Journal of Computational Physics 478 (2023) 111926

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Robust second-order approximation of the compressible Euler equations with an arbitrary equation of state ☆

Bennett Clayton [a], Jean-Luc Guermond [a,*], Matthias Maier [a], Bojan Popov [a], Eric J. Tovar [b]

[a] Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843, USA
[b] X Computational Physics, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM, 87545, USA

ARTICLE INFO

Article history:
Received 25 July 2022
Received in revised form 15 December 2022
Accepted 9 January 2023
Available online 20 January 2023

Keywords:
Euler equations
Gas dynamics
Equation of state
Tabulated equation of state
Invariant-domain preserving
Convex limiting

ABSTRACT

This paper is concerned with the approximation of the compressible Euler equations supplemented with an arbitrary or tabulated equation of state. The proposed approximation technique is robust, formally second-order accurate in space, invariant-domain preserving, and works for every equation of state, tabulated or analytic, provided the pressure is nonnegative. An entropy surrogate functional that grows across shocks is proposed. The numerical method is verified with novel analytical solutions and then validated with several computational benchmarks seen in the literature including problems with composite waves.

© 2023 Elsevier Inc. All rights reserved.

1. Introduction

The objective of this paper is to continue the work started in Clayton et al. [9] where we introduced an explicit and invariant-domain preserving method to approximate the Euler equations equipped with an equation of state that is either tabulated or is given by an expression that is so involved that elementary Riemann problems cannot be solved exactly or efficiently. The method proposed in [9] is invariant-domain preserving in the sense that it guarantees that the density and the internal energy of the approximation are positive. A key ingredient of the method is a local approximation of the equation of state using a covolume ansatz (a.k.a. Noble-Abel equation of state) from which upper bounds on the maximum wave speed are derived for every elementary Riemann problem. The downside is that the method is first-order accurate in space. In the present paper we propose a technique that increases the accuracy in space to second-order and preserves

☆ This material is based upon work supported in part by the National Science Foundation grants DMS-1912847 (MM), DMS-2045636 (MM), DMS-2110868 (JLG, BP), by the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397 (JLG, BP), the Army Research Office, under grant number W911NF-19-1-0431 (JLG, BP), and the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contracts B640889, B641173 (JLG, BP). ET acknowledges the support from Los Alamos National Laboratory's (LANL) Advanced Simulation and Computing Program, Integrated Codes (IC) and Physics & Engineering Models (PEM) sub-programs, operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). ET also acknowledges support from the U.S. Department of Energy's Office of Applied Scientific Computing Research (ASCR) and Center for Nonlinear Studies (CNLS) at LANL under the Mark Kac Postdoctoral Fellowship in Applied Mathematics. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. http://www.tacc.utexas.edu.

* Corresponding author.
E-mail address: guermond@tamu.edu (J.-L. Guermond).

https://doi.org/10.1016/j.jcp.2023.111926
0021-9991/© 2023 Elsevier Inc. All rights reserved.

the invariant-domain properties. We also introduce a functional that can be used as an entropy surrogate. The functional in question is shown to be increasing across shocks in the Riemann problem involved in the construction of the local maximum wave speed. We show how to use this functional to limit the internal energy from below. This feature is useful for equations of state that are tabulated or interpolated from experimental data since in this case no natural notion of entropy is available.

It is sometimes possible, though possibly expensive, to solve Riemann problems when the equation of state is analytic. For instance this is done in Colella and Glaz [10, §1], Ivings et al. [28], Quartapelle et al. [40]. This cannot be done with tabulated equations of state because the information on the pressure is incomplete. Several attempts to develop methods working with an arbitrary or tabulated equation of state have been reported in the literature. One way to do so consists of using approximate Riemann solvers like those found in Dukowicz [11], [10, §2], Roe and Pike [43], Pike [39], and Lee et al. [31]. One can also simplify the Riemann problem by using flux splitting techniques as those reported in Toro et al. [51]. We also refer to Saurel et al. [44], Banks [2], Dumbser and Casulli [12], Dumbser et al. [13] where approximation techniques are developed using approximate Riemann solvers for various equations of state. Some of these techniques guarantee positivity of the density, but little else is guaranteed in general. The method introduced in Clayton et al. [9] is based on a graph viscosity technique using upper bounds on the maximum wave speed in the Riemann problem. Instead of using the two-shock approximation of the Riemann solution, as done in most methods based on approximate solvers, the method proposed in [9] approximates the pressure in the Riemann fan by the covolume equation of state and subsequently estimates guaranteed upper bounds on the maximum wave speed. This in turn ensures that the internal energy in the proposed algorithm is positive in addition to the density being positive. If it happens that the equation of state is of covolume type, then the method also preserves the minimum principle on the specific entropy. A technique based on similar principles is reported in Wang and Li [52] where the authors use a stiffened gas equation of state to approximate the pressure in the Riemann fan. To the best of our knowledge, the method proposed in the present paper is among the very first ones that are provably invariant-domain preserving for complex or tabulated equations of state and high-order accurate in space.

The paper is organized as follows. In Section 2, we introduce the mathematical model of interest and the corresponding notation. We also briefly discuss the assumptions we make on the equation of state and recall results from [9] that are needed for this work (i.e., we recall the first-order approximation of the Euler equations in §2.3). In Section 3, we construct a provisional update that is higher-order accurate in space. This update is based on a high-order graph viscosity using an entropy commutator and an activation function. Then, in Section 4, we apply a novel convex limiting technique that corrects the invariant-domain violations of the provisional higher-order method. The final result is an approximation technique that is robust, formally second-order accurate in space, provably invariant-domain preserving, and works for every equation of state (tabulated or analytic) that satisfies the mild assumptions stated in §2.2. Finally in Section 5, the method is verified with analytical solutions and published benchmarks and is validated with experiments. A short conclusion is given in §6.

2. Preliminaries

We formulate the problem and introduce notation in this section. We also recall essential results from [9] for completeness.

2.1. The Euler equations

Let D be a bounded polyhedron in $\mathbb{R}^d$. Given some initial data $\boldsymbol{u}_0(\boldsymbol{x}) := (\rho_0, \boldsymbol{m}_0, E_0)(\boldsymbol{x})$ over D at the initial time t_0, we look for $\boldsymbol{u}(\boldsymbol{x}, t) := (\rho, \boldsymbol{m}, E)(\boldsymbol{x}, t)$ solving the compressible Euler equations in the weak sense:

$$\partial_t \rho + \nabla\cdot(\boldsymbol{v}\rho) = 0 \qquad \text{a.e. } t > t_0, \boldsymbol{x} \in D, \tag{2.1a}$$

$$\partial_t \boldsymbol{m} + \nabla\cdot\big(\boldsymbol{v}\otimes\boldsymbol{m} + p(\boldsymbol{u})\mathbb{I}_d\big) = \boldsymbol{0} \qquad \text{a.e. } t > t_0, \boldsymbol{x} \in D, \tag{2.1b}$$

$$\partial_t E + \nabla\cdot\big(\boldsymbol{v}(E + p(\boldsymbol{u}))\big) = 0 \qquad \text{a.e. } t > t_0, \boldsymbol{x} \in D. \tag{2.1c}$$

The components of the dependent variable $\boldsymbol{u} := (\rho, \boldsymbol{m}, E)^\intercal \in \mathbb{R}^{d+2}$ (considered to be a column vector) are the density, ρ, the momentum, $\boldsymbol{m}$, and the total mechanical energy, E. We also introduce the velocity $\boldsymbol{v}(\boldsymbol{u}) := \rho^{-1}\boldsymbol{m}$ and the specific internal energy, $e(\boldsymbol{u}) := \rho^{-1}E - \frac{1}{2}\|\boldsymbol{v}(\boldsymbol{u})\|_{\ell^2}^2$. To simplify the notation later on we introduce the flux $\mathbb{f}(\boldsymbol{u}) := (\boldsymbol{m}, \boldsymbol{v}(\boldsymbol{u})\otimes\boldsymbol{m} + p(\boldsymbol{u})\mathbb{I}_d, \boldsymbol{v}(\boldsymbol{u})(E+p))^\intercal \in \mathbb{R}^{(d+2)\times d}$, where $\mathbb{I}_d$ is the $d\times d$ identity matrix.

2.2. Equation of state

In (2.1) the function $p(\boldsymbol{u})$ denotes the pressure which we assume to be given by an *oracle*. This means we assume to have no a priori knowledge of the function $p(\boldsymbol{u})$ itself apart from some mild structural assumptions stated below. We only assume that for a given state $\boldsymbol{u}$ we are able to retrieve a pressure $p(\boldsymbol{u})$ in a suitable way, for example by evaluating an arbitrary analytic expression, or by deriving a value from tabulated experimental data. For instance, in the numerical illustrations reported in §5, we consider oracles given by analytic functions (e.g., the ideal gas, van der Waals, Jones–Wilkins–Lee and Mie-Gruneisen equations of state) and we also use the SESAME database developed at Los Alamos National Laboratory [33] to test the method with experimental tabulated data.

Throughout the paper we assume that the domain of definition of the oracle $p(\boldsymbol{u})$ is the set $\mathcal{B}(b) \subset \mathbb{R}^{d+2}$ given by

$$\mathcal{B}(b) := \left\{ \boldsymbol{u} := (\rho, \boldsymbol{m}, E) \in \mathbb{R}^{d+2} \mid 0 < \rho,\ 0 < 1 - b\rho,\ 0 < e(\boldsymbol{u}) \right\}. \tag{2.2}$$

We henceforth refer to $\mathcal{B}(b)$ as the admissible set. One of the objectives of the paper is to guarantee that the approximation is high-order accurate and leaves $\mathcal{B}(b)$ invariant. The inequality $\rho < \frac{1}{b}$ found in the definition of $\mathcal{B}(b)$ is the so called *maximum compressibility* condition. The constant b can be set to zero if the user has no a priori knowledge about the maximum compressibility of the fluid under consideration. We recall, however, that a large class of analytic equations of state, such as the Noble-Abel, the Mie-Gruneisen (with the Hugoniot locus as the reference curve), and the Noble-Abel-Stiffened-Gas equations of state all involve a maximum compressibility constant. Finally, we assume that the oracle $p(\boldsymbol{u})$ returns a non-negative pressure,

$$p : \mathcal{B}(b) \to \mathbb{R}_{\geq 0}. \tag{2.3}$$

The assumption can be weakened, but for the sake of simplicity, we refrain from doing so in this paper. We leave this extension for future works.

2.3. First-order time and space approximation

The time and space discretization proposed in Clayton et al. [9] is based on [20]. This method is in some sense a discretization-agnostic generalization of an algorithm introduced by Lax [30, p. 163]. We denote by t^n the current time, $n \in \mathbb{N}$, and we denote by τ the current time step size; that is $t^{n+1} := t^n + \tau$. Without going into the details of the space approximation, we assume that the current approximation of $\boldsymbol{u}(\cdot, t^n)$ is a collection of states $\{\mathbf{U}_i^n\}_{i\in\mathcal{V}}$, where the index set $\mathcal{V}$ is used to enumerate all the degrees of freedom of the approximation, and $\mathbf{U}_i^n$ is in $\mathbb{R}^{d+2}$ for all $i \in \mathcal{V}$. The update at t^{n+1} is obtained as follows:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{L},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{J}(i)} \mathbb{f}(\mathbf{U}_j^n)\,\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{J}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) = \mathbf{0}. \tag{2.4}$$

The quantity m_i is called the lumped mass and we assume that $m_i > 0$ for all $i \in \mathcal{V}$. The vector $\boldsymbol{c}_{ij} \in \mathbb{R}^d$ encodes the space discretization. The index set $\mathcal{J}(i)$ is called the local stencil. This set collects only the degrees of freedom in $\mathcal{V}$ that interact with i (i.e., $j \notin \mathcal{J}(i) \Rightarrow \boldsymbol{c}_{ij} = \mathbf{0}$). We assume that the $\mathbb{R}^d$-valued coefficients $\boldsymbol{c}_{ij}$ are such that $\sum_{j\in\mathcal{J}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}$ approximates $\nabla\cdot(\mathbb{f}(\boldsymbol{u}(\cdot, t^n)))$ at some grid point $\boldsymbol{a}_i$, and the consistency error in space scales optimally with respect to the mesh size for the considered approximation setting. We require that the method be conservative; more precisely, we assume that

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji} \quad \text{and} \quad \sum_{j\in\mathcal{J}(i)} \boldsymbol{c}_{ij} = \mathbf{0}. \tag{2.5}$$

Concrete expressions for $\boldsymbol{c}_{ij}$ and m_i are given in [23, §4] for continuous and discontinuous finite elements as well as for finite volumes. Many approximation techniques found in the literature can be put in the above generic form. The computations reported at the end of this paper are done with piecewise linear continuous finite elements. But to stay general, we continue with the abstract discretization-agnostic notation introduced above.

Given $i \in \mathcal{V}$ and $j \in \mathcal{J}(i)$, we set $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2}^{-1}$. Then, the low-order graph viscosity coefficient $d_{ij}^{\mathrm{L},n}$ is defined as follows:

$$d_{ij}^{\mathrm{L},n} := \max(\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \widehat{\lambda}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{2.6}$$

where the quantity $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is a positive number larger than or equal to the maximum wave speed $\lambda(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ in a local Riemann problem. Notice that by computing both "directions" (i, j) and (j, i) in the definition (2.6) implies that the graph viscosity $d_{ij}^{\mathrm{L},n}$ is symmetric which is needed for conservation. For completeness, we now recall how the maximum wave speed $\lambda(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is defined in [9]. For $Z \in \{i, j\}$, we set $\mathbf{U}_Z := (\varrho_Z, \mathbf{M}_Z, \mathsf{E}_Z)^\mathsf{T}$, $\mathsf{p}_Z := p(\mathbf{U}_Z^n)$, $\mathsf{e}_Z := e(\mathbf{U}_Z^n)$, $\Gamma_Z := \varrho_Z + \frac{\mathsf{p}_Z(1-b\varrho_Z)}{\mathsf{e}_Z}$, and $\mathcal{E}_Z := \mathsf{E}_Z - \frac{\|\mathbf{M}_Z - (\mathbf{M}_Z\cdot\boldsymbol{n}_{ij})\boldsymbol{n}_{ij}\|_{\ell^2}^2}{2\varrho_Z}$. The non-negativity assumption on the pressure (2.3) and the assumptions on the density ($0 < \varrho_Z < \frac{1}{b}$) implies that

$$\gamma_Z := \frac{\Gamma_Z}{\varrho_Z} \geq 1. \tag{2.7}$$

Then we consider the following Riemann problem:

$$\partial_t \begin{pmatrix} \rho \\ m \\ \mathcal{E} \\ \Gamma \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + p_{\mathrm{cov}} \\ \frac{m}{\rho}(\mathcal{E} + p_{\mathrm{cov}}) \\ \frac{m}{\rho}\Gamma \end{pmatrix} = 0, \quad \text{with} \quad p_{\mathrm{cov}}(\rho, m, \mathcal{E}, \Gamma) := \frac{\frac{\Gamma}{\rho} - 1}{1 - b\rho}\left(\mathcal{E} - \frac{m^2}{2\rho}\right), \tag{2.8}$$

with left data $(\rho_i, \boldsymbol{m}_i\cdot\boldsymbol{n}_{ij}, \mathcal{E}_i, \Gamma_i)^\mathsf{T}$ and right data $(\rho_j, \boldsymbol{m}_j\cdot\boldsymbol{n}_{ij}, \mathcal{E}_j, \Gamma_j)^\mathsf{T}$. It is essential to notice here that $p_{\mathrm{cov}}(\rho_Z, m_Z, \mathcal{E}_Z, \Gamma_Z) = p_Z$; that is, the function $p_{\mathrm{cov}}(\rho, m, \mathcal{E}, \Gamma)$ interpolates the left and right pressures. The problem (2.8) is well-posed because $\gamma_Z \geq 1$. Its complete solution is given in Clayton et al. [9, §4]. The maximum wave speed in this problem is denoted $\lambda(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)$.

Let $\mathcal{A}$ be a nonempty convex subset of $\mathcal{B}(b)$. We say that $\mathcal{A}$ is an invariant set for (2.8) if for every pair of Riemann data in $\mathcal{A}$ the solution of (2.8) takes values in $\mathcal{A}$. We then have:

Theorem 2.1 *([9, Thm. 4.6]). Let $i \in \mathcal{V}$. Let $\mathcal{A} \subset \mathcal{B}(b)$ be a convex invariant set for (2.8). Assume that $\mathbf{U}_j^n \in \mathcal{A}$ for all $j \in \mathcal{J}(i)$. For all $j \in \mathcal{J}(i)$, let $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ be any positive number larger than or equal to $\lambda(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$. Let*

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

Assume that $\tau \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2d_{ij}^{\mathrm{L},n}}{m_i} \leq 1$. Let $\mathbf{U}_i^{\mathrm{L},n+1}$ be defined by (2.4). Then $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{A} \subset \mathcal{B}(b)$.

Remark 2.2 *(Invariant domains).* Theorem 2.1 asserts that every convex invariant set in $\mathcal{B}(b)$ is invariant by the update procedure (2.4) with the graph viscosity $d_{ij}^{\mathrm{L},n}$ defined in (2.6). This means that if $\mathbf{U}_i^n \in \mathcal{A}$ for all $i \in \mathcal{V}$, then $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathcal{A}$ for all $i \in \mathcal{V}$. We say that the method is invariant-domain preserving *for* $\mathcal{A}$. Notice in particular that the method is invariant-domain preserving for the admissible set $\mathcal{B}(b)$. The admissible set $\mathcal{B}(b)$ may not be the smallest invariant domain. For instance, if the oracle admits a mathematical entropy s, then the approximation defined above is also invariant-domain preserving for the set $\mathcal{A}(\boldsymbol{u}_0) := \{\boldsymbol{v} \in \mathcal{B}(b) \mid s(\boldsymbol{v}) \geq \operatorname{ess\,inf}_{\boldsymbol{x}\in D} s(\boldsymbol{u}_0(\boldsymbol{x}))\}$; see e.g., [20, Cor. 4.2]. By slight abuse of terminology and provided the context is unambiguous we will simply call a method invariant-domain preserving without quantifying the precise convex set.

A source code to compute $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is available online [8]. This code has been used in all the numerical simulations reported at the end of the paper. A notable drawback of the graph viscosity (2.6) is that it reduces the space accuracy of the method to first-order. The remainder of the paper is concerned with constructing a higher-order approximation and a respective novel convex limiting technique that is invariant-domain preserving.

Remark 2.3 *(Pressure approximation).* Notice that the oracle is only invoked to compute the left and right values of the pressure in (2.8). The pressure in the Riemann fan is approximated by two covolume equations of state. There is one covolume equation of state for each side of the contact discontinuity since $\gamma = \gamma_L$ on the left of the contact wave and $\gamma = \gamma_R$ on the right.

3. Provisional higher-order method

In this section, we introduce a provisional higher-order method that extends the update (2.4) to second-order when using linear finite elements for the discretization. A convex limiting technique for this provisional update is introduced in §4. The provisional method is based on the work reported in [22,34].

3.1. The method

Higher-order accuracy in space requires using the consistent mass matrix instead of the lumped mass matrix for the discretization of the time derivative. By reducing dispersive errors, the consistent mass matrix is known to yield superconvergence at the grid points; see e.g., Christon et al. [7], [18], [34, Sec. 3.4], Thompson [50]. Let $\mathbb{M}$ be the mass matrix with entries $\{m_{ij}\}_{i,j\in\mathcal{V}}$, then $\mathbb{M}^{-1}$ can be approximated by $\mathbb{I} + \mathbb{B}$ where $b_{ij} = \delta_{ij} - \frac{m_{ij}}{m_j}$, δ_{ij} is the Kronecker symbol, and $m_i := \sum_{j\in\mathcal{J}(i)} m_{ij}$. This approximation has been shown in [7,18] to be superconvergent for piecewise linear continuous finite elements. Let $\mathsf{R} \in \mathbb{R}^I$ with $I := \mathrm{card}(\mathcal{V})$, then using this approximation we have $(\mathbb{M}^{-1}\mathsf{R})_i \approx \mathsf{R}_i + \sum_{j\in\mathcal{J}(i)} \left(b_{ij}\mathsf{R}_j - b_{ji}\mathsf{R}_i\right)$. Notice that $\sum_{j\in\mathcal{J}(i)} b_{ji}\mathsf{R}_i = 0$, because $\sum_{j\in\mathcal{J}(i)} b_{ji} = 0$. The skew-symmetry of the summand in $\sum_{j\in\mathcal{J}(i)} \left(b_{ij}\mathsf{R}_j - b_{ji}\mathsf{R}_i\right)$ is used in §4 for limiting purposes.

Let $\mathbf{U}^{\mathrm{H},n+1}$ denote the high-order update (here the superindex $^{\mathrm{H}}$ reminds us that the update is higher-order accurate). Then, for every $i \in \mathcal{V}$, the provisional high-order update is given by:

$$\frac{m_i}{\tau}\left(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n\right) = \boldsymbol{R}_i^n + \sum_{j\in\mathcal{J}(i)} \left(b_{ij}\boldsymbol{R}_j^n - b_{ji}\boldsymbol{R}_i^n\right), \tag{3.1a}$$

$$\text{with} \qquad \boldsymbol{R}_i^n := \sum_{j\in\mathcal{J}(i)} \left(-\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} + d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)\right). \tag{3.1b}$$

The high-order graph viscosity coefficient $d_{ij}^{\mathrm{H},n}$, defined in Section 3.2, shares the same properties as its low-order counterpart which are necessary for conservation:

$$d_{ij}^{\mathrm{H},n} = d_{ji}^{\mathrm{H},n}, \quad 0 \le d_{ij}^{\mathrm{H},n} \le d_{ij}^{\mathrm{L},n}, \quad \text{for every } j \in \mathcal{J}(i),\ i \in \mathcal{V}. \tag{3.2}$$

3.2. Entropy commutator

Using the entropy-viscosity methodology introduced in Guermond et al. [21], we now discuss the construction of the high-order graph viscosity coefficient $d_{ij}^{\mathrm{H},n}$ that is used in the high-order update (3.1b). A complicating factor in this construction is that the pressure is given by an oracle. We thus do not have a priori knowledge on the equation of state and entropies of the system.

Recall that $(\eta(\boldsymbol{u}), \boldsymbol{F}(\boldsymbol{u}))$ is an entropy pair for the Euler equations if

$$\nabla\cdot\boldsymbol{F}(\boldsymbol{u}) = (\nabla_{\boldsymbol{u}}\eta(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot\mathbb{f}(\boldsymbol{u}), \qquad \forall \boldsymbol{u} \in \mathcal{B}(b), \tag{3.3}$$

where $\mathbb{f}(\boldsymbol{u})$ is the flux of the system (2.1). Since one may not have access to entropies for tabulated equations of state, we are going to use at every $i \in \mathcal{V}$ and every time t^n one entropy pair associated with the following flux

$$\mathbb{f}^{i,n}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{m} \\ \boldsymbol{v}\otimes\boldsymbol{m} + p_{\mathrm{cov}}^{i,n}(\boldsymbol{u})\mathbb{I}_d \\ \boldsymbol{v}(E + p_{\mathrm{cov}}^{i,n}(\boldsymbol{u})) \end{pmatrix}, \qquad p_{\mathrm{cov}}^{i,n}(\boldsymbol{u}) := (\gamma_i^{\min,n} - 1)\frac{\rho e(\boldsymbol{u})}{1-b\rho}. \tag{3.4}$$

Here, $\gamma_i^{\min,n} := \min_{j\in\mathcal{J}(i)}\gamma_j^n$ with $\gamma_j^n := 1 + \frac{\mathrm{p}_j^n(1-b\varrho_j^n)}{\varrho_j^n \mathrm{e}_j^n}$ (see (2.7)). We use the following shifted Harten entropy pair in the numerical tests reported in §5:

$$\eta^{i,n}(\boldsymbol{u}) := \left(\frac{\rho^2 e(\boldsymbol{u})}{(1-b\rho)^{1-\gamma_i^{\min,n}}}\right)^{\frac{1}{\gamma_i^{\min,n}+1}} - \frac{\rho}{\varrho_i^n}\eta_{\mathrm{ref}}^{i,n}, \qquad \boldsymbol{F}^{i,n}(\boldsymbol{u}) := \frac{\boldsymbol{m}}{\rho}\eta^{i,n}(\boldsymbol{u}), \tag{3.5}$$

where $\eta_{\mathrm{ref}}^{i,n} := \left(\frac{(\varrho_i^n)^2 e(\boldsymbol{u}_i^n)}{(1-b\varrho_i^n)^{1-\gamma_i^{\min,n}}}\right)^{\frac{1}{\gamma_i^{\min,n}+1}}$ is the entropy shift; see Remark 3.1. Then, we estimate "entropy production" by inserting the approximate solution $\boldsymbol{u}_h^n$ into a discrete counterpart of (3.3) which we write as follows:

$$\int_D \Big(\nabla\cdot\boldsymbol{F}^{i,n}(\boldsymbol{u}) - (\nabla_{\boldsymbol{u}}\eta^{i,n}(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot\mathbb{f}^{i,n}(\boldsymbol{u})\Big)\varphi_i\,\mathrm{d}x = 0, \tag{3.6}$$

where $\{\varphi_i\}_{i\in\mathcal{V}}$ are the global shape functions associated with the underlying approximation. Notice that the above identity holds true for every smooth function $\boldsymbol{u} \in \mathcal{B}(b)$ and for every $i \in \mathcal{V}$. Substituting $\boldsymbol{u}_h := \sum_{i\in\mathcal{V}}\mathbf{U}_i^n\varphi_i$ into (3.6), we estimate the local entropy residual as follows:

$$N_i^n := \sum_{j\in\mathcal{J}(i)}\Big(\boldsymbol{F}^{i,n}(\mathbf{U}_j^n) - (\nabla_{\boldsymbol{u}}\eta^{i,n}(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}^{i,n}(\mathbf{U}_j^n)\Big)\cdot\boldsymbol{c}_{ij}. \tag{3.7}$$

The residual (3.7) can be thought of as a measure of how well the discrete solution $\boldsymbol{u}_h^n$ satisfies (3.6) in each local stencil $\mathcal{J}(i)$. We then define the normalized entropy residual as follows:

$$R_i^n := \frac{|N_i^n|}{D_i + \epsilon D^{\max}}, \qquad D^{\max} = \max_{i\in\mathcal{V}} D_i, \qquad \epsilon = 10^{-2}, \tag{3.8a}$$

$$D_i^n := \Big|\sum_{j\in\mathcal{J}(i)}\boldsymbol{F}^{i,n}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\Big| + \Big|\sum_{j\in\mathcal{J}(i)}(\nabla_{\boldsymbol{u}}\eta^{i,n}(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}^{i,n}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\Big|. \tag{3.8b}$$

By definition, the normalized residual R_i^n has values in $[0,1]$ and, for piecewise linear finite elements, behaves like $\mathcal{O}(h)$ where h is the typical meshsize. Finally we set

$$d_{ij}^{\mathrm{H},n} := d_{ij}^{\mathrm{L},n}\max(\psi(R_i^n), \psi(R_j^n)), \tag{3.9}$$

where the activation function ψ is defined as follows:

$$\psi(x) := \frac{4x_0^3 - (x+x_0)(x-2x_0)\big\{(x-2x_0) - \mathrm{ReLU}(x-2x_0)\big\}}{4x_0^3}, \tag{3.10}$$

and $\mathrm{ReLU}(x) = (x+|x|)/2$ is the rectified linear activation function. Notice that $\psi(0) = 0$, $\psi(x_0) = \frac{1}{2}$ and $\psi(x) = 1$ for all $x \in [2x_0, 1]$ for a chosen $x_0 \in [0, 0.5]$. As a result, one recovers $d_{ij}^{\mathrm{H},n} = d_{ij}^{\mathrm{L},n}$ if the entropy residual is larger than $2x_0$. We

have $\psi(x) \sim \frac{3}{4}\frac{x^2}{x_0^2} + \mathcal{O}(x^3)$. The numerical tests reported in §5 are done with $x_0 = 0.4$. An activation function with the same purpose is used in Persson and Peraire [37, Eq. (8)]. Up to a translation, the activation function therein behaves like $\frac{1}{2} + \frac{1}{2}\sin(\pi\frac{x-x_0}{2x_0})$ in the interval $[0, 2x_0]$.

Remark 3.1 *(Entropy shift).* The entropy shift considered in (3.5) is motivated by the observation that the numerator N_i^n is unchanged by the change of variable $\eta(\boldsymbol{u}) \mapsto \eta(\boldsymbol{u}) - \lambda\rho$ for all $\lambda \in \mathbb{R}$ and for every entropy η associated with the flux (3.4). The constant λ is chosen in (3.5) so that $\eta^{i,n}(\boldsymbol{u}_i^n) = 0$. This entropy shift was first is introduced in [22, §3.4] to minimize the effects of the mass conservation approximation errors on the entropy commutator.

4. Convex limiting

The high-order update (3.1) is not guaranteed to be invariant-domain preserving; in particular, it is not guaranteed to stay in the admissible set $\mathcal{B}(b)$. In order to correct this defect we now discuss a new convex limiting technique that re-establishes invariant-domain preservation for the final high-order update, i.e., $\mathbf{U}_i^{n+1} \in \mathcal{B}(b)$.

4.1. Key observation

A classical key observation is that one can rewrite (2.4) as follows:

$$\mathbf{U}_i^{\mathrm{L},n+1} = \Big(1 - \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathbf{U}}_{ij}^n, \tag{4.1a}$$

$$\text{with} \qquad \overline{\mathbf{U}}_{ij}^n := \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - \frac{1}{2d_{ij}^{\mathrm{L},n}}\Big(\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n)\Big)\boldsymbol{c}_{ij}. \tag{4.1b}$$

That is, the low-order update $\mathbf{U}_i^{\mathrm{L},n+1}$ is a convex combination (under the appropriate CFL condition) of the local state $\mathbf{U}_i^n$ and the auxiliary states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{V}}$, i.e.,

$$\mathbf{U}_i^{\mathrm{L},n+1} \in \mathrm{Conv}\{\overline{\mathbf{U}}_{ij}^n \mid j \in \mathcal{J}(i)\}. \tag{4.2}$$

The main result established in [9, Thm. 4.6] (and summarized in Theorem 2.1) is that under the CFL condition stated in Theorem 2.1 and the definition (2.6) for $d_{ij}^{\mathrm{L},n}$, the states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{J}(i)}$ are in $\mathcal{B}(b)$ provided this is already the case of the states $\{\mathbf{U}_j^n\}_{j\in\mathcal{J}(i)}$. This is done by proving that the states $\overline{\mathbf{U}}_{ij}^n$ are space averages of the solution to a Riemann problem. We refer the reader to [20, Sec. 3.3], [22, Sec. 3.2] and [23, Sec. 3.2] where this is discussed in detail. We are going to use the states $\overline{\mathbf{U}}_{ij}^n$ to define local bounds in space and time to perform the limiting of the high-order states $\{\mathbf{U}_i^{\mathrm{H},n+1}\}_{i\in\mathcal{V}}$.

Remark 4.1 *(Literature).* The importance of the auxiliary states $\overline{\mathbf{U}}_{ij}^n$ in the construction of the low-order approximation has been recognized in Nessyahu and Tadmor [36, Eq. (2.6)]. That these states are averages of Riemann solution provided $d_{ij}^{\mathrm{L},n}$ is large enough is well documented in Harten et al. [27, §3.A] (see also Thm. 3.1 in [27], p. 375, line 12 in Tadmor [47] and Perthame and Shu [38, §5].)

4.2. Entropy surrogate

Our goal is to use the methodology introduced in [21] to perform the convex limiting of the update $\mathbf{U}_i^{\mathrm{H},n+1}$. In this context the use of an oracle with little a priori knowledge on the equation of state poses a significant challenge as it makes it impossible to properly define an entropy. We resolve the impasse by introducing an artificial *surrogate* entropy that has the right mathematical properties for the convex limiting methodology to be applied; see Theorem 4.3.

For any admissible state $\boldsymbol{u} \in \mathcal{B}(b)$ and $\gamma \geq 1$, we define

$$S(\boldsymbol{u}, \gamma) := \frac{(\rho e)(\boldsymbol{u})}{\rho^\gamma}(1 - b\rho)^{\gamma-1}, \tag{4.3}$$

where $\rho(\boldsymbol{u})$ and $e(\boldsymbol{u})$ are the density and specific internal energy of the state $\boldsymbol{u}$. We note in passing that for each γ, the quantity defined in (4.3) is convex. Furthermore, for every index $i \in \mathcal{V}$, we set

$$\gamma_i^n := 1 + \frac{\mathsf{p}_i^n(1 - b\varrho_i^n)}{\varrho_i^n \mathsf{e}_i^n}, \qquad \Gamma_i^n := \varrho_i^n\,\gamma_i^n, \tag{4.4}$$

where $\varrho_i^n := \rho(\mathbf{U}_i^n)$, $\mathsf{p}_i^n := p(\mathbf{U}_i^n)$ and $\mathsf{e}_i^n := e(\mathbf{U}_i^n) = \frac{1}{\varrho_i^n}\big(\mathsf{E}_i^n - \frac{\|\mathbf{M}_i^n\|_{\ell^2}^2}{2\varrho_i^n}\big)$. The following result is the key motivation for the definition of an entropy surrogate.

Lemma 4.2. *For all $i \in \mathcal{V}$, assume that $\mathbf{U}_i^n \in \mathcal{B}(b)$. For all $i \in \mathcal{V}$, all $j \in \mathcal{I}(i)$, all $\gamma_{ij} \in [1, \min(\gamma_i^n, \gamma_j^n)]$, all left data $(\varrho_i^n, \mathbf{M}_i^n \cdot \boldsymbol{n}_{ij}, \mathcal{E}_i^n, \Gamma_i^n)^\mathsf{T}$ and right data $(\varrho_j^n, \mathbf{M}_j^n \cdot \boldsymbol{n}_{ij}, \mathcal{E}_j^n, \Gamma_j^n)^\mathsf{T}$ in the extended Riemann problem (2.8), and all $\boldsymbol{u} \in \mathcal{B}(b)$, we set*

$$\Psi_{ij}(\boldsymbol{u}) := \rho e(\boldsymbol{u}) - S_{ij}^{\min} \rho^{\gamma_{ij}} (1 - b\rho)^{1-\gamma_{ij}}, \quad \text{where } S_{ij}^{\min} := \min\big(S(\mathbf{U}_i^n, \gamma_{ij}), S(\mathbf{U}_j^n, \gamma_{ij})\big). \tag{4.5}$$

Then, $\Psi_{ij}(\boldsymbol{u})$ increases across shocks in the solution of the extended Riemann problem (2.8) (if a shock wave exists).

Proof. We omit the superscript ^{n} in the proof to simplify the notation. The solution to the extended Riemann problem (2.8), $(\rho, m, \mathcal{E}, \Gamma)^\mathsf{T}(x,t)$, is given in Clayton et al. [9]. In particular, we have that $\gamma(x,t) = \gamma_i$ if $x/t < v^*$ and $\gamma(x,t) = \gamma_j$ if $x/t > v^*$. Here, $\gamma(x,t) = \Gamma(x,t)/\rho(x,t)$, and v^* is the speed of the contact wave. Let $Z \in \{i, j\}$, with the convention that the index i is for the left state and j is for the right state. Assume that the Z-wave is a shock wave. Let $\boldsymbol{u}_Z \in \mathcal{B}(b)$ be the state before the shock and let $\boldsymbol{u} \in \mathcal{B}(b)$ be an arbitrary state connected to $\boldsymbol{u}_Z$ through a shock curve. With the notation $\tau := \frac{1}{\rho}$ for the specific volume (not to be confused with the time step), and since $\gamma = \gamma_Z$ along the wave curve, the Rankine-Hugoniot condition implies that (see Godlewski and Raviart [16, Eq. (4.8), p. 144]), $e(\boldsymbol{u}) - \mathsf{e}_Z + \frac{1}{2}\big(p_{\mathrm{cov}}(\tau, e(\boldsymbol{u})) + p_{\mathrm{cov}}(\tau_Z, \mathsf{e}_Z)\big)(\tau - \tau_Z) = 0$, where by slight abuse of notation we renamed the pressure in (2.8) by setting $p_{\mathrm{cov}}(\tau, e) := \frac{\gamma_Z - 1}{\tau - b} e$. We then infer that $e(\boldsymbol{u})$ only depends on ρ along the shock curve; more precisely, we have

$$e(\boldsymbol{u}) = \mathsf{e}_Z \frac{1 - \frac{(\gamma_Z - 1)(\tau - \tau_Z)}{2(\tau_Z - b)}}{1 + \frac{(\gamma_Z - 1)(\tau - \tau_Z)}{2(\tau - b)}} =: r(\tau).$$

Notice that this function is well defined only on the interval (τ_Z^∞, ∞) with

$$\tau_Z^\infty := \frac{(\gamma_Z - 1)\tau_Z + 2b}{\gamma_Z + 1}, \tag{4.6}$$

and it is nonnegative on the interval $(\tau_Z^\infty, \tau_Z^0)$ with $\tau_Z^0 := \frac{(\gamma_Z + 1)\tau_Z - 2b}{\gamma_Z - 1}$. Notice that $b < \tau_Z^\infty < \tau_Z < \tau_Z^0$ since we assumed that $\mathbf{U}_Z \in \mathcal{B}(b)$. We now show that the function $\mathcal{B}(b) \ni \boldsymbol{u} \mapsto \rho e(\boldsymbol{u}) - c\rho^\gamma (1 - b\rho)^{1-\gamma}$ is nonnegative and increasing on the shock curve for all $\gamma \in (1, \gamma_Z]$ and $c \in (0, \mathsf{e}_Z(\tau_Z - b)^{\gamma - 1}]$. This will then prove the assertion for the choices $\gamma := \gamma_{ij}$ and $c := S_{ij}^{\min}$, because $\gamma_{ij} \in (1, \gamma_Z]$ and

$$0 < S_{ij}^{\min} \le \varrho_Z^{1-\gamma_{ij}} \mathsf{e}_Z (1 - b\varrho_Z)^{\gamma_{ij} - 1} = \mathsf{e}_Z (\tau_Z - b)^{\gamma - 1},$$

i.e., $S_{ij}^{\min} \in (0, \mathsf{e}_Z(\tau_Z - b)^{\gamma - 1}]$. Setting $q(\tau) := \frac{\gamma_Z - 1}{\tau - b} r(\tau)$, we have $q'(\tau) = \frac{-4\mathsf{e}_Z \gamma_Z (\gamma_Z - 1)}{(\tau + \tau_Z - 2b + \gamma_Z(\tau - \tau_Z))^2} < 0$; hence, the pressure, q, is a monotone increasing function of ρ along shock curves. As the pressure increases on shock curves when starting from the state $\mathbf{U}_Z$, we conclude that ρ also increases along shock curves, i.e., $\rho \in [\varrho_Z, \frac{1}{b})$ or $\tau \in (b, \tau_Z]$. Actually, the pressure is finite only in the range $\tau \in (\tau_Z^\infty, \tau_Z]$; hence, showing that $\mathcal{B}(b) \ni \boldsymbol{u} \mapsto \rho e(\boldsymbol{u}) - c\rho^\gamma (1 - b\rho)^{1-\gamma}$ is nonnegative increasing on shock curves, is equivalent to showing that

$$(\tau_Z^\infty, \tau_Z] \ni \tau \mapsto g(\tau) := \tau^{-1} r(\tau) - c\tau^{-1} (\tau - b)^{1-\gamma}$$

is nonnegative decreasing on shock curves. Recall that the specific entropy for the covolume equation of state is given by $s(\tau, \mathsf{e}) = \log(\mathsf{e}^{\frac{1}{\gamma_Z - 1}}(\tau - b))$. A fundamental property of the specific entropy is that it is an increasing function along shocks. That is, $s(\tau, r(\tau))$ is a decreasing function over the interval $\tau \in (\tau_Z^\infty, \tau_Z]$. This also means that $(\tau_Z^\infty, \tau_Z] \ni \tau \mapsto \varsigma(\tau) := \exp\big((\gamma_Z - 1) s(\tau, r(\tau))\big) = r(\tau)(\tau - b)^{\gamma_Z - 1}$ is a decreasing function. We now have

$$g(\tau) = \tau^{-1} r(\tau) - c\tau^{-1} (\tau - b)^{1-\gamma} = \varsigma(\tau) \tau^{-1} (\tau - b)^{1-\gamma_Z} - c\tau^{-1} (\tau - b)^{1-\gamma}.$$

Let $\widetilde{\varsigma}$ be the defined by $\widetilde{\varsigma}(\tau) := \varsigma(\tau)(\tau - b)^{\gamma - \gamma_Z}$. Then we have

$$g(\tau) = \frac{(\tau - b)^{1-\gamma}}{\tau} (\widetilde{\varsigma}(\tau) - c).$$

Computing the derivative of g, we see,

$$g'(\tau) = \frac{(\tau - b)^{1-\gamma}}{\tau} \widetilde{\varsigma}'(\tau) - \frac{(\gamma - 1)\tau + (\tau - b)}{\tau^2 (\tau - b)^\gamma} (\widetilde{\varsigma}(\tau) - c).$$

Note that

$$\widetilde{\varsigma}'(\tau) = (\tau - b)^{\gamma - \gamma_Z} \varsigma'(\tau) + (\gamma - \gamma_Z)(\tau - b)^{\gamma - \gamma_Z - 1} \varsigma(\tau).$$

Then $\widetilde{\varsigma}'(\tau) \le 0$ as $\varsigma'(\tau) \le 0$ for $\tau \in (\tau_Z^\infty, \tau_Z]$, $\gamma \le \gamma_Z$, and $0 \le \varsigma(\tau)$, and $\inf_{\tau \in (b, \tau_Z]} \widetilde{\varsigma}(\tau) = \widetilde{\varsigma}(\tau_Z) = \mathsf{e}_Z (\tau_Z - b)^{\gamma - 1} \ge c$ as $\widetilde{\varsigma}$ is a decreasing function. Thus, by the choice of the constants c and γ, we have that $g(\tau) \ge 0$ and $g'(\tau) \le 0$. Hence, $[\varrho_Z, \frac{1}{\tau_Z^\infty}) \ni \rho \mapsto \rho e(\boldsymbol{u}) - c\rho^\gamma (1 - b\rho)^{1-\gamma}$ is nonnegative increasing on shock curves. This completes the proof. □

We now define the surrogate entropy. For all $i \in \mathcal{V}$, we set

$$\gamma_i^{\min,n} := \min_{j\in\mathcal{J}(i)} \gamma_j^n, \qquad S_i^{\min,n} := \min(\min_{j\in\mathcal{J}(i)} S(\mathbf{U}_j^n; \gamma_i^{\min,n}), \min_{j\in\mathcal{J}(i)} S(\overline{\mathbf{U}}_{ij}^n; \gamma_i^{\min,n})), \tag{4.7a}$$

$$\Psi_i^s(\boldsymbol{u}) := \rho e(\boldsymbol{u}) - S_i^{\min,n} \rho^{\gamma_i^{\min,n}} (1-b\rho)^{1-\gamma_i^{\min,n}}. \tag{4.7b}$$

The following result summarizes the content of this section.

Theorem 4.3 *(Surrogate entropy). The following holds true with the same assumptions and definitions as in Lemma 4.2:*

(i) *The function $\Psi_i^s : \mathcal{B}(b) \to \mathbb{R}$ is concave.*

(ii) *Let $\mathbf{U}_i^{\mathrm{L},n+1}$ be the update defined in (2.4). We have $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}) \geq 0$ under the CFL condition $\tau \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2d_{ij}^{\mathrm{L},n}}{m_i} \leq 1$.*

(iii) *Let $j \in \mathcal{J}(i)$. Consider the extended Riemann problem (2.8) with left data $(\varrho_i^n, \mathbf{M}_i^n\cdot\boldsymbol{n}_{ij}, \mathcal{E}_i^n, \Gamma_i^n)^\intercal$ and right data $(\varrho_j^n, \mathbf{M}_j^n\cdot\boldsymbol{n}_{ij}, \mathcal{E}_j^n, \Gamma_j^n)^\intercal$. If the solution to (2.8) has shock waves, then the function $\Psi_i^s(\boldsymbol{u})$ increases across the shocks.*

Proof. (i) Recall that $\boldsymbol{u} \mapsto \rho e(\boldsymbol{u})$ is concave. Moreover, the function $\phi(x) = x(\frac{1}{x}-b)^{\gamma_i^{\min,n}}$ is convex because $\gamma_i^{\min,n} \geq 1$. As a result, $-S_i^{\min,n}\rho^{\gamma_i^{\min,n}}(1-b\rho)^{1-\gamma_i^{\min,n}} = -S_i^{\min,n}\phi(\rho)$ is concave because $S_i^{\min,n} \geq 0$. This proves that $\Psi_i^s : \mathcal{B}(b) \to \mathbb{R}$ is concave.

(ii) Recalling from Theorem 2.1 that under the CFL condition $\tau \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2d_{ij}^{\mathrm{L},n}}{m_i} \leq 1$, the state $\mathbf{U}_i^{\mathrm{L},n+1}$ is in the convex hull of the states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{J}(i)}$, the concavity of the functional Ψ_i^s implies that $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}) \geq \min_{j\in\mathcal{J}(i)} \Psi_i^s(\overline{\mathbf{U}}_{ij}^n)$. But for all $j \in \mathcal{J}(i)$ we have

$$\begin{aligned}\Psi_i^s(\overline{\mathbf{U}}_{ij}^n) &= (\rho e)(\overline{\mathbf{U}}_{ij}^n) - S_i^{\min,n}(\overline{\varrho}_{ij}^n)^{\gamma_i^{\min,n}}(1-b\overline{\varrho}_{ij}^n)^{1-\gamma_i^{\min,n}} \\ &= (S(\overline{\mathbf{U}}_{ij}^n, \gamma_i^{\min,n}) - S_i^{\min,n})(\overline{\varrho}_{ij}^n)^{\gamma_i^{\min,n}}(1-b\overline{\varrho}_{ij}^n)^{1-\gamma_i^{\min,n}}.\end{aligned}$$

And we conclude that $\Psi_i^s(\overline{\mathbf{U}}_{ij}^n) \geq 0$ by definition of $S_i^{\min,n}$. The assertion readily follows.
(iii) The assertion is a consequence of Lemma 4.2. □

4.3. Limiting on the density

Limiting on the density is performed exactly as done in [22, §4.4]. First, we define the local bounds

$$\varrho_i^{\min,n} := \min_{j\in\mathcal{J}(i)} \overline{\varrho}_{ij}^n \quad \text{and} \quad \varrho_i^{\max,n} := \max_{j\in\mathcal{J}(i)} \overline{\varrho}_{ij}^n, \tag{4.8}$$

where $\overline{\varrho}_{ij}^n$ is the density of the auxiliary state $\overline{\mathbf{U}}_{ij}^n$.

Second, we relax these bounds to ensure that second-order accuracy is maintained in the maximum norm. The relaxation is done as in [22, §4.7] with a modification of $\varrho_i^{\max,n}$ to accommodate the covolume constraint $1-b\rho > 0$. This modification is justified in the following lemma.

Lemma 4.4 *(Maximum density bound). The following holds true.*

(i) *The density in the Riemann problem (2.8) satisfies the following upper bound*

$$\rho \leq \max_{Z\in\{i,j\}} \frac{1}{\tau_Z^\infty} = \max_{Z\in\{i,j\}} \frac{(\gamma_Z+1)\rho_Z}{(\gamma_Z-1)+2b\rho_Z}. \tag{4.9}$$

(ii) *Under the CFL condition stated in Theorem 2.1, the low-order update satisfies the following:*

$$\varrho_i^{\mathrm{L},n+1} \leq \frac{(\gamma_i^{\min,n}+1)\varrho_i^{\max,n}}{(\gamma_i^{\min,n}-1)+2b\varrho^{\max,n}}. \tag{4.10}$$

Proof. (i) Let $Z \in \{i,j\}$, with the convention that the index i is for the left state and j is for the right state. Recall that the pressure in the Riemann problem (2.8) is defined by the function $p_{\mathrm{cov}}(\mathbf{U})$ (with a slight abuse of notation) and we also have $\mathrm{p}_Z = p_{\mathrm{cov}}(\mathbf{U}_Z)$. If the elementary Z-wave is an expansion then ρ decreases along the expansion wave and we have $\rho \leq \varrho_Z$. If instead the elementary Z-wave is a shock, we have established in the proof of Lemma 4.2 that $\tau \in (\tau_Z^\infty, \tau_Z]$, i.e., $\rho \in [\rho_Z, \frac{1}{\tau_Z^\infty})$; see (4.6). Whence the assertion.
(ii) Using (4.1a), we observe that

$$\varrho_i^{\mathrm{L},n+1} \le \max_{j\in\mathcal{J}(i)} \overline{\varrho}_{ij}^n \le \max_{j\in\mathcal{J}(i)} \frac{\gamma_j+1}{(\gamma_j-1)\tau_j+2b} = \max_{j\in\mathcal{J}(i)} \frac{(\gamma_j+1)\varrho_j}{(\gamma_j-1)+2b\varrho_j}.$$

As the function $\frac{(\gamma+1)\rho}{(\gamma-1)+2b\rho}$ is monotone increasing with respect to ρ and monotone decreasing with respect to γ, the inequality (4.10) follows readily. □

Proceeding as in [22, §4.7], we estimate the local curvature of the density by

$$\Delta^2\varrho_i^n := \frac{\sum_{j\in\mathcal{J}(i)\backslash\{i\}} \beta_{ij}(\varrho_i^n-\varrho_j^n)}{\sum_{j\in\mathcal{J}(i)\backslash\{i\}}\beta_{ij}}, \qquad \overline{\Delta^2\varrho_i^n} := \frac{1}{2(\mathrm{card}(\mathcal{J}(i))-1)} \sum_{j\in\mathcal{J}(i)\backslash\{i\}} (\tfrac{1}{2}\Delta^2\varrho_i^n + \tfrac{1}{2}\Delta^2\varrho_j^n), \tag{4.11}$$

where $\beta_{ij} = \int_D \nabla\varphi_j\cdot\nabla\varphi_i \,\mathrm{d}x$ are the stiffness coefficients of the Laplace operator and we recall that $\{\varphi_i\}_{i\in\mathcal{V}}$ are the global shape functions. We note in passing that $\sum_{j\in\mathcal{J}(i)\backslash\{i\}}\beta_{ij} = -\beta_{ii} = -\int_D(\nabla\varphi_i)^2\,\mathrm{d}\boldsymbol{x} \neq 0$. The relaxed local density bounds are then defined as follows:

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

$$\overline{\varrho_i^{\max,n}} := \min(\varrho_i^{\max,n} + \overline{\Delta^2\varrho_i^n},\, (1+r_h)\varrho_i^{\max,n},\, \tfrac{(1+\gamma_i^{\min,n})\varrho_i^{\max,n}}{\gamma_i^{\min,n}-1+2b\varrho_i^{\max,n}}), \tag{4.13}$$

where $r_h := (\frac{m_i}{|D|})^{\frac{1.5}{d}}$. The argumentation for the presence of the terms involving $1-r_h$ and $1+r_h$ is given in Remark 4.15 in [22, §4.7].

The actual limiting on the density guaranteeing that $\overline{\varrho_i^{\min,n}} \le \varrho_i^{n+1} \le \overline{\varrho_i^{\max,n}}$ is done as explained in [22, §4.4].

4.4. Limiting on the surrogate entropy

After limiting the density, we limit the surrogate entropy. Recall that we established in Theorem 4.3 that $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}) \ge 0$. We want the limiting operation to guarantee that $\Psi_i^s(\mathbf{U}_i^{n+1}) \ge 0$ as well. This limiting in turn implies a positive lower bound on the internal energy. In this section we use the notation $\boldsymbol{v} := (\rho(\boldsymbol{v}), \boldsymbol{m}(\boldsymbol{v}), E(\boldsymbol{v}))$ for all $\boldsymbol{v}\in\mathbb{R}^{d+2}$; that is, $\rho(\boldsymbol{v})$ is the first component of $\boldsymbol{v}$, $\boldsymbol{m}(\boldsymbol{v})\in\mathbb{R}^d$ is composed of the components 2 to $d+1$, and $E(\boldsymbol{v})$ is the last component of $\boldsymbol{v}$.

The limiting on Ψ_i^s is done with the convex limiting explained in [22, Sec. 4.2]. Given some state $\mathbf{P}_{ij}\in\mathbb{R}^{d+2}$, $j\in\mathcal{J}(i)\backslash\{i\}$, and $\ell_0\in[0,1]$ so that $0<\rho(\mathbf{U}_i^{\mathrm{L},n+1}+\ell_0\mathbf{P}_{ij})<b$, one has to find the largest ℓ in $[0,\ell_0]$ so that $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})\ge 0$. Setting $\mathcal{A}(b) := \{\boldsymbol{v}\in\mathbb{R}^{d+2} \mid 0<\rho(\boldsymbol{v})<b^{-1}\}$ and observing that $\mathcal{A}(b)$ is convex, we notice that $\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij}\in\mathcal{A}(b)$ for all $\ell\in[0,\ell_0]$. This implies that $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})$ is well-defined for all $\ell\in[0,\ell_0]$. One sets $\ell=\ell_0$ if $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell_0\mathbf{P}_{ij})\ge 0$; otherwise, one solves the equation $g(\ell)=0$ where $g(\ell) := \Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})$. This equation has at least one solution because Ψ_i^s is continuous. The solution set is connected because Ψ_i^s is concave, and the solution set is a singleton if Ψ_i^s is strictly concave.

Lemma 4.5 *(Internal energy). Let $\mathbf{U}^{n+1}$ be the final stage obtained after limiting on the density and the surrogate entropy. Then the specific internal energy of this state satisfies the following lower bound for all $i\in\mathcal{V}$:*

$$\mathsf{e}_i^{n+1} \ge S_i^{\min,n}\big(\tfrac{1}{\varrho_i^n}-b\big)^{1-\gamma_i^{\min,n}} \ge 0. \tag{4.14}$$

Remark 4.6 *(Entropy for the Covolume EOS).* If the oracle coincides with the covolume equation of state, then limiting the entropy surrogate is equivalent to limiting the physical entropy.

We now propose two ways to find ℓ^* so that $g(\ell^*)=0$ assuming that this root is unique. Both methods are iterative and are guaranteed to return an answer $\tilde{\ell}$ that is such that $\tilde{\ell}\uparrow\ell^*$ (hence $g(\tilde{\ell})\ge 0$ for every termination criteria). The first method consists of proceeding as described in full details in Section 7.5.4 in [23] (see also the end of Section 4.6 in [22]). The algorithm is based on the Newton-Secant method. One uses the secant method on the left of ℓ^* and the Newton method on the right. The convergence rate is between 1 and 2. The second method is based on the quadratic Newton method and its convergence rate is cubic. Instead of solving $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})=0$ one defines $\Phi_i^s(\boldsymbol{v}) := \rho(\boldsymbol{v})\Psi_i^s(\boldsymbol{v})$ and solves $\Phi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})=0$. The solution sets of the two equations $\Psi_i^s(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})=0$ and $\Phi_i^s(\boldsymbol{v}) := \rho(\boldsymbol{v})\Psi_i^s(\boldsymbol{v})$ are identical since we assumed that $0<\rho(\mathbf{U}_i^{\mathrm{L},n+1}+\ell\mathbf{P}_{ij})$ for all $\ell\in[0,\ell_0]$. A key to the method is the following result.

Lemma 4.7. *Let $\boldsymbol{u}\in\mathcal{A}(b) := \{\boldsymbol{v}\in\mathbb{R}^{d+2} \mid 0<\rho(\boldsymbol{v})<b^{-1}\}$ and $\boldsymbol{p}\in\mathbb{R}^{d+2}$. Assume that $\boldsymbol{u}+\ell_0\boldsymbol{p}\in\mathcal{A}(b)$. Let $f:[0,\ell_0]\ni\ell\mapsto f(\ell) := \Phi_i^s(\boldsymbol{u}+\ell\boldsymbol{p})$. The sign of $f'''(l)$ is constant over $[0,\ell_0]$.*

Proof. By definition we have $\Phi_i^s(\boldsymbol{v}) = \rho(\boldsymbol{v})E(\boldsymbol{v}) - \frac{1}{2}\|\boldsymbol{m}(\boldsymbol{v})\|_\ell^2 - c\rho(\boldsymbol{v})^{\gamma+1}(1-b\rho(\boldsymbol{v}))^{1-\gamma}$, with $c := S_i^{\min,n}$ and $\gamma := \gamma_i^{\min,n}$. Notice that $\boldsymbol{u}+\ell\boldsymbol{p} \in \mathcal{A}(b)$ for all $\ell \in [0,\ell_0]$ since $\mathcal{A}(b)$ is convex; as a result, $f(\ell)$ is well-defined for all $\ell \in [0,\ell_0]$. A direct computation shows that

$$f'''(\ell) = (\rho(\boldsymbol{p}))^3 \partial_\rho^3 \Phi_i^s(\boldsymbol{u}+\ell\boldsymbol{p}).$$

But

$$\partial_\rho^3 \Phi_i^s(\boldsymbol{v}) = -c\gamma(\gamma^2-1)\rho(\boldsymbol{v})^{\gamma-2}(1-b\rho(\boldsymbol{v}))^{-\gamma-2}.$$

Hence $f'''(\ell)$ has the same sign as $-\rho(\boldsymbol{p})$ for all $\ell \in [0,\ell_0]$. □

We now show how the quadratic Newton algorithm can be implemented to estimate ℓ^* from below. We initialize the iterative process by setting $\ell_L = 0$ and $\ell_R = \ell_0$. Then, let ℓ_L and ℓ_R be the current left and right estimates of ℓ^* with $0 \le \ell_L < \ell^* < \ell_R \le \ell_0$. We now construct a quadratic polynomial $P_L(\ell)$ such that $P_L(\ell_L) = f(\ell_L)$, $P_L(\ell_R) = f(\ell_R)$ and $P_L'(\ell_L) = f'(\ell_L)$. We similarly define the quadratic polynomial $P_R(\ell)$ such that $P_R(\ell_L) = f(\ell_L)$, $P_R(\ell_R) = f(\ell_R)$ and $P_R'(\ell_R) = f'(\ell_R)$. Using the divided difference notation we have:

$$P_L(\ell) := f(\ell_L) + f'(\ell_L)(\ell-\ell_L) + f[\ell_L,\ell_L,\ell_R](\ell-\ell_L)^2, \tag{4.15a}$$

$$P_R(\ell) := f(\ell_R) + f'(\ell_R)(\ell-\ell_R) + f[\ell_L,\ell_R,\ell_R](\ell-\ell_R)^2. \tag{4.15b}$$

Lemma 4.8. *The following holds true:*

(i) *The polynomials $P_L(\ell)$ and $P_R(\ell)$ bound the function $f(\ell)$ in the following sense:*

$$\min\big(P_L(\ell), P_R(\ell)\big) < f(\ell) < \max\big(P_L(\ell), P_R(\ell)\big), \qquad \forall \ell \in (\ell_L,\ell_R). \tag{4.16}$$

(ii) *$P_L(\ell)$ and $P_R(\ell)$ have each a unique zero over the interval (ℓ_L,ℓ_R) respectively given by*

$$\ell^L(\ell_L,\ell_R) := \ell_L - \frac{2f(\ell_L)}{f'(\ell_L) + \sqrt{f'(\ell_L)^2 - 4f(\ell_L)f[\ell_L,\ell_L,\ell_R]}}, \tag{4.17a}$$

$$\ell^R(\ell_L,\ell_R) := \ell_R - \frac{2f(\ell_R)}{f'(\ell_R) + \sqrt{f'(\ell_R)^2 - 4f(\ell_R)f[\ell_L,\ell_R,\ell_R]}}. \tag{4.17b}$$

(iii) *Properties (i) and (ii) imply that*

$$\min\big(\ell^L(\ell_L,\ell_R), \ell^R(\ell_L,\ell_R)\big) < \ell^* < \max\big(\ell^L(\ell_L,\ell_R), \ell^R(\ell_L,\ell_R)\big). \tag{4.18}$$

Proof. (i) The argumentation is largely the same as in the proof Lemma 4.5 in [19] and relies on the sign of $f'''(\ell)$ being constant over (ℓ_L,ℓ_R), as established in Lemma 4.7, which implies that one of the quadratic polynomials ($P_L(\ell)$ or $P_R(\ell)$) is above $f(\ell)$ and the other one is below $f(\ell)$ for $\ell \in (\ell_L,\ell_R)$. (ii) Moreover, both polynomials have exactly one zero in the interval (ℓ_L,ℓ_R) given by (4.17) even in the degenerated case when one or both of them are linear functions. □

Lemma 4.8 implies that $\min\big(\ell^L(\ell_L,\ell_R), \ell^R(\ell_L,\ell_R)\big)$ is always a lower bound on the root ℓ^*. The quadratic Newton algorithm consists of replacing ℓ_L by $\min\big(\ell^L(\ell_L,\ell_R), \ell^R(\ell_L,\ell_R)\big)$ and ℓ_R by $\max\big(\ell^L(\ell_L,\ell_R), \ell^R(\ell_L,\ell_R)\big)$ and looping until some threshold criterion is reached. The convergence rate of this algorithm is cubic.

Remark 4.9 *(Relaxation on the surrogate entropy bound).* To ensure that second-order accuracy is obtained in the maximum norm for smooth solutions, one must relax the surrogate entropy minimum bound $S_i^{\min,n}$ defined in (4.7) (see: Khobalatte and Perthame [29, Sec. 3.3]). The relaxation process we use is described in [22, Sec. 4.7.2].

Remark 4.10 *(Higher-order accuracy).* We have observed empirically that at each time step the convex limiting procedure is only active in a small set of degrees of freedom when the solution is smooth. After limiting, the solution retains the order of accuracy in space of the provisional high-order update (3.1) (in the L^q-norm $q \in [1,\infty]$). It is essential to relax the density and the surrogate entropy bounds as described above to maintain optimality on the convergence rates.

5. Numerical illustrations

In this section we illustrate the proposed method with several benchmarks and experiments.

5.1. Preliminaries

Two independent codes have been written to ascertain reproducibility. The first one, henceforth referred to as the `TAMU` code, is serial, does not use any particular software and is written in Fortran 95/2003. It is based on Lagrange elements on simplices and is dimension-independent. The second code is based on continuous $\mathbb{Q}_1$ finite elements on quadrangular meshes and use the `Ryujin` [34,24] software, a high-performance finite-element solver based on the `deal.II` library. The time-stepping in both codes is done with the three stage, third-order strong stability preserving Runge-Kutta method, SSPRK(3,3). The time step is defined by the expression $\tau = \frac{1}{2}\text{CFL}\times \min_{i\in\mathcal{I}} \frac{m_i}{\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\text{L},n}}$ with $d_{ij}^{\text{L},n}$ defined in (2.6) and $\text{CFL}\in (0,1)$ is fixed by the user. We refer the reader to [25, Sec. 4] for a discussion on the implementation of boundary conditions for the Euler Equations. In this work, we have modified the non-reflecting boundary conditions described in [25, Sec. 4.3.2] to account for an arbitrary of state; for brevity, we skip the discussion of the details of said boundary conditions. All the computations involving dimensional quantities are done in the `SI` unit system unless otherwise specified. We have verified on all the benchmark tests reported below that the two codes give the same results. For brevity, in each test case, we only show the results from one code. The outputs for the one-dimensional tests come from the `TAMU` code. The outputs for the two-dimensional tests come from `Ryujin`. The tests accessing the `SESAME` database [33] are only done with the TAMU code.

The numerical Schlieren representations of the density field shown in the numerical illustrations are computed with the discrete version of [3, Eq. (35)] with $\beta = 15$; see the Supplemental Material of Guermond et al. [23] for more details.

5.2. Equations of state

In this section we list the equations of state that we use in the numerical illustrations below. In all the tests reported, the equations of state are used as an oracle. We make no assumptions on the physical validity of the equations of state and only require that they provide a positive pressure.

Noble-Abel equation of state The caloric Noble-Abel (or covolume) equation of state reads

$$p(\rho,e) := (\gamma-1)\frac{\rho e}{1-b\rho}. \tag{5.1}$$

Van der Waals equation of state The caloric Van der Waals equation of state is given by

$$p(\rho,e) := (\gamma-1)\frac{\rho e + a\rho^2}{1-b\rho} - a\rho^2. \tag{5.2}$$

Mie-Gruneisen with linear Hugoniot locus The Mie-Gruneisen equation of state with a linear Hugoniot locus as the reference curve is defined by

$$p(\rho,e) := p_{\text{ref}}(\rho) + \rho\Gamma(\rho)\big(e - e_{\text{ref}}(\rho)\big), \tag{5.3a}$$

$$\text{with}\quad p_{\text{ref}}(\rho) := P_0 + \rho_0 c_0^2 \frac{1-\frac{\rho_0}{\rho}}{\Big(1-s(1-\frac{\rho_0}{\rho})\Big)^2},\quad e_{\text{ref}}(\rho) := e_0 + \frac{P_0 + p_{\text{ref}}(\rho)}{2\rho_0}(1-\frac{\rho_0}{\rho}), \tag{5.3b}$$

where $s>1$. We refer the reader to Menikoff [35, Sec. 4.4] for a discussion of this particular equation of state and respective parameters. For simplicity, we take $\Gamma(\rho) := \Gamma_0$ and $P_0 = 0$, $e_0 = 0$.

Jones-Wilkins-Lee equation of state The pressure given by the Jones-Wilkins-Lee (JWL) equation of state is defined as follows:

$$p(\rho,e) := A\Big(1-\frac{\omega}{R_1}\frac{\rho}{\rho_0}\Big)\exp\Big(-R_1\frac{\rho_0}{\rho}\Big) + B\Big(1-\frac{\omega}{R_2}\frac{\rho}{\rho_0}\Big)\exp\Big(-R_2\frac{\rho_0}{\rho}\Big) + \omega\rho e. \tag{5.4}$$

The JWL equation of state was first introduced in Lee et al. [32, Eq. (I.3)]. We refer the reader to Segletes [46] for a discussion of the various forms of this equation of state seen in the literature. We note that (5.4) can be recast in "Mie-Gruneisen" form as follows:

$$p(\rho,e) := p_{\text{ref}}(\rho) + \omega\rho\big(e - e_{\text{ref}}(\rho)\big),$$

$$\text{with}\quad p_{\text{ref}}(\rho) := Ae^{-R_1\frac{\rho_0}{\rho}} + Be^{-R_2\frac{\rho_0}{\rho}},\quad e_{\text{ref}}(\rho) := \frac{A}{R_1}\frac{1}{\rho_0}e^{-R_1\frac{\rho_0}{\rho}} + \frac{B}{R_2}\frac{1}{\rho_0}e^{-R_2\frac{\rho_0}{\rho}}.$$

Table 1
$\delta_\infty(T)$ error (5.5) and corresponding convergence rates for the one-dimensional smooth traveling wave problem with exact solution (5.6) under uniform refinement of the interval $D = (0, 1)$.

	Ideal		VdW		JWL		MG	
$\lvert\mathcal{V}\rvert$	$\delta_\infty(T)$		$\delta_\infty(T)$		$\delta_\infty(T)$		$\delta_\infty(T)$	
101	1.94e-02	–	1.24e-01	–	7.93e-02	–	1.24e-05	–
201	4.03e-03	2.27	6.24e-03	4.30	2.53e-02	1.65	2.56e-06	2.28
401	7.91e-04	2.35	9.92e-04	2.65	3.61e-03	2.81	5.03e-07	2.35
801	1.44e-04	2.46	1.75e-04	2.51	1.31e-04	4.78	9.17e-08	2.46
1601	2.75e-05	2.39	3.29e-05	2.41	2.51e-05	2.38	1.75e-08	2.39
3201	5.18e-06	2.41	6.17e-06	2.41	4.73e-06	2.41	3.29e-09	2.41
6401	9.69e-07	2.42	1.16e-06	2.42	8.87e-07	2.42	6.22e-10	2.41

5.3. Convergence tests

We now verify the accuracy of the proposed method. We define a consolidated error indicator at time t by accumulating the relative error in the L^q-norm ($q \in [1, \infty]$):

$$\delta_q(t) := \frac{\|\rho_h(t) - \rho(t)\|_q}{\|\rho(t)\|_q} + \frac{\|\boldsymbol{m}_h(t) - \boldsymbol{m}(t)\|_q}{\|\boldsymbol{m}(t)\|_q} + \frac{\|E_h(t) - E(t)\|_q}{\|E(t)\|_q}, \tag{5.5}$$

where $\rho(t)$, $\boldsymbol{m}(t)$, $E(t)$ are the exact states at time t, and $\rho_h(t)$, $\boldsymbol{m}_h(t)$, $E_h(t)$ are the finite element approximations at time t for the respective conserved variables.

5.3.1. 1D – Smooth traveling wave

We consider a one-dimensional test proposed in [22, Sec. 5.2] consisting of a smooth traveling wave. The goal of this test is to show that we achieve (at least) second-order accuracy in space with any equation of state used for the pressure oracle. The smooth traveling wave is an exact solution to the Euler equations where the primitive variables are set as follows:

$$\rho(x,t) = \begin{cases} \rho_0 + 2^6(x_1 - x_0)^{-6}(x - v_0 t - x_0)^3(x_1 - x + v_0 t)^3 & \text{if } x_0 \le x - v_0 t \le x_1, \\ \rho_0 & \text{otherwise,} \end{cases} \tag{5.6a}$$

$$v(x,t) = v_0, \qquad p(x,t) = p_0, \tag{5.6b}$$

where x_0 and x_1 are arbitrary constants such that $x_0 < x_1$. This solution is compactly supported in space, has two continuous derivatives and its third derivative is uniformly bounded and has bounded variations. Just as in [22], we set the constants to $x_0 = 0.1$ and $x_1 = 0.3$. Notice that by fixing the pressure to be constant, the solution is independent of the equation of state. The internal energy is initiated by using the respective equations of state defined in §5.2. The constants p_0, ρ_0, v_0 are chosen to accommodate for the material in question.

We consider the following four configurations (here, T is the final time of the simulation):

- Ideal EOS: (5.1) with $b = 0$, $\gamma = 1.4$, $\rho_0 = p_0 = v_0 = 1$, $T = 0.6$;
- Van der Waals EOS: (5.2) with $a = 1$, $b = 0.075$, $\gamma = 1.4$, $\rho_0 = p_0 = v_0 = 1$, $T = 0.6$;
- Jones-Wilkins-Lee EOS: (5.4) with $A = 1$, $B = -1$, $R_1 = 2$, $R_2 = \omega = \rho_0 = p_0 = v_0 = 1$, $T = 0.6$;
- Mie-Gruneisen EOS: (5.3) with $\tilde{\rho}_0 = 2790$, $c_0 = 5330$, $s = 1.34$, $\Gamma_0 = 2.00$, $\rho_0 = 3500$, $p_0 = 1 \times 10^{11}$, $v_0 = 1 \times 10^4$, $T = 6 \times 10^{-5}$.

The tests are performed on uniform meshes with the domain $D = (0, 1)$ using the `TAMU` code. The first mesh is composed of 100 cells. The other meshes are obtained by uniform refinement via bisection. We use CFL $= 0.1$ and set Dirichlet boundary conditions for all tests. We report in Table 1 the quantity $\delta_\infty(T)$ and the respective convergence rates for the equations of state used. We observe that the convergence rate is greater than 2 with each EOS (this is a well known super convergence effect observed on uniform meshes [7,18,50]).

5.3.2. 2D – Isentropic vortex with Van der Waals EOS

To demonstrate higher-order convergence in $\mathbb{R}^2$, we consider a novel exact solution of the Euler equations (2.1a)–(2.1c) using the Van der Waals equation of state (5.2) with $b := 0$. The exact solution is a modified version of the isentropic vortex problem with an ideal gas equation of state (see Yee et al. [53]). For completeness, a derivation of the solution is presented in Appendix A.

Recalling that a and γ are the parameters of the Van der Waals equation of state (5.2), the isentropic vortex solution is given by

$$\rho(\boldsymbol{x},t) = \Big[\frac{3C}{8a} - \frac{1}{2}\sqrt{\frac{9C^2}{16a^2} + \frac{2}{a}\Big(F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\Big)}\Big]^2, \tag{5.7a}$$

Table 2
Error quantity (5.5) and convergence rates for the isentropic vortex problem with the Van der Waals equation of state. The exact solution is given by (5.7).

$\lvert\mathcal{V}\rvert$	$\delta_1(T)$		$\delta_2(T)$		$\delta_\infty(T)$	
289	1.17e-01	–	2.01e-01	–	6.82e-01	–
1089	1.18e-02	3.46	2.65e-02	3.06	1.05e-01	2.82
4225	7.92e-04	3.98	1.96e-03	3.84	7.87e-03	3.82
16641	5.57e-05	3.87	1.32e-04	3.93	5.50e-04	3.88
66049	5.07e-06	3.48	1.20e-05	3.48	7.79e-05	2.83
263169	7.55e-07	2.76	2.25e-06	2.42	2.03e-05	1.95
1050625	1.64e-07	2.20	5.51e-07	2.04	5.52e-06	1.88
4198401	4.08e-08	2.01	1.38e-07	2.00	1.51e-06	1.87

Table 3
JWL parameters for Woodward-Colella interacting blast wave benchmark.

	A	B	R_1	R_2	ω	ρ_0	Final time T
Case 1	6.321×10^3	-4.472	11.3	1.13	0.8938	1	0.038
Case 2	7.7828×10^{11}	7.071428×10^9	4.2	1.00	0.3000	1891	8.2×10^{-4}

$$\boldsymbol{v}(\boldsymbol{x},t)=\boldsymbol{v}_\infty+\psi(\bar{\boldsymbol{x}})\big(-\bar{x}_2,\bar{x}_1\big), \tag{5.7b}$$

$$p(\boldsymbol{x},t)=C(\gamma-1)\rho(\boldsymbol{x},t)^\gamma-a\rho(\boldsymbol{x},t)^2, \tag{5.7c}$$

where $\bar{\boldsymbol{x}}:=\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t=(\bar{x}_1,\bar{x}_2)$, $C:=(p_\infty+a\rho_\infty^2)/\rho_\infty^{3/2}$, $F:=-a\rho_\infty-3p_\infty/\rho_\infty$. Here ρ_∞ and p_∞ are the density and pressure in the far field, and $\psi(\boldsymbol{x}):=\frac{\beta}{2\pi}\exp\Big(\frac{1}{2}(1-\frac{1}{r_0^2}\|\boldsymbol{x}\|_{\ell^2}^2)\Big)$. We set the far field conditions to $\rho_\infty=0.1$, $p_\infty=1$ and $\boldsymbol{v}_\infty=(1,1)$. We also set $\gamma=\frac{3}{2}$ and $a=1$. This gives $C=\frac{101}{\sqrt{10}}$ and $F=-\frac{301}{10}$. The rest of the constants are set as follows: $\boldsymbol{x}^0=(-1,-1)$, $r_0=1$, $\beta=20$. We perform the convergence tests using the `Ryujin` code. The computational domain is set to $D=(-5,5)\times(-5,5)$. We run the simulations until the final time $T=2$ with CFL $=0.1$ and Dirichlet boundary conditions. In Table 2, we report the $\delta_q(T)$ errors and the respective convergence rates for $q\in\{1,2,\infty\}$. The tests are done over eight quadrilateral grids and seven levels of uniform refinements. We observe second-order convergence in all the norms.

5.4. Benchmark configurations

In this section, we consider common benchmarks seen in the literature and modify them appropriately using different equations of state. Our goal is to compare the present method with the state of the art in the literature.

5.4.1. 1D – Woodward-Colella blast wave

We demonstrate the robustness of the proposed method by reproducing the Woodward-Colella interacting blast wave benchmark using the Jones-Wilkins-Lee equation of state (5.4). For this benchmark, we consider two different cases. The first is that seen in Toro et al. [51, Sec. 5.1] and the second consists of parameters found in Lee et al. [32, Tab. 2("HMX")]. The parameters for both cases are given in Table 3.

The initial state for the blast wave problem is given as follows:

$$(\rho_0(x),v_0(x),p_0(x))=\begin{cases}(1,0,10^3) & \text{if } x\in[0,0.1],\\ (1,0,10^{-2}) & \text{if } x\in(0.1,0.9),\\ (1,0,10^2) & \text{if } x\in[0.9,1].\end{cases} \tag{5.8}$$

The simulations are performed on the domain $D=(0,1)$ with the `TAMU` code. All the tests use CFL $=0.9$ and slip boundary conditions. We show in Fig. 1 the density profiles for both cases at their respective final times using three different meshes composed of 1600, 6400, and 25600 elements. The results compare well with what is available in the literature; see e.g., Toro et al. [51, Fig. 2] for Case 1.

5.4.2. 1D – Riemann problem with SESAME database

We now demonstrate the method's ability to handle tabulated data as the pressure oracle. In particular, we consider a Riemann problem modeling the impact of a right-moving aluminum slab into a stationary aluminum slab at high velocities. To simulate the material aluminum, we use the Material ID `3720` in the `SESAME` database [33] (access to the database can be obtained by contacting `sesame@lanl.gov`). We let the density of the two aluminum slabs be $\rho_0=3000\,\mathrm{kg\,m^{-3}}$. The pressure at this density and at room temperature (293 K) is $p_0=1.004\,489\times10^{10}\,\mathrm{Pa}$ (this value is obtained from the `SESAME` database). For completeness, given these pressure and density values, the initial specific internal energy is $e_0=1.106\,117\times10^5\,\mathrm{J\,kg^{-1}}$.

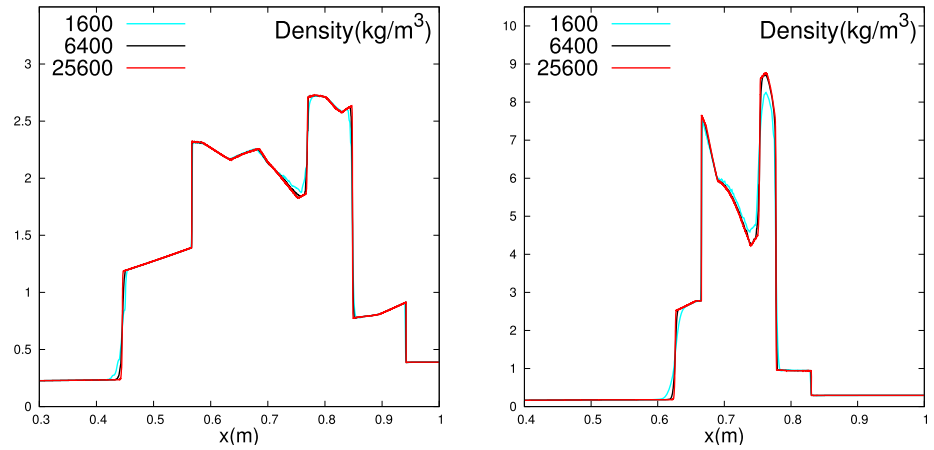

Fig. 1. Density profile for the Woodward-Colella blast wave problem using the Jones-Wilkins-Lee equation of state. Left: Case 1 at $T = 0.038$. Right: Case 2 at $T = 8.2 \times 10^{-4}$.

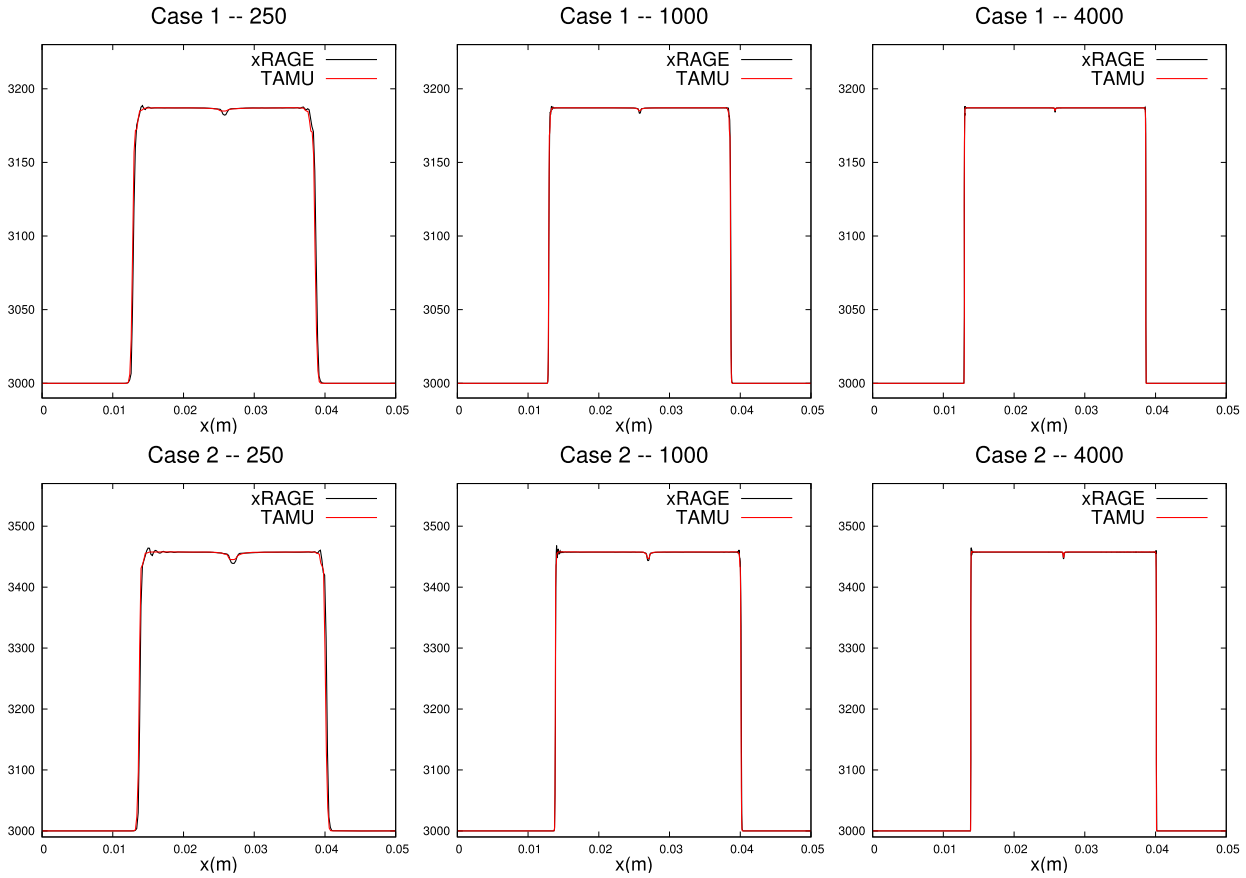

Fig. 2. Density comparison for the aluminum impact problem with the TAMU and xRAGE codes using the SESAME database.

The computational domain is set to $D = (0, 0.05\,\mathrm{m})$ where the two aluminum slabs are separated at $x = 0.025\,\mathrm{m}$. We consider two cases: we assume in the first case that the left slab initially moves with velocity $800\,\mathrm{m\,s^{-1}}$ (Case 1), and in the second case we assume that the velocity is $2000\,\mathrm{m\,s^{-1}}$ (Case 2). The simulations are run until final time $T = 2 \times 10^{-6}\,\mathrm{s}$ and performed with 250, 1000 and 4000 mesh elements to show convergence. We use CFL $= 0.5$ and set Dirichlet conditions on the left boundary and slip conditions on the right boundary. For verification, we run the same configuration with the xRAGE code developed at Los Alamos National Laboratory (see: [15] and [17]). In Fig. 2, we show the density output comparison between the two codes. We see very good agreement between the codes. The small oscillations seen on the xRAGE results

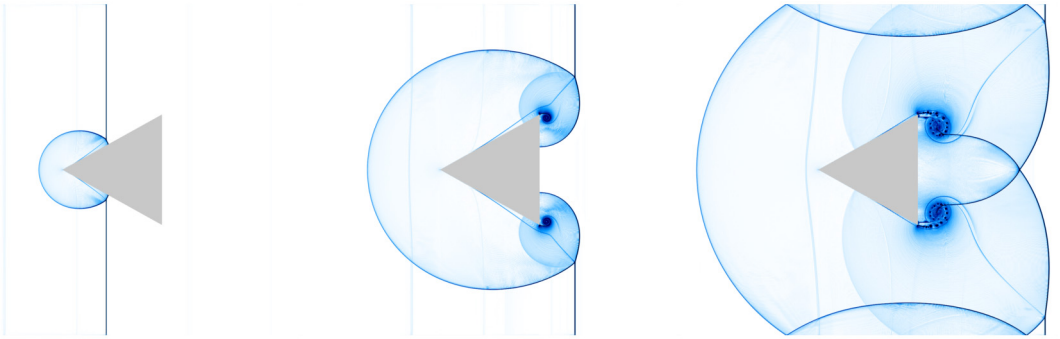

Fig. 3. Shock wave interacting with triangular obstacle at $t = \{1 \times 10^{-3}, 1.6 \times 10^{-3}, 2.2 \times 10^{-3}\}$.

are also observed with the TAMU code when relaxation on the density and the surrogate entropy is active. The relaxation is disabled in the results shown here. This test clearly demonstrates the method's ability to handle tabulated data.

5.5. 2D – Shock collision with triangular obstacle

We now reproduce a 2D problem proposed in Toro et al. [51, Sec. 5.4] which investigates traveling shock waves colliding with a triangular obstacle. This configuration is commonly known in the literature as Schardin's problem [45]. We refer the reader to Chang and Chang [6] where the authors experimentally reproduced Schardin's original experiments and give a detailed description of the experiments.

This test is performed with the Van der Waals equation of state (5.2) with $\gamma = 864.7/577.8$, $a = 0.14$, $b = 3.258 \times 10^{-5}$ and initialized as follows. The relative Mach speed is $M_S = 1.3$ with the primitive state in front of the shock set to $\boldsymbol{u}_R = (1.225, 0, 0, 101325)^\intercal$. Using the Rankine-Hugoniot conditions to derive the post shock state, the complete initial state is given as follows:

$$\big(\rho_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x}), p_0(\boldsymbol{x})\big) = \begin{cases} (1.82039, 148.597, 0, 185145), & \text{if } x \leq -0.55, \\ (1.225, 0, 0, 101325), & \text{if } x > -0.55. \end{cases} \tag{5.9}$$

The computations are performed with the Ryujin code. The computational domain is defined as $D = (-0.65, 0.5) \times (-0.5, 0.5) \backslash K$ where K is the triangle formed by the vertices $(-0.2, 0.0)$, $(0.1, 1/6)$, and $(0.1, -1/6)$. The simulations are run until $T = 2.2 \times 10^{-3}$ with a CFL of 0.6. The mesh is composed of 7,347,200 $\mathbb{Q}_1$-nodes. Dirichlet conditions are imposed on the left boundary, dynamic outflow conditions on the right boundary and slip conditions on the rest of the boundaries.

We show in Fig. 3 the Schlieren output of the shock wave interacting with the triangular obstacle at three time snapshots $t = \{1 \times 10^{-3}, 1.6 \times 10^{-3}, 2.2 \times 10^{-3}\}$. The results match up well with the experimental photos shown in [6]. In particular, we see vortices developing along the slip layer behind the back vertices of the triangle (see: [6, Fig. 8]); these so-called vortexlets are not apparent in [51, Fig. 6] likely due to a lack of mesh resolution.

5.6. 2D – Shock bubble interaction

In this section, we consider a single material shock-bubble interaction benchmark proposed in Wang and Li [52, Sec 5.2.2] using the Jones-Wilkins-Lee equation of state. For more details on shock-bubble interaction problems, we refer to Haas and Sturtevant [26] for experimental results and to Quirk and Karni [41] for the description of the simulation setup. This test demonstrates the robustness of our method by being able to reproduce the complex interactions of the shock hitting the bubble.

Let $\mathfrak{B}$ denote the bubble centered at $(0.5, 0.5)$ with radius 0.25. The primitive states at the initial time for the ambient fluid and bubble are respectively defined as follows:

$$(\rho_R, \boldsymbol{v}_R, p_R) = (1.0 \times 10^3, 0, 5.0 \times 10^{10}), \tag{5.10}$$

$$(\rho_{\mathfrak{B}}, \boldsymbol{v}_{\mathfrak{B}}, p_{\mathfrak{B}}) = (2.0 \times 10^3, 0, 5.0 \times 10^{10}). \tag{5.11}$$

The pressure across the shock is prescribed to be $p_L = 4.369 \times 10^{11}$, and the remaining state variables, ρ_L and $\boldsymbol{v}_L$, are computed using the Rankine-Hugoniot conditions. Thus, the initial primitive state for the problem is given by

$$(\rho_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x}), p_0(\boldsymbol{x})) = \begin{cases} (3778.85, 16867.6, 0, 4.369 \times 10^{11}), & \text{if } x < 0.05, \\ (1000, 0, 0, 5 \times 10^{10}), & \text{if } x \geq 0.05 \text{ and } \boldsymbol{x} \notin \mathfrak{B}, \\ (2000, 0, 0, 5 \times 10^{10}), & \text{if } \boldsymbol{x} \in \mathfrak{B}. \end{cases} \tag{5.12}$$

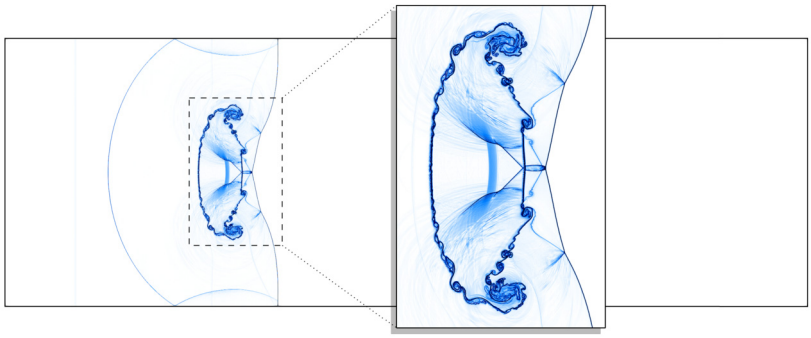

Fig. 4. Schlieren plots for the shock-bubble interaction benchmark for $t = 40\,\mu$s.

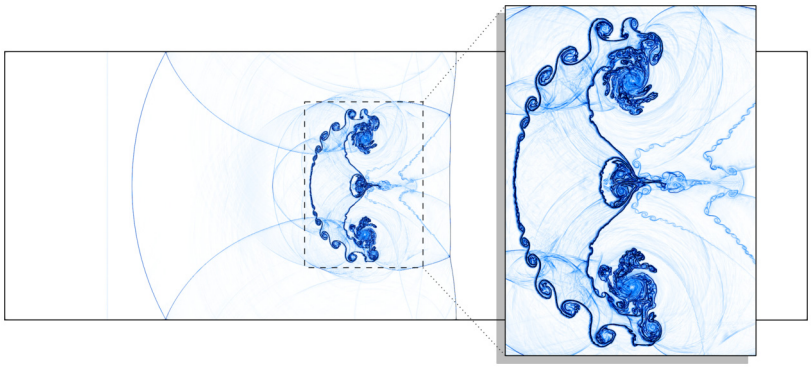

Fig. 5. Schlieren plots for the shock-bubble interaction benchmark for $t = 70\,\mu$s.

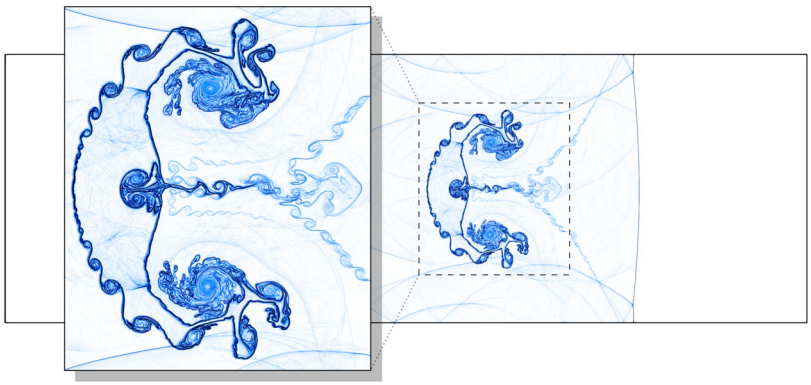

Fig. 6. Schlieren plots for the shock-bubble interaction benchmark for $t = 100\,\mu$s (bottom).

We perform the tests with the `Ryujin` code. The parameters for the JWL EOS (5.4) are set to

$$A = 8.545 \times 10^{11},\ B = 2.05 \times 10^{10},\ R_1 = 4.6,\ R_2 = 1.35,\ \omega = 0.25,\ \rho_0 = 1.84 \times 10^3.$$

The computational domain is set to $D = (0, 3) \times (0, 1)$ and the mesh is composed of 50,348,033 $\mathbb{Q}_1$-nodes. The simulations are performed until the final time $T = 100\,\mu$s with CFL$= 0.5$. Slip boundary conditions are applied on all boundaries. In Figs. 4 to 6, we show the Schlieren plots of the density at times $t \in \{40\,\mu\text{s}, 70\,\mu\text{s}, 100\,\mu\text{s}\}$. The results compare well with Figs. 5.6 and 5.7 shown in Wang and Li [52].

5.7. Composite wave diffraction around a backward facing step

We now consider the phenomenon of a composite wave diffracting around a 90° angle. Experimental results for shock waves for a number of different fluids and geometries have been reported in Bazhenova et al. [4]. Shock wave diffraction (at least for the case of an ideal gas equation of state) is a frequently studied problem with results reported for a variety

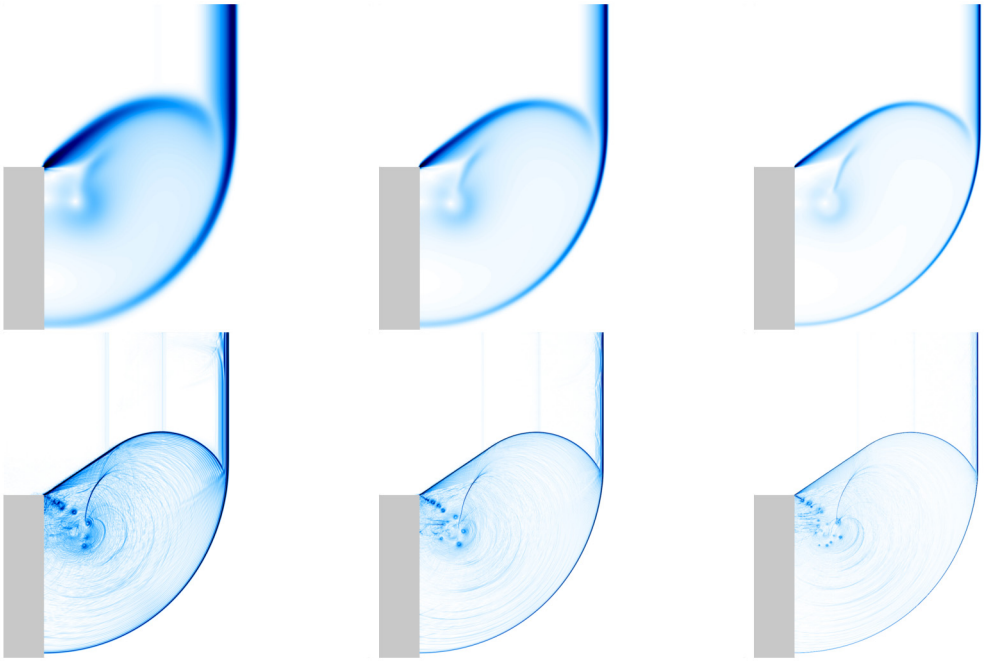

Fig. 7. Comparison of numerical Schlieren plots for first-order accurate (top row) and second-order accurate (bottom row) solutions at time $T = 0.02\,\mathrm{s}$. The mesh resolution increases from left to right as follows: 1,116,289, 4,460,801, and 17,834,497 $\mathbb{Q}_1$-nodes.

of different Mach numbers; we refer the reader to Artebrant and Schroll [1], Ripley et al. [42] and to a collection of posters from the 18th International Symposium on Shock Waves (ISSW) published in Takayama and Inoue [48]. However, only a few results are available in the literature for non-ideal fluids. Brown and Argrow [5] have simulated the shock wave diffraction problem for Bethe-Zeldovich-Thompson fluids using the Van der Waals equation of state.

We recreate Problem TD3 reported in [5] with the notable difference that we retain the dimensional form of the Van der Waals equation of state instead of nondimensionalizing the equations. We also modify the initial data so that it satisfies the Rankine-Hugoniot conditions, which is not the case of the initial states reported in [5]. More precisely, we use the Van der Waals EOS (5.2) with $a = 2.2\,\mathrm{Pa\,m^6/kg^2}$, $b = 7.25 \times 10^{-4}\,\mathrm{m^3/kg}$, and $\gamma = 1.0125$. Note that $\gamma < 1.06$ which ensures the existence of a thermodynamic region where the fundamental derivative is negative, see [49, §3.1] or [14, §6.3]. The right state is given by $(\rho_R, \boldsymbol{v}_R, p_R) = (128.76\,\mathrm{kg/m^3}, 0\,\mathrm{m/s}, 89\,910.6\,\mathrm{Pa})$. We prescribe the left pressure to be $p_L = 156\,568\,\mathrm{Pa}$ and compute the remaining states through the Rankine-Hugoniot conditions. The shock Mach number is $M_{\mathcal{S}} = \mathcal{S}/a_R \approx 1.22$, where $\mathcal{S} \approx 28.06\,\mathrm{m/s}$ is the shock speed and a_R is the material speed of sound.

The computational domain is given by $D = ([0, 0.125] \times [0.5, 1]) \cup ([0.125, 1.125] \times [0, 1])$. For the boundary conditions, we enforce Dirichlet on $\{x = 0\}$, dynamic outflow on $\{x = 1.125\}$, and slip on the remaining boundaries. The simulations are done with the `Ryujin` code using three meshes with 1,116,289, 4,460,801, and 17,834,497 $\mathbb{Q}_1$-nodes to observe convergence. We run until final time $T = 0.02\,\mathrm{s}$ with $\mathrm{CFL} = 0.5$. The numerical Schlieren plots at the final time are shown in Fig. 7. One important property of this test is that the structure created at the initial time is a composite wave composed of a shock followed by an expansion (see the vertical structure normal to the top of the channel). Another feature of this problem is that, immediately upon diffracting around the corner, a second composite wave is created; this wave is composed of an expansion followed by a shock. In Fig. 7 the wave in question is the oblique structure attached to the corner. The simulations have been done with the low-order method and the second-order method described in the paper to convince the reader and ourselves that the composite waves observed here are physical. We indeed observe that the same composite waves are present in both series of simulations.

6. Conclusion

We have developed an invariant-domain preserving second-order accurate method for the Euler equations with arbitrary or tabulated equation of states. The work presented is the continuation and extension of Guermond et al. [22] and Clayton et al. [9]. We proposed a surrogate entropy functional that increases across shocks in the associated 1D Riemann problem to work around the lack of a general mathematical entropy valid for arbitrary equations of state. A convex limiting procedure was performed on this surrogate entropy functional to enforce a local minimum principle. This in turn implies a positive local lower bound on the internal energy. Numerical evidence of higher-order accuracy was demonstrated with convergence tests and several computational benchmarks.

CRediT authorship contribution statement

All the authors have equally contributed to this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Isentropic vortex with van der Waals equation of state

We present a derivation of the isentropic vortex solution with the van der Waals equation of state and give some necessary conditions for the existence of this solution.

Theorem A.1. *Consider the van der Waals equation of state (5.2) with $a>0$, $b:=0$, and $\gamma:=\frac{3}{2}$ or $\gamma:=2$. Let $\boldsymbol{x}^0\in\mathbb{R}^2$, $\beta>0$, $r_0>0$. Let $\rho_\infty>0$, $\boldsymbol{v}_\infty\in\mathbb{R}^2$, $p_\infty>0$ and assume that*

$$p_\infty>\tfrac{1}{3}a\rho_\infty^2,\qquad a\rho_\infty+\frac{3p_\infty}{\rho_\infty}>\frac{\beta^2\exp(1)}{8r_0^2\pi^2} \tag{A.1}$$

if $\gamma=\frac{3}{2}$. The following density, velocity, and pressure fields solve the compressible Euler equations (2.1a)–(2.1c) with the van der Waals equation of state:

$$\rho(\boldsymbol{x},t):=\begin{cases}\Big(\frac{3C}{4a}-\frac{1}{2}\sqrt{\frac{9C^2}{4a^2}+\frac{2}{a}\big(F+\frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\big)}\Big)^2, & \text{if } \gamma=\frac{3}{2},\\ \rho_\infty-\frac{\rho_\infty^2}{4p_\infty r_0^2}\psi(\overline{\boldsymbol{x}}), & \text{if } \gamma=2,\end{cases} \tag{A.2a}$$

$$\boldsymbol{v}(\boldsymbol{x},t):=\boldsymbol{v}_\infty+\psi(\overline{\boldsymbol{x}})(-\bar{x}_2,\bar{x}_1)^{\mathsf{T}}, \tag{A.2b}$$

$$p(\boldsymbol{x},t):=C\rho(\boldsymbol{x},t)^\gamma-a\rho(\boldsymbol{x},t)^2, \tag{A.2c}$$

with $\psi(\overline{\boldsymbol{x}}):=\frac{\beta}{2\pi}\exp(\frac{1}{2}(1-\frac{1}{r_0^2}\|\overline{\boldsymbol{x}}\|_{\ell^2}^2))$, $(\bar{x}_1,\bar{x}_2)=\overline{\boldsymbol{x}}:=\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t$, $C=(p_\infty+a\rho_\infty^2)/\rho_\infty^{3/2}$, and $F=-a\rho_\infty-3p_\infty/\rho_\infty$.

Proof. The derivation of the isentropic vortex begins with the additional assumption that the velocity field is divergence free. That is, $\nabla\cdot\boldsymbol{v}=0$. Under this assumption the Euler equations take the following simplified form:

$$\partial_t\rho(\boldsymbol{x},t)+\boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla\rho(\boldsymbol{x},t)=0,\qquad \boldsymbol{x}\in\mathbb{R}^2,\ t>0, \tag{A.3}$$

$$\partial_t\boldsymbol{v}(\boldsymbol{x},t)+(\boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla)\boldsymbol{v}(\boldsymbol{x},t)=-\frac{1}{\rho(\boldsymbol{x},t)}\nabla p(\boldsymbol{x},t),\qquad \boldsymbol{x}\in\mathbb{R}^2,\ t>0 \tag{A.4}$$

$$\partial_t e(\boldsymbol{x},t)+\boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla e(\boldsymbol{x},t)=0,\qquad \boldsymbol{x}\in\mathbb{R}^2,\ t>0, \tag{A.5}$$

with $\boldsymbol{x}:=(x_1,x_2)$, boundary conditions, $(\rho_\infty,\boldsymbol{v}_\infty:=(v_{1,\infty},v_{2,\infty})^{\mathsf{T}},p_\infty)$ and yet to be determined initial conditions $(\rho_0(\boldsymbol{x}),\boldsymbol{v}_0(\boldsymbol{x}),p_0(\boldsymbol{x}))$. To keep things general, we make no assumption on the equation of state for $p=p(\rho,e)$.

We write the solution as a perturbation of the far-field state; i.e., we define $\boldsymbol{v}:=\boldsymbol{v}_\infty+\delta\boldsymbol{v}$ with

$$\delta\boldsymbol{v}(\boldsymbol{x},t):=\begin{pmatrix}\partial_{x_2}\psi(\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t)\\ -\partial_{x_1}\psi(\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t)\end{pmatrix}, \tag{A.6}$$

with the stream function $\psi(\boldsymbol{x}):=\frac{\beta}{2\pi}\exp(\frac{1}{2}(1-\frac{\|\boldsymbol{x}\|_{\ell^2}^2}{r_0^2}))$. Here $\boldsymbol{x}^0:=(x_1^0,x_2^0)\in\mathbb{R}^2$, β, and r_0 are free parameters. To further simplify notation, define $(\bar{x}_1,\bar{x}_2)=\overline{\boldsymbol{x}}:=\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t$ and $r^2:=\|\overline{\boldsymbol{x}}\|_{\ell^2}^2$. Note the following identities which will be used later on:

$$\partial_{x_i}\psi(\overline{\boldsymbol{x}})=-\frac{\bar{x}_i}{r_0^2}\psi(\overline{\boldsymbol{x}}), \tag{A.7}$$

$$\partial_{x_ix_j}\psi(\overline{\boldsymbol{x}})=\frac{1}{r_0^2}\Big(-\delta_{ij}+\frac{\bar{x}_i\bar{x}_j}{r_0^2}\Big)\psi(\overline{\boldsymbol{x}}), \tag{A.8}$$

$$\partial_{tx_i}\psi(\overline{\boldsymbol{x}}) = \frac{1}{r_0^2}\Big(v_{i,\infty} - \frac{\bar{x}_i \boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\Big)\psi(\overline{\boldsymbol{x}}), \tag{A.9}$$

where δ_{ij} is the Kronecker symbol and $i, j \in \{1, 2\}$.

Using that $\boldsymbol{v} = \boldsymbol{v}_\infty + \delta\boldsymbol{v}$, the left hand side of (A.4) becomes,

$$\partial_t \boldsymbol{v} + \boldsymbol{v}\cdot\nabla\boldsymbol{v} = \partial_t(\delta\boldsymbol{v}) + (\boldsymbol{v}_\infty\cdot\nabla)\delta\boldsymbol{v} + (\delta\boldsymbol{v}\cdot\nabla)\delta\boldsymbol{v}. \tag{A.10}$$

From the definition of $\delta\boldsymbol{v}$ and the identities (A.7), (A.8) and (A.9), we have,

$$(\delta\boldsymbol{v}\cdot\nabla)\delta\boldsymbol{v} = \begin{bmatrix} (\partial_{x_2}\psi)(\partial^2_{x_1x_2}\psi) - (\partial_{x_1}\psi)(\partial^2_{x_2}\psi) \\ -(\partial_{x_2}\psi)(\partial^2_{x_1}\psi) + (\partial_{x_1}\psi)(\partial^2_{x_1x_2}\psi) \end{bmatrix} = -\frac{\overline{\boldsymbol{x}}}{r_0^4}\psi(\overline{\boldsymbol{x}})^2 \tag{A.11}$$

$$(\boldsymbol{v}_\infty\cdot\nabla)\delta\boldsymbol{v} = \frac{1}{r_0^2}\Big(-\begin{bmatrix} v_{2,\infty} \\ -v_{1,\infty}\end{bmatrix} + \frac{\boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\begin{bmatrix}\bar{x}_2 \\ -\bar{x}_1\end{bmatrix}\Big)\psi(\overline{\boldsymbol{x}}) \tag{A.12}$$

$$\partial_t(\delta\boldsymbol{v}) = \frac{1}{r_0^2}\Big(\begin{bmatrix} v_{2,\infty} \\ -v_{1,\infty}\end{bmatrix} - \frac{\boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\begin{bmatrix}\bar{x}_2 \\ -\bar{x}_1\end{bmatrix}\Big)\psi(\overline{\boldsymbol{x}}) \tag{A.13}$$

Thus equation (A.4) becomes $-\frac{\overline{\boldsymbol{x}}}{r_0^4}\psi(\overline{\boldsymbol{x}})^2 = -\frac{1}{\rho}\nabla p$. This identity is furthermore written as,

$$-\frac{1}{2r_0^2}\nabla(\psi(\overline{\boldsymbol{x}})^2) = \frac{1}{\rho(t,\boldsymbol{x})}\nabla p(\rho(t,\boldsymbol{x})). \tag{A.14}$$

Up to this point, we have not made any assumption on the equation of state. We recover the well known isentropic vortex solution if we assume the pressure is given by the ideal gas law; i.e., $p(\rho) = C\rho^\gamma$ for the isentropic flow where $C = p_\infty/\rho_\infty^\gamma$. We now proceed with the van der Waals equation of state. For isentropic flows we have

$$p(\rho) = \frac{C\rho^\gamma}{(1-b\rho)^\gamma} - a\rho^2, \tag{A.15}$$

where C is some constant. (Note, we work with an arbitrary b to keep things general in the beginning.) Following the same process as in the ideal gas case, we compute the indefinite integral, $\int \frac{1}{\rho}\partial_{x_i}p(\rho)\,dx_i$:

$$\begin{aligned} -\frac{1}{2r_0^2}\int \partial_{x_i}\psi(\overline{\boldsymbol{x}})^2\,dx_i &= \int \frac{1}{\rho}\partial_{x_i}p(\rho)\,dx_i = \frac{p(\rho)}{\rho} + \int \frac{p(\rho)}{\rho^2}\partial_{x_i}\rho\,dx_i \\ &= \frac{p(\rho)}{\rho} + \int\Big(\frac{C\rho^{\gamma-2}}{(1-b\rho)^\gamma} - a\Big)\rho_{x_i}dx_i = \frac{p(\rho)}{\rho} + \int\frac{\partial}{\partial x_i}\Big[\frac{C}{\gamma-1}\Big(\frac{\rho}{1-b\rho}\Big)^{\gamma-1} - a\rho\Big]dx_i \\ &= \frac{C\rho^{\gamma-1}(\gamma-b\rho)}{(\gamma-1)(1-b\rho)^\gamma} - 2a\rho + F. \end{aligned}$$

Hence, $\rho(\overline{\boldsymbol{x}})$ can be found by solving the equation,

$$-\frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2 = \frac{C\rho^{\gamma-1}(\gamma-b\rho)}{(\gamma-1)(1-b\rho)^\gamma} - 2a\rho + F. \tag{A.16}$$

We have two immediate cases for solutions that can be found explicitly.

Case 1: $\gamma = 3/2$ **and** $b = 0$**:** In this case, (A.16) becomes a quadratic equation for $\sqrt{\rho}$,

$$\rho - \frac{3C}{2a}\sqrt{\rho} - \frac{1}{2a}\Big(F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\Big) = 0. \tag{A.17}$$

The constants C and F are determined by applying the far field condition to (A.15) and (A.17):

$$C = \frac{p_\infty + a\rho_\infty^2}{\rho_\infty^{3/2}} \quad\text{and}\quad F = -a\rho_\infty - \frac{3p_\infty}{\rho_\infty}. \tag{A.18}$$

However, care must be taken in the choice of p_∞ and ρ_∞ so that the sound speed remains real. Recall that the sound speed for the van der Waals EOS is $c(\rho,p) = \sqrt{\gamma\frac{p+a\rho^2}{\rho(1-b\rho)} - 2a\rho}$. The hypothesis $p_\infty > \frac{1}{3}a\rho_\infty^2$ guarantees that $c(\rho_\infty, p_\infty)^2 > 0$.

The physical root for equation (A.17) is $\sqrt{\rho} = \frac{3C}{4a} - \frac{1}{2}\sqrt{\frac{9C^2}{4a^2} + \frac{2}{a}\big(F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\big)}$. Furthermore, for the root to be real we require that $-F > \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2$ for all $\overline{\boldsymbol{x}}\in\mathbb{R}^2$. In particular,

$$a\rho_\infty + \frac{3p_\infty}{\rho_\infty} > \frac{\beta^2 \exp(1)}{8r_0^2\pi^2}. \tag{A.19}$$

Lastly, we must justify that the system remains hyperbolic; that is, the sound speed is real for all $(\boldsymbol{x}, t) \in \mathbb{R}^2 \times [0, \infty)$. Since the flow is isentropic, the sound speed for the van der Waals EOS (with $\gamma = 3/2$ and $b = 0$) is, $f(\rho) := c(p(\rho), \rho)^2 = \frac{3}{2}C\sqrt{\rho} - 2a\rho$. Note that $\lim_{\rho\to 0^+} f(\rho) = 0$, $f'(\rho) = \frac{3C}{4\sqrt{\rho}} - 2a$, and $\lim_{\rho\to 0^+} f'(\rho) = \infty$. Therefore, $f(\rho)$ has a maximum at $\rho = \left(\frac{3C}{8a}\right)^2$ and hence $f(\rho) > 0$ for $\rho \in (0, \left(\frac{3C}{8a}\right)^2)$. From the definition of ρ, (A.2a), we see that $0 < \rho < \left(\frac{3C}{4a}\right)^2$. Thus the sound speed is always real.

Case 2: $\gamma = 2$ and $b = 0$: For these choices of parameters, (A.16) becomes,

$$2(C - a)\rho + F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}}) = 0. \tag{A.20}$$

Using the far field boundary conditions for (A.2c) and (A.20) we find that $C = \frac{p_\infty}{\rho_\infty^2} + a$ and $F = -2p_\infty/\rho_\infty$, respectively. Solving for ρ in (A.20) we have,

$$\rho = \rho_\infty - \frac{\rho_\infty^2}{4p_\infty r_0^2}\psi(\overline{\boldsymbol{x}}). \tag{A.21}$$

Note the sound speed is $c(p(\rho), \rho)^2 = 2(C - a)\rho = \frac{2p_\infty}{\rho_\infty}\rho > 0$. □

References

[1] R. Artebrant, H.J. Schroll, Limiter-free third order logarithmic reconstruction, SIAM J. Sci. Comput. 28 (1) (2006) 359–381.
[2] J.W. Banks, On exact conservation for the Euler equations with complex equations of state, Commun. Comput. Phys. 8 (2010) 995–1015.
[3] J.W. Banks, W.D. Henshaw, D.W. Schwendeman, A.K. Kapila, A study of detonation propagation and diffraction with compliant confinement, Combust. Theory Model. 12 (4) (2008) 769–808.
[4] T. Bazhenova, L. Gvozdeva, M. Nettleton, Unsteady interactions of shock waves, Prog. Aerosp. Sci. 21 (1984) 249–331.
[5] B.P. Brown, B.M. Argrow, Nonclassical dense gas flows for simple geometries, AIAA J. 36 (10) (1998) 1842–1847.
[6] S.-M. Chang, K.-S. Chang, On the shock–vortex interaction in schardin's problem, Shock Waves 10 (5) (2000) 333–343.
[7] M.A. Christon, M.J. Martinez, T.E. Voth, Generalized Fourier analyses of the advection-diffusion equation-part I: one-dimensional domains, Int. J. Numer. Methods Fluids 45 (8) (2004) 839–887.
[8] B. Clayton, J.-L. Guermond, B. Popov, Upper bound on the maximum wave speed in Riemann problems for the Euler equations with tabulated equation of state, https://doi.org/10.5281/zenodo.4685868, April 2021.
[9] B. Clayton, J.-L. Guermond, B. Popov, Invariant domain-preserving approximations for the Euler equations with tabulated equation of state, SIAM J. Sci. Comput. 44 (1) (2022) A444–A470.
[10] P. Colella, H.M. Glaz, Efficient solution algorithms for the Riemann problem for real gases, J. Comput. Phys. 59 (2) (1985) 264–289.
[11] J.K. Dukowicz, A general, noniterative Riemann solver for Godunov's method, J. Comput. Phys. 61 (1) (1985) 119–137.
[12] M. Dumbser, V. Casulli, A conservative, weakly nonlinear semi-implicit finite volume scheme for the compressible Navier-Stokes equations with general equation of state, Appl. Math. Comput. 272 (part 2) (2016) 479–497.
[13] M. Dumbser, U. Iben, C.-D. Munz, Efficient implementation of high order unstructured weno schemes for cavitating flows, Comput. Fluids 86 (2013) 141–168.
[14] M. Fossati, L. Quartapelle, The Riemann problem for hyperbolic equations under a nonconvex flux with two inflection points, https://arxiv.org/abs/1402.5906, 2014.
[15] M. Gittings, R. Weaver, M. Clover, T. Betlach, N. Byrne, R. Coker, E. Dendy, R. Hueckstaedt, K. New, W.R. Oakes, D. Ranta, R. Stefan, The RAGE radiation-hydrodynamic code, Comput. Sci. Discov. 1 (1) (Nov 2008) 015005.
[16] E. Godlewski, P.-A. Raviart, Numerical Approximation of Hyperbolic Systems of Conservation Laws, Applied Mathematical Sciences, vol. 118, Springer-Verlag, New York, 1996.
[17] J.W. Grove, The xrage hydrodynamic solver, Technical report, 2019, https://www.osti.gov/biblio/1532686.
[18] J.-L. Guermond, R. Pasquetti, A correction technique for the dispersive effects of mass lumping for transport problems, Comput. Methods Appl. Mech. Eng. 253 (2013) 186–198.
[19] J.-L. Guermond, B. Popov, Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations, J. Comput. Phys. 321 (2016) 908–926.
[20] J.-L. Guermond, B. Popov, Invariant domains and first-order continuous finite element approximation for hyperbolic systems, SIAM J. Numer. Anal. 54 (4) (2016) 2466–2489.
[21] J.-L. Guermond, R. Pasquetti, B. Popov, Entropy viscosity method for nonlinear conservation laws, J. Comput. Phys. 230 (11) (2011) 4248–4267.
[22] J.-L. Guermond, M. Nazarov, B. Popov, I. Tomas, Second-order invariant domain preserving approximation of the Euler equations using convex limiting, SIAM J. Sci. Comput. 40 (5) (2018) A3211–A3239.
[23] J.-L. Guermond, B. Popov, I. Tomas, Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems, Comput. Methods Appl. Mech. Eng. 347 (2019) 143–175.
[24] J.-L. Guermond, M. Maier, B. Popov, I. Tomas, Second-order invariant domain preserving approximation of the compressible Navier-Stokes equations, Comput. Methods Appl. Mech. Eng. 375 (1) (2021) 113608.
[25] J.-L. Guermond, M. Kronbichler, M. Maier, B. Popov, I. Tomas, On the implementation of a robust and efficient finite element-based parallel solver for the compressible Navier–Stokes equations, Comput. Methods Appl. Mech. Eng. 389 (2022) 114250.
[26] J.-F. Haas, B. Sturtevant, Interaction of weak shock waves with cylindrical and spherical gas inhomogeneities, J. Fluid Mech. 181 (1987) 41–76.
[27] A. Harten, P.D. Lax, B. van Leer, On upstream differencing and Godunov-type schemes for hyperbolic conservation laws, SIAM Rev. 25 (1) (1983) 35–61.
[28] M.J. Ivings, D.M. Causon, E.F. Toro, On Riemann solvers for compressible liquids, Int. J. Numer. Methods Fluids 28 (3) (1998) 395–418.
[29] B. Khobalatte, B. Perthame, Maximum principle on the entropy and second-order kinetic schemes, Math. Comput. 62 (205) (1994) 119–131.
[30] P.D. Lax, Weak solutions of nonlinear hyperbolic equations and their numerical computation, Commun. Pure Appl. Math. 7 (1954) 159–193.

[31] B.J. Lee, E.F. Toro, C.E. Castro, N. Nikiforakis, Adaptive Osher-type scheme for the Euler equations with highly nonlinear equations of state, J. Comput. Phys. 246 (2013) 165–183.
[32] E.L. Lee, H.C. Hornig, J.W. Kury, Adiabatic expansion of high explosive detonation products, Technical Report UCRL-50422, Lawrence Radiation Laboratory, University of California, Livermore, 2 May 1968, https://www.osti.gov/biblio/4783904.
[33] S.P. Lyon, SESAME: the Los Alamos National Laboratory equation of state database, Los Alamos National Laboratory report LA-UR-92-3407, 1992.
[34] M. Maier, M. Kronbichler, Efficient parallel 3d computation of the compressible Euler equations with an invariant-domain preserving second-order finite-element scheme, ACM Trans. Parallel Comput. 8 (3) (2021) 16.
[35] R. Menikoff, Empirical Equations of State for Solids, Springer Berlin Heidelberg, 2007, pp. 143–188.
[36] H. Nessyahu, E. Tadmor, Nonoscillatory central differencing for hyperbolic conservation laws, J. Comput. Phys. 87 (2) (1990) 408–463.
[37] P.-O. Persson, J. Peraire, Sub-cell shock capturing for discontinuous Galerkin methods, in: 44th AIAA Aerospace Sciences Meeting and Exhibit, 2006, number AIAA paper no. 2015-2006-112 in Aerospace Sciences Meetings.
[38] B. Perthame, C.-W. Shu, On positivity preserving finite volume schemes for Euler equations, Numer. Math. 73 (1) (1996) 119–130.
[39] J. Pike, Riemann solvers for perfect and near-perfect gases, AIAA J. 31 (10) (1993) 1801–1808.
[40] L. Quartapelle, L. Castelletti, A. Guardone, G. Quaranta, Solution of the Riemann problem of classical gasdynamics, J. Comput. Phys. 190 (1) (2003) 118–140.
[41] J.J. Quirk, S. Karni, On the dynamics of a shock-bubble interaction, J. Fluid Mech. 318 (1996) 129–163.
[42] R. Ripley, F.-S. Lien, M. Yovanovich, Numerical simulation of shock diffraction on unstructured meshes, Comput. Fluids 35 (10) (2006) 1420–1431.
[43] P.L. Roe, J. Pike, Efficient construction and utilisation of approximate Riemann solutions, in: Proceedings of the Sixth International Symposium on Computing Methods in Applied Sciences and Engineering, VI, North-Holland Publishing Co., Netherlands, 1985, pp. 499–518.
[44] R. Saurel, E. Franquet, E. Daniel, O. Le Metayer, A relaxation-projection method for compressible flows. Part I: the numerical equation of state for the Euler equations, J. Comput. Phys. 223 (2) (2007) 822–845.
[45] H. Schardin, High frequency cinematography in the shock tube, J. Photogr. Sci. 5 (2) (1957) 17–19.
[46] S.B. Segletes, An examination of the JWL equation of state, Technical Report AD1055483, Army Research Lab Aberdeen Proving Ground, MD, United States, 2018.
[47] E. Tadmor, Numerical viscosity and the entropy condition for conservative difference schemes, Math. Comput. 43 (168) (1984) 369–381.
[48] K. Takayama, O. Inoue, Shock wave diffraction over a 90 degree sharp corner—posters presented at 18th ISSW, Shock Waves 1 (4) (1991) 301–312.
[49] P.A. Thompson, K.C. Lambrakis, Negative shock waves, J. Fluid Mech. 60 (1973) 187–208.
[50] 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) 229–248.
[51] E.F. Toro, C.E. Castro, B.J. Lee, A novel numerical flux for the 3D Euler equations with general equation of state, J. Comput. Phys. 303 (2015) 80–94.
[52] Y. Wang, J. Li, Stiffened gas approximation and grp resolution for fluid flows of real materials, arXiv preprint, arXiv:2108.13780, 2021.
[53] H. Yee, N. Sandham, M. Djomehri, Low-dissipative high-order shock-capturing methods using characteristic-based filters, J. Comput. Phys. 150 (1) (1999) 199–238.