{\ell^2} &= \left\| \sum_{j\in\mathcal{I}(i)} \big(\boldsymbol{F}(\mathbf{U}_j^n) - \boldsymbol{F}(\mathbf{U}_i^n) - \eta'(\mathbf{U}_i^n)^\mathsf{T}(\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\big)\cdot\boldsymbol{c}_{ij} \right\|_{\ell^2} \\ &\le \frac{1}{2}(F''_{\max} + \eta'(\mathbf{U}_i^n) f''_{\max}) \max_{j\in\mathcal{I}(i)} \|\boldsymbol{c}_{ij}\|_{\ell^2} \sum_{j\in\mathcal{I}(i)} \|\mathbf{U}_j^n - \mathbf{U}_i^n\|_{\ell^2}^2.\end{aligned}$$

Assuming that η' is not zero over D_i and denoting by $\eta'_{\min}$ the minimum of $\|\eta'\|_{\ell^2}$ over $\mathrm{conv}(\mathbf{U}_j^n)_{j\in\mathcal{I}(i)}$, the quantity $\eta'_{\min}\sum_{j\in\mathcal{I}(i)} \|\mathbf{U}_j^n - \mathbf{U}_i^n\|_{\ell^2}$ is a lower bound for the denominator in (3.15). The conclusion follows readily. ∎

Remark 3.4 (choice of high-order graph viscosity). One advantage we see in the EV (3.16) over the smoothness-based viscosity (3.11) is that, in addition to being consistent for any polynomial degree, it is also consistent with at least one entropy inequality. That is to say the viscosity is large when the entropy production is large and it is small otherwise. In any case, we have observed that (3.16) always gives a scheme that is more robust than (3.11) albeit being slightly more oscillatory. We refer the reader to Guermond and Popov [15, sect. 6.5], where this issue is discussed in detail. ∎

Remark 3.5 (entropy). We have found in our numerical experiments that using $p^{\frac{1}{\gamma}}$ for polytropic gases is a very good choice to construct the entropy residual since

it minimizes dissipation across contact discontinuities. Recall that $p^{\frac{1}{\gamma}}$ is indeed an entropy in the case of polytropic gases. All of the numerical tests reported at the end of the paper are done with this entropy. ∎

Remark 3.6 (entropy ansatz). In realistic applications the equation of state is often tabulated, and the entropy $\Phi(\mathbf{U})$ may be either unavailable or costly to estimate. We have found that the pressure may be used as an ansatz for the entropy; one can then replace (3.15) by setting $N_i^n := \sum_{j\in\mathcal{I}(i)} \mathbf{V}_i^n\cdot\boldsymbol{c}_{ij}(\mathsf{P}_j^n - D\mathsf{P}(\mathbf{U}_i^n)^\mathsf{T}\mathbf{U}_j^n)$, $D_i^n := \max(|\mathsf{P}_i^{\max} - \mathsf{P}_i^{\min}|,\epsilon_i)\sum_{j\in\mathcal{I}(i)}|\mathbf{V}_i^n\cdot\boldsymbol{c}_{ij}|$, with $\epsilon_i = \epsilon\max(|\mathsf{P}_i^{\max}|,|\mathsf{P}_i^{\min}|)$ and $\mathbf{V}_i := \mathbf{M}_i/\rho_i$. ∎

3.5. Slip boundary condition. Since some of the numerical examples reported at the end of the paper involve using the slip boundary condition $\boldsymbol{v}\cdot\boldsymbol{n} = 0$, we explain in this section how this condition is enforced with Lagrange finite elements.

The slip boundary condition can be imposed weakly by adding a penalty term to the system as done in traditional discontinuous Galerkin formulations, but our experience is that weakly imposing the slip boundary condition may cause severe time-step restriction for explicit schemes; see, for example, Nazarov and Larcher [37, sect. 4]. Therefore, we impose the slip boundary condition strongly. Let ∂D_0 be the part of the boundary where the condition $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ has to be enforced, i.e., $\partial D_0 = \{\boldsymbol{x}\in\partial D \mid \boldsymbol{v}(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) = 0,\ \text{a.e. } t>0\}$. Let $\mathcal{I}^\partial$ be the set of the indices of the boundary Lagrange nodes, and let $\mathcal{I}_0^\partial = \{j\in\mathcal{I}^\partial \mid \varphi_{j|\partial D_0}\not\equiv 0\}$. Let $\mathcal{F}_h^\partial$ be the boundary faces of the mesh. We assume that the mesh $\mathcal{T}_h$ is constructed so that there exists a subset $\mathcal{F}_{h0}^\partial$ of $\mathcal{F}_h^\partial$ that exactly covers ∂D_0. For any $j\in\mathcal{I}_0^\partial$ let $(F_l)_{l\in L_j}$ be the list of the boundary faces in $\mathcal{F}_{h0}^\partial$ such that $\boldsymbol{x}_j$ belongs to F_l for all $l\in L_j$, and let $\boldsymbol{n}_l^j := \boldsymbol{n}_{|F_l}(\boldsymbol{x}_j)$ for all $l\in L_j$; then we set $\widetilde{\boldsymbol{n}}_j = \mathrm{card}(L_j)^{-1}\sum_{l\in L_j}\boldsymbol{n}_l^j$ and $\boldsymbol{n}_j := \widetilde{\boldsymbol{n}}_j/\|\widetilde{\boldsymbol{n}}_j\|_{\ell^2}$. The vector $\boldsymbol{n}_j$ is an approximation of the outward pointing unit normal at the Lagrange nodes $\boldsymbol{x}_j$ for any $j\in\mathcal{I}_0^\partial$. Let us now approximate the flux term $\int_D\nabla\cdot\mathbb{f}(\boldsymbol{u})\varphi_i\,\mathrm{d}\boldsymbol{x}$. One integration by parts gives the identity $\int_D\nabla\cdot\mathbb{f}(\boldsymbol{u}^n)\varphi_i\,\mathrm{d}\boldsymbol{x} = -\int_D\mathbb{f}(\boldsymbol{u}^n)\nabla\varphi_i\,\mathrm{d}\boldsymbol{x} + \int_{\partial D}(\mathbb{f}(\boldsymbol{u}^n)\boldsymbol{n})\varphi_i\,\mathrm{d}\boldsymbol{x}$, which we approximate as follows:

$$\int_D\nabla\cdot\mathbb{f}(\boldsymbol{u}^n)\varphi_i\,\mathrm{d}\boldsymbol{x} \approx \sum_{j\in\mathcal{I}(i)} -\mathbb{f}(\mathbf{U}_j^n)\int_D\varphi_j\nabla\varphi_i\,\mathrm{d}\boldsymbol{x} + \mathbb{f}(\mathbf{U}_j^n)\sum_{F\in\mathcal{F}_h\setminus\mathcal{F}_{h0}^\partial}\int_F\varphi_j\varphi_i\boldsymbol{n}\,\mathrm{d}\boldsymbol{x} + \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{n}_j\sum_{F\in\mathcal{F}_{h0}^\partial}\int_F\varphi_j\varphi_i\,\mathrm{d}\boldsymbol{x}.$$

We then slightly change the definition of $\boldsymbol{c}_{ij}$ given in (2.10) and set instead

$$\boldsymbol{c}_{ij} := -\int_D\varphi_j\nabla\varphi_i\,\mathrm{d}\boldsymbol{x} + \sum_{F\in\mathcal{F}_h\setminus\mathcal{F}_{h0}^\partial}\int_F\varphi_j\varphi_i\boldsymbol{n}\,\mathrm{d}\boldsymbol{x} + \boldsymbol{n}_j\sum_{F\in\mathcal{F}_{h0}^\partial}\int_F\varphi_j\varphi_i\,\mathrm{d}\boldsymbol{x}. \tag{3.18}$$

The expressions (2.10) and (3.18) are identical if $\boldsymbol{x}_i$ or $\boldsymbol{x}_j$ are internal nodes, or if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are boundary nodes and these two nodes sit on a flat portion of the boundary. Notice also that (3.18) is an approximation of (2.10) that is consistent with the polynomial degree of the finite element space when the boundary is smooth.

The boundary conditions are enforced at the end of each Euler substep of the SSP RK algorithm. In particular, denoting by $\boldsymbol{m}_h^{n+1}$ the approximation of the momentum at the time step t^{n+1}, the slip boundary condition is enforced by replacing $\mathbf{M}_j^{n+1}$ by $\mathbf{M}_j^{n+1} - (\mathbf{M}_j^{n+1}\cdot\boldsymbol{n}_j)\boldsymbol{n}_j$ for all the indices j in $\mathcal{I}_0^\partial$.

LEMMA 3.7 (conservation). *Let $\boldsymbol{c}_{ij}$ be defined by* (3.18). *Assume that $\mathcal{I}^\partial = \mathcal{I}_0^\partial$, and let $m_j^\partial := \int_{\partial D} \varphi_j \,\mathrm{d}s$. Assume that $\mathbf{M}_j^n \cdot \boldsymbol{n}_j = 0$ for all $j \in \mathcal{I}^\partial$. Let $\boldsymbol{U}^{n+1}$ be the update given by one of the following schemes:* (3.1), (3.12), *or* (3.17); *then*

$$\sum_{j\in\mathcal{I}} m_j \rho_j^{n+1} = \sum_{j\in\mathcal{I}} m_j \rho_j^n, \quad \sum_{j\in\mathcal{I}} m_j \mathsf{E}_j^{n+1} = \sum_{j\in\mathcal{I}} m_j \mathsf{E}_j^n,$$
$$\sum_{j\in\mathcal{I}} m_j \mathbf{M}_j^{n+1} = \sum_{j\in\mathcal{I}} m_j \mathbf{M}_j^n - \tau \sum_{j\in\mathcal{I}} m_j^\partial \mathsf{P}_j^n \boldsymbol{n}_j.$$

Proof. Summing (3.1), (3.12), or (3.17) over i in $\mathcal{I}$ and using the partition of unity property together with (3.18) gives

$$\sum_{i\in\mathcal{I}} m_i \mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{I}} m_i \mathbf{U}_i^{n+1} - \tau \sum_{j\in\mathcal{I}^\partial} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{n}_j \sum_{F\in\mathcal{F}_h^\partial} \int_F \varphi_j \,\mathrm{d}\boldsymbol{x}.$$

The rest of the proof follows from the definition of $\mathbb{f}$ and the boundary condition $\mathbf{M}_j \cdot \boldsymbol{n}_j = 0$ for all $j \in \mathcal{I}^\partial$. ∎

4. Quasiconcavity-based limiting. In this section we discuss the bounds we want the numerical solution $\mathbf{U}_i^{n+1}$ to satisfy, and we develop a novel limiting technique that is convexity-based and does not invoke arguments like linearization, worst-case scenario estimates, a posteriori fixes, or auxiliary discontinuous spaces as is often seen in the literature. This technique takes its roots in Khobalatte and Perthame [27], Perthame and Qiu [39], and Perthame and Shu [40]. We also refer to Zhang and Shu [51, 52] and Jiang and Liu [26] for extensions in the context of the discontinuous Galerkin approximation.

4.1. Bounds and quasi-concavity. Since the high-order update $\mathbf{U}_i^{H,n+1}$ (using either (3.12) or (3.17)) is not guaranteed to be oscillation free, and to preserve physical bounds, some form of limiting must applied. The question is now the following: What should be limited and how? Whichever representation is chosen for the dependent variable (conserved, primitive, or characteristic variables), the Euler equations are not known to satisfy any maximum or minimum principle, with the exception of the minimal principle on the specific entropy. Despite this fundamental negative result and with varying levels of success, a number of techniques have been proposed over the years in the finite element literature to enforce some kinds of discrete maximum principles (see, for instance, Boris and Book [5], Zalesak [49], Löhner et al. [35], Kuzmin and Möller [30], Zalesak [50], and Lohmann and Kuzmin [34]). Some of these limiting techniques enforce properties that are not necessarily satisfied by the Euler equations, or in the best case scenario, satisfied by the first-order method of choice (usually a Lax–Friedrichs-like first-order scheme).

In the present paper, we take a different point of view. In addition to the local minimum principle on the specific entropy, the strategy that we propose consists of enforcing bounds that are naturally satisfied by the low-order solution. More precisely, let us set

$$\rho_i^{\min} := \min_{j\in\mathcal{I}(i)} (\overline{\rho}_{ij}^{n+1}, \rho_j^n), \qquad \rho_i^{\max} := \max_{j\in\mathcal{I}(i)} (\overline{\rho}_{ij}^{n+1}, \rho_j^n), \tag{4.1}$$

$$E_i^{\min} := \min_{j\in\mathcal{I}(i)} (\overline{E}_{ij}^{n+1}, E_j^n), \qquad E_i^{\max} := \max_{j\in\mathcal{I}(i)} (\overline{E}_{ij}^{n+1}, E_j^n), \tag{4.2}$$

$$s_i^{\min} := \min_{j\in\mathcal{I}(i)} \Phi(\mathbf{U}_j^n). \tag{4.3}$$

We have already established in section 3.2 that $\rho_i^{\min} \le \rho_i^{L,n+1} \le \rho_i^{\max}$, $E_i^{\min} \le E_i^{L,n+1} \le E_i^{\max}$, and $s_i^{\min} \le \Phi(\mathbf{U}_i^{L,n+1})$. In the next section we are going to modify the graph viscosity so that the resulting high-order update $\mathbf{U}_i^{n+1}$ satisfies $\rho_i^{\min} \le \rho_i^{n+1} \le \rho_i^{\max}$ and $s_i^{\min} \le \Phi(\mathbf{U}_i^{n+1})$ (and possibly $E_i^{\min} \le E_i^{n+1} \le E_i^{\max}$ if one wishes).

In general, the equation for the specific entropy may not be explicitly available and, therefore, limiting the specific entropy may not be possible. An alternative strategy consists of limiting the internal energy ρe. Using the Frechet derivative notation, a straightforward computation shows that $D^2(\rho e(\boldsymbol{u}))((\varrho,\boldsymbol{q},a),(\varrho,\boldsymbol{q},a)) = -\frac{1}{\rho}(\frac{\varrho}{\rho}\boldsymbol{m}-\boldsymbol{q})^2$ for all directions $(\varrho,\boldsymbol{q},a)^\intercal \in \mathbb{R}^{d+2}$ and all points $\boldsymbol{u}=(\rho,\boldsymbol{m},E)^\intercal \in \mathbb{R}^{d+2}$, thereby showing that the internal energy is concave with respect to the conserved variables irrespective of the equation of state. Hence, the concavity of (ρe) along with the convex combination (3.9) implies that the low-order solution $\mathbf{U}_i^{L,n+1}$ satisfies the following discrete minimum principle:

$$(\rho e)(\mathbf{U}_i^{L,n+1}) \ge \varepsilon_i^{\min} := \min\Big(\min_{j\in\mathcal{I}(i)}(\rho e)(\mathbf{U}_j^n), \min_{j\in\mathcal{I}(i)}(\rho e)(\overline{\mathbf{U}}_{ij}^{n+1})\Big). \tag{4.4}$$

In order to unify into one single framework all of the bounds that we want to enforce, we are going to rely on the notion of quasi-concavity, the definition of which we now recall.

DEFINITION 4.1 (quasi-concavity). *Given a convex set $\mathcal{A}\subset\mathbb{R}^m$, we say that a function $\Psi:\mathcal{A}\to\mathbb{R}$ is quasi-concave if every* upper *level set of Ψ is* convex*; that is, the set $L_\lambda(\Psi):=\{\mathbf{U}\in\mathcal{A}\mid\Psi(\mathbf{U})\ge\lambda\}$ is convex for any $\lambda\in\mathbb{R}$ in the range of Ψ.*

Note in passing that concavity implies quasi-concavity. We are going to use the above definition in the following three settings: (i) $\mathcal{A}=\mathbb{R}^{d+2}$ and $\Psi(\mathbf{U})=\rho-\rho_i^{\min}$ or $\Psi(\mathbf{U})=\rho_i^{\max}-\rho$. Note that in both cases the upper level sets are half spaces (i.e., these sets are obviously convex). (ii) $\mathcal{A}=\{\mathbf{U}:=(\rho,\boldsymbol{m},E)\mid\rho>0\}$ and $\Psi(\mathbf{U})=(\rho e)-\varepsilon_i^{\min}$. We have shown above that $(\rho e)(\mathbf{U})$ is concave provided $\rho>0$ (the Hessian of ρe is nonpositive); then it follows that $\{\mathbf{U}:=(\rho,\boldsymbol{m},E)\mid\rho>0,e>0\}$ is convex. (iii) $\mathcal{A}=\{\mathbf{U}:=(\rho,\boldsymbol{m},E)\mid\rho>0,e>0\}$, $\Psi(\mathbf{U})=\Phi(\mathbf{U})-s_i^{\min}$. The quasi-convexity of $\Psi:\mathcal{A}\to\mathbb{R}$ is proved in Serre [42, Thm. 8.2.2].

Remark 4.2 (concavity versus quasi-concavity). Note that the two sets

$$\{(\rho,\boldsymbol{m},E)\mid\rho>0,e>0,s\ge r\},\qquad \{(\rho,\boldsymbol{m},E)\mid\rho>0,\rho e>0,\rho(s-r)\ge 0\}$$

are identical. In the first case, quasi-concavity is invoked to prove that the upper level sets are convex, whereas in the second case one just has to rely on concavity since the three functions ρ, $\rho e(\boldsymbol{u})$, and $\rho(\Phi(\boldsymbol{u})-r)$ are concave. It is easier to impose concave (or convex) constraints than quasi-concave ones. More precisely, in practice it is simpler to apply Newton's method on a concave function than on a quasi-concave function; in the first case Newton's method is guaranteed to converge under appropriate assumptions on the initial guess, whereas it may not in the second case. ∎

4.2. An abstract limiting scheme. Simple linear constraints, such as $\overline{\rho}_i^{\min}\le\rho_i^{n+1}\le\overline{\rho}_i^{\max}$ and $\overline{E}_i^{\min}\le E_i^{n+1}\le\overline{E}_i^{\max}$, can be easily enforced by using the Flux Transport Corrected paradigm of Zalesak [49] (see also Boris and Book [5]). However, to the best of our knowledge, the Zalesak's grouping methodology cannot be (easily) extended to handle general convex constraints like the minimum principle on the specific entropy without losing second-order accuracy. We introduce in this section a methodology that does exactly that.

We start as in the FCT methodology by estimating the difference $\mathbf{U}_i^{H,n+1}-\mathbf{U}_i^{L,n+1}$. Subtracting (3.1) from (3.12) (or (3.17)) we obtain the following identity satisfied by the low-order and the (provisional) high-order solution:

$$\sum_{j\in\mathcal{I}(i)} m_{ij}(\mathbf{U}_j^{H,n+1}-\mathbf{U}_j^n)-\tau d_{ij}^{H,n}(\mathbf{U}_j^n-\mathbf{U}_i^n)-\delta_{ij}m_i(\mathbf{U}_i^{L,n+1}-\mathbf{U}_i^n)+\tau d_{ij}^{L,n}(\mathbf{U}_j^n-\mathbf{U}_i^n)=\mathbf{0}.$$

This equality is rewritten in the following form better suited for postprocessing:

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

where we have set $\Delta_{ij}:=m_i\delta_{ij}-m_{ij}$. The above identity can be rewritten as follows:

$$\begin{cases} m_i(\mathbf{U}_i^{H,n+1}-\mathbf{U}_i^{L,n+1})=\displaystyle\sum_{j\in\mathcal{I}(i)}\boldsymbol{A}_{ij}^n, \\ \boldsymbol{A}_{ij}^n:=\Delta_{ij}\big(\mathbf{U}_j^{H,n+1}-\mathbf{U}_j^n-(\mathbf{U}_i^{H,n+1}-\mathbf{U}_i^n)\big)+\tau(d_{ij}^{H,n}-d_{ij}^{L,n})(\mathbf{U}_j^n-\mathbf{U}_i^n), \end{cases} \tag{4.5}$$

where we used $\sum_{j\in\mathcal{I}(i)}\Delta_{ij}=0$. Observe that the matrix $\boldsymbol{A}^n$ is skew-symmetric; the immediate consequence is that $\sum_{i\in\mathcal{I}} m_i\mathbf{U}_i^{H,n+1}=\sum_{i\in\mathcal{I}} m_i\mathbf{U}_i^{L,n+1}$, i.e., the total mass of the provisional high-order solution is the same at that of the low-order solution.

The next step consists of introducing symmetric limiting parameters $\ell_{ij}=\ell_{ji}\in[0,1]$ and estimating ℓ_{ij} so that the new quantity $\mathbf{U}_i^{n+1}=\mathbf{U}_i^{L,n+1}+\frac{1}{m_i}\sum_{j\in\mathcal{I}(i)}\ell_{ij}\boldsymbol{A}_{ij}^n$ satisfies the expected bounds. Note again that the skew-symmetry of $\boldsymbol{A}^n$ together with the symmetry of the limiter implies that $\sum_{i\in\mathcal{I}} m_i\mathbf{U}_i^{n+1}=\sum_{i\in\mathcal{I}} m_i\mathbf{U}_i^{L,n+1}$ for any choice of limiter ℓ_{ij}, i.e., the limiting process is conservative. Using the notation introduced at the end of section 4.1, we seek ℓ_{ij} so that $\Psi(\mathbf{U}_i^{n+1})\geq 0$.

We now depart from the FCT algorithm as described in [49] by introducing $\lambda_j:=\frac{1}{\mathrm{card}(\mathcal{I}(i))-1}$, $j\in\mathcal{I}(i)\setminus\{i\}$, and rewriting (4.5) as follows:

$$\mathbf{U}_i^{n+1}=\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\lambda_j(\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij}),\qquad\text{with}\qquad\mathbf{P}_{ij}:=\frac{1}{m_i\lambda_j}\mathbf{A}_{ij}^n. \tag{4.6}$$

Note that $\mathbf{U}_i^{n+1}=\mathbf{U}_i^{L,n+1}$ if $\ell_{ij}=0$ and $\mathbf{U}_i^{n+1}=\mathbf{U}_i^{H,n+1}$ if $\ell_{ij}=1$. The following lemma is the driving force of the limiting technique that we propose.

LEMMA 4.3. *Let $\Psi(\boldsymbol{u}):\mathcal{A}\to\mathbb{R}$ be a quasi-concave function. Assume that the limiting parameters $\ell_{ij}\in[0,1]$ are such that $\Psi(\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij})\geq 0$ for all $j\in\mathcal{I}(i)\setminus\{i\}$; then the following inequality holds true:*

$$\Psi\left(\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\lambda_j(\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij})\right)\geq 0.$$

Proof. Let $L_0=\{\mathbf{U}\in\mathcal{A}\mid\psi(\mathbf{U})\geq 0\}$. By definition all of the limited states $\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij}$ are in L_0 for all $i\neq j\in\mathcal{I}(i)$. Since Ψ is quasi-concave, the upper level set L_0 is convex. As a result, the convex combination $\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\lambda_j(\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij})$ is in L_0, i.e., $\Psi\big(\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\lambda_j(\mathbf{U}_i^{L,n+1}+\ell_{ij}\mathbf{P}_{ij})\big)\geq 0$, which concludes the proof. ∎

LEMMA 4.4. *Let ℓ_j^i be defined by*

$$\ell_j^i=\begin{cases}1 & \text{if } \Psi(\mathbf{U}_i^{L,n+1}+\mathbf{P}_{ij})\geq 0,\\ \max\{\ell\in[0,1]\mid\Psi(\mathbf{U}_i^{L,n+1}+\ell\mathbf{P}_{ij})\geq 0\} & \text{otherwise}\end{cases} \tag{4.7}$$

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

Proof. (i) First, if $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) \geq 0$, we observe that $\Psi(\mathbf{U}_i^{L,n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for any $\ell \in [0,1]$ because $\mathbf{U}_i^{L,n+1} \in L_0(\Psi)$, $\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij} \in L_0(\Psi)$, and $L_0(\Psi)$ is convex. Second, if $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) < 0$, we observe that ℓ_j^i is uniquely defined, and for any $\ell \in [0, \ell_j^i]$ we have $\Psi(\mathbf{U}_i^{L,n+1} + \ell \mathbf{P}_{ij}) \geq 0$ because $\mathbf{U}_i^{L,n+1} \in L_0(\Psi)$, $\mathbf{U}_i^{L,n+1} + \ell_j^i \mathbf{P}_{ij} \in L_0(\Psi)$, and $L_0(\Psi)$ is convex. (ii) Since $\ell_{ij} = \min(\ell_j^i, \ell_i^j) \leq \ell_j^i$, the above construction implies that $\Psi(\mathbf{U}_i^{L,n+1} + \ell_{ij}\mathbf{P}_{ij}) \geq 0$. Note, finally, that $\ell_{ij} = \min(\ell_j^i, \ell_i^j) = \ell_{ji}$. ∎

Remark 4.5 (extension to general hyperbolic systems). Notice that Lemmas 4.3 and 4.4 are not specific to the Euler equations. These results can be used to limit solutions of arbitrary hyperbolic systems where the invariant domain is described by quasi-concave constraints. ∎

4.3. Application to the Euler equations. In this section we explain how to use Lemmas 4.3 and 4.4 to enforce the quasi-concave constraints described in section 4.1. The algorithm goes as follows:

(i) Given the state $\mathbf{U}^n$, which we assume to be admissible, we compute $\mathbf{U}^{L,n+1}$ and $\mathbf{U}^{H,n+1}$ as explained in section 3.1 as well as section 3.3 or 3.4.

(ii) The density is limited by invoking Lemmas 4.3 and 4.4 and the bounds described in section 4.1 to enforce the quasi-concave constraints $\Psi(\mathbf{U}) = \rho - \rho_i^{\min} \geq 0$ and $\Psi(\mathbf{U}) = \rho_i^{\max} - \rho \geq 0$. The resulting limiter is denoted by ℓ_{ij}^{ρ} and the details on the computation of ℓ_{ij}^{ρ} are given in section 4.4.

(iii) The internal energy $\rho e := E - \frac{\boldsymbol{m}^2}{2\rho}$ is limited by invoking Lemmas 4.3 and 4.4 to enforce the quasi-concave constraint $\Psi(\mathbf{U}) = E - \frac{\boldsymbol{m}^2}{2\rho} - \varepsilon_i^{\min} \geq 0$. The corresponding limiter is denoted $\ell_{ij}^e \leq \ell_{ij}^{\rho}$ and the details on the computation of ℓ_{ij}^e are given in section 4.5.

(iv) The minimum principle on the specific entropy is enforced by using $\Psi(\mathbf{U}) = \Phi(\mathbf{U}) - s_i^{\min}$. The details on the computation of the corresponding limiter $\ell_{ij}^s \leq \ell_{ij}^e$ are given in section 4.6.

(v) Finally, upon setting $\ell_{ij} := \ell_{ij}^s$, the update $\mathbf{U}^{n+1}$ is computed by setting $\mathbf{U}^{n+1} = \mathbf{U}^{L,n+1} + \frac{1}{m_i} \sum_{j \in \mathcal{I}(i)} \ell_{ij} \boldsymbol{A}_{ij}^n$. This type of limiting can be iterated a few times by observing that

$$\mathbf{U}^{H,n+1} = \mathbf{U}^{L,n+1} + \frac{1}{m_i} \sum_{j \in \mathcal{I}(i)} \ell_{ij} \boldsymbol{A}_{ij}^n + \frac{1}{m_i} \sum_{j \in \mathcal{I}(i)} (1 - \ell_{ij}) \boldsymbol{A}_{ij}^n.$$

Then setting $\mathbf{U}^{(0)} := \mathbf{U}^{L,n+1}$ and $\boldsymbol{A}_{ij}^{(0)} := \boldsymbol{A}_{ij}^n$, the iterative algorithm proceeds as shown in Algorithm 1. In the numerical simulations reported at the end of this paper we have taken $k_{\max} = 1$.

Remark 4.6 (other quantities). As observed in section 4.1 it is also possible to impose additional limiting based on quasi-concave constraints. For example, one could limit the total energy from below and from above. Numerical experiments reveal that this extra limiting does not improve the performance of the scheme. All of the tests reported in section 5 are done by limiting the density and the specific entropy as described above. We have found that limiting the internal energy delivers second-

Algorithm 1. Iterative limiting.

Input: $\mathbf{U}^{L,n+1}$, $\boldsymbol{A}^n$, and $k_{\max}$
Output: $\mathbf{U}^{n+1}$
1: Set $\mathbf{U}^{(0)} := \mathbf{U}^{L,n+1}$ and $\boldsymbol{A}^{(0)} := \boldsymbol{A}^n$
2: **for** $k = 0$ **to** $k_{\max} - 1$ **do**
3: Compute limiter matrix $\ell^{(k)}$
4: Update $\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} + \frac{1}{m_i} \sum_{j \in \mathcal{I}(i)} \ell_{ij}^{(k)} \boldsymbol{A}_{ij}^{(k)}$
5: Update $\boldsymbol{A}_{ij}^{(k+1)} = (1 - \ell_{ij}^{(k)}) \boldsymbol{A}_{ij}^{(k)}$
6: **end for**
7: $\mathbf{U}^{n+1} = \mathbf{U}^{(k_{\max})}$

order accuracy in the maximum norm, but has a tendency to overdissipate contact discontinuities. Note that limiting the specific entropy amounts in effect to limiting the internal energy. ∎

Remark 4.7 (equation of state). So far, everything we have described is independent of the equation of state. ∎

4.4. Limiting on the density. The limiting on the density as specified by (4.7) proceeds as follows. To avoid divisions by zero, we introduce the small parameter $\epsilon := 10^{-14}$ and we set $\epsilon_i = \epsilon \rho_i^{\max}$ for all $i \in \mathcal{I}$. Let us denote by P_{ij}^{ρ} the ρ-component of $\boldsymbol{P}_{ij}$, and let us set

$$\ell_j^{i,\rho} = \begin{cases} \min(\frac{|\rho_i^{\min} - \rho_i^{L,n+1}|}{|P_{ij}^{\rho}| + \epsilon_i}, 1) & \text{if } \rho_i^{L,n+1} + P_{ij}^{\rho} < \rho_i^{\min}, \\ 1 & \text{if } \rho_i^{\min} \le \rho_i^{L,n+1} + P_{ij}^{\rho} \le \rho_i^{\max}, \\ \min(\frac{|\rho_i^{\max} - \rho_i^{L,n+1}|}{|P_{ij}^{\rho}| + \epsilon_i}, 1) & \text{if } \rho_i^{\max} < \rho_i^{L,n+1} + P_{ij}^{\rho}. \end{cases} \tag{4.8}$$

Setting $\Psi_+(\mathbf{U}) = \rho - \rho_i^{\min}$ and $\Psi_-(\mathbf{U}) = \rho_i^{\max} - \rho$, we have the following result whose proof is left to the reader.

LEMMA 4.8. *The definition* (4.8) *implies that* $\Psi_{\pm}(\mathbf{U}_i^{L,n+1} + \ell \mathbf{P}_{ij}) \ge 0$ *for all* $\ell \in [0, \ell_j^{i,\rho}]$.

Remark 4.9 (co-volume EOS). In the case of the co-volume equation of state, $p(1 - b\rho) = (\gamma - 1)\rho e$, it is known that $\mathcal{A} = \{(\rho, \boldsymbol{m}, E) \mid \rho > 0, e > 0, s \ge s^{\min}, b\rho < 1\}$ is an invariant domain; see Guermond and Popov [14, Prop. A.1]. The above method will enforce the additional affine constraint $1 - b\rho > 0$ automatically. ∎

Remark 4.10 (total energy). The limiting on the total energy can be done exactly as for the density. Let us emphasize though that we have not found this operation to be useful, and it is not done in the numerical tests reported at the end of this paper. ∎

4.5. Local minimum on the internal energy ρe. In this section we explain how to compute the limiter to enforce the local minimum principle on the internal energy ρe as stated in (4.4).

Upon setting $\Psi(\mathbf{U}) := (\rho e)(\mathbf{U}) - \varepsilon_i^{\min}$ with $\mathbf{U} := (\rho, \boldsymbol{m}, E)$ and $\varepsilon_i^{\min}$ defined in (4.4), by virtue of Lemmas 4.3 and 4.4, we have to estimate $\ell_j^{i,e} \in [0, \ell_j^{i,\rho}]$ so that $\Psi(\mathbf{U}_i^{L,n+1} + \ell \mathbf{P}_{ij}) \ge 0$ for all $\ell \in [0, \ell_j^{i,e}]$. In order to facilitate this computation, we

define the auxiliary function $\psi : \{\mathbf{U} \mid \rho > 0\} \to \mathbb{R}$,

$$\psi(\mathbf{U}) := \rho\Psi(\mathbf{U}) = (\rho^2 e)(\mathbf{U}) - \varepsilon_i^{\min}\rho = \rho E - \frac{1}{2}\boldsymbol{m}^2 - \varepsilon_i^{\min}\rho.$$

Then the above problem is equivalent to seeking $\ell_j^{i,e} \in [0, \ell_j^{i,\rho}]$ so that $\psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,e}]$. The key observation is that now ψ is a quadratic functional with

$$D\psi(\mathbf{U}) = \begin{pmatrix} E - \varepsilon_i^{\min} \\ -\boldsymbol{m} \\ \rho \end{pmatrix}, \qquad D^2\psi(\mathbf{U}) = \begin{pmatrix} 0 & \mathbf{0}^\mathsf{T} & 1 \\ \mathbf{0} & -\mathbb{I}_d & \mathbf{0} \\ 1 & \mathbf{0}^\mathsf{T} & 0 \end{pmatrix}.$$

Then upon setting $a := \frac{1}{2}\mathbf{P}_{ij}^\mathsf{T} D^2\psi \mathbf{P}_{ij}$, $b := D\psi(\mathbf{U}_i^{L,n+1})\cdot\mathbf{P}_{ij}$, and $c := \psi(\mathbf{U}_i^{L,n+1})$, we have

$$\psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = at^2 + bt + c.$$

Let t_0 be the smallest positive root of the equation $at^2+bt+c = 0$, with the convention that $t_0 = 1$ if the equation has no positive root. Then we choose $\ell_j^{i,e}$ to be such that

$$\ell_j^{i,e} = \min(t_0, \ell_j^{i,\rho}). \tag{4.9}$$

LEMMA 4.11. *The definition* (4.9) *implies that* $\Psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ *for all* $\ell \in [0, \ell_j^{i,e}]$.

Proof. If there is no positive root to the equation $at^2+bt+c = 0$ and since we have established that $c = \psi(\mathbf{U}_i^{L,n+1}) \geq 0$ (see (4.4)), we have $at^2 + bt + c \geq 0$ for all $t \geq 0$; that is, $\Psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ for all $\ell \geq 0$, and in particular this is true for all $\ell \in [0, \ell_j^{i,e}]$. Otherwise, if there is at least one positive root to the equation $at^2+bt+c = 0$, then denoting by t_0 the smallest positive root, we have $at^2+bt+c \geq 0$ for all $t \in [0, t_0]$ (if not, there would exist $t_1 \in (0, t_0)$ s.t. $at_1^2 + bt_1 + c < 0$, and the intermediate value theorem would imply the existence a root $t^* \in (0, t_1)$ which contradict that t_0 is the smallest positive root). This implies that $\Psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, t_0]$, and in particular this is true for all $\ell \in [0, \ell_j^{i,e}]$ owing to (4.9). ∎

Remark 4.12 (equation of state). Observe that the proposed limiting on ρe is independent of the equation of state. ∎

4.6. Minimum principle on the specific entropy. We now describe how to compute the limiter to enforce the local minimum principle on the specific entropy. Khobalatte and Perthame [27] is the first paper we are aware of where this type of limiting is done.

By virtue of Lemmas 4.3 and 4.4, we have to estimate $\ell_j^{i,s} \in [0, \ell_j^{i,e}]$ so that $\Psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,s}]$, with $\Psi(\mathbf{U}) := \Phi(\mathbf{U}) - s_i^{\min}$, where we recall that $\Phi(\mathbf{U}) := s(\rho, e(\mathbf{U}))$ is the specific entropy as a function of the conserved variables.

LEMMA 4.13. *Let* t_0 *be defined as follows:* (i) *If* $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) \geq 0$, *then we set* $t_0 = 1$. (ii) *If* $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) < 0$ *and* $\Psi(\mathbf{U}_i^{L,n+1}) > 0$, *we set* t_0 *to be the unique positive root to the equation* $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$. (iii) *If* $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) < 0$ *and* $\Psi(\mathbf{U}_i^{L,n+1}) = 0$, *the equation* $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ *has exactly two roots (possibly equal) and we take* t_0 *to be the largest nonnegative root. More precisely, if* $D\psi(\mathbf{U}_i^{L,n+1})\cdot\mathbf{P}_{ij} \leq 0$, *then* $t_0 = 0$, *and if* $D\psi(\mathbf{U}_i^{L,n+1})\cdot\mathbf{P}_{ij} > 0$, *then* $t_0 > 0$ *is the unique positive root of* $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ *and has to be computed. Then setting* $\ell_j^{i,s} = \min(t_0, \ell_j^{i,e})$, *we have* $\Psi(\mathbf{U}_i^{L,n+1} + \ell\mathbf{P}_{ij}) \geq 0$ *for all* $\ell \in [0, \ell_j^{i,s}]$.

Proof. Let us first observe that the equation $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ has at most two roots (possibly equal) because the upper level set $L_0 = \{\mathbf{U} \mid \Psi(\mathbf{U}) \geq 0\}$ is convex, and any line that intersects the upper level set crosses the boundary at two points, at most, say $t_- \leq t_+$ ($t_- = t_+$ when the line is tangential to the boundary of the upper level set). Note that if there are two roots, then $t_- \leq 0 \leq t_+$, since $\Psi(\mathbf{U}_i^{L,n+1}) \geq 0$. (i) If $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) \geq 0$, then $t_+ \geq 1$ and the entire segment $\{\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij} \mid t \in [0, t_0 = 1]\}$ is in L_0 by convexity. (ii) If $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) < 0$ and $\Psi(\mathbf{U}_i^{L,n+1}) > 0$, then $t_+ \in (0, 1)$, and upon setting $t_0 = t_+$, the entire segment $\{\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij} \mid t \in [0, t_0]\}$ is in L_0 by convexity. (iii) If $\Psi(\mathbf{U}_i^{L,n+1} + \mathbf{P}_{ij}) < 0$ and $\Psi(\mathbf{U}_i^{L,n+1}) = 0$, there are two possibilities: (i) $D\psi(\mathbf{U}_i^{L,n+1}) \cdot \mathbf{P}_{ij} \leq 0$; and (ii) $D\psi(\mathbf{U}_i^{L,n+1}) \cdot \mathbf{P}_{ij} > 0$. If $D\psi(\mathbf{U}_i^{L,n+1}) \cdot \mathbf{P}_{ij} \leq 0$, then by convexity $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) < 0$ for all $t > 0$. Hence, $t_+ = 0$ is the largest nonnegative root of the equation $\Psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ and, therefore, $t_0 = t_+ = 0$. In the remaining case, $D\psi(\mathbf{U}_i^{L,n+1}) \cdot \mathbf{P}_{ij} > 0$, we have that $0 < t_+ < 1$ and $t_0 = t_+$. ∎

Let us now explain how the above line search can be done efficiently. Thermodynamic principles imply that there exists a smooth function $g : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ such that $\rho e = g(s, \rho)$. Note that the identity $\partial_s e(\rho, s) = \frac{1}{\partial_e s(\rho, e)}$ together with the fundamental thermodynamic inequality $\partial_e s(\rho, e) > 0$, which is equivalent to the temperature being positive, implies $\partial_s g(\rho, s) > 0$. Since $\partial_s g > 0$, the minimum principle on the specific entropy $\Phi(\mathbf{U}) - s_i^{\min} \geq 0$ is equivalent to enforcing $g(s, \rho) = \rho e \geq g(s_i^{\min}, \rho)$; i.e., $\Psi(\mathbf{U}) := \rho e(\mathbf{U}) - g(s_i^{\min}, \rho) \geq 0$. When the function $g(s, \rho)$ satisfies $\partial_{\rho\rho} g \geq 0$, the function $\Psi(\mathbf{U})$ is concave with respect to the conserved variables; as a result, the line search $h(t) := \psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ can be done efficiently because h is concave and $h(0) \geq 0$. If $h(1) \geq 0$, we set $t_0 = 1$, and if $h(1) < 0$, we use a combination of secant and Newton methods to find the unique $0 \leq t_0 < 1$ such that $h(t_0) = 0$. For example, the co-volume equation of state falls into this category since in this case we have $\rho e = \frac{\rho^\gamma}{(1-b\rho)^{\gamma-1}} \exp((\gamma - 1)(s - s_0)) =: g(s, \rho)$. This is also the case for the stiffened gas equation of state, $\rho e = e_0\rho + p_\infty(1 - b\rho) + \frac{\rho^\gamma}{(1-b\rho)^{\gamma-1}} \exp((\gamma - 1)(s - s_0)) =: g(s, \rho)$, where e_0, s_0, and p_∞ are are constant coefficients characteristic of the thermodynamic properties of the fluid; see Metayer and Saurel [36] for details.

In the general case, i.e., when $g(s, \rho)$ does not satisfy $\partial_{\rho\rho} g \geq 0$, we can use a different strategy for imposing $\Phi(\mathbf{U}) - s_i^{\min} \geq 0$. Namely, using that $\rho > 0$ and using again a change of notation, we transform the constraint to $\Psi(\mathbf{U}) := \rho\Phi(\mathbf{U}) - s_i^{\min}\rho \geq 0$. Note that the function $-\rho\Phi(\mathbf{U})$ is a mathematical entropy for the Euler system, and under the standard assumptions (hyperbolicity and positive temperature) it is convex; see Harten et al. [21, Thm. 2.1]. Therefore, the line search $h(t) := \psi(\mathbf{U}_i^{L,n+1} + t\mathbf{P}_{ij}) = 0$ can be done efficiently because h is concave and $h(0) \geq 0$. If $h(1) \geq 0$, we set $t_0 = 1$, and if $h(1) < 0$, we use a combination of secant and Newton methods to find the unique $0 \leq t_0 < 1$ such that $h(t_0) = 0$.

4.7. Relaxation. It is observed in Khobalatte and Perthame [27, sect. 3.3] that strictly enforcing the minimum principle on the specific entropy 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 S." We have also observed this phenomenon. Moreover, it is well known that when applied to scalar conservation equations, limiting (in some broad sense) reduces the accuracy to first-order near maxima and minima of the solution. One typical way to address this issue in the finite volume literature consists of relaxing

the slope reconstructions; see Harten and Osher [20] and Schmidtmann, Abgrall, and Torrilhon [41, sect. 2.1]. In the present context, since we do not have any slope to reconstruct, we are going to relax the constraints so that the violation of the constraint is second-order accurate.

4.7.1. Relaxation on the density and the internal energy. Let us denote by ϱ one of the quantities that we may want to limit from below, excluding the specific entropy, say ρ, $-\rho$, or (ρe), and let $\varrho^{\min}$ be the corresponding bound given by the technique described in section 4.1. (Recall that limiting $-\rho$ from below is equivalent to limiting ρ from above.) For each $i \in \mathcal{I}$, we set $\Delta^2\varrho_i^n := \sum_{i\neq j\in\mathcal{I}(i)} \varrho_i^n - \varrho_j^n$, and we define

$$\overline{\Delta^2\varrho_i^n} := \frac{1}{2\mathrm{card}(\mathcal{I}(i))} \sum_{i\neq j\in\mathcal{I}(i)} \left(\frac{1}{2}\Delta^2\varrho_i^n + \frac{1}{2}\Delta^2\varrho_j^n\right), \tag{4.10}$$

$$\widetilde{\Delta^2\varrho_i^n} := \mathrm{minmod}\Big\{\frac{1}{2}\Delta^2\varrho_j^n \mid j \in \mathcal{I}(i)\Big\}, \tag{4.11}$$

where the minmod function of a finite set is defined to be zero if there are two numbers of different sign in this set, and it is equal to the number whose absolute value is the smallest otherwise. Then we propose two types of relaxation defined as follows:

$$\overline{\varrho_i^{\min}} = \max((1-r_h)\varrho_i^{\min}, \varrho_i^{\min} - |\overline{\Delta^2\varrho_i^n}|), \tag{4.12}$$

$$\widetilde{\varrho_i^{\min}} = \max((1-r_h)\varrho_i^{\min}, \varrho_i^{\min} - |\widetilde{\Delta^2\varrho_i^n}|), \tag{4.13}$$

where $r_h = (\frac{m_i}{|D|})^{\frac{1.5}{d}}$. When applying limiting we use either $\overline{\varrho_i^{\min}}$ or $\widetilde{\varrho_i^{\min}}$ instead of $\varrho_i^{\min}$. It is shown in the numerical section that both relaxations are robust.

Remark 4.14 (relaxation versus no relaxation). We have observed numerically that the proposed method is second-order accurate in the L^1-norm without relaxation if limiting is done on the density and the internal energy. Relaxation is necessary only to get second-order accuracy in the L^∞-norm. We have observed though that the minmod relaxation is slightly more restrictive than the other one since it does not deliver second-order accuracy in the maximum norm; only the averaging relaxation (4.12) has been found to give second-order in the L^∞-norm. ∎

Remark 4.15 (positivity). Note that the somewhat ad hoc threshold $(1-r_h)$ in the above definitions is positive and, when applied to the density or the internal energy, guarantees positivity of the density and the internal energy. The exponent 1.5 is somewhat ad hoc; in principle, one could take $r_h = (\frac{m_i}{|D|})^{\frac{\delta}{d}}$ with $\delta < 2$. Taking $\delta = 2$ would require one to multiply the tolerance $|\overline{\Delta^2\varrho_i^n}|$ by a problem-dependent constant. Taking $\delta = 1.5$ gives uniform performance of the method for all meshes without tuning any constant. ∎

4.7.2. Relaxation on the specific entropy. We proceed as in Khobalatte and Perthame [27, sect. 3.3] and relax the lower bound on the specific entropy more aggressively than on the density since in smooth regions this function is constant. Let ϱ be the quantity associated with the constraint on the specific entropy; it could be s or $\exp(s)$ (i.e., $\rho e/\rho^\gamma$ in the case of polytropic gases), depending on the way one chooses to enforce the minimum principle on the specific entropy (see section 4.6). Let $\boldsymbol{x}_{ij} = \frac{1}{2}(\boldsymbol{x}_i + \boldsymbol{x}_j)$. We measure the local variations of ϱ by setting $\Delta\varrho_i^n = \max_{i\neq j\in\mathcal{I}(i)}(\varrho^n(\boldsymbol{x}_{ij}) - \varrho_i^{\min})$ and we relax $\varrho_i^{\min}$ by setting

$$\overline{\varrho_i^{\min}} = \max((1-r_h)\varrho_i^{\min}, \varrho_i^{\min} - \Delta\varrho_i^n). \tag{4.14}$$

Note that contrary to the appearances, and as already observed in [27], the size of the relaxation is $\mathcal{O}(h^2)$: In the vicinity of shocks, there is no need for relaxation since the first-order viscosity takes over and thereby makes the solution minimum principle preserving on the specific entropy. In smooth regions, i.e., isentropic regions, the specific entropy is constant, and $\Delta\varrho_i^n$ measures the local curvature of s induced by the nonlinearity of s, which is to say $\Delta\varrho_i^n$ is $\mathcal{O}(h^2)$.

Remark 4.16 (positivity). The threshold $(1-r_h)\varrho_i^{\min}$ guarantees positivity of the internal energy. It also enforces a weak minimum principle on the specific entropy when using $\varrho = \frac{\rho e}{\rho^\gamma}$ in the case of polytropic gases. ∎

5. Numerical illustrations. We report in this section numerical tests we have done to illustrate the performance of the proposed method. All of the tests are done with the equation of state $p = (\gamma-1)\rho e$, i.e., $s(\rho, e) = \frac{1}{\gamma-1}\log(\rho e/\rho^\gamma)$.

5.1. Technical details. Three different codes implementing the method described in the paper have been written to ensure reproducibility. Limiting is done only once in the three codes, i.e., $k_{\max} = 1$. The three codes use the same low-order method (GMS-GV1) described in section 3.1. For two of these codes the high-order method is the entropy viscosity (EV) method as described in section 3.4; the third code uses the smoothness-based viscosity described in section 3.3.

The first code, henceforth called Code 1, does not use any particular software. It is based on Lagrange elements on simplices. This code has been written to be dimension-independent, i.e., the same data structure and subroutines are used in one dimension and in two dimensions. The two-dimensional meshes used for Code 1 are nonuniform triangular Delaunay meshes. All of the computations reported in the paper are done with continuous $\mathbb{P}_1$ elements. The high-order method uses the entropy viscosity commutator described in (3.15)–(3.16) with the entropy $p^{\frac{1}{\gamma}}$. Limiting on the density is enforced by using the technique described in section 4.4. The bounds on the density are relaxed by using the averaging technique described in section 4.7. The minimum principle on the specific entropy $\exp((\gamma-1)s) \geq \exp((\gamma-1)s^{\min})$ is enforced by using the method described in section 4.6 with the constraint $\Psi(\mathbf{U}) := \rho e - \varrho^{\min}\rho^\gamma \geq 0$, where we recall that $\rho e/\rho^\gamma = \exp((\gamma-1)s)$. The lower bound on the specific entropy is defined by using $\varrho_i^{\min} = \min_{j\in\mathcal{I}(i)} \rho_i^n e_i^n/(\rho_i^n)^\gamma$ (instead of (4.3)), and $\varrho_i^{\min}$ is relaxed by using (4.14) with $\varrho = \rho e/\rho^\gamma$. No limiting on the internal energy is applied in Code 1; the positivity of the internal energy is guaranteed by the minimum principle on the specific entropy.

The second code, henceforth called Code 2, uses the open-source finite element library FEniCS (see, e.g., Logg, Mardal, and Wells [33]) and the computations are done on simplices. The implementation in FEniCS is independent of the space dimension and the polynomial degree of the approximation. The library is fully parallel. All of the numerical integrations are done exactly by automatically determining the quadrature degree with respect to the complexity of the underlying integrand and the polynomial space. The results reported in the paper use the EV method with the entropy $p^{\frac{1}{\gamma}}$. The limiting and bound relaxation is done on the density, and the specific entropy is exactly like in Code 1. We refer the reader to the description of Code 1 for the details.

The last code, henceforth called Code 3, is based on the open-source finite element library deal.II; see Arndt et al. [1] and Bangerth, Hartmann, and Kanschat [2]. The tests reported in the paper are done with continuous $\mathbb{Q}_1$ (quadrilateral) elements. The code is written in a dimension-independent fashion, and all of the computational

tasks (e.g., assembly, linear solvers, and output) are implemented in parallel via message passing interface (MPI). This code implements the smoothness-based high-order graph viscosity described in section 3.3 with $\beta_{ij} = \int_D \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x$ for all i, j. All the computations reported in the paper are done with $g(\boldsymbol{u}) = S(\boldsymbol{u})$ and $\psi(\alpha) = \alpha^4$. (We have verified that the other choices for $g(\boldsymbol{u})$ mentioned in section 3.3 produce comparable results.) As stated at the beginning of section 3.3, this method introduces additional diffusion close to local extrema (whether smooth or not). While this does not affect the second-order decay rates in the L^1-norm, it degrades the accuracy in the L^∞-norm. Limiting is done on the density and the specific entropy exactly as explained above for Code 1. The minimum principle on the specific entropy is relaxed as explained in section 4.7.2, but no relaxation is applied on the density bounds.

The time stepping is done in the three codes by using the SSP(3,3) method (three stages, third-order); see Shu and Osher [43, eq. (2.18)] and Kraaijevanger [28, Thm. 9.4]. The time step is recomputed at each time step by using the formula $\tau = \mathrm{CFL} \times \min_{i\in\mathcal{I}} \frac{m_i}{|d_{ii}^{L,n}|}$ with $d_{ii}^{L,n} = -\sum_{ij} d_{ij}^{L,n}$ given in (3.2).

When working with manufactured solutions, for $q \in [1, \infty]$, we compute a consolidated error indicator at time t by adding the relative error in the L^q-norm of the density, the momentum, and the total energy as follows:

$$\delta_q(t) := \frac{\|\rho_h(t) - \rho(t)\|_{L^q(D)}}{\|\rho(t)\|_{L^q(D)}} + \frac{\|\boldsymbol{m}_h(t) - \boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}}{\|\boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}} + \frac{\|E_h(t) - E(t)\|_{L^q(D)}}{\|E(t)\|_{L^q(D)}}. \tag{5.1}$$

As some tests may exhibit superconvergence effects we also consider the fully discrete consolidated error indicator $\delta_q^\Pi(t)$ defined as above with $\rho(t)$, $\boldsymbol{m}(t)$, and $E(t)$ replaced by $\Pi_h\rho(t)$, $\Pi_h\boldsymbol{m}(t)$, and $\Pi_h E(t)$, where Π_h is the Lagrange interpolation operator.

5.2. One-dimensional smooth wave. We start with a one-dimensional test, the purpose of which is to estimate the convergence rate of the method with a very smooth solution. We consider the following exact solution to the Euler equations: $v(x,t) = 1$, $p(x,t) = 1$, and

$$\rho(x,t) = \begin{cases} 1 + 2^6(x_1 - x_0)^{-6}(x - t - x_0)^3(x_1 - x + t)^3 & \text{if } x_0 \le x - t < x_1, \\ 1 & \text{otherwise} \end{cases} \tag{5.2}$$

with $x_0 = 0.1$, $x_1 = 0.3$, and $\gamma = \frac{7}{5}$. The computational domain is $D = (0, 1)$, and the computation is done from $t = 0$ to $t = 0.6$. The consolidated error indicator in the maximum norm $\delta_\infty(t)$ is reported in Table 1. Note that we report the discrete error indicator $\delta_\infty^\Pi(t)$ for Code 1 (based on the EV method) in order to show that we obtain $\mathcal{O}(h^3)$ superconvergence in compliance with the theoretical result stated in Guermond and Pasquetti [12, Prop 2.2]. Code 3, which we recall is based on the smoothness of the mathematical entropy, delivers $\mathcal{O}(h^{1.5})$ as expected due to clipping effects induced by the smoothness indicator.

5.3. Rarefaction wave. We now consider a Riemann problem with a solution whose components are all continuous and whose derivatives have bounded variations. The best-approximation error in the L^1-norm on quasi-uniform meshes is then $\mathcal{O}(h^2)$. The Riemann problem in question has the following data: $(\rho_L, v_L, p_L) = (3, c_L, 1)$, $(\rho_R, v_R, p_R) = (\frac{1}{2}, v_L + \frac{2}{\gamma-1}(c_L - c_R), p_L(\frac{\rho_R}{\rho_L})^\gamma)$, where $c_L = \sqrt{\gamma p_L/\rho_L}$, $c_R = \sqrt{\gamma p_R/\rho_R}$. The equation of state is a gamma-law with $\gamma = \frac{7}{5}$. The exact solution to this problem is a rarefaction wave which can be constructed analytically; see, e.g., Toro [46, sect. 4.4]. The solution is given in Table 2. In this table, the ratio

TABLE 1
One-dimensional smooth wave, $\mathbb{P}_1$ meshes. Convergence tests with Code 1 and Code 3, CFL = 0.25.

# nodes	Code 1		Code 3	
	$\delta_\infty^{\Pi}(t)$	Rate	$\delta_\infty(t)$	Rate
100	9.02E-03		2.32E-01	
200	1.34E-04	6.07	8.30E-02	1.48
400	1.01E-05	3.72	2.87E-02	1.53
800	1.12E-06	3.18	9.66E-03	1.57
1600	1.23E-07	3.19	3.22E-03	1.58
3200	1.33E-08	3.21	1.07E-03	1.58
6400	1.42E-09	3.22	3.74E-04	1.52

$\xi := \frac{x-x_0}{t}$ is the self-similar variable, where x_0 is the location of the discontinuity at $t = 0$. This problem is quite challenging for any method enforcing the minimum principle on the specific entropy. We have observed that the convergence rate on this problem reduces to $\mathcal{O}(h)$ if the minimum on the specific entropy is not relaxed (see also Khobalatte and Perthame [27, sect. 3.3]). This test is meant to validate the relaxation technique described in (4.14).

TABLE 2
Solution to the rarefaction wave.

	$\xi \le v_L - c_L$	$v_L - c_L < \xi \le v_R - c_R$	$v_R - c_R < \xi$
ρ	ρ_L	$\rho_L\big(\frac{2}{\gamma+1} + \frac{\gamma-1}{\gamma+1}\frac{v_L-\xi}{c_L}\big)^{\frac{2}{\gamma-1}}$	ρ_R
v	v_L	$\frac{2}{\gamma+1}(c_L + \frac{\gamma-1}{2}v_L + \xi)$	v_R
p	p_L	$p_L\big(\frac{2}{\gamma+1} + \frac{\gamma-1}{\gamma+1}\frac{v_L-\xi}{c_L}\big)^{\frac{2\gamma}{\gamma-1}}$	p_R

We run Code 1, Code 2, and Code 3 on the computational domain $D = (0, 1)$ with $x_0 = 0.2$, and the initial time is $t_0 = \frac{0.2}{v_R - c_R}$. The initial data is the exact solution at $t = \frac{0.2}{v_R - c_R}$ given in Table 2. The simulations are run until $t = 0.5$. The consolidated error indicator $\delta_1(t)$ defined in (5.1) is reported in Table 3 for Code 1 and Code 3 only for brevity. This series of tests shows that the proposed method, with limiting of the density and the specific entropy as described in sections 4.4–4.6, converges with rate at least $\mathcal{O}(h^{1.5})$ on the rarefaction wave problem. We observe that the low-order method is indeed asymptotically first-order (the rate is 0.96 for 12800 grid points).

TABLE 3
Rarefaction wave, $\mathbb{P}_1$ meshes. Convergence tests with Code 1 and Code 3, CFL = 0.25.

# nodes	Code 1		Code 3		Galerkin		Low-order	
	$\delta_1(t)$	Rate	$\delta_1(t)$	Rate	$\delta_1(t)$	Rate	$\delta_1(t)$	Rate
100	1.30E-03		3.33E-03		1.44E-03		5.10E-02	
200	4.06E-04	1.68	1.08E-03	1.61	4.38E-04	1.71	2.96E-02	0.78
400	1.40E-04	1.54	3.57E-04	1.61	1.42E-04	1.62	1.68E-02	0.82
800	5.00E-05	1.48	1.18E-04	1.60	4.73E-05	1.59	9.23E-03	0.86
1600	1.78E-05	1.49	3.96E-05	1.58	1.60E-05	1.57	4.96E-03	0.89
3200	6.24E-06	1.51	1.31E-05	1.59	5.47E-06	1.55	2.62E-03	0.92
6400	2.11E-06	1.57	4.32E-06	1.61	1.82E-06	1.59	1.37E-03	0.94
12800	6.80E-07	1.63	1.38E-06	1.64	5.83E-07	1.64	7.05E-04	0.96

When comparing the results from Code 1 with the Galerkin solution, we observe that the extra dissipation induced in Code 1 by the entropy viscosity and limiting

is of the order of the truncation error, which is optimal. That is, the method does not introduce any extraneous dissipation on smooth solutions. Note in passing that the maximum norm error indicator $\delta_\infty(t)$ has also been computed for this test (not reported here for brevity), yielding the rate $\mathcal{O}(h)$ for Code 1 and $\mathcal{O}(h^{0.75})$ for Code 3. The rate $\mathcal{O}(h)$ is optimal since the solution is in $\boldsymbol{W}^{1,\infty}(D)$.

5.4. Leblanc shocktube. We continue with a Riemann problem that is known in the literature as the Leblanc shocktube. The data are as follows: $(\rho_L, v_L, p_L) = (1, 0, (\gamma-1)10^{-1})$ and $(\rho_R, v_R, p_R) = (10^{-3}, 0, (\gamma-1)10^{-10})$ and the equation of state is a gamma-law with $\gamma = \frac{5}{3}$. The exact solution is described in Table 4. Denoting by x_0 the location of the discontinuity at $t = 0$, the quantity $\xi = \frac{x-x_0}{t}$ is the self-similar variable and the other numerical values in the table are given with 15 digit accuracy by $\rho_L^* = 5.40793353493162\times10^{-2}$, $\rho_R^* = 3.99999806043000\times10^{-3}$, $v^* = 0.621838671391735$, $p* = 0.515577927650970\times10^{-3}$, $\lambda_1 = 0.495784895188979$, $\lambda_3 = 0.829118362533470$.

TABLE 4
Solution of the Leblanc shocktube.

	$\xi \le -\frac{1}{3}$	$-\frac{1}{3} < \xi \le \lambda_1$	$\lambda_1 < \xi \le v^*$	$v^* < \xi \le \lambda_3$	$\lambda_3 < \xi$
ρ	ρ_L	$(0.75 - 0.75\xi)^3$	ρ_L^*	ρ_R^*	ρ_R
v	v_L	$0.75(\frac{1}{3} + \xi)$	v^*	v^*	v_R
p	p_L	$\frac{1}{15}(0.75 - 0.75\xi)^5$	p^*	p^*	p_R

We do simulations with $x_0 = 0.33$ until $t = 2/3$ with Code 1 and Code 3. The error indicator $\delta_1(t)$ is reported in Table 5. The convergence rate on $\delta_1(t)$ is close to $\mathcal{O}(h^{0.9})$ for Code 1 and $\mathcal{O}(h)$ for Code 3, which is optimal for this problem.

TABLE 5
Leblanc shocktube, $\mathbb{P}_1$ meshes. Convergence tests with Code 1 and Code 3, CFL = 0.25.

# nodes	Code 1		Code 3		Low-order	
	$\delta_1(t)$	rate	$\delta_1(t)$	rate	$\delta_1(t)$	rate
100	1.21E-01		1.49E-01		2.61E-01	
200	7.56E-02	0.68	9.01E-02	0.72	1.94E-01	0.43
400	4.50E-02	0.75	4.92E-02	0.87	1.41E-01	0.46
800	2.64E-02	0.77	2.61E-02	0.91	9.95E-02	0.50
1600	1.49E-02	0.82	1.34E-02	0.96	6.74E-02	0.56
3200	8.35E-03	0.84	6.83E-03	0.97	4.40E-02	0.62
6400	4.55E-03	0.88	3.42E-03	0.99	2.78E-02	0.66
12800	2.49E-03	0.87	1.70E-03	1.00	1.73E-02	0.68

5.5. Sod, Lax, blast wave. We now illustrate the method on a series of traditional problems without giving the full tables for the convergence rates for the sake of brevity. We consider the Sod shocktube, the Lax shocktube, and the Woodward–Collela blast wave. We refer the reader to the literature for the initial data for these test cases. The computations are done on the domain $D = (0, 1)$ with CFL = 0.5 on four different grids with Code 1. The final times are $t = 0.225$ for the Sod shocktube, $t = 0.15$ for the Lax shocktube, and $t = 0.038$ for the Woodward–Collela blast wave. We show the graph of the density for these three cases and for the four meshes in Figure 1. We have observed the convergence rate to be between $\mathcal{O}(h^{0.9})$ and $\mathcal{O}(h)$ on $\delta_1(t)$ for both the Sod and the Lax shocktubes, which is near optimal (results not reported for the sake of brevity).

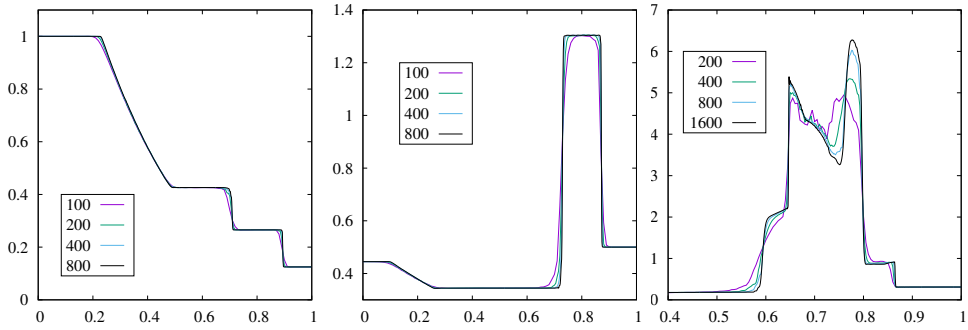

FIG. 1. *Code* 1, $CFL = 0.5$. *Left: Sod shocktube,* $t = 0.225$. *Center: Lax shocktube,* $t = 0.15$. *Right: Woodward–Collela blast wave,* $t = 0.038$.

5.6. Two-dimensional isentropic vortex. We now consider a two-dimensional problem introduced in Yee, Sandham, and Djomehri [48]. This test case is often used to assess the accuracy of numerical schemes. The flow field is isentropic; i.e., the solution is smooth and does not involve any steep gradients or discontinuities. Let $\rho_\infty = P_\infty = T_\infty = 1$, $\boldsymbol{u}_\infty = (u_\infty, v_\infty)^\mathsf{T}$, $u_\infty = 1$, $v_\infty = 1$, be free stream values; then the exact solution that we consider is a passive convection of a vortex with mean velocity $\boldsymbol{u}_\infty$:

$$\rho(\boldsymbol{x},t) = (T_\infty + \delta T)^{1/(\gamma-1)}, \qquad \boldsymbol{u}(\boldsymbol{x},t) = \boldsymbol{u}_\infty + \delta\boldsymbol{u}, \qquad p(\boldsymbol{x},t) = \rho^\gamma, \tag{5.3}$$

$$\delta\boldsymbol{u}(\boldsymbol{x},t) = \frac{\beta}{2\pi} e^{\frac{1-r^2}{2}}(-\bar{x}_2, \bar{x}_1), \qquad \delta T(\boldsymbol{x},t) = -\frac{(\gamma-1)\beta^2}{8\gamma\pi^2} e^{1-r^2}, \tag{5.4}$$

with $\bar{\boldsymbol{x}} = (x_1 - x_1^0 - u_\infty t, x_2 - x_2^0 - v_\infty t)$, and $r^2 = \|\bar{\boldsymbol{x}}\|_{\ell^2}^2$. It is standard to take $\gamma = \frac{7}{5} = 1.4$ and $\beta = 5$ for this test, $(x_1^0, x_2^0) = (4, 4)$.

We perform the numerical computations in the rectangle $D = (-5, 15)\times(-5, 15)$ from $t = 0$ until $t = 2$. The computations are done with Code 2 only. The initial mesh is uniform and consists of 20×20 squares each divided into two triangles; then the mesh is refined uniformly to compute finer solutions. The consolidated error indicators $\delta_1(t)$, $\delta_2(t)$, and $\delta_\infty(t)$ are reported in Table 6. Two series of computations are done: one with limiting and one without limiting. Here again we observe second-order accuracy in the maximum norm thereby confirming the accuracy of the proposed method. When comparing the results with limiting and without limiting we observe that limiting does not deteriorate the accuracy of the method. Contour lines of the density computed on two coarse meshes and contour lines of the exact solution are shown in Figure 2.

5.7. Supersonic flow around circular cylinder. In this example we consider a supersonic flow past a circular cylinder of radius 0.25 centered at the point $\boldsymbol{x} = (0.6, 1)$ in a two-dimensional wind tunnel of size $(x_1, x_2) \in [0, 4] \times [0, 2]$. We use the γ-law with $\gamma = \frac{7}{5}$. The initial data is $\rho = 1.4$, $p = 1$, $\boldsymbol{v} = (3, 0)^\mathsf{T}$. The Mach 3 flow enters from the left boundary $\{x_1 = 0\}$, where Dirichlet boundary conditions are prescribed. The slip boundary condition is applied at the walls $\{x_2 = 0\}$ and $\{x_2 = 2\}$ and on the cylinder. Since the flow is supersonic, no boundary condition is applied at the outflow $\{x_1 = 4\}$.

TABLE 6

Isentropic vortex, $\mathbb{P}_1$ meshes. Convergence tests with limiting and without limiting, $t = 2$. Code 2, CFL $= 0.1$.

	# nodes	$\delta_1(t)$	Rate	$\delta_2(t)$	Rate	$\delta_\infty(t)$	Rate
Limiting	441	2.46E-02	–	9.61E-02	–	1.31E+00	–
	1681	8.50E-03	1.59	3.36E-02	1.57	4.45E-01	1.62
	6561	1.99E-03	2.13	7.45E-03	2.21	9.40E-02	2.28
	25921	4.16E-04	2.28	1.50E-03	2.33	2.05E-02	2.21
	103041	8.11E-05	2.37	2.92E-04	2.37	4.68E-03	2.14
No limiting	441	2.47E-02	–	9.61E-02	–	1.31E+00	–
	1681	8.48E-03	1.60	3.34E-02	1.58	4.52E-01	1.59
	6561	1.99E-03	2.13	7.45E-03	2.20	9.40E-02	2.31
	25921	4.16E-04	2.28	1.50E-03	2.33	2.05E-02	2.21
	103041	8.11E-05	2.37	2.92E-04	2.37	4.68E-03	2.14

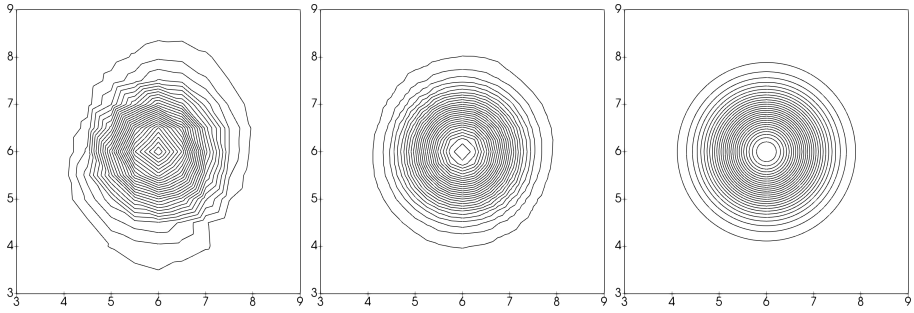

FIG. 2. *Two-dimensional isentropic vortex, $t = 2$, contour lines of density for 1681 $\mathbb{P}_1$ grid points (left), 6561 $\mathbb{P}_1$ grid points (middle), and the exact solution (right).*

The flow enters the wind tunnel and hits the cylinder; then a strong bow shock develops at the front of the cylinder, while two attached oblique shocks develop from the back side of the cylinder. The flow separates at the points where the oblique shocks start. After separation the flow fluctuates and creates small scale vortices traveling downstream. The strong shocks travel towards the wall boundaries, reflect back in the tunnel, and pass through the small scale vortices.

We show in Figure 3 the density and the entropy residual defined in (3.15). The computation is done with Code 2 with a mesh composed of 252820 unstructured $\mathbb{P}_1$ points and CFL $= 0.3$. The entropy residual captures the area with shocks and steep gradients, while it is almost zero in the smooth regions. As in the Mach 3 flow around a forward facing step, we observe two shock wave triple points. The Kelvin–Helmholtz instability deforming the two contact lines emerging from these triple points is clearly visible. This instability is usually hard to capture when the numerical diffusion is too large. The dynamics of the flow are captured as well as in Nazarov and Larcher [37, sect. 5.6], where the same problem is solved with about 500k $\mathbb{P}_2$ nodes.

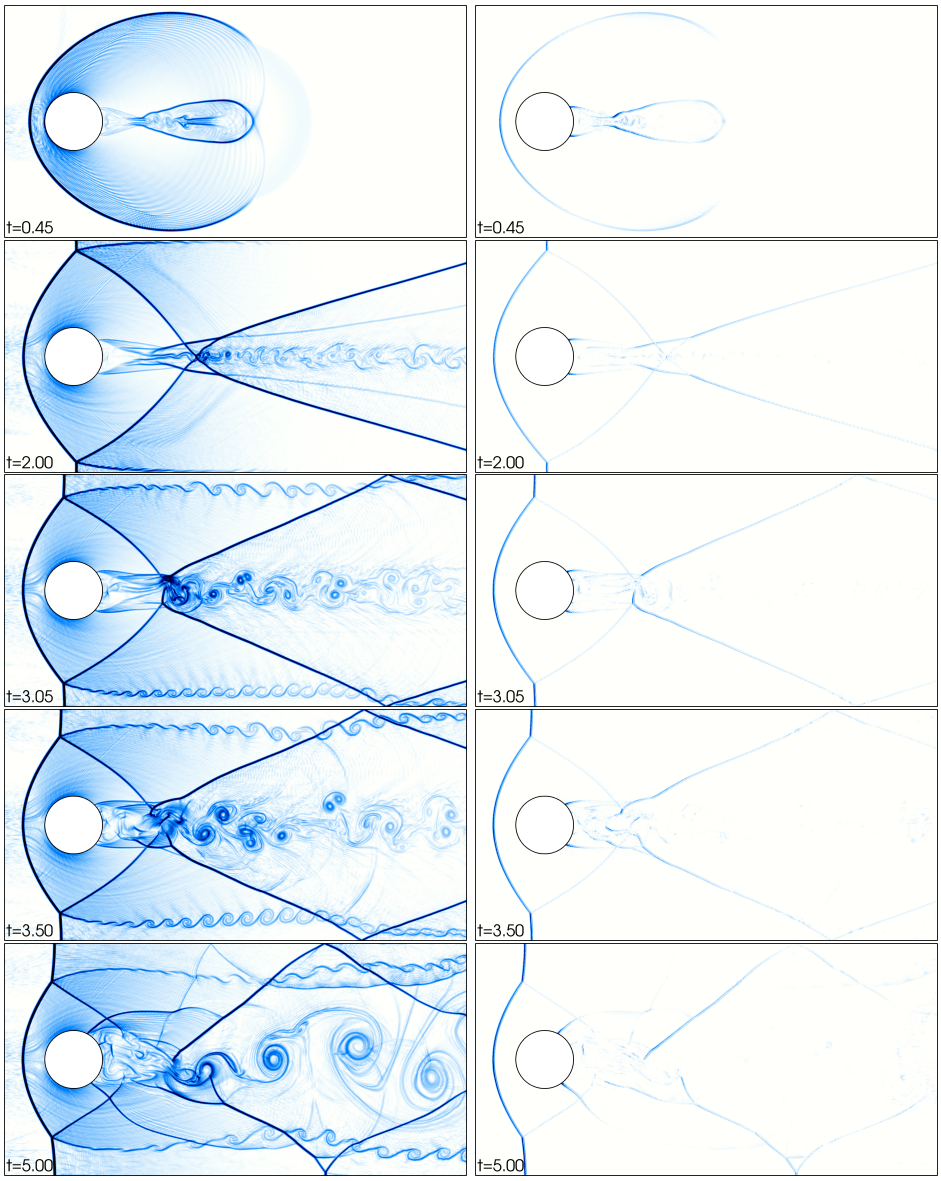

FIG. 3. *Supersonic Mach* 3 *flow around a circular cylinder;* 252820 *unstructured* $\mathbb{P}_1$ *nodes. The density* ρ_h *is plotted at different time levels in the left panels; the corresponding residual defined in* (3.15) *is shown in the right panels.*

Supplementary material. Additional tests on the forward facing step at Mach 3 and the two-dimensional double Mach reflection at Mach 10 are reported in the supplementary material section of this paper (supplementary.pdf [local/web 2.34MB]).

REFERENCES

[1] D. ARNDT, W. BANGERTH, D. DAVYDOV, T. HEISTER, L. HELTAI, M. KRONBICHLER, M. MAIER, J.-P. PELTERET, B. TURCKSIN, AND D. WELLS, *The* `deal.II` *library, version* 8.5, J. Numer.

Math., 25 (2017), pp. 137–145, https://doi.org/10.1515/jnma-2016-1045.
[2] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.*II – *a general purpose object oriented finite element library*, ACM Trans. Math. Softw., 33 (2007), 24.
[3] M. Berger, M. J. Aftosmis, and S. M. Murman, *Analysis of slope limiters on irregular grids*, AIAA Paper 2005-0490, American Institute for Aeronautics and Astronautics, Reno, NV, 2005, also NAS Technical Report NASA TM NAS-05-007.
[4] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math. (2), 161 (2005), pp. 223–342.
[5] J. P. Boris and D. L. Book, *Flux-corrected transport.* I. *SHASTA, a fluid transport algorithm that works [J. Comput. Phys.* 11 (1973)*,* 38–69*]*, J. Comput. Phys., 135 (1997), pp. 170–186.
[6] M. A. Christon, M. J. Martinez, and T. E. Voth, *Generalized Fourier analyses of the advection-diffusion equation-part* I: *One-dimensional domains*, Internat. J. Numer. Methods Fluids, 45 (2004), pp. 839–887.
[7] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.
[8] M. S. Floater, *Generalized barycentric coordinates and applications*, Acta Numer., 24 (2015), pp. 161–214.
[9] H. Frid, *Maps of convex sets and invariant regions for finite-difference systems of conservation laws*, Arch. Ration. Mech. Anal., 160 (2001), pp. 245–269.
[10] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.
[11] P. Gresho, R. Sani, and M. Engelman, *Incompressible Flow and the Finite Element Method: Advection-Diffusion and Isothermal Laminar Flow*, John Wiley and Sons, New York, 1998.
[12] J.-L. Guermond and R. Pasquetti, *A correction technique for the dispersive effects of mass lumping for transport problems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 186–198.
[13] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489, https://doi.org/10.1137/16M1074291.
[14] J.-L. Guermond and B. Popov, *Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations*, J. Comput. Phys., 321 (2016), pp. 908–926.
[15] J.-L. Guermond and B. Popov, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017), pp. 3120–3146, https://doi.org/10.1137/16M1106560.
[16] J.-L. Guermond, R. Pasquetti, and B. Popov, *Entropy viscosity method for nonlinear conservation laws*, J. Comput. Phys., 230 (2011), pp. 4248–4267.
[17] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 52 (2014), pp. 2163–2182, https://doi.org/10.1137/130950240.
[18] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., in review, https://arxiv.org/abs/1807.02563, 2018.
[19] A. Harten, *On the symmetric form of systems of conservation laws with entropy*, J. Comput. Phys., 49 (1983), pp. 151–164.
[20] A. Harten and S. Osher, *Uniformly high-order accurate nonoscillatory schemes.* I, SIAM J. Numer. Anal., 24 (1987), pp. 279–309, https://doi.org/10.1137/0724022.
[21] A. Harten, P. D. Lax, C. D. Levermore, and W. J. Morokoff, *Convex entropies and hyperbolicity for general Euler equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2117–2127, https://doi.org/10.1137/S0036142997316700.
[22] J. S. Hesthaven, *From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex*, SIAM J. Numer. Anal., 35 (1998), pp. 655–676, https://doi.org/10.1137/S003614299630587X.
[23] D. Hoff, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610.
[24] A. Jameson, *Origins and further development of the Jameson-Schmidt-Turkel scheme*, AIAA J., 55 (2017), pp. 1487–1510.
[25] A. Jameson, W. Schmidt, and E. Turkel, *Numerical solution of the Euler equations by finite volume. Methods using Runge-Kutta time-stepping schemes*, in Proceedings of the 14th AIAA Fluid and Plasma Dynamics Conference, 1981, AIAA Paper 1981–1259.
[26] Y. Jiang and H. Liu, *An invariant-region-preserving (irp) limiter for compressible euler equations*, in Proceedings of the Hyp2016 Conference, Numerics and Applications of Hyperbolic Problems II. HYP 2016, Springer Proc. Math. Stat. 237, Springer, Cham, 2017, pp. 71–83.

[27] B. Khobalatte and B. Perthame, *Maximum principle on the entropy and second-order kinetic schemes*, Math. Comp., 62 (1994), pp. 119–131.
[28] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.
[29] A. Kurganov, G. Petrova, and B. Popov, *Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws*, SIAM J. Sci. Comput., 29 (2007), pp. 2381–2401, https://doi.org/10.1137/040614189.
[30] D. Kuzmin and M. Möller, *Algebraic flux correction.* II. *Compressible Euler equations*, in Flux-Corrected Transport, Sci. Comput., Springer, Berlin, 2005, pp. 207–250.
[31] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558.
[32] D. Kuzmin, R. Löhner, and S. Turek, *Flux-Corrected Transport*, Scientific Computation, Springer-Verlag, Berlin, Heidelberg, 2005.
[33] A. Logg, K.-A. Mardal, and G. N. Wells, eds., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, Heidelberg, 2012.
[34] C. Lohmann and D. Kuzmin, *Synchronized flux limiting for gas dynamics variables*, J. Comput. Phys., 326 (2016), pp. 973–990.
[35] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati, *Finite element flux-corrected transport (FEM-FCT) for the euler and Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 7 (1987), pp. 1093–1109.
[36] O. Metayer and R. Saurel, *The noble-abel stiffened-gas equation of state*, Phys. Fluids, 28 (2016), 046102.
[37] M. Nazarov and A. Larcher, *Numerical investigation of a viscous regularization of the Euler equations by entropy viscosity*, Comput. Methods Appl. Mech. Engrg., 317 (2017), pp. 128–152.
[38] H. Nessyahu and E. Tadmor, *Nonoscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.
[39] B. Perthame and Y. Qiu, *A variant of Van Leer's method for multidimensional systems of conservation laws*, J. Comput. Phys., 112 (1994), pp. 370–381.
[40] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130.
[41] B. Schmidtmann, R. Abgrall, and M. Torrilhon, *On third-order limiter functions for finite volume methods*, Bull. Braz. Math. Soc. (N.S.), 47 (2016), pp. 753–764.
[42] D. Serre, *Systems of Conservation Laws.* 2. *Geometric Structures, Oscillations, and Initial-Boundary Value Problems*, translated from the 1996 French original by I. N. Sneddon, Cambridge University Press, Cambridge, 2000.
[43] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[44] M. A. Taylor, B. A. Wingate, and R. E. Vincent, *An algorithm for computing Fekete points in the triangle*, SIAM J. Numer. Anal., 38 (2000), pp. 1707–1720, https://doi.org/10.1137/S0036142998337247.
[45] T. Thompson, *A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation*, J. Comput. Appl. Math., 303 (2016), pp. 229–248.
[46] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction*, 3rd ed., Springer-Verlag, Berlin, 2009.
[47] T. Warburton, *An explicit construction of interpolation nodes on the simplex*, J. Engrg. Math., 56 (2006), pp. 247–262.
[48] H. C. Yee, N. D. Sandham, and M. J. Djomehri, *Low-dissipative high-order shock-capturing methods using characteristic-based filters*, J. Comput. Phys., 150 (1999), pp. 199–238.
[49] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.
[50] S. T. Zalesak, *The design of flux-corrected transport (FCT) algorithms for structured grids*, in Flux-Corrected Transport, Sci. Comput., Springer, Berlin, 2005, pp. 29–78.
[51] X. Zhang and C.-W. Shu, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.
[52] X. Zhang and C.-W. Shu, *A minimum entropy principle of high order schemes for gas dynamics equations*, Numer. Math., 121 (2012), pp. 545–563.