INVARIANT DOMAIN-PRESERVING APPROXIMATIONS FOR THE EULER EQUATIONS WITH TABULATED EQUATION OF STATE[*]

BENNETT CLAYTON[†], JEAN-LUC GUERMOND[†], AND BOJAN POPOV[†]

Abstract. This paper is concerned with the approximation of the compressible Euler equations supplemented with an equation of state that is either tabulated or is given by an expression that is so involved that solving elementary Riemann problems is hopeless. A robust first-order approximation technique that guarantees that the density and the internal energy are positive is proposed. A key ingredient of the method is a local approximation of the equation of state using a covolume ansatz from which upper bounds on the maximum wave speed are derived for every elementary Riemann problem.

Key words. compressible Euler equation, tabulated equation of state, maximum wave speed, Riemann problem, invariant domain preserving approximation, composite waves

AMS subject classifications. 65M60, 65M12, 65M22, 35L65

DOI. 10.1137/21M1414097

1. Introduction.

In many important applications, the compressible Euler equations are supplemented with an equation of state that is either tabulated or given by a complicated analytic expression. Throughout this paper, we refer to this type of equation of state as the "oracle." In this case, approximating the Euler equations while guaranteeing positivity of the density and positivity of the internal energy is problematic since no exact solution of elementary Riemann problems can be a priori inferred. Solving a Riemann problem when the equation of state is analytically well defined is feasible, though possibly expensive (see, e.g., Colella and Glaz [6, sect. 1], Ivings, Causon, and Toro [19], Quartapelle et al. [31]). This cannot be efficiently done with an oracle for this requires interpolating/approximating the equation of state, and to the best of our knowledge, there is no clear technique to do so. Various methods to avoid this problem have been proposed in the literature. For instance, one can use approximate Riemann solvers like in Dukowicz [8], [6, sect. 2], Roe and Pike [33], Pike [29], or simplify the Riemann problem by using flux splitting techniques like in Toro, Castro, and Lee [35]. However, for most of these techniques very little is guaranteed besides positivity of the density, which is not difficult to achieve. The objective of this paper is to address these questions. More precisely, we propose an approximation method to solve the Euler equations equipped with an oracle. This is done by adapting the technique from Guermond and Popov [14] where invariant-domain properties are obtained by ascertaining that they hold true for elementary Riemann problems. The key is to augment each elementary Riemann system with an additional scalar equation and replace the oracle by a covolume equation of state where the coefficient γ is variable and obtained as the solution to the additional equation. This idea is

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section April 21, 2021; accepted for publication (in revised form) November 22, 2021; published electronically February 17, 2022.
https://doi.org/10.1137/21M1414097
Funding: This work was partially supported by National Science Foundation grant DMS-1619892, by the Air Force Office of Scientific Research, USAF, under grant/contract FA9550-18-1-0397, by the Army Research Office, under grant W911NF-15-1-0517, and by the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract B640889.

[†] Department of Mathematics, Texas A & M University, College Station, TX 77843-3368 USA (bgclayto@tamu.edu, guermond@math.tamu.edu, popov@math.tamu.edu).

adapted from Abgrall and Karni [2]. A variation of this idea is also employed in [6, eq. (37)] and Pantano, Saurel, and Schmitt [28, eq. (22)]. The proposed algorithm is explicit in time and preserves the positivity of the density and the internal energy under an appropriate CFL restriction on the time step. Additional properties can be preserved depending on the nature of the oracle. As in Guermond, Popov, and Tomas [17], the method is agnostic to the space approximation. An interesting feature of the method is that it automatically recovers the standard covolume behavior if the oracle is indeed a covolume equation of state. In compliance with Godunov's theorem, the method is only first-order accurate in space. Achieving higher-order accuracy in space can be done by implementing the convex limiting technique described in [16, 17]. We do not discuss this point here since it is beyond the scope of this paper. This work is in progress and will be reported elsewhere.

This paper is organized as follows. The problem and the notation are introduced in section 2. The space and time approximation method from [14] is also briefly recalled in this section. The main motivation of the paper is given at the end of section 2.2. We introduce an extended Riemann problem in section 3. The key point of section 3 is summarized in Remark 3.1. An exact weak solution to the extended Riemann problem is constructed in section 4. It is also shown in section 4 that this weak solution satisfies the expected invariant-domain properties. The main results of section 4 are Lemmas 4.4 and 4.5 and Theorem 4.6. An upper bound on the maximum wave speed for the extended Riemann problem is derived in section 5. This upper bound is the key piece of information that is needed for practitioners who may have little interest in the Riemann problem theory (see sections 5.2–5.5). The fact that this estimate of the maximum wave speed is a guaranteed upper bound implies that the proposed numerical algorithm satisfies the invariant-domain properties stated in Theorem 4.6. The technique introduced in this paper is illustrated in section 6 with continuous finite elements and various equations of states, including tabulated equations of state for various materials using the SESAME database [25]. Finally, this paper is supplemented with an appendix collecting technical results. Various pieces of software are made publicly available to guarantee reproducibility (Clayton, Guermond, and Popov [4, 5]).

2. Formulation of the problem.

We formulate the problem and introduce notation in this section. The main motivation for the theory developed in this paper is given at the end of section 2.2.

2.1. The Euler equations.

We consider a compressible inviscid fluid occupying a bounded, polyhedral domain D in $\mathbb{R}^d$. Here d is the space dimension. We assume that the dynamics of the system is modeled by the compressible Euler equations equipped with an equation of state that can be either tabulated or given by a very complicated analytic expression. The dependent variable is $\boldsymbol{u} := (\rho, \boldsymbol{m}, E)^\mathsf{T} \in \mathbb{R}^{d+2}$, where ρ is the density, $\boldsymbol{m}$ the momentum, and E the total mechanical energy. In this paper $\boldsymbol{u}$ is considered to be a column vector. The velocity is given by $\boldsymbol{v} := \rho^{-1}\boldsymbol{m}$. The quantity $e(\boldsymbol{u}) := \rho^{-1}E - \frac{1}{2}\|\boldsymbol{v}\|_{\ell^2}^2$ is the specific internal energy. To simplify the notation later we introduce the flux $\mathbb{f}(\boldsymbol{u}) := (\boldsymbol{m}, \boldsymbol{v} \otimes \boldsymbol{m} + p(\boldsymbol{u})\mathbb{I}_d, \boldsymbol{v}(E+p))^\mathsf{T} \in \mathbb{R}^{(d+2)\times d}$, where $\mathbb{I}_d$ is the $d\times d$ identity matrix. The convention adopted in this paper is that for any vectors $\boldsymbol{a}$, $\boldsymbol{b}$, with entries $\{a_k\}_{k\in\{1:d\}}$, $\{b_k\}_{k\in\{1:d\}}$, the following holds: $(\boldsymbol{a}\otimes\boldsymbol{b})_{kl} = a_k b_l$ and $\nabla{\cdot}\boldsymbol{a} = \sum_{k=\in\{1:d\}} \partial_{x_k} a_k$. Moreover, for any second-order tensor $\mathbb{g}$ with entries $\{\mathbb{g}_{kl}\}_{k\in\{1:d+2\}}^{l\in\{1:d\}}$, we define $(\nabla{\cdot}\mathbb{g})_k = \sum_{l\in\{1:d\}} \partial_{x_l}\mathbb{g}_{kl}$.

Given some initial time t_0 and initial data $\boldsymbol{u}_0(\boldsymbol{x}) := (\rho_0, \boldsymbol{m}_0, E_0)(\boldsymbol{x})$, we look for

$\boldsymbol{u}(\boldsymbol{x},t) := (\rho, \boldsymbol{m}, E)(\boldsymbol{x},t)$, solving the following system in some weak sense:

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

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

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

where $p : \mathcal{A} \to \mathbb{R}$ is the pressure, and $\mathcal{A}$ is the admissible set:

$$\mathcal{A} := \big\{\boldsymbol{u} = (\rho, \boldsymbol{m}, E) \in \mathbb{R}^{d+2} \mid \rho > 0,\ e(\boldsymbol{u}) > 0\big\}. \tag{2.2}$$

We refer to the mapping $p : \mathcal{A} \to \mathbb{R}$ as the *oracle*. For all $\beta \geq 0$, we introduce the following convex subset of $\mathcal{A}$:

$$\mathcal{B}(\beta) := \big\{\boldsymbol{u} = (\rho, \boldsymbol{m}, E) \in \mathbb{R}^{d+2} \mid \rho > 0,\ 1 - \beta\rho > 0,\ e(\boldsymbol{u}) > 0\big\}. \tag{2.3}$$

We further assume in this paper that the oracle is such that there exists a number $b \geq 0$, henceforth called the covolume constant, so that the following holds for all $\boldsymbol{u} \in \mathcal{B}(b)$:

$$p(\boldsymbol{u}) > 0. \tag{2.4}$$

The inverse of the covolume constant b is the maximal density the fluid can reach. We take $b = 0$ if this constant is not a priori known.

Our goal in this paper is to approximate (2.1) by adapting the technique described in Guermond and Popov [14]. As explained in the next section, this is done by constructing an artificial viscosity that ensures that some relevant invariant-domain properties can be established, thereby guaranteeing that the approximation technique is robust (i.e., satisfies physical bounds under a reasonable CFL condition). The key difficulty that arises in this endeavor is that constructing solutions to elementary Riemann problems is nearly impossible (or at least highly nontrivial; see, e.g., Quartapelle et al. [31], Fossati and Quartapelle [10]), since the equation of state is either not available or too complicated. We propose a solution to this problem in sections 3 and 4. Taking inspiration from Colella and Glaz [6], Abgrall and Karni [2], and Pantano, Saurel, and Schmitt [28], we introduce a technique consisting of approximating the oracle by a covolume γ-law, where γ solves an additional conservation equation.

Remark 2.1 (pressure). In practice there are many equations of state that cannot guarantee (2.4) over the entire set $\mathcal{B}(b)$, but the algorithm proposed in this paper works properly as long as the numerical states stay in a subset of $\mathcal{B}(b)$ where the pressure stays positive. This situation occurs in many realistic applications. Up to minor adjustments, the positivity assumption on the pressure can be relaxed to mimic equations of state authorizing negative pressures like for stiffened gases. □

2.2. Space and time approximation.

Let us first recall the space and time approximation technique described in [14]. This method is in some sense a discretization-independent extension of the scheme by Lax [23, p. 163]. Without going into the details, we assume that we have at hand a fully discrete scheme where time is approximated by using the forward Euler time stepping, and space is approximated by using some "centered" approximation of (2.1) (i.e., without any artificial viscosity to stabilize the approximation). We denote by t^n the current time, $n \in \mathbb{N}$, and we denote by τ the current time step size; that is, $t^{n+1} := t^n + \tau$. Let us assume that the current approximation is a collection of states $\{\mathbf{U}_i^n\}_{i\in\mathcal{V}}$, where the index set $\mathcal{V}$ is used to enumerate all the degrees of freedom of the approximation. Here $\mathbf{U}_i^n \in \mathbb{R}^{d+2}$ for all $i \in \mathcal{V}$. We assume that the centered update is given by $\mathbf{U}_i^{\mathrm{G},n+1}$ with

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{G},n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} = \mathbf{0}. \tag{2.5}$$

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 $\mathcal{I}(i)$ is called local stencil. This set collects only the degrees of freedom in $\mathcal{V}$ that interact with i (i.e., $j \notin \mathcal{I}(i) \Rightarrow \boldsymbol{c}_{ij} = \mathbf{0}$). We view $\frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}$ as a Galerkin (or centered or inviscid) approximation of $\nabla{\cdot}\mathbb{f}(\boldsymbol{u})$ at time t^n at some grid point (or cell) $i \in \mathcal{V}$. The superindex $^{\mathrm{G}}$ is meant to remind us that (2.5) is a Galerkin (or inviscid or centered) approximation of (2.1). That is, we assume that the consistency error in space in (2.5) scales optimally with respect to the mesh-size for the considered approximation setting. We do not need to be more specific at this point. The only requirement that we make on the coefficients $\boldsymbol{c}_{ij}$ is that the method is conservative; that is, we assume that

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji} \quad \text{and} \quad \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij} = \mathbf{0}. \tag{2.6}$$

An immediate consequence of this assumption is that the the following global conservation property holds: $\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{G},n+1} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^n$. Notice that for every $i \in \mathcal{V}$, the update (2.5) invokes the oracle card$(\mathcal{I}(i))$ times, because computing $\mathbb{f}(\mathbf{U}_j^n)$ requires computing $p(\mathbf{U}_j^n)$ for all $j \in \mathcal{I}(i)$.

Of course, the approximation (2.5) is, in general, not appropriate if the solution to (2.1) is not smooth. To recover some sort of stability (we are going to make a more precise stability statement later in Theorem 4.6), we modify the scheme by adding an artificial graph viscosity based on the stencil $\mathcal{I}(i)$; that is, we compute the stabilized update $\mathbf{U}_i^{n+1}$ by setting the following:

$$\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n) = \mathbf{0}. \tag{2.7}$$

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

$$d_{ij}^n = d_{ji}^n > 0 \quad \text{if} \quad i \neq j. \tag{2.8}$$

The symmetry assumption is essential for the method to be conservative. The question addressed in this paper is the following: How large should d_{ij}^n be for the scheme to preserve invariant sets (and possibly be entropy satisfying for some finite collection of entropies)?

One key observation is that one can rewrite (2.7) as follows:

$$\mathbf{U}_i^{n+1} = \Big(1 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^n}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^n}{m_i}\overline{\mathbf{U}}_{ij}^n, \tag{2.9}$$

with the auxiliary states $\overline{\mathbf{U}}_{ij}^n$ defined as follows:

$$\overline{\mathbf{U}}_{ij}^n := \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\boldsymbol{n}_{ij}\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}. \tag{2.10}$$

Hence, if the time step is small enough, (2.9) shows that $\mathbf{U}_i^{n+1}$ is a convex combination of the following states $(\overline{\mathbf{U}}_{ij}^n)_{j\in\mathcal{I}(i)}$ (with the convention $\overline{\mathbf{U}}_{ii}^n := \mathbf{U}_i^n$). Hence if one can prove that the auxiliary states $\overline{\mathbf{U}}_{ij}^n$ are in the set $\mathcal{B}(b)$ for all $j \in \mathcal{I}(i)$, then the update $\mathbf{U}_i^{n+1}$ is also in $\mathcal{B}(b)$, thereby establishing one important invariant-domain property. (Notice in passing that it is essential here to assume $d_{ij}^n \neq 0$.)

The main objective of this paper is to describe a technique to estimate d_{ij}^n that guarantees that $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}(b)$ provided both states $\mathbf{U}_i^n$ and $\mathbf{U}_j^n$ are in $\mathcal{B}(b)$. This is done by showing that $\overline{\mathbf{U}}_{ij}^n$ is a space average of a solution to a Riemann problem, and by showing that this solution does satisfy the invariant-domain property we are seeking. Then d_{ij}^n is defined so that $d_{ij}^n \geq \lambda_{ij,\max}\|\boldsymbol{c}_{ij}\|_{\ell^2}$, where $\lambda_{ij,\max}$ is any upper bound on the maximum wave speed in the aforementioned Riemann problem.

Remark 2.2 (literature). The interested reader is referred to [14, 16] for realizations of the algorithm (2.5) with continuous finite elements. Realizations of the algorithm with discontinuous elements and with finite volumes are described in [17]. Lumping the mass is essential to arrive at (2.9). It can also be shown for scalar conservation equations that it is impossible to construct an explicit continuous finite element method that is stabilized with artificial viscosity and satisfies the maximum principle if the time derivative is approximated with the consistent mass matrix (see [15, Thm. 4.3]). □

3. The extended Riemann problem.

An important step in [14] toward proving that the auxiliary state $\overline{\mathbf{U}}_{ij}^n$ defined in (2.10) is a "good" state consists of realizing that $\overline{\mathbf{U}}_{ij}^n$ is a space average of the exact solution to the one-dimensional (1D) Riemann problem with flux $\mathbb{f}(\boldsymbol{v})\boldsymbol{n}_{ij}$, left data $\mathbf{U}_i$, and right data $\mathbf{U}_j$ if $\lambda_{ij,\max}$ is an upper bound on the maximum wave speed in the Riemann problem in question. The main difficulty in the present situation is that there is no analytical way to estimate an upper bound $\lambda_{ij,\max}$ since the pressure is given by an oracle. We show in this section how to get around this difficulty.

3.1. Extension of the system and 1D reduction.

To avoid having to refer to particular states $\mathbf{U}_i^n$ and $\mathbf{U}_j^n$, we now assume that we are given left and right admissible states, $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$. We also denote $\boldsymbol{n}_{ij}$ by $\boldsymbol{n}$. Instead of considering the Riemann problem where the pressure is given by the oracle, we now consider an extended Riemann problem. First, we make a change of basis and introduce $\boldsymbol{t}_1,\ldots,\boldsymbol{t}_{d-1}$ so that $\{\boldsymbol{n},\boldsymbol{t}_1,\ldots,\boldsymbol{t}_{d-1}\}$ forms an orthonormal basis of $\mathbb{R}^d$. With this new basis we have $\boldsymbol{m} = (m, \boldsymbol{m}^\perp)^\mathsf{T}$, where $m := \boldsymbol{m}\cdot\boldsymbol{n}$, $\boldsymbol{m}^\perp := \rho(\boldsymbol{v}\cdot\boldsymbol{t}_1,\ldots,\boldsymbol{v}\cdot\boldsymbol{t}_{d-1}) := \rho\boldsymbol{v}^\perp$. We additionally introduce $v := m/\rho = \boldsymbol{v}\cdot\boldsymbol{n}$. Second, we augment the system by introducing a new scalar variable Γ (and $\gamma := \frac{\Gamma}{\rho}$), an augmented state $\widetilde{\boldsymbol{u}} := (\boldsymbol{u},\Gamma)^\mathsf{T}$, and an extended flux as follows:

$$\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}}) := (\boldsymbol{m}, \boldsymbol{v}\otimes\boldsymbol{m} + \widetilde{p}(\widetilde{\boldsymbol{u}})\mathbb{I}_d, \boldsymbol{v}(E+\widetilde{p}(\widetilde{\boldsymbol{u}})), \boldsymbol{v}\Gamma)^\mathsf{T} = (\mathbb{f}(\widetilde{\boldsymbol{u}}), \boldsymbol{v}\Gamma)^\mathsf{T}, \tag{3.1}$$

with the new pressure

$$\widetilde{p}(\widetilde{\boldsymbol{u}}) := \frac{(\Gamma-\rho)e(\boldsymbol{u})}{1-b\rho} = (\gamma-1)\frac{\rho e(\boldsymbol{u})}{1-b\rho}, \tag{3.2}$$

where $e(\boldsymbol{u}) := \frac{1}{\rho}(E - \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho})$. Here b is either given because this parameter can be measured, or b is set to be zero if one does not have any a priori knowledge on the nature of the fluid. Notice that Γ is the last component of the extended variable $\widetilde{\boldsymbol{u}}$;

neither Γ nor $\gamma = \rho^{-1}\Gamma$ are assumed to be constant. The extended Riemann problem consists of seeking $\widetilde{\boldsymbol{u}} := (\boldsymbol{u}, \Gamma)^{\mathsf{T}} = (\rho, \boldsymbol{m}, E, \Gamma)^{\mathsf{T}}$ so that

$$\partial_t \widetilde{\boldsymbol{u}} + \partial_x(\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n}) = \boldsymbol{0}, \quad \widetilde{\boldsymbol{u}} = \begin{pmatrix} \rho \\ m \\ \boldsymbol{m}^{\perp} \\ E \\ \Gamma \end{pmatrix}, \qquad \widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n} = \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + \widetilde{p}(\widetilde{\boldsymbol{u}}) \\ \frac{m}{\rho}\boldsymbol{m}^{\perp} \\ \frac{m}{\rho}(E + \widetilde{p}(\widetilde{\boldsymbol{u}})) \\ \frac{m}{\rho}\Gamma \end{pmatrix}, \tag{3.3}$$

with left data and right data $(\rho_Z, \boldsymbol{m}_Z{\cdot}\boldsymbol{n}, \boldsymbol{m}_Z^{\perp}, E_Z, \Gamma_Z)^{\mathsf{T}}$, where $Z \in \{L, R\}$, and Γ_Z is defined so that $\widetilde{p}(\widetilde{\boldsymbol{u}}_Z) = p(\boldsymbol{u}_Z) =: p_Z$, i.e., $\Gamma_Z := \rho_Z + \frac{p_Z(1-b\rho_Z)}{e_Z}$ (notice that this means $\gamma_Z := 1 + \frac{p_Z(1-b\rho_Z)}{\rho_Z e_Z}$).

As is usually done in the literature, the above problem can be solved in two steps. First, one solves

$$\partial_t \begin{pmatrix} \rho \\ m \\ \mathcal{E} \\ \Gamma \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + \mathsf{p} \\ \frac{m}{\rho}(\mathcal{E} + \mathsf{p}) \\ \frac{m}{\rho}\Gamma \end{pmatrix} = 0, \quad \text{with} \quad \mathsf{p}(\rho, m, \mathcal{E}, \Gamma) := \frac{\gamma - 1}{1 - b\rho}\left(\mathcal{E} - \tfrac{m^2}{2\rho}\right), \tag{3.4}$$

with left data and right data $(\rho_Z, \boldsymbol{m}_Z{\cdot}\boldsymbol{n}, \mathcal{E}_Z, \Gamma_Z)^{\mathsf{T}}$, where $\mathcal{E} := E - \frac{\|\boldsymbol{m}^{\perp}\|_{\ell^2}^2}{2\rho}$. Notice in passing that $E - \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho} = \mathcal{E} - \frac{m^2}{2\rho}$, i.e., the internal energy does not depend on the change of basis. This, together with the definition of γ_Z, implies that $\mathsf{p}_Z := \frac{\gamma_Z - 1}{1 - b\rho_Z}(\mathcal{E}_Z - \frac{m_Z^2}{2\rho_Z}) = \frac{\gamma_Z - 1}{1 - b\rho_Z}(E_Z - \frac{\|\boldsymbol{m}_Z\|_{\ell^2}^2}{2\rho_Z}) = p_Z$. Second, one obtains the full solution to the Riemann problem (3.3) by determining $\boldsymbol{m}^{\perp}$. This field is obtained by solving $\partial_t \boldsymbol{m}^{\perp} + \partial_x(v\boldsymbol{m}^{\perp}) = 0$ with the appropriate left and right data. Just like in the case of the Euler equations, one never solves the second step since it does not affect the maximum wave speed and the structure of the Riemann problem. In the rest of this paper we solely focus our attention on the system (3.4).

Remark 3.1 (invariant-domain properties). At this point, it is important to notice that the following identity holds: $\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}}_Z) = (\mathbb{f}(\boldsymbol{u}_Z), \boldsymbol{v}_Z\Gamma_Z)^{\mathsf{T}}$ because, as already mentioned above, $\widetilde{p}(\widetilde{\boldsymbol{u}}_Z) = p_Z = p(\boldsymbol{u}_Z)$. Then, recalling (2.10) and setting $\lambda := \frac{d_{ij}^n}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}$ and $\overline{\boldsymbol{u}}_{LR} := \overline{\mathbf{U}}_{ij}$, the extended auxiliary state based on the extended flux $\widetilde{\mathbb{f}}$, say $\overline{\widetilde{\boldsymbol{u}}}_{LR}$, satisfies the following identity:

$$\overline{\widetilde{\boldsymbol{u}}}_{LR} = \begin{pmatrix} \overline{\boldsymbol{u}}_{LR} \\ \frac{1}{2}(\Gamma_L + \Gamma_R) - \frac{1}{2\lambda}(\boldsymbol{v}_R\Gamma_R - \boldsymbol{v}_L\Gamma_L){\cdot}\boldsymbol{n} \end{pmatrix}. \tag{3.5}$$

That is, the density, the momentum, and the total energy of the states $\overline{\widetilde{\boldsymbol{u}}}_{LR}$ and $\overline{\boldsymbol{u}}_{LR}$ are identical. This implies that these two states have the same density and the same internal energy. As a result, if one can prove that the density and the internal energy of the state $\overline{\widetilde{\boldsymbol{u}}}_{LR}$ are both positive, then this conclusion automatically carries over to the state $\overline{\boldsymbol{u}}_{LR}$. This remark is essential, and it is the main motivation for introducing the extended Riemann problem. □

3.2. The invariant-domain preserving properties. We will use the technique of Lax, consisting of piecing together elementary waves to construct a weak solution to the extended Riemann problem (3.4). We will show that this weak solution preserves the positivity of the density and the internal energy (see Remark 3.1).

We will also show that the local gamma constant is uniformly bounded from below: $\gamma \geq \min(\gamma_L, \gamma_R)$. The key tool we are going to invoke is the following lemma.

Lemma 3.2 (Riemann average). *Let m be a positive integer. Let $\mathcal{A}$ be a subset of $\mathbb{R}^m$. Let $\boldsymbol{g} \in C^1(\mathcal{A};\mathbb{R}^m)$ be a 1D flux. Let $\boldsymbol{w}_L, \boldsymbol{w}_R \in \mathcal{A}$. Assume that the following Riemann problem has a weak solution $\boldsymbol{w}$ in $L^\infty(\mathbb{R}\times(0,\infty);\mathbb{R}^m) \cap C^0([0,\infty); L^1_{\text{loc}}(\mathbb{R};\mathbb{R}^m))$:*

$$\partial_t \boldsymbol{w} + \partial_x \boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{0}, \quad \boldsymbol{w}(x,0) = \begin{cases} \boldsymbol{w}_L & x < 0, \\ \boldsymbol{w}_R & x > 0. \end{cases} \tag{3.6}$$

Assume that this Riemann solution has a finite maximum wave speed (meaning there exists $\lambda_{\max} > 0$ s.t. $\boldsymbol{w}(x,t) = \boldsymbol{w}_L$ if $x < -\lambda_{\max} t$ and $\boldsymbol{w}(x,t) = \boldsymbol{w}_R$ if $x > \lambda_{\max} t$). Let $\mathcal{B}$ be a convex subset of $\mathcal{A}$ and assume that $\boldsymbol{w}(x,t) \in \mathcal{B}$ for a.e. $x \in \mathbb{R}$ and all $t > 0$. Let $\overline{\boldsymbol{w}} := \int_{-\frac12}^{\frac12} \boldsymbol{w}(x,t)\,\mathrm{d}x$. Then the following hold true for all $t \in (0, \frac{1}{2\lambda_{\max}})$:

(i) $\overline{\boldsymbol{w}}(t) = \frac12(\boldsymbol{w}_L + \boldsymbol{w}_R) - (\boldsymbol{g}(\boldsymbol{w}_R) - \boldsymbol{g}(\boldsymbol{w}_L))t$.

(ii) $\overline{\boldsymbol{w}}(t) \in \mathcal{B}$.

(iii) Let $\Psi \in C^1(\mathcal{B};\mathbb{R})$ be a quasiconcave functional. Assume that $\Psi(\boldsymbol{w}(x,t)) \geq 0$ for a.e. $x \in \mathbb{R}$ and all $t > 0$. Then $\Psi(\overline{\boldsymbol{w}}(t)) \geq 0$.

(iv) Let $\Psi \in C^1(\mathcal{B};\mathbb{R})$ be a concave functional. Assume that $\Psi(\boldsymbol{w}(x,t)) \geq 0$ for a.e. $x \in \mathbb{R}$ and all $t > 0$. Assume that there exists $\lambda_\flat, \lambda_\sharp \in [-\lambda_{\max}, \lambda_{\max}]$, $\lambda_\flat < \lambda_\sharp$, so that $\Psi(\boldsymbol{w}(x,t)) > 0$ for a.e. $\frac{x}{t} \in (\lambda_\flat, \lambda_\sharp)$. Then $\Psi(\overline{\boldsymbol{w}}(t)) > 0$.

Proof. For the entire proof t is a fixed real number in $(0, \frac{1}{2\lambda_{\max}})$.

(i) Let $w_1, \ldots, w_m$ be the m components of $\boldsymbol{w}$, and let $g_1, \ldots, g_m$ be the m components of the flux $\boldsymbol{g}$. Let $l \in \{1{:}m\}$. Since $\boldsymbol{w}$ is a weak solution to (3.6), we have

$$0 = \int_{-\infty}^{\infty}\int_0^{\infty} (-w_l \partial_\tau \phi - g_l(\boldsymbol{w})\partial_x \phi)\,\mathrm{d}\tau\,\mathrm{d}x - w_{l,L}\int_{-\infty}^0 \phi(x,0)\,\mathrm{d}x - w_{l,R}\int_0^\infty \phi(x,0)\,\mathrm{d}x$$

for all $\phi \in W^{1,\infty}(\mathbb{R}\times[0,\infty);\mathbb{R})$ with compact support in $\mathbb{R}\times[0,\infty)$. Here $w_{l,Z}$ is the lth component of $\boldsymbol{w}_Z$. Now we define a sequence of smooth functions $(\phi_\epsilon)_{\epsilon>0}$ with $\phi_\epsilon(x,\tau) = \phi_{1,\epsilon}(|x|)\phi_{2,\epsilon}(\tau)$,

$$\phi_{1,\epsilon}(x) = \begin{cases} 1, & 0 \leq x \leq \frac12, \\ \frac{1}{\epsilon}(-x + \frac12 + \epsilon), & \frac12 \leq x \leq \frac12 + \epsilon, \\ 0, & \frac12 + \epsilon \leq x, \end{cases} \quad \phi_{2,\epsilon}(\tau) = \begin{cases} 1, & 0 \leq \tau \leq t, \\ \frac{1}{\epsilon}(-\tau + t + \epsilon), & t \leq \tau \leq t + \epsilon, \\ 0, & t + \epsilon \leq \tau. \end{cases}$$

Using that $w_l \in C^0([0,\infty); L^1_{\text{loc}}(\mathbb{R}))$, we infer that $\int_{-\infty}^{\infty}\int_0^\infty -w_l \partial_\tau \phi_\epsilon\,\mathrm{d}x\,\mathrm{d}\tau \to \int_{-\frac12}^{\frac12} w_l(x,t)\,\mathrm{d}x$ as $\epsilon \to 0$. Likewise, we have $\int_{-\infty}^{\infty}\int_0^\infty -g_l(\boldsymbol{w})\partial_x \phi_\epsilon\,\mathrm{d}\tau\,\mathrm{d}x \to \int_0^t (g_l(\boldsymbol{w}_R) - g_l(\boldsymbol{w}_L))\,\mathrm{d}\tau = (g_l(\boldsymbol{w}_R) - g_l(\boldsymbol{w}_L))t$ as $\epsilon \to 0$. Finally, $-w_{l,L}\int_{-\infty}^0 \phi_\epsilon(x,0)\,\mathrm{d}x - w_{l,R}\int_0^\infty \phi_\epsilon(x,0)\,\mathrm{d}x \to -\frac12(w_{l,L} + w_{l,R})$ as $\epsilon \to 0$. In conclusion, we have established that

$$0 = \overline{\boldsymbol{w}}(t) + (\boldsymbol{g}(\boldsymbol{w}_R) - \boldsymbol{g}(\boldsymbol{w}_L))t - \tfrac12(\boldsymbol{w}_L + \boldsymbol{w}_R).$$

(ii) Since $\mathcal{B}$ is convex, $\boldsymbol{w}(x,t) \in \mathcal{B}$ for a.e. $x \in \mathbb{R}$ and all $t > 0$, and the length of the interval $[-\frac12, \frac12]$ is 1, we infer that $\overline{\boldsymbol{w}}(t) \in \mathcal{B}$.

(iii) Let $\Psi \in C^1(\mathcal{B};\mathbb{R})$ be a quasiconcave functional. The quasiconcavity implies that $\Psi(\overline{\boldsymbol{w}}(t)) \geq \operatorname{ess\,inf}_{x\in(-\frac12,\frac12)} \Psi(\boldsymbol{w}(x,t)) \geq 0$.

(iv) Let $\Psi \in C^1(\mathcal{B};\mathbb{R})$ be a concave functional. Jensen's inequality implies

$$\Psi(\overline{\boldsymbol{w}}(t)) \geq \int_{-\frac{1}{2}}^{\frac{1}{2}} \Psi(\boldsymbol{w}(x,t))\,\mathrm{d}x \geq \int_{\lambda_\flat t}^{\lambda_\sharp t} \Psi(\boldsymbol{w}(x,t))\,\mathrm{d}x > 0,$$

where we used $-\frac{1}{2} \leq \lambda_\flat t < \lambda_\sharp t \leq \frac{1}{2}$. This concludes the proof. □

Remark 3.3 (weak solution). Notice that Lemma 3.2 only requires us to have access to a weak solution of (3.6) that satisfies an invariant-domain property (i.e., $\boldsymbol{w}(x,t) \in \mathcal{B}$ for a.e. $x \in \mathbb{R}$ and all $t > 0$). No entropy inequality or additional smoothness condition is needed. □

4. Solution of the extended Riemann problem.

We now construct a weak solution to the extended Riemann problem (3.4) using the technique described in Lax [24] (we also refer to Holden and Risebro [18, Chap. 5], Godlewski and Raviart [11, Chap. 1], and Toro [34, Chap. 4] for further details on the Riemann problem). No originality is claimed in the construction of each elementary wave, but the full construction of the Riemann solution involving two different covolume equations of state is original.

4.1. Definition of the star states.

We first notice that the Jacobian matrix of (3.4) is diagonalizable and has three distinct eigenvalues. The eigenvalue $\frac{m}{\rho}$ has multiplicity 2. Then, as usual, we postulate that the solution to (3.4) is self-similar and composed of three waves, hereafter called the L-wave, C-wave, and R-wave. The L-wave and the R-wave are either shocks or expansions. The L-wave is generated using the covolume equation of state with γ_L, and the R-wave is generated by using the covolume equation of state with γ_R. The C-wave is a contact discontinuity for the density and Γ. Compared to the technique described in Toro [34, Chap. 4], the only new feature here is that the dependent variable has a fourth component Γ. The purpose of this section is to introduce quantities that are useful to define the three waves in question: the intermediate densities ρ_L^*, ρ_R^*, the intermediate velocities v_L^*, v_R^*, v^*, and the intermediate pressure p^*. The actual construction of the solution is done in sections 4.2 and 4.3.

In the rest of this section we use the primitive variables: density ρ, velocity v, pressure p, and $\gamma := \Gamma/\rho$. We use the symbol $\mathfrak{p}$ to denote the pressure defined in (3.4). Notice that the oracle is only invoked to compute the two states p_L and p_R. We define the primitive state $\boldsymbol{c} := (\rho, v, \mathfrak{p}, \gamma)^\mathsf{T}$ and set $\boldsymbol{c}_Z := (\rho_Z, v_Z, p_Z, \gamma_L)^\mathsf{T}$. Recalling that we have defined $\gamma_Z := 1 + \frac{p_Z(1-b\rho_Z)}{\rho_Z e_Z}$, the oracle assumption (2.4) implies that $\min(\gamma_L, \gamma_R) > 1$.

We define the covolume sound speed $a_Z := \sqrt{\frac{\gamma_Z p_Z}{\rho_Z(1-b\rho_Z)}}$, the parameters $A_Z := \frac{2(1-b\rho_Z)}{(\gamma_Z+1)\rho_Z}$ and $B_Z := \frac{\gamma_Z-1}{\gamma_Z+1}p_Z$ corresponding to the Z state (see, e.g., Toro [34, sect. 4.7], [13]), and introduce the following function:

$$f_Z(p) := \begin{cases} f_Z^R(p) := \frac{2a_Z(1-b\rho_Z)}{\gamma_Z-1}\left(\left(\frac{p}{p_Z}\right)^{\frac{\gamma_Z-1}{2\gamma_Z}} - 1\right) & \text{if } 0 \leq p < p_Z, \\ f_Z^S(p) := (p-p_Z)\left(\frac{A_Z}{p+B_Z}\right)^{\frac{1}{2}} & \text{if } p_Z \leq p. \end{cases} \tag{4.1}$$

The definition of $f_Z(p)$ makes sense because $1 < \gamma_Z$ and $0 \leq B_Z$. It is shown in Toro [34, sect. 4.3.1] that the function $f_Z(p)$ is in $C^2(\mathbb{R}_+;\mathbb{R})$, monotone increasing, and concave.

We also define the function $\phi \in C^2(\mathbb{R}_+;\mathbb{R})$,

$$\phi(p) := f_L(p) + f_R(p) + v_R - v_L, \qquad p \in [0,\infty). \tag{4.2}$$

Notice in passing that assuming $\phi(0) < 0$ is equivalent to assuming that the following holds true:

$$v_R - v_L < \frac{2a_L(1-b\rho_L)}{\gamma_L - 1} + \frac{2a_R(1-b\rho_R)}{\gamma_R - 1}. \tag{4.3}$$

This condition is known in the literature as the nonvacuum condition (see Toro [34, (4.40), p. 127]).

LEMMA 4.1. *If* (4.3) *holds, then ϕ has a unique positive root p^*.*

Proof. Since $\phi(0) = v_R - v_L - \frac{2a_L(1-b\rho_L)}{\gamma_L-1} - \frac{2a_R(1-b\rho_R)}{\gamma_R-1}$, the assumption (4.3) means that $\phi(0) < 0$. We then conclude that ϕ has a unique positive root since $\phi(p) \in C^2(\mathbb{R}_+;\mathbb{R})$ is strictly monotone increasing (and concave). ◻

DEFINITION 4.2 (p^*, ρ_L^*, ρ_R^*, v_L^*, v_R^*, v^*). (i) *If the nonvacuum condition* (4.3) *holds, we denote by p^* the unique root of ϕ, and we set $v_L^* := v_L - f_L(p^*)$, $v_R^* := v_R + f_R(p^*)$, $v^* := v_L^* = v_R^*$.*

(ii) *If instead there is vacuum, we define $p^* := 0$ and set $v_L^* := v_L - f_L(0)$, $v_R^* := v_R + f_R(0)$.*

(iii) *If $p^* \neq 0$, we set $\rho_Z^* := \big(b + \frac{1-b\rho_Z}{\rho_Z}(\frac{p_Z}{p^*})^{\frac{1}{\gamma_Z}}\big)^{-1}$, $Z \in \{L,R\}$, and we extend this definition by continuity by setting $\rho_L^* := \rho_R^* := 0$ if $p^* = 0$.*

Notice that the definition of v^* makes sense if the nonvacuum condition (4.3) holds since in this case $\phi(p^*) = 0 = v_R^* - v_L^*$. To fully describe our weak solution, we introduce the following wave speeds:

$$\lambda_L^-(p^*) := v_L - a_L\left(1 + \frac{\gamma_L+1}{2\gamma_L}\left(\frac{p^*-p_L}{p_L}\right)_+\right)^{\frac{1}{2}},$$

$$\lambda_L^+(p^*) := \begin{cases} v_L - f_L(p^*) - a_L\frac{1-b\rho_L}{1-b\rho_L^*}\left(\frac{p^*}{p_L}\right)^{\frac{\gamma_L-1}{2\gamma_L}} & \text{if } p^* < p_L, \\ \lambda_L^-(p^*) & \text{if } p_L \leq p^*, \end{cases}$$

$$\lambda_R^+(p^*) := v_R + a_R\left(1 + \frac{\gamma_R+1}{2\gamma_R}\left(\frac{p^*-p_R}{p_R}\right)_+\right)^{\frac{1}{2}},$$

$$\lambda_R^-(p^*) := \begin{cases} v_R + f_R(p^*) + a_R\frac{1-b\rho_R}{1-b\rho_R^*}\left(\frac{p^*}{p_R}\right)^{\frac{\gamma_R-1}{2\gamma_R}} & \text{if } p^* < p_R, \\ \lambda_R^+(p^*) & \text{if } p_R \leq p^*. \end{cases}$$

LEMMA 4.3 (wave speeds). *Assume $1 < \min(\gamma_L,\gamma_R)$ and $0 < a_L, a_R$. Then, the following holds true:*

$$\lambda_L^-(p^*) \leq \lambda_L^+(p^*) \leq v_L^* \leq v_R^* \leq \lambda_R^-(p^*) \leq \lambda_R^+(p^*). \tag{4.4}$$

Proof. We will only consider the case $Z = L$; the case $Z = R$ is analogous. There are two possibilities: either $p^* < p_L$ or $p_L \leq p^*$. In the first case, $p^* < p_L$, we have

$$\lambda_L^-(p^*) = v_L - a_L < \lambda_L^+(p^*) = v_L - f_L(p^*) - a_L\frac{1-b\rho_L}{1-b\rho_L^*}\left(\frac{p^*}{p_L}\right)^{\frac{\gamma_L-1}{2\gamma_L}} \leq v_L - f_L(p^*) = v_L^*,$$

where we used above that $f_L(p^*) < 0$, $1 < \gamma_L$, $0 < a_L$, $\rho_L^* < \rho_L$, $0 \le p^* < p_L$, and $0 < \frac{1-b\rho_L}{1-b\rho_L^*} \le 1$. In the second case, $p_L \le p^*$, we have

$$\lambda_L^+(p^*) = \lambda_L^-(p^*) = v_L - a_L \left(1 + \frac{\gamma_L+1}{2\gamma_L}\left(\frac{p^*-p_L}{p_L}\right)\right)^{\frac{1}{2}}$$

and

$$v_L^* = v_L - f_L(p^*) = v_L - (p^*-p_L)\left(\frac{A_L}{p^*+B_L}\right)^{\frac{1}{2}}.$$

Then proving the inequality $\lambda_L^+(p^*) < v_L^*$ is equivalent to showing that

$$\left(\frac{p^*}{p_L}-1\right)\left(\frac{2(1-b\rho_L)}{\gamma_L(\gamma_L+1)}\frac{\gamma_L p_L}{\rho_L}\frac{1}{\frac{p^*}{p_L}+\frac{\gamma_L-1}{\gamma_L+1}}\right)^{\frac{1}{2}} < a_L\left(1+\frac{\gamma_L+1}{2\gamma_L}\left(\frac{p^*-p_L}{p_L}\right)\right)^{\frac{1}{2}}.$$

Using the substitution $x := \frac{p^*}{p_L} - 1$ and that $a_L = \sqrt{\frac{\gamma_L p_L}{\rho_L(1-b\rho_L)}}$, we derive that the above inequality is equivalent to proving that

$$\left(\frac{2}{\gamma_L(\gamma_L+1)}\right)^{\frac{1}{2}} x(1-b\rho_L) < \left(\left(x+\frac{2\gamma_L}{\gamma_L+1}\right)\left(\frac{\gamma_L+1}{2\gamma_L}x+1\right)\right)^{\frac{1}{2}}$$

for all $x > 0$. Squaring both sides and recalling that $x > 0$, we observe that the above is equivalent to the inequality

$$0 < \left(\frac{\gamma_L+1}{2\gamma_L} - \frac{2(1-b\rho_L)^2}{\gamma_L(\gamma_L+1)}\right)x^2 + 2x + \frac{2\gamma_L}{\gamma_L+1}.$$

This inequality holds true for all $x \ge 0$ since we assumed that $1 < \gamma_L$ and $0 \le 1 - b\rho_L \le 1$. ∎

4.2. Definition of the L-wave and R-wave without vacuum. We assume in this section that the nonvacuum condition (4.3) holds. The solution with vacuum is given in section 4.3. The main result of this section is Lemma 4.4.

Recalling the notation from Definition 4.2, the proposed solution to (3.4) is self-similar and has the following form:

$$\boldsymbol{c}(x,t) := \begin{cases} \boldsymbol{c}_L & \text{if } \frac{x}{t} < \lambda_L^-, \\ \boldsymbol{c}_{LL}(\frac{x}{t}) & \text{if } \lambda_L^- \le \frac{x}{t} < \lambda_L^+, \\ \boldsymbol{c}_L^* & \text{if } \lambda_L^+ \le \frac{x}{t} < v*, \\ \boldsymbol{c}_R^* & \text{if } v* \le \frac{x}{t} < \lambda_R^-, \\ \boldsymbol{c}_{RR}(\frac{x}{t}) & \text{if } \lambda_R^- \le \frac{x}{t} < \lambda_R^+, \\ \boldsymbol{c}_R & \text{if } \lambda_R^+ \le \frac{x}{t}, \end{cases} \tag{4.5}$$

with $\boldsymbol{c}_L^* := (\rho_L^*, v^*, p^*, \gamma_L)^\mathsf{T}$ and $\boldsymbol{c}_R^* := (\rho_R^*, v^*, p^*, \gamma_R)^\mathsf{T}$. The parameters p^*, v^*, ρ_L^*, and ρ_R^* are defined in Definition 4.2. The two functions $\boldsymbol{c}_{LL}$, $\boldsymbol{c}_{RR}$ are going to be defined to make sure that (4.5) is indeed a weak solution to (3.4). Notice that $\boldsymbol{c}$ is uniquely defined owing to Lemma 4.3 (i.e., the waves are well ordered).

Let us first construct the L-wave, i.e., we construct the function $\boldsymbol{c}_{LL}(\xi)$ where $\lambda_L^- \le \xi < \lambda_L^+$ and $\xi := \frac{x}{t}$. If $p_L \le p^*$, then $\lambda_L^-(p^*) = \lambda_L^+(p^*)$, and the L-wave is a shock. In this case one does not need to define $\boldsymbol{c}_{LL}$ since the interval $[\lambda_L^-, \lambda_L^+)$ is empty. If $p^* < p_L$, we postulate that the γ-component of $\boldsymbol{c}_{LL}$ is constant and equal to γ_L. This means that the L-wave can be computed by assuming that the equation of state is a standard covolume γ-law $p(1-b\rho) = (\gamma_L-1)\rho e$ (with $e = \frac{1}{\rho}(\mathcal{E} - \frac{m^2}{2\rho})$). In this case the L-wave is an expansion. The construction of this wave is well established; we refer, for instance, to Toro [34, Chap. 4]. More precisely, the self-similarity parameter $\xi = \frac{x}{t}$ (which is the eigenvalue of the Jacobian of the flux, $v - a$) can be expressed in terms of the parameter p:

$$\xi_L(p) := v_L - f_L(p) - a_L \frac{1-b\rho_L}{1-b\rho(p)} \left(\frac{p}{p_L}\right)^{\frac{\gamma_L-1}{2\gamma_L}}, \qquad p \in [p^*, p_L], \tag{4.6}$$

where $\rho(p)$ is defined as follows:

$$\frac{1}{\rho(p)} - b := \left(\frac{1}{\rho_L} - b\right)\left(\frac{p_L}{p}\right)^{\frac{1}{\gamma_L}}.$$

To simplify the notation we use the symbol $\xi(p)$ instead of $\xi_L(p)$ when the context is unambiguous. Notice that $\lambda_L^-(p^*) = \xi(p_L)$ and $\lambda_L^+(p^*) = \xi(p^*)$. Since the function ξ is strictly deceasing in the interval $p \in [p^*, p_L]$, the inverse function theorem implies that p can be uniquely expressed in terms of ξ. We abuse the notation and denote by $p(\xi)$ the inverse function. Over the interval $\xi \in [\xi(p_L), \xi(p^*)] = [\lambda_L^-(p^*), \lambda_L^+(p^*)]$, we have (see Toro [34, sect. 4.7.1])

$$\boldsymbol{c}_{LL}(\xi) := \left(\rho_L\left(b\rho_L + (1-b\rho_L)\left(\frac{p_L}{p(\xi)}\right)^{\frac{1}{\gamma_L}}\right)^{-1}, v_L - f_L(p(\xi)), p(\xi), \gamma_L\right)^{\mathsf{T}}. \tag{4.7}$$

Now we define $\boldsymbol{c}_L^*$. If $p^* < p_L$, the L-wave is an expansion, and $\boldsymbol{c}_L^*$ is defined to be the end point of the L-wave: $\boldsymbol{c}_L^* := \boldsymbol{c}_{LL}(\xi(p^*))$. If $p_L \le p^*$, the L-wave is a shock. We still postulate that the γ-component of $\boldsymbol{c}$ is equal to γ_L for $\frac{x}{t} \le \lambda_L^+(p^*)$. In this case we define $\boldsymbol{c}_L^*$ so that the Rankine–Hugoniot relation holds between the two states $\boldsymbol{c}_L$ and $\boldsymbol{c}_L^*$ (see Toro [34, sect. 4.7.1]). In conclusion, we have

$$\boldsymbol{c}_L^* := \begin{cases} \boldsymbol{c}_{LL}(\xi(p^*)) & \text{if } p^* < p_L, \\ \left(\frac{\rho_L\left(\frac{p^*}{p_L} + \frac{\gamma_L-1}{\gamma_L+1}\right)}{\frac{\gamma_L-1+2b\rho_L}{\gamma_L+1}\frac{p^*}{p_L} + \frac{\gamma_L+1-2b\rho_L}{\gamma_L+1}}, v_L - f_L(p^*), p^*, \gamma_L\right)^{\mathsf{T}} & \text{if } p_L \le p^*. \end{cases} \tag{4.8}$$

We define $\boldsymbol{c}_{RR}(\xi)$ similarly. If $p^* < p_R$, the R-wave is an expansion, otherwise it is a shock. Assuming that $p^* < p_R$, the self-similarity parameter $\xi = \frac{x}{t}$ can be expressed in terms of the parameter $p \in [p^*, p_R]$:

$$\xi_R(p) := v_R + f_R(p) + a_R \frac{1-b\rho_R}{1-b\rho(p)} \left(\frac{p}{p_R}\right)^{\frac{\gamma_R-1}{2\gamma_R}}, \tag{4.9}$$

where we have defined

$$\frac{1}{\rho(p)} - b := \left(\frac{1}{\rho_R} - b\right)\left(\frac{p_R}{p}\right)^{\frac{1}{\gamma_R}}.$$

To simplify the notation, we use the symbol $\xi(p)$ instead of $\xi_R(p)$ when the context is unambiguous. Notice that in this case $\lambda_R^- = \xi(p^*)$, $\lambda_R^+ = \xi(p_R)$, and ξ is a strictly increasing function over the interval $[p^*, p_R]$. Over the interval $\xi \in [\xi(p^*), \xi(p_R)]$, we have

$$\boldsymbol{c}_{RR}(\xi) := \left(\rho_R\Big(b\rho_R + (1-b\rho_R)\Big(\frac{p_R}{p(\xi)}\Big)^{\frac{1}{\gamma_R}}\Big)^{-1}, v_R + f_R(p(\xi)), p(\xi), \gamma_R\right)^{\mathsf{T}}. \tag{4.10}$$

Now we define $\boldsymbol{c}^*_R$. If $p^* < p_R$, the R-wave is an expansion, and $\boldsymbol{c}^*_R$ is defined to be the end point of the wave: $\boldsymbol{c}^*_R = \boldsymbol{c}_{RR}(\xi(p^*))$. If $p_R \leq p^*$, the R-wave is a shock. We still postulate that the γ-component of $\boldsymbol{c}$ is equal to γ_R for $v^* \leq \frac{x}{t} < \lambda^+_R$. In this case we define $\boldsymbol{c}^*_R$ so that the Rankine–Hugoniot relation holds between the two states $\boldsymbol{c}_R$ and $\boldsymbol{c}^*_R$. In conclusion, we have

$$\boldsymbol{c}^*_R = \begin{cases} \boldsymbol{c}_{RR}(\xi(p^*)) & \text{if } p^* < p_R, \\ \left(\dfrac{\rho_R\left(\frac{p^*}{p_R} + \frac{\gamma_R - 1}{\gamma_R + 1}\right)}{\frac{\gamma_R - 1 + 2b\rho_R}{\gamma_R+1}\frac{p^*}{p_R} + \frac{\gamma_R + 1 - 2b\rho_R}{\gamma_R + 1}}, v_R + f_R(p^*), p^*, \gamma_R\right)^{\mathsf{T}} & \text{if } p_R \leq p^*. \end{cases} \tag{4.11}$$

The key result of this section is summarized in the following lemma.

Lemma 4.4. *Assume that the nonvacuum condition* (4.3) *holds.* (i) *The field* $(\rho, \boldsymbol{m}, E, \Gamma)^{\mathsf{T}}$ *defined by* (4.5) *is a weak solution to* (3.4). (ii) *This field takes values in* $\mathcal{B}(b)$.

Proof. (i) In the domain $\{x < v^* t\}$, we have $\gamma = \gamma_L$; hence, $\Gamma = \gamma_L \rho$. This implies that the last equation in (3.4) is equivalent to the first equation (the conservation of mass). Moreover, the first three equations in (3.4) hold true in the weak sense since the field $(\rho, m, \mathcal{E})$ defined in (4.5) is by construction a weak solution to the regular Euler equations with the pressure law $p(1 - b\rho) := (\gamma_L - 1)\big(\mathcal{E} - \frac{m^2}{2\rho}\big)$.

Similarly, in the domain $\{x > v^* t\}$, we have $\gamma = \gamma_R$; hence, $\Gamma = \gamma_R \rho$, and the last equation in (3.4) is equivalent to the the conservation of mass equation. The first three equations in (3.4) hold true in the weak sense because the field $(\rho, m, \mathcal{E})$ defined in (4.5) is by construction a weak solution to the regular Euler equations with a pressure law $p(1 - b\rho) := (\gamma_R - 1)\big(\mathcal{E} - \frac{m^2}{2\rho}\big)$.

To be able to conclude the proof, we now have to make sure that the two states that are separated by the line $\{x = v^* t\}$ satisfy the Rankine–Hugoniot relation. Let $\boldsymbol{c}^*_L = (\rho^*_L, v^*_L, p^*_L, \gamma^*_L)$ and $\boldsymbol{c}^*_R = (\rho^*_L, v^*_L, p^*_L, \gamma^*_L)$ be the two constant states defined above. Recall that the construction of $\boldsymbol{c}^*_L$ and $\boldsymbol{c}^*_R$ is such that that $p^*_L = p^*_R = p^*$ (see (4.8) and (4.11)). We have to show that

$$\begin{aligned} \rho^*_L v^*_L - \rho^*_R v^*_R &= v^*(\rho^*_L - \rho^*_R), \\ \rho^*_L (v^*_L)^2 + p_L - \rho^*_R (v^*_R)^2 - p_R &= v^*(\rho^*_L v^*_L - \rho^*_R v^*_R), \\ v^*_L(E^*_L - p^*_L) - v^*_R(E^*_R - p^*_R) &= v^*(E^*_L - E^*_R), \\ v^*_L \gamma_L - v^*_R \gamma_R &= v^*(\gamma_L - \gamma_R). \end{aligned}$$

Since the nonvacuum condition (4.3) holds, we have $v^* := v^*_L = v^*_R$ (see Definition 4.2). Thus it follows that the above four equations indeed hold true. Therefore, the field defined in (4.5) is a weak solution to (3.4).

(ii) By construction the waves defined above do not contain any vacuum state and the density and the internal energy are both positive. Moreover, by applying Proposition A.1 in Guermond and Popov [13] we infer that the density is bounded from above by $\frac{1}{b}$. As a result, the solution takes values in $\mathcal{B}(b)$. ∎

4.3. Definition of the L-wave and R-wave when a vacuum is present. When (4.3) fails, the solution contains a vacuum state. In this case both the L-wave and the R-wave are expansions. Recall that in Definition 4.2 we have set

$$p^* := 0,\ v_L^* := v_L - f_L(0) = v_L + \frac{2a_L(1-b\rho_L)}{\gamma_L - 1},\ v_R^* := v_R + f_R(0) = v_R - \frac{2a_R(1-b\rho_R)}{\gamma_R - 1}. \tag{4.12}$$

The solution to the extended Riemann problem (3.4) we propose is as follows:

$$\boldsymbol{c}(x,t) = \begin{cases} \boldsymbol{c}_L & \text{if } \frac{x}{t} < v_L - a_L, \\ \boldsymbol{c}_{LL}(\frac{x}{t}) & \text{if } \lambda_L^- \le \frac{x}{t} < v_L^*, \\ \frac{v_R^* - \frac{x}{t}}{v_R^* - v_L^*}\boldsymbol{c}_L^* + \frac{\frac{x}{t} - v_L^*}{v_R^* - v_L^*}\boldsymbol{c}_R^* & \text{if } v_L^* \le \frac{x}{t} < v_R^*, \\ \boldsymbol{c}_{RR}(\frac{x}{t}) & \text{if } v_R^* \le \frac{x}{t} < v_R + a_R, \\ \boldsymbol{c}_R & \text{if } v_R + a_R \le \frac{x}{t}. \end{cases} \tag{4.13}$$

The definitions of the expansion waves $\boldsymbol{c}_{LL}$ and $\boldsymbol{c}_{RR}$ are the same as in the nonvacuum case. We define the states $\boldsymbol{c}_L^*$ and $\boldsymbol{c}_R^*$ as in section 4.2 by setting $\boldsymbol{c}_L^* := \boldsymbol{c}_{LL}(v_L^*) = (0, v_L^*, 0, \gamma_L)^\mathsf{T}$ and $\boldsymbol{c}_R^* := \boldsymbol{c}_{RR}(v_R^*) = (0, v_R^*, 0, \gamma_R)^\mathsf{T}$. The key result of this section is the following lemma.

LEMMA 4.5. *Assume that the vacuum condition holds, i.e., $p^* = 0$.* (i) *The field $(\rho, \boldsymbol{m}, E, \Gamma)^\mathsf{T}$ defined by* (4.13) *is a weak solution to* (3.4). (ii) *This field takes values in $\mathcal{B}(b)$ for $\frac{x}{t} \in (v_L - a_L, v_L^*) \cup (v_R^*, v_R + a_R)$.* (iii) *The field takes values in $\overline{\mathcal{B}}(b)$.*

Proof. (i) We have already established that once expressed as a conserved variable, (4.13) is a weak solution to (3.4) in the regions $\{x < v_L^* t\} \cup \{v_R^* t < x\}$. In the region $\{v_L^* t < x < v_R^* t\}$, all the conserved variables are zero by construction. Hence, (4.13) rewritten as conserved variables is also weak solution to (3.4) in the region $\{v_L^* t < x < v_R^* t\}$. Let us now verify that the field defined in (4.13) is continuous across the line $\{x = v_L^* t\}$. Denoting $\xi_L(p)$ the function defined in (4.6), we obtain $\xi_L(0) = v_L - f_L(0) =: v_L^*$, i.e., $p(v_L^*) = 0$. Hence $\lim_{\xi \uparrow v_L^*} \boldsymbol{c}_{LL}(\xi) = (0, v_L^*, 0, \gamma_L)$. Moreover, $\lim_{\xi \downarrow v_L^*} \frac{v_R^* - \xi}{v_R^* - v_L^*}\boldsymbol{c}_L^* + \frac{\xi - v_L^*}{v_R^* - v_L^*}\boldsymbol{c}_R^* = (0, v_L^*, 0, \gamma_L)$. This proves the assertion. This in turn establishes that the conserved field is also continuous across $\{x = v_L^* t\}$. The argument to prove continuity across $\{x = v_R^* t\}$ is similar. The conclusion follows readily.

(ii) By proceeding as in the proof of Lemma 4.4, one verifies that the solution takes values in $\mathcal{B}(b)$.

(iii) The specific internal energy decreases along the expansion wave and reaches zero when $x/t = v_Z^*$. By convention we define the specific internal energy in a vacuum to be zero. This is consistent with the definition (4.13). Hence, the solution takes values in $\{\boldsymbol{u} = (\rho, \boldsymbol{m}, E) \in \mathbb{R}^{d+2} \mid \rho \ge 0,\ 1 - \beta\rho > 0,\ e(\boldsymbol{u}) \ge 0\} \subset \overline{\mathcal{B}(b)}$. □

4.4. Summary. In sections 4.2 and 4.3 we have defined a weak solution to the extended Riemann problem (3.4). Notice that this weak solution satisfies the assumption of Lemma 3.2, i.e., it is in $L^\infty(\mathbb{R}\times(0,\infty);\mathbb{R}^m) \cap C^0([0,\infty); L^1_{\text{loc}}(\mathbb{R};\mathbb{R}^m))$ with $m = d+2$, and the maximum wave speed $\lambda_{\max} = \max(|\lambda_L^-(p^*)|, |\lambda_R^+(p^*)|) = \max(-\lambda_L^-(p^*), \lambda_R^+(p^*))$ is finite. As a result, we can invoke Lemma 3.2 for any quasiconcave functional. The following theorem is the main result of section 4.

THEOREM 4.6. (i) *Let $\mathbf{U}_i^n$, $\mathbf{U}_j^n$ be two states in $\mathcal{B}(b)$ (with $\mathcal{B}(b)$ defined in* (2.3)*). Let p^* be defined as in Definition* 4.2 *with left state $\mathbf{U}_i^n$ and right state $\mathbf{U}_j^n$. Let $\widehat{p}^*$ be* *any upper bound on p^* (i.e., $\widehat{p}^* \geq p^*$). Let*

$$\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) := \max(-\lambda_L^-(\widehat{p}^*), \lambda_R^+(\widehat{p}^*)), \tag{4.14a}$$

$$d_{ij}^n := \max(\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \widehat{\lambda}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}). \tag{4.14b}$$

Let $\overline{\mathbf{U}}_{ij}^n$ be defined by (2.10). *Then $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}(b)$.*

(ii) *Let $i \in \mathcal{V}$. Assume that $\mathbf{U}_j^n \in \mathcal{B}(b)$ for all $j \in \mathcal{I}(i)$. Assume that d_{ij}^n is defined as above in* (4.14b) *for all $j \in \mathcal{I}(i)$. Assume that τ is small enough so that $\tau \sum_{j \in \mathcal{I}(i)\setminus\{i\}} \frac{2d_{ij}^n}{m_i} \leq 1$. Let $\mathbf{U}_i^{n+1}$ be the update defined in* (2.7). *Then $\mathbf{U}_i^{n+1} \in$* $\mathrm{Conv}\{\overline{\mathbf{U}}_{ij}^n \mid j \in \mathcal{I}(i)\} \subset \mathcal{B}(b)$.

Proof. (i) We first notice that $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \geq \max(-\lambda_L^-(p^*), \lambda_R^+(p^*))$ since the functions $-\lambda_L^-$ and λ_R^+ are monotone increasing and $\widehat{p}^* \geq p^*$. We now apply Lemma 3.2 with the flux $\boldsymbol{g}(\widetilde{\boldsymbol{w}}) = \widetilde{\mathbb{f}}(\widetilde{\boldsymbol{w}})\boldsymbol{n}$ and the Riemann data $\widetilde{\mathbf{U}}_i^n$, $\widetilde{\mathbf{U}}_j^n$. We observe that the Riemann solution defined in (4.5) and (4.13) has nonnegative density and nonnegative internal energy (recall that the internal energy ρe is equal to $\frac{1}{(\gamma-1)}(1-b\rho)p$). Notice also that the only way to have zero density and zero internal energy on a set of nonzero measure is when a vacuum is present in the solution and $v_L^* < v_R^*$; in this case, $\lambda_L^- < \lambda_L^+$ and $\lambda_R^- < \lambda_R^+$, and the density and the internal energy are positive in the regions $\frac{x}{t} \in [\lambda_L^-, \lambda_L^+)$, $\frac{x}{t} \in (\lambda_R^-, \lambda_R^+]$. Consider the concave functionals $\widetilde{\Psi}_1 : \widetilde{\boldsymbol{u}} \mapsto \rho$, $\widetilde{\Psi}_2 : \widetilde{\boldsymbol{u}} \mapsto 1 - b\rho$, and $\widetilde{\Psi}_3 : \widetilde{\boldsymbol{u}} \mapsto \rho e$. Notice that $\widetilde{\Psi}_l(\widetilde{\mathbf{U}}_j^n) > 0$ for all $j \in \mathcal{I}(i)$ and all $l \in \{1{:}3\}$, whether or not a vacuum occurs, because we assume that $\mathbf{U}_i^n, \mathbf{U}_j^n \in \mathcal{B}(b)$. We conclude that $\widetilde{\Psi}_l(\overline{\widetilde{\mathbf{U}}}_{ij}^n) > 0$ for all $l \in \{1{:}3\}$ by invoking item (3.2) in Lemma 3.2. But the identity (3.5) shows that the density and the internal energy of the states $\overline{\widetilde{\mathbf{U}}}_{ij}^n$ and $\overline{\mathbf{U}}_{ij}^n$ are identical; as a result, defining $\Psi_1 : \boldsymbol{u} \mapsto \rho$, $\Psi_2 : \boldsymbol{u} \mapsto 1 - b\rho$, and $\Psi_3 : \boldsymbol{u} \mapsto \rho e$, we infer that $\Psi_l(\overline{\mathbf{U}}_{ij}^n) = \widetilde{\Psi}_l(\overline{\widetilde{\mathbf{U}}}_{ij}^n) > 0$ for all $l \in \{1{:}3\}$. This establishes that $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}(b)$.

(ii) The assertion follows from (i), the convexity of $\mathcal{B}(b)$, and the observation that (2.9) implies that $\mathbf{U}_i^{n+1}$ is in the convex hull of $\{\overline{\mathbf{U}}_{ij}^n \mid j \in \mathcal{I}(i)\}$ if $\tau \sum_{j \in \mathcal{I}(i)\setminus\{i\}} \frac{2d_{ij}^n}{m_i} \leq 1$. This completes the proof. □

Theorem 4.6 says that the algorithm (2.7) is invariant-domain preserving under the appropriate CFL condition. To make this theorem useful, we now need to derive a computable upper bound on the maximum wave speed in the extended Riemann problem (3.4). This task is achieved in section 5.

5. Upper bound on the maximum wave speed.

Setting $\lambda_{\max}(p) := \max(-\lambda_L^-(p), \lambda_R^+(p))$, we recall that the maximum wave speed in the Riemann problem (3.4) is given by $\lambda_{\max}(p^*)$. Recall also that $p \mapsto \lambda_{\max}(p)$ is a nondecreasing function. Since we only need an upper bound on $\lambda_{\max}(p^*)$, we derive in this section an explicit upper bound on p^*.

5.1. Motivation and notation.

We recall that $p^* = 0$ if vacuum is present, and the maximum speed of propagation is then $\lambda_{\max}(0) = \max(|v_L - a_L|, |v_R + a_R|)$. The L-wave and the R-wave are both expansions in this case.

If the nonvacuum condition holds (see (4.3)), p^* solves

$$\phi(p) = f_L(p) + f_R(p) + v_R - v_L = 0, \qquad p \in (0, \infty).$$

As proved in Guermond and Popov [13, Lem. 4.2], a simple upper bound for p^* can be obtained by using the so-called double-rarefaction approximation (see also Pike [29]), which consists of finding the unique root of the modified equation $\phi_{RR}(p) = 0$, where

$$\phi_{RR}(p) := \frac{2a_L(1-b\rho_L)}{\gamma_L - 1}\left(\left(\frac{p}{p_L}\right)^{\frac{\gamma_L-1}{2\gamma_L}} - 1\right) + \frac{2a_L(1-b\rho_R)}{\gamma_R - 1}\left(\left(\frac{p}{p_R}\right)^{\frac{\gamma_R-1}{2\gamma_R}} - 1\right) + v_R - v_L. \tag{5.1}$$

It can be shown that $\phi_{RR}(p) \le \phi(p)$ for all $p \in [\min(p_L, p_R), \infty)$ if $\max(\gamma_L, \gamma_R) \in (1, \frac{5}{3}]$. Using the notation from (4.1), this result is proved in [13, Lem. 4.2] by showing that $f_Z^S(p) \ge f_Z^R(p)$ for all $p > p_Z$ if $\gamma_Z \in (1, \frac{5}{3}]$. We revisit this idea in the rest of section 5 and remove the assumption $\max(\gamma_L, \gamma_R) \in (1, \frac{5}{3}]$. More precisely, we use a result from Theorem A.2, proved in Appendix A: there exists a function $c(\gamma_Z)$ (defined in (A.3)) so that $f_Z^S(p) \ge c(\gamma_Z) f_Z^R(p)$ for all $p > p_Z$. This function is equal to 1 over the range $\gamma_Z \in (1, \frac{5}{3}]$ and decreases monotonically to $\frac{1}{\sqrt{2}}$ as γ_Z grows to infinity. To simplify the notation, let us set $\alpha_Z := \frac{2a_Z(1-b\rho_Z)}{\gamma_Z - 1}$, $\widehat{\alpha}_Z := c(\gamma_Z)\alpha_Z$. In the rest of section 5 we extract lower bounds on ϕ to derive an explicit upper bound on p^*.

To simplify the notation in many of the expressions used below, we introduce two indices in the set $\{L, R\}$, denoted by "min" and "max" and defined as follows:

$$\min := \begin{cases} L & \text{if } p_L \le p_R, \\ R & \text{if } p_L > p_R, \end{cases} \qquad \max := \begin{cases} R & \text{if } p_L \le p_R, \\ L & \text{if } p_L > p_R. \end{cases} \tag{5.2}$$

Notice that $p_{\min} = \min(p_L, p_R)$ and $p_{\max} = \max(p_L, p_R)$. For instance, $a_{\min} = a_Z$ and $\gamma_{\min} = \gamma_Z$ if $p_{\min} = p_Z$, and $a_{\max} = a_Z$ and $\gamma_{\max} = \gamma_Z$ if $p_{\max} = p_Z$. We also introduce the two indices $m \in \{L, R\}$ and $M \in \{L, R\}$ defined as follows:

$$m := \begin{cases} L & \text{if } \gamma_L \le \gamma_R, \\ R & \text{if } \gamma_L > \gamma_R, \end{cases} \qquad M := \begin{cases} R & \text{if } \gamma_L \le \gamma_R, \\ L & \text{if } \gamma_L > \gamma_R. \end{cases} \tag{5.3}$$

Notice that $\gamma_m = \min(\gamma_L, \gamma_R)$ and $\gamma_M := \max(\gamma_L, \gamma_R)$. However, $\gamma_{\min}$ and $\gamma_{\max}$ may not coincide with the values γ_m and γ_M, respectively. We now propose an upper bound on p^* based on the signs of $\phi(p_{\min})$ and $\phi(p_{\max})$.

5.2. Case 0: vacuum.

If the vacuum condition holds, i.e., $v_R - v_L \ge \frac{2a_L(1-b\rho_L)}{\gamma_L-1} + \frac{2a_R(1-b\rho_R)}{\gamma_R-1}$, we have $p^* = 0 =: \widehat{p}^*$. Notice in passing that $\lambda_1^-(0) = v_L - a_L$ and $\lambda_3^+(0) = v_R + a_R$.

5.3. Case 1: $0 < p^*$ and $0 < \phi(p_{\min})$.

This case corresponds to the L-wave and the R-wave both being expansion waves. In this case $p^* < p_{\min}$, which means that we do not need to compute p^* as we have $\lambda_1^-(p^*) = v_L - a_L$ and $\lambda_3^+(p^*) = v_R + a_R$. But, if for some reason an upper bound for p^* is needed, one can use the root of the function

$$\widehat{\phi}_{RR}(p) := \alpha_L\left(\left(\frac{p}{p_L}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + \alpha_R\left(\left(\frac{p}{p_R}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_R - v_L. \tag{5.4}$$

Note that $\widehat{\phi}_{RR}(p) \le \phi_{RR}(p) = \phi(p)$ for all $p \in [0, p_{\min}]$. We give the root for completeness:

$$\widetilde{p}^* = \left(\frac{(\alpha_R + \alpha_L - (v_R - v_L))_+}{\alpha_R p_R^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_L p_L^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}}. \tag{5.5}$$

We have that $p^* \le \widetilde{p}^*$. In conclusion, an upper bound on p^* is $\widehat{p}^* := \min(p_{\min}, \widetilde{p}^*)$. This implies that $0 < p^* \le \widehat{p}^*$. Taking the positive part of $\alpha_R + \alpha_L - (v_R - v_L)$ in (5.5) allows us to combine Cases 0 and 1 into one single case. Notice in passing that $\lambda_1^-(\widehat{p}^*) = \lambda_1^-(0) = v_L - a_L$ and $\lambda_3^+(\widehat{p}^*) = \lambda_3^+(0) = v_R + a_R$.

5.4. Case 2: $\boldsymbol{\phi(p_{\min}) < 0 < \phi(p_{\max})}$. In this case the min-wave is a shock and the max-wave is an expansion. Here we have $p_{\min} < p^* < p_{\max}$, and so for $p \in (p_{\min}, p_{\max})$ we have that

$$\widehat{\phi}_{RR}(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p}{p_{\min}}\right)^{\frac{\gamma_{\min}-1}{2\gamma_{\min}}} - 1\right) + \alpha_{\max}\left(\left(\frac{p}{p_{\max}}\right)^{\frac{\gamma_{\max}-1}{2\gamma_{\max}}} - 1\right) + v_R - v_L. \tag{5.6}$$

We consider two cases to derive a lower bound on $\widehat{\phi}_{RR}(p)$. If $\gamma_{\min} = \gamma_m$, we define

$$\widehat{\phi}_1(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p}{p_{\min}}\right)^{\frac{\gamma_M-1}{2\gamma_M}} r - 1\right) + \alpha_{\max}\left(\left(\frac{p}{p_{\max}}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_R - v_L,$$

$$\widehat{\phi}_2(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p}{p_{\min}}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + \alpha_{\max}\left(\left(\frac{p}{p_{\max}}\right)^{\frac{\gamma_m-1}{2\gamma_m}} r - 1\right) + v_R - v_L,$$

where $r := (\frac{p_{\min}}{p_{\max}})^{\frac{\gamma_M-\gamma_m}{2\gamma_m\gamma_M}}$. We have $\max(\widehat{\phi}_1(p), \widehat{\phi}_2(p)) \le \widehat{\phi}_{RR}(p)$ for all $p \in (p_{\min}, p_{\max})$. Solving $\widehat{\phi}_1(p) = 0$ and $\widehat{\phi}_2(p) = 0$ gives

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_R - v_L)}{r\widehat{\alpha}_{\min} p_{\min}^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_{\max} p_{\max}^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}}, \tag{5.7}$$

$$\widetilde{p}_2^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_R - v_L)}{\widehat{\alpha}_{\min} p_{\min}^{-\frac{\gamma_m-1}{2\gamma_m}} + r\alpha_{\max} p_{\max}^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}}.$$

Hence, an upper bound on p^* is $\widehat{p}^* := \min(p_{\max}, \widetilde{p}_1^*, \widetilde{p}_2^*)$ if $\gamma_{\min} = \gamma_m$. This implies that $p_{\min} < p^* \le \widehat{p}^*$. In the other case, $\gamma_{\min} = \gamma_M$, we have $\gamma_{\max} = \gamma_m$, and two lower bounds on $\widehat{\phi}(p)$ are given by

$$\widehat{\phi}_1(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p}{p_{\min}}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + \alpha_{\max}\left(\left(\frac{p}{p_{\max}}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + v_R - v_L,$$

$$\widehat{\phi}_2(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p}{p_{\min}}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + \alpha_{\max}\left(\left(\frac{p}{p_{\max}}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_R - v_L.$$

Again, the equations $\widehat{\phi}_1(p) = 0$, $\widehat{\phi}_2(p) = 0$ are linear (up to a change of variable). The roots are

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_R - v_L)}{\widehat{\alpha}_{\min} p_{\min}^{-\frac{\gamma_m-1}{2\gamma_m}} + \alpha_{\max} p_{\max}^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}}, \tag{5.8}$$

$$\widetilde{p}_2^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_R - v_L)}{\widehat{\alpha}_{\min} p_{\min}^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_{\max} p_{\max}^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}}.$$

An upper bound on p^* is $\widehat{p}^* := \min(p_{\max}, \widetilde{p}_1^*, \widetilde{p}_2^*)$ if $\gamma_{\min} = \gamma_M$. Hence $p_{\min} < p^* \le \widehat{p}^*$.

5.5. Case 3: $\boldsymbol{\phi(p_{\max}) < 0}$. In this case we have $p_{\max} < p^*$ and the L-wave and the R-wave are shocks. We bound $\phi(p)$ from below by the function

$$\widehat{\phi}_{RR}(p) := \widehat{\alpha}_L\bigg(\Big(\frac{p}{p_L}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\bigg) + \widehat{\alpha}_R\bigg(\Big(\frac{p}{p_R}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\bigg) + v_R - v_L. \tag{5.9}$$

The corresponding root for $\widehat{\phi}_{RR}(p) = 0$ is

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_L + \widehat{\alpha}_R - (v_R - v_L)}{\widehat{\alpha}_L p_L^{-\frac{\gamma_m-1}{2\gamma_m}} + \widehat{\alpha}_R p_R^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}}. \tag{5.10}$$

Another possibility consists of observing that ϕ is the sum of two shock curves plus the constant $v_R - v_L$. Observing that $B_Z \le B_Z\, p\, p_{\max}^{-1}$ for all $p \in (p_{\max}, \infty)$, we infer that the graph of the following function is also below the graph of ϕ:

$$\widehat{\phi}_{SS}(p) := \frac{p - p_L}{\sqrt{p}}\left(\frac{A_L}{1 + \frac{B_L}{p_{\max}}}\right)^{\frac{1}{2}} + \frac{p - p_R}{\sqrt{p}}\left(\frac{A_R}{1 + \frac{B_R}{p_{\max}}}\right)^{\frac{1}{2}} + v_R - v_L. \tag{5.11}$$

Let $x_Z := \big(\frac{A_Z}{1 + B_Z p_{\max}^{-1}}\big)^{\frac{1}{2}}$, $a := x_L + x_R$, $b := v_R - v_L$, $c := -p_L x_L - p_R x_R$. Then the only positive root of $\widehat{\phi}_{SS}$ is

$$\widetilde{p}_2^* = \left(\frac{-b + (b^2 - 4ac)^{\frac{1}{2}}}{2a}\right)^2. \tag{5.12}$$

An upper bound on p^* is $\widehat{p}^* := \min(\widetilde{p}_1^*, \widetilde{p}_2^*)$. Hence $p_{\max} < p^* \le \widehat{p}^*$.

Algorithm 1 Computing $\lambda_{\max}(\widehat{p}^*)$.

```
Require: $\boldsymbol{u}_L$, $\boldsymbol{u}_R$, $\boldsymbol{n}_{LR}$, $p_L$, $p_R$
  compute $\phi(p_{\min})$, $\phi(p_{\max})$ from (4.2)
  if $0 \le \phi(p_{\min})$ then
      compute $\widetilde{p}^*$ from (5.5) and set $\widehat{p}^* = \max(p_{\min}, \widetilde{p}^*)$
  else if $0 \le \phi(p_{\max})$ then
      if $\gamma_{\min} = \gamma_m$ then
          compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (5.7)
      else
          compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (5.8)
      end if
      $\widehat{p}^* = \min\{p_{\max}, \widetilde{p}_1^*, \widetilde{p}_1^*\}$
  else
      compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (5.10) and (5.12) respectively
      $\widehat{p}^* = \min\{\widetilde{p}_1^*, \widetilde{p}_2^*\}$
  end if
  return $\lambda_{\max}(\widehat{p}^*) := \max\{-\lambda_L^-(\widehat{p}^*), \lambda_R^+(\widehat{p}^*)\}$
```

5.6. Iterative solution. Another possibility to estimate p^* as above consists of solving $\phi(p) = 0$ by using the iterative quadratic Newton method described in Guermond and Popov [13, Alg. 1]. The method is guaranteed to be convergent since the function ϕ defined in (4.2) is concave. Using the lower and upper bounds provided in sections 5.2–5.5 to initialize the algorithm, the method is also guaranteed to deliver an upper bound on p^* for every termination threshold since $\phi'''(\xi) > 0$ for all $\xi > 0$ (see the proof of Lemma 4.5 in [13]). A source code for this method based on Algorithm 1 is publicly available at [4].

6. Numerical results.

We numerically illustrate in this section the algorithm (2.7) with the viscosity defined in Theorem 4.6 using the explicit upper bound $\widehat{p}^*$ defined in sections 5.2–5.5.

6.1. Convergence tests.

We use the van der Waals equation of state as the oracle to verify the method. More precisely, we consider the solution to a Riemann problem and compare it to the numerical approximation (2.7), where the viscosity d_{ij}^n is defined in (4.14b) with $\widehat{p}^*$ being the upper bound on p^* derived in sections 5.2–5.5. Recall that for the van der Waals equation of state, the pressure is given by $p(\rho,e) := (\gamma-1)\frac{\rho e + a\rho^2}{1-b\rho} - a\rho^2$, where γ, a, and b are constants depending on the nature of the fluid (see, e.g., Callen [3, sect. 3.5], Fossati and Quartapelle [10, sect. 6.3]). We select the parameters γ, a, b so that the problem is hyperbolic and the solution exhibits a composite wave structure: we use $\gamma = 1.02$, $a = 1$, $b = 1$. With these parameters the isentropes in the $(p, \frac{1}{\rho})$ diagram are nonconvex. The loss of convexity is necessary for the existence of composite waves. The initial left and right states we choose are

$$\begin{aligned}(\rho_L, v_L, p_L) &:= (0.10, -0.475504638574729, 0.022084258693080),\\ (\rho_R, v_R, p_R) &:= (0.39, -0.121375781741349, 0.039073167077590).\end{aligned} \tag{6.1}$$

The exact solution is a 3-wave composed of an expansion fan, a shock, and another expansion fan. The details of the construction of the solution can be found in Cramer and Sen [7], Lai [21], and Fossati and Quartapelle [10, sect. 6.4]. For completeness and reproducibility, the construction of the exact solution is given in the accompanying supplementary material file supplementary.pdf [local/web 319KB] and a code computing the exact solution is available at Clayton, Guermond, and Popov [5].

We approximate the solution with $\mathbb{P}_1$ continuous finite elements in one dimension. The computational domain is $D := (-1, 1)$. The estimation of the maximum wave speed (see (4.14a)) is done by using $\widehat{p}^*$ as explained in sections 5.2–5.5. The time step size τ is computed at each time step by using the expression

$$\tau := \tfrac{\mathrm{CFL}}{2} \min_{i\in\mathcal{V}} \frac{m_i}{\sum_{j\in\mathcal{I}^*(i)} d_{ij}^n}, \tag{6.2}$$

where $\mathcal{I}^*(i) := \mathcal{I}(i) \setminus \{i\}$. We use CFL = 0.5 in this test. A series of computations is done on nested uniform meshes to estimate the convergence rate of the method. Denoting by $(\rho_h(t), \boldsymbol{m}_h(t), E_h(t))$ the approximation at time t, we compute a consolidated error indicator by adding the relative error in the L^q-norm on the density, the momentum, and the total energy as follows:

$$\delta_q(t) := \frac{\|\rho_h(t)-\rho(t)\|_{L^q(D)}}{\|\rho(t)\|_{L^q(D)}} + \frac{\|\boldsymbol{m}_h(t)-\boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}}{\|\boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}} + \frac{\|E_h(t)-E(t)\|_{L^q(D)}}{\|E(t)\|_{L^q(D)}}. \tag{6.3}$$

The results of the convergence tests are reported in Table 1. The number of grid points is reported in the leftmost column. The errors are computed at $t = 0.5$. We observe that the method is convergent, and the convergence rates are consistent with the approximation being formally first-order accurate.

TABLE 1
Consolidated errors and convergence rates. Solution computed at $t = 5.0$.

#dof	$\delta_1(t)$	Rate	$\delta_2(t)$	Rate
101	2.14E-01	–	2.67E-01	–
201	1.44E-01	0.58	2.07E-01	0.37
401	9.40E-02	0.62	1.58E-01	0.39
801	5.96E-02	0.66	1.20E-01	0.40
1601	3.66E-02	0.70	8.96E-02	0.42
3201	2.18E-02	0.75	6.66E-02	0.43
6401	1.27E-02	0.78	4.93E-02	0.43
12801	7.26E-03	0.81	3.66E-02	0.43
25601	4.09E-03	0.83	2.72E-02	0.43

6.2. The two-expansion-wave-speed estimate. It is often reported in the literature that, for practical purposes, one can use the two expansion wave speeds, $v_L - c_L$, $v_R + c_R$, to estimate the maximum wave speed. Using the covolume equation of state, we have shown in [13, App. B] that $\max(|v_L - c_L|, |v_R + c_R|)$ is not an upper bound on the maximum wave speed in the Riemann problem. But the reader could legitimately be skeptical about this kind of theoretical result and may wonder whether these academic arguments have any impact on practical computations. We now illustrate that the two-expansion-wave-speed estimate is not robust: it can either lead to an underestimation or to an overestimation of the viscosity, with severe consequences in both cases.

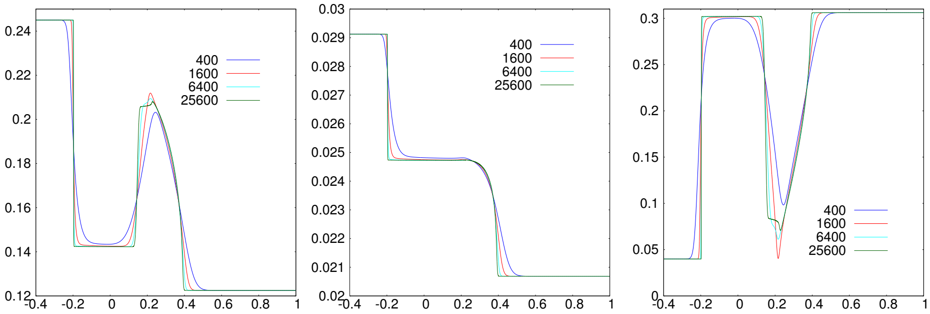

FIG. 1. *Test with the data* (6.4), $t = 1.25$, *computed with present method (results using expansion wave speeds are not shown because the simulations crash after a few time steps). From left to right: density, pressure, sound speed.*

We start by showing that $\max(|v_L - c_L|, |v_R + c_R|)$ can lead to an underestimation of the viscosity and therefore lead to violations of important properties. Our oracle is the van der Waals equation of state with $a = 1$, $b = 1$, $\gamma = 1.02$. We solve two Riemann problems. The first one is equipped with the following data set:

$$\begin{aligned}(\rho_L, v_L, p_L) &:= (0.2450, 0, 2.9123894332846005\times 10^{-2}),\\ (\rho_R, v_R, p_R) &:= (0.1225, 0, 2.0685894810791836\times 10^{-2}),\end{aligned} \tag{6.4}$$

which gives the sound speeds $(c_L, c_R) \approx (0.00399, 0.306)$. The second one is equipped with the following data set:

$$\begin{aligned}(\rho_L, v_L, p_L) &:= (2.5\times 10^{-1}, 0, 3\times 10^{-2}),\\ (\rho_R, v_R, p_R) &:= (4.9\times 10^{-5}, 0, 5\times 10^{-8}),\end{aligned} \tag{6.5}$$

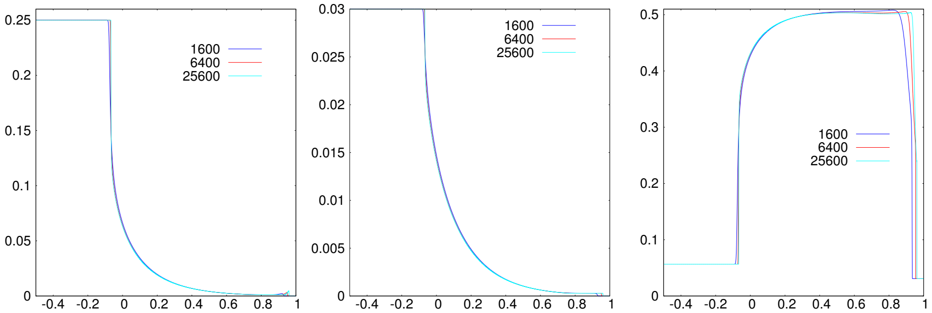

FIG. 2. *Test with the data* (6.5), $t = 0.4$ *(results using expansion wave speeds are not shown because the simulations crash after a few time steps). From left to right: density, pressure, sound speed.*

which gives the sound speeds $(c_L, c_R) \approx (0.057, 0.031)$. For each data set, we perform two series of computations on the domain $D = (-0.5, 1)$. The computations are done up to $t = 1.25$ for the first data set and up to $t = 0.4$ for the second data set. In both cases we use CFL $= 0.5$. One series of computations is done with the estimation of the maximum wave speed (see (4.14a)) using $\widehat{p}^*$ as explained in sections 5.2–5.5 (no iteration is done). The other series is done using the two-expansion-wave-speed estimate $\max(|v_L - c_L(p_L, \rho_L)|, |v_R + c_R(p_L, \rho_L)|)$ with $c(p, \rho) = (\gamma \frac{p+a\rho^2}{\rho(1-b\rho)} - 2a\rho)^{\frac{1}{2}}$. It turns out that the computations done with the two-expansion-wave-speed estimate violate the invariant-domain property after a few time steps for both data sets: one obtains a complex sound speed for the first data set and one obtains a negative internal energy for the second data set. We have verified that these violations occur no matter how small the CFL number is. The computations done with the method proposed in the paper run without any problem. We show in Figure 1 the density, the pressure and the sound speed profiles for various mesh sizes $(\frac{1.5}{100}, \frac{1.5}{400}, \frac{1.5}{1600}, \frac{1.5}{25600})$ for the data set (6.4). The results for the second data set (6.5) are shown in Figure 2 with the mesh sizes $\frac{1.5}{1600}, \frac{1.5}{6400}$. Notice that in both cases the R-wave is a composite wave composed of an expansion followed by a shock.

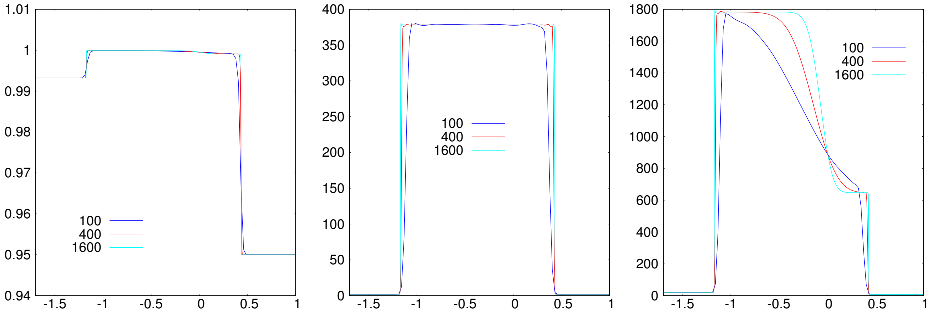

FIG. 3. *Test with the data* (6.6), $t = 0.005$ *(results using expansion wave speeds are not shown). From left to right: density, pressure, sound speed.*

We now show that the two-expansion-wave-speed estimate can lead to a local overestimation of the viscosity and thereby to a reduction of the admissible range of time step sizes. We use again the van der Waals equation of state with the same parameters as above for the oracle. We consider the Riemann problem with the following data:

$$
\begin{aligned}
(\rho_L, v_L, p_L) &:= (0.9932,\ 3, 2),\\
(\rho_R, v_R, p_R) &:= (0.9500, -3, 2).
\end{aligned}
\tag{6.6}
$$

The corresponding sound speeds are $(c_L, c_R) \approx (21.2, 7.77)$. The computational domain is $D = (-1.7, 1)$, and the computations are done up to $t = 0.005$. For the computation with the two-expansion-wave-speed, the CFL number needed to avoid producing negative internal energy is about 0.06. The maximal admissible CFL number for the present method is about 0.71 (i.e., below this CFL number the sound speed is real and the internal energy is positive at every grid point and for every time step). As a result the computational cost of the method using the two-expansion-wave-speed estimate is almost 12 times higher than that of the present method. We show in Figure 3 the density, the pressure, and the sound speed for various meshes using the present method. The results obtained with the two-expansion-wave-speed estimate are almost identical (and are thus not shown).

6.3. Cubic equation of state. We continue by illustrating the proposed method by using a cubic equation of state as the oracle; see Redlich and Kwong [32] and Valderrama [36]. We refer the reader to Dumbser and Casulli [9], where series of tests are done with this type of equation of state. For a general cubic equation of state, the pressure is given by

$$
p(\rho, e) = \frac{R\rho T(\rho, e)}{1 - b\rho} - \frac{\alpha\rho^2}{\sqrt{T(\rho, e)}(1 - br_1\rho)(1 - br_2\rho)},
\tag{6.7}
$$

where $T(\rho, e)$ solves the following cubic equation:

$$
e = c_v T + \frac{3\alpha}{2b\sqrt{T}} \frac{1}{r_1 - r_2} \log\left(\frac{1 - br_1\rho}{1 - br_2\rho}\right).
\tag{6.8}
$$

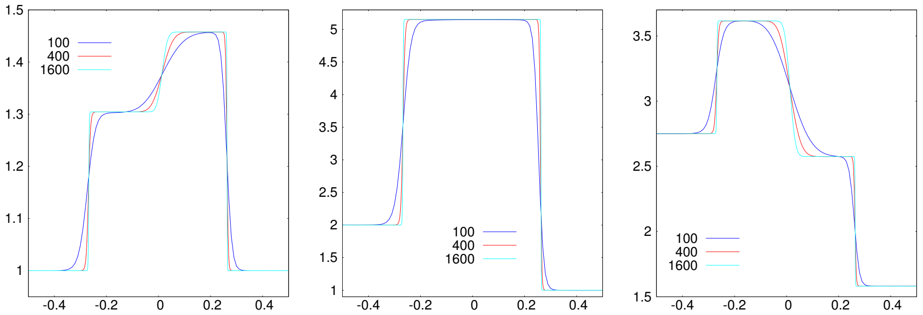

FIG. 4. *Test with the data* (6.9), $t = 0.1$. *From left to right: density, pressure, temperature.*

We take $r_1 = 0$ and $r_2 = -1$ (this corresponds to the so-called Redlich–Kwong equation). We solve two of the problems from [9, sect. 3.3] where $R = 0.4$, $\alpha = 0.5$, $b = 0.5$. These are two Riemann problems. For the first problem we take $c_v = 1$, and the initial data are

$$
\begin{aligned}
(\rho_L, v_L, p_L) &:= (1, 1, 2),\\
(\rho_R, v_R, p_R) &:= (1, -1, 1).
\end{aligned}
\tag{6.9}
$$

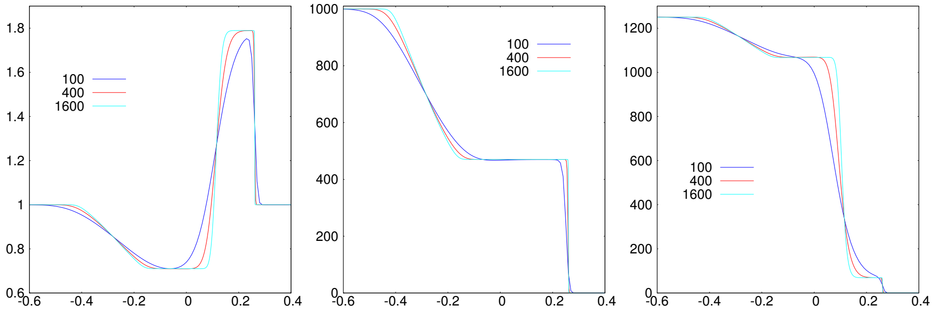

FIG. 5. *Test with the data* (6.10), $t = 0.008$. *From left to right: density, pressure, temperature.*

The computational domain is $(-0.5, 0.5)$ and the final time is $t = 0.1$. For the second problem we take

$$(\rho_L, v_L, p_L) := (1, 0, 1000),\\ (\rho_R, v_R, p_R) := (1, 0, 0.01), \tag{6.10}$$

with $c_v = 1.5$ (we suspect there is a typo in [9, sect. 3.3], since the authors say that they use $c_v = 1$ with the above data, but this gives a negative internal energy for the right state). The computational domain is $D = (-0.6, 0.4)$ and the final time is $t = 0.008$. In both cases, we take the covolume constant in (3.2) to be $b = 0.5$ (using $b = 0$ in (3.2) gives similar results, not shown). The CFL number is 0.5. The results obtained with various meshes are displayed in Figure 4, for the first case, and in Figure 5, for the second case. In each case, we show the density, the pressure, and the temperature. These results are similar to those reported in [9, sect. 3.3].

6.4. Tabulated equation of state. To demonstrate that the proposed method works with a tabulated equation of state, we present three simulations using the SESAME database [25], which was developed by the Physics and Chemistry of Materials Group in the Theoretical Division at the Los Alamos National Lab. The SESAME database is accessed through the use of the open source software EOSPAC6 [1]. We select a single material through the use of a material identification number and specify a "table type." The "table type" indicates the thermodynamic variables to be used in the tabulated data. A tabulated equation of state is then precomputed at the beginning of the simulation. In our case, we provide $\{(\rho_i^n, e_i^n)\}_{i\in\mathcal{V}}$ and EOSPAC6 returns $\{p_i^n\}_{i\in\mathcal{V}}$ at each time step. The units in the SESAME database are as follows: $\mathrm{Mg\,m^{-3}}$ for the density, $\mathrm{MJ\,kg^{-1}}$ for the specific internal energy, and GPa for the pressure. We refer the reader to the user manual Pimentel [30] for the technical details.

We solve three 1D Riemann problems using various grid sizes. In the first Riemann problem we use dry air (material ID: 5030 [12]), in the second we use helium (material ID: 5760 [22]), and in the third we use cesium (material ID: 3510 [20]).

The first example is done with dry air. We choose the initial data in order to have a wave profile similar to that observed in the Sod shocktube:

$$(\rho_L, v_L, e_L) := (0.01\,\mathrm{Mg\,m^{-3}}, 0\,\mathrm{ms^{-1}}, 4000\,\mathrm{MJ\,kg^{-1}}),\\ (\rho_R, v_R, e_R) := (0.003\,\mathrm{Mg\,m^{-3}}, 0\,\mathrm{ms^{-1}}, 3400\,\mathrm{MJ\,kg^{-1}}). \tag{6.11}$$

The final time is $t = 3.8 \times 10^{-4}\,\mathrm{s}$. The density, the pressure, and γ are shown in Figure 6 for three meshes (100, 400, and 1600 grid points). The approximate

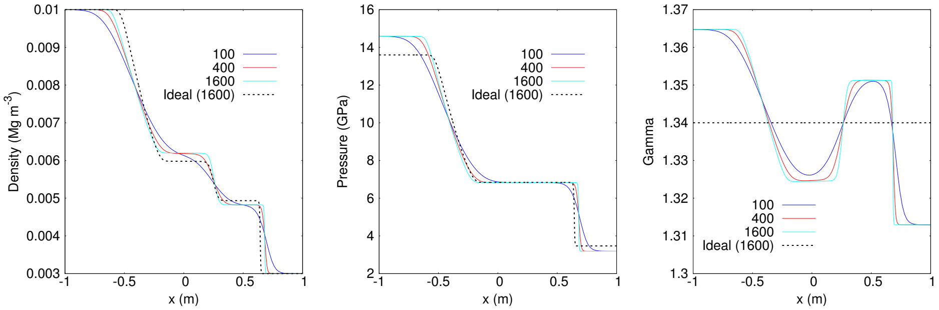

Fig. 6. *Test with the data* (6.11)*,* $t = 3.8 \times 10^{-4}$ s*. From left to right: density, pressure, gamma.*

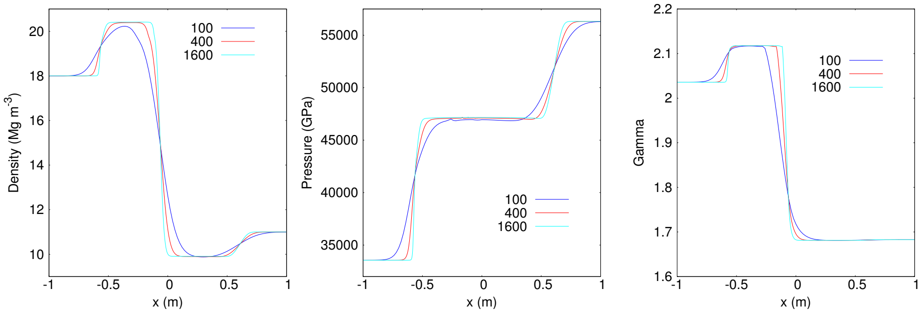

Fig. 7. *Test with the data* (6.12)*,* $t = 2.21 \times 10^{-4}$ s*. From left to right: density, pressure, gamma.*

solutions are compared with a solution using the ideal gas law, $p(\rho, e) = (\gamma - 1)\rho e$, with $\gamma = 1.34$.

In the second example, we use helium with the following initial data:

$$\begin{aligned}(\rho_L, v_L, e_L) &:= (18\,\mathrm{Mg\,m^{-3}}, 0\,\mathrm{ms^{-1}}, 1800\,\mathrm{MJ\,kg^{-1}}),\\ (\rho_R, v_R, e_R) &:= (11\,\mathrm{Mg\,m^{-3}}, 0\,\mathrm{ms^{-1}}, 7500\,\mathrm{MJ\,kg^{-1}}).\end{aligned} \tag{6.12}$$

The final time is $t = 2.21 \times 10^{-4}$ s. In this case γ is in the range $[1.7, 2.2]$. These values are larger than $\frac{5}{3}$. This example demonstrates the necessity of estimating the max wave speed with $\gamma > \frac{5}{3}$ (see Appendix A). The density, the pressure, and γ are shown in Figure 7 for three meshes (400, 1600, and 6400 grid points).

For the last example we use cesium (material ID: 3510) for a Riemann problem with data,

$$\begin{aligned}(\rho_L, v_L, e_L) &:= (33\,\mathrm{Mg\,m^{-3}}, -3000\,\mathrm{m\,s^{-1}}, 75\,\mathrm{MJ\,kg^{-1}}),\\ (\rho_R, v_R, e_R) &:= (15.25\,\mathrm{Mg\,m^{-3}}, 3000\,\mathrm{m\,s^{-1}}, 12.2\,\mathrm{MJ\,kg^{-1}}).\end{aligned} \tag{6.13}$$

The final time is $t = 2.0 \times 10^{-4}$ s. This final tabulated example is used to illustrate a double expansion problem with a solution near a vacuum. The minimum density (with 6401 grid points) is approximately $1.68 \times 10^{-3}\,\mathrm{Mg\,m^{-3}}$. The plots are shown in Figure 8.

6.5. Two-dimensional illustration.

To demonstrate that the proposed method is actually independent of the space dimension, we illustrate it by doing two-dimensional simulations. We use a finite element code documented in Maier and Tomas [27]. We

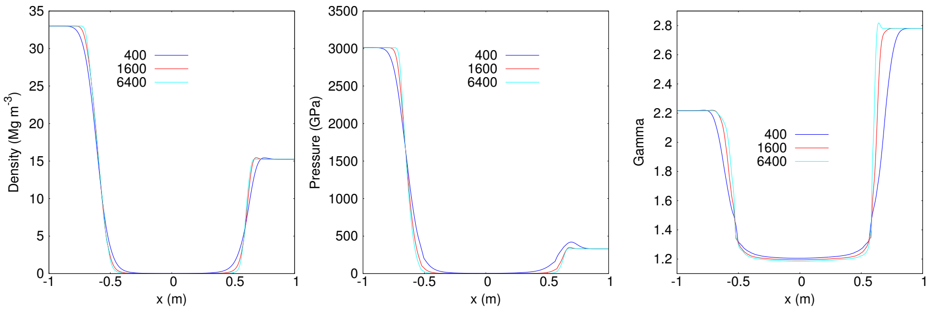

Fig. 8. *Test with the data* (6.13), $t = 2.0 \times 10^{-4}$ s. *From left to right: density, pressure, gamma.*

replace the estimation of $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)$ used in this code (and described in [13]) by the estimation (4.14a) with $\widehat{p}^*$ computed as explained in sections 5.2–5.5. The oracle is the van der Waals equation of state with $\gamma = 1.4$, $a = 0.3215$, and $b = 0.1$. The computation of $\widehat{p}^*$ is done with the assumption that $b = 0$. That is, we assume that the covolume constant b is unknown.

We simulate the flow around a cylinder in a two-dimensional channel. The computational domain is $D = (-0.9, 3.1) \times (-1, 1) \backslash C$, with C being the disk of radius 0.15 centered at $(0, 0)$. We enforce the density, the momentum, and the total energy at the inflow boundary, $\{x = -0.9\}$: $(\rho, \boldsymbol{m}, E) = (1.4, (4.2, 0)^{\mathsf{T}}, 9.154375)$. The primitive variable corresponding to these data are $\boldsymbol{v} = (3, 0)^{\mathsf{T}}$ and $p = 1$. The corresponding Mach number is 3. The slip boundary condition is enforced at the top and at the bottom of the channel. Nothing is done at the outflow boundary condition (this is a supersonic outflow boundary). We use continuous $\mathbb{Q}_1$ finite elements. We refer the reader to [27] for the implementation details.

We show in Figure 9 the density computed at time $t = 4$ using a Schlieren-like representation. Letting $\sum_{i \in \mathcal{V}} \rho_i^n \varphi_i$ be the approximation of the density, we approximate the Euclidean norm of the gradient of the density as follows: $r_i^n := m_i^{-1} \| \sum_{j \in \mathcal{I}(D_i)} \boldsymbol{c}_{ij} \rho_j^n \|_{\ell^2}$ for all $i \in \mathcal{V}$. The values of the Schlieren field are defined at the grid points by $\exp(-\beta(r_i^n - \min_{j \in \mathcal{I}(i)} r_j^n)/(\max_{j \in \mathcal{I}(i)} r_j^n - \min_{j \in \mathcal{I}(i)} r_j^n))$ where $\beta = 10$. For comparison, we also show in the right panel of Figure 9 the density obtained at the same time using the ideal gas equation of state. The inflow boundary data is $(\rho, \boldsymbol{m}, E) := (1.4, (4.2, 0)^{\mathsf{T}}, 8.8)$ and $\gamma = 1.4$. This corresponds to the same primitive state, $\boldsymbol{v} = (3, 0)^{\mathsf{T}}$ and $p = 1$, as the simulation done with the van der Waals equation of state. The mesh used for these computations has 1.4×10^6 grid points.

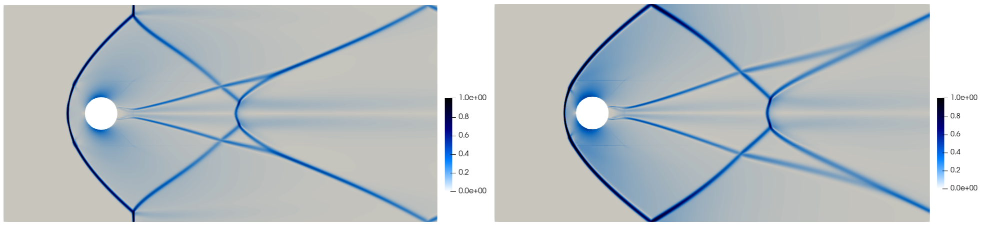

Fig. 9. *Cylinder at Mach 3 in a channel. Density at $t = 4$. Left: the oracle is the van der Waals equation of state. Right: the oracle is the ideal gas equation of state with $\gamma = 1.4$.*

Of course, these simulations are first-order accurate in space. Making the approximation higher-order accurate can be done by implementing the convex limiting technique described in [16, 17]. This, however, requires developing surrogate entropy functionals for the oracle in order to be able to compute entropy commutators and to enforce locally a minimum principle on the surrogate entropies. This task is under way, and the results of this work will be reported elsewhere. We are currently implementing the technique in the massively parallel code `Ryujin` documented in Maier and Kronbichler [26].

7. Conclusions.

We have proposed in the paper an approximation technique for the compressible Euler equations where the equation of state is given by an oracle. The key feature is an artificial graph viscosity using an estimate on the maximum wave speed on each elementary Riemann problem that guarantees the positivity of the density and of the internal energy. This estimate also guarantees an upper bound on the density when a covolume constant in known. The main theoretical result of the paper is Theorem 4.6. The guaranteed bounds developed in sections 5.2–5.5 are easy to compute. These upper bounds can be used in any algorithm that is based on approximate Riemann solvers. A computer code implementing all these bounds is freely available at Clayton, Guermond, and Popov [4]. All the simulations reported in this paper have been done with this code.

Appendix A. Improvement on the $\gamma > \frac{5}{3}$ estimates.

The objective of this appendix is to prove that $\phi_{RR}(p) \leq \phi(p)$ for all $p \in [\min(p_L, p_R), \infty)$, where we recall that the function ϕ is defined in (4.2), and the function ϕ_{RR} is defined in (5.1). For future reference we also recall that

$$f_Z^S(p) := (p - p_Z)\sqrt{\frac{2}{(\gamma_Z+1)\rho_Z}}\left(p + \frac{\gamma_Z - 1}{\gamma_Z + 1}p_Z\right)^{-\frac{1}{2}}\sqrt{1 - b\rho_Z}, \tag{A.1}$$

$$f_Z^R(p) := \frac{2\sqrt{\frac{\gamma_Z p_Z}{\rho_Z}}}{\gamma_Z - 1}\left(\left(\frac{p}{p_Z}\right)^{\frac{\gamma_Z - 1}{2\gamma_Z}} - 1\right)\sqrt{1 - b\rho_Z}. \tag{A.2}$$

The functions $f_Z^S(p)$ and f_Z^R are, respectively, the shock and rarefaction curves introduced in (4.1). The following lemma is one of the main result established in Guermond and Popov [13].

Lemma A.1 (see [13, Lem. 4.2]). *Let $p_Z > 0$, ρ_Z be such that $0 < 1 - b\rho_Z < 1$, and let $\gamma_Z \in (1, \infty)$. Assume that $\gamma \in (1, \frac{5}{3}]$. Then $f_R(p) < f_S(p)$ for all $p \in (p_Z, \infty)$ and $f_R(p_Z) = f_S(p_Z)$, i.e., the shock curve is above the rarefaction curve.*

Theorem A.2. *Assume $\gamma \in (1, \frac{5}{3}]$. Let $p_{\min}$ and $p_{\max}$ be defined as in section* 5.1. *For any $p \geq 0$, the graph of $\phi(p)$ is above the graph of $\phi_{RR}(p)$; more precisely, $\phi_{RR}(p) = \phi(p)$ for all $p \in [0, p_{\min}]$, and $\phi_{RR}(p) < \phi(p)$ for all $p \in (p_{\min}, \infty)$.*

Proof. Note that the two curves $(p, \phi(p))$ and $(p, \phi_{RR}(p))$ coincide if $p \leq p_{\min}$ because both ϕ and ϕ_{RR} are the sum of the two rarefaction curves plus the constant $v_R - v_L$. If $p_{\min} < p \leq p_{\max}$, then the function $\phi(p)$ is the sum of one rarefaction curve and one shock curve plus the constant $v_R - v_L$. We then conclude by invoking Lemma A.1 with $(p_Z, \rho_Z) = (p_{\min}, \rho_{\min})$. If $p_{\max} \leq p$, then the function $\phi(p)$ is the sum of two shock curves plus the constant $v_R - v_L$. Now we invoke Lemma A.1 twice to complete the proof, once with $(p_Z, \rho_Z) = (p_{\min}, \rho_{\min})$ and once with $(p_Z, \rho_Z) = (p_{\max}, \rho_{\max})$. ∎

The assertion in Lemma A.1 is false when $\frac{5}{3} < \gamma_Z$. To remedy this deficiency, we now define a new function that is guaranteed to always be under $\phi(p)$ for all

$\gamma_Z \in (1,\infty)$ and all $p \in (p_{\min},\infty)$. Consider

$$c(\gamma_Z) := \begin{cases} 1 & \text{if } 1 < \gamma_Z \leq \frac{5}{3}, \\ (\frac{1}{2} + \frac{4}{3(\gamma_Z+1)})^{\frac{1}{2}} & \text{if } \frac{5}{3} \leq \gamma_Z \leq 3, \\ (\frac{1}{2} + \frac{2}{\gamma_Z-1} 3^{\frac{4-2\gamma_Z}{\gamma_Z-1}})^{\frac{1}{2}} & \text{if } 3 \leq \gamma_Z. \end{cases} \tag{A.3}$$

Notice that $(1,\infty) \ni \gamma_z \mapsto c(\gamma_Z)$ is continuous and $c(\gamma_Z) \in (\frac{1}{2}, 1]$.

Lemma A.3. *Let $p_Z > 0$, ρ_Z be such that $0 < 1 - b\rho_Z < 1$, and let $\gamma_Z \in (1,\infty)$. Then $c(\gamma_Z) f^R_Z(p_Z) = f^S_Z(p_Z) = 0$ and $c(\gamma_Z) f^R_Z(p) < f^S_Z(p)$ for all $p \in (p_Z,\infty)$.*

Proof. The proof of the assertion is in the accompanying supplementary material file supplementary.pdf [local/web 319KB]. □

Acknowledgments. The authors thank Matthias Maier and Eric Tovar for stimulating discussions and the help they provided at various stages of this project. The authors also thank David Pimentel for his invaluable help with the setup and use of EOSPAC6 and the SESAME database.

REFERENCES

[1] EOSPAC6. https://github.com/KineticTheory/eospac6.

[2] R. Abgrall and S. Karni, *Computations of compressible multifluids*, J. Comput. Phys., 169 (2001), pp. 594–623.

[3] H. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed., John Wiley & Sons, New York, 1985.

[4] B. Clayton, J.-L. Guermond, and 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 (13 April 2021).

[5] B. Clayton, J.-L. Guermond, and B. Popov, *Solution to the Double Sonic Shock Riemann Problem with the van der Waals Equation of State*, https://doi.org/10.5281/zenodo.4685958, 2021.

[6] P. Colella and H. M. Glaz, *Efficient solution algorithms for the Riemann problem for real gases*, J. Comput. Phys., 59 (1985), pp. 264–289.

[7] M. S. Cramer and R. Sen, *Exact solutions for sonic shocks in van der Waals gases*, Phys. Fluids, 30 (1987), pp. 377–385.

[8] J. K. Dukowicz, *A general, noniterative Riemann solver for Godunov's method*, J. Comput. Phys., 61 (1985), pp. 119–137.

[9] M. Dumbser and 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 (2016), pp. 479–497.

[10] M. Fossati and L. Quartapelle, *The Riemann Problem for Hyperbolic Equations under a Nonconvex Flux with Two Inflection Points*, https://arxiv.org/abs/1402.5906, 2014.

[11] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.

[12] H. C. Graboske, *A New EOS for Air*, UCID 16901, Material 5030 in the SESAME Database, Lawrence Livermore National Laboratory, September 1975.

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

[14] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489, https://doi.org/10.1137/16M1074291 .

[15] J.-L. Guermond, B. Popov, and Y. Yang, *The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations*, J. Sci. Comput., 70 (2017), pp. 1358–1366.

[16] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Comput., 40 (2018), pp. A3211–A3239, https://doi.org/10.1137/17M1149961.

[17] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 143–175.

[18] H. Holden and N. H. Risebro, *Front tracking for hyperbolic conservation laws*, 2nd ed. Appl. Math. Sci. 152, Springer, Heidelberg, 2015.

[19] M. J. Ivings, D. M. Causon, and E. F. Toro, *On Riemann solvers for compressible liquids*, Internat. J. Numer. Methods Fluids, 28 (1998), pp. 395–418.

[20] J. Johnson and S. Lyon, *SESAME EOS Table* 3510 *for Cesium*, Material 3510 in the SESAME Database, October 1994.

[21] G. Lai, *Interactions of composite waves of the two-dimensional full Euler equations for van der Waals gases*, SIAM J. Math. Anal., 50 (2018), pp. 3535–3597, https://doi.org/10.1137/17M1144660.

[22] Lawrence Livermore National Laboratory, *Helium*, UCIR-740, Material 5760 in the SESAME Database, Lawrence Livermore National Laboratory, 1974.

[23] P. D. Lax, *Weak solutions of nonlinear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math., 7 (1954), pp. 159–193.

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

[25] S. P. Lyon, *SESAME: The Los Alamos National Laboratory Equation of State Database*, TINC; SESAME Database Documentation, Los Alamos National Laboratory (USDOE), Report LA-UR-92-3407, 2006, http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-92-3407.

[26] M. Maier and M. Kronbichler, *Efficient parallel* 3*d computation of the compressible Euler equations with an invariant-domain preserving second-order finite-element scheme*, ACM Trans. Parallel Comput., 8 (2021), 16.

[27] M. Maier and I. Tomas, *tamiko/step*-69*: step*-69 *v*20200305, https://doi.org/10.5281/zenodo.3698223, March 2020.

[28] C. Pantano, R. Saurel, and T. Schmitt, *An oscillation free shock-capturing method for compressible van der Waals supercritical fluid flows*, J. Comput. Phys., 335 (2017), pp. 780–811.

[29] J. Pike, *Riemann solvers for perfect and near-perfect gases*, AIAA J. 31 (1993), pp. 1801–1808.

[30] D. A. Pimentel, *Eospac User's Manual: Version* 6.4, https://www.osti.gov/biblio/1489917, 2019.

[31] L. Quartapelle, L. Castelletti, A. Guardone, and G. Quaranta, *Solution of the Riemann problem of classical gasdynamics*, J. Comput. Phys., 190 (2003), pp. 118–140.

[32] O. Redlich and J. N. S. Kwong, *On the thermodynamics of solutions. V. An equation of state. Fugacities of gaseous solutions*, Chem. Rev., 44 (1949), pp. 233–244.

[33] P. L. Roe and 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, 1985, pp. 499–518.

[34] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 3rd ed., Springer-Verlag, Berlin, 2009.

[35] E. F. Toro, C. E. Castro, and B. J. Lee, *A novel numerical flux for the* 3*D Euler equations with general equation of state*, J. Comput. Phys., 303 (2015), pp. 80–94.

[36] J. O. Valderrama, *The state of the cubic equations of state*, Ind. Eng. Chem. Res., 42 (2003), pp. 1603–1618.