Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 375 (2021) 113608
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Second-order invariant domain preserving approximation of the compressible Navier–Stokes equations☆

Jean-Luc Guermond[a,*], Matthias Maier[a], Bojan Popov[a], Ignacio Tomas[b]

[a] Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA
[b] Sandia National Laboratories[1], P.O. Box 5800, MS 1320, Albuquerque, NM 87185-1320, USA

Received 28 August 2020; received in revised form 10 November 2020; accepted 23 November 2020
Available online 23 December 2020

Abstract

We present a fully discrete approximation technique for the compressible Navier–Stokes equations that is second-order accurate in time and space, semi-implicit, and guaranteed to be invariant domain preserving. The restriction on the time step is the standard hyperbolic CFL condition, i.e. $\tau \lesssim \mathcal{O}(h)/V$ where V is some reference velocity scale and h the typical meshsize.

MSC: 35L65; 65M60; 65M12; 65N30

Keywords: Conservation equations; Hyperbolic systems; Navier–Stokes equations; Euler equations; Invariant domains; High-order method; Convex limiting; Finite element method

1. Introduction

The objective of this paper is to present a fully-discrete approximation technique for the compressible Navier–Stokes equations that is implicit–explicit, second-order accurate in time and space, and guaranteed to be invariant domain preserving. The restriction on the time-step size is the standard hyperbolic CFL condition, i.e., $\tau \lesssim \mathcal{O}(h)/V$, where V is some reference velocity scale and h is the typical meshsize. To the best of our knowledge, this method is the first one that is guaranteed to be invariant domain preserving under the standard hyperbolic CFL condition and be second-order accurate in time and space.

Of course there are countless papers in the literature describing techniques to approximate the time-dependent compressible Navier–Stokes equations, but there are very few papers establishing invariant domain properties. Among the latest results in this direction we refer the reader to Grapsas et al. [1] where a first-order method using

☆ This material is based upon work supported in part by the National Science Foundation, USA grants DMS 1619892, DMS 1620058 and DMS 1912847, by the Air Force Office of Scientific Research, USA, USAF, under grant/contract number FA9550-18-1-0397, and by the Army Research Office, USA under grant/contract number W911NF-15-1-0517.

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

[1] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This document describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

https://doi.org/10.1016/j.cma.2020.113608
0045-7825/

upwinding and staggered grid is developed (see Eq. (3.1) therein). The authors prove positivity of the density and the internal energy (Lem. 4.4 therein). Unconditional stability is obtained by solving a nonlinear system involving the mass conservation equation and the internal energy equation. One important aspect of this method is that it is robust in the low Mach regime. A similar technique is developed in Gallouët et al. [2] for the compressible barotropic Navier–Stokes equations (see §3.6 therein). We also refer to Zhang [3] where a fully explicit dG scheme is proposed with positivity on the internal energy enforced by limiting. The invariant domain properties are proved there under the parabolic time step restriction $\tau \lesssim \mathcal{O}(h^2)/\mu$, where μ is some reference viscosity scale.

The key idea of the present paper is to build on [4,5] and use an operator splitting technique to treat separately the hyperbolic part and the parabolic part of the problem. The hyperbolic sub-step is treated explicitly and the parabolic sub-step is treated implicitly. This idea is not new and we refer for instance to Demkowicz et al. [6] for an early attempt in this direction. The novelty of our approach is that each sub-step is guaranteed to be invariant domain preserving. In addition, the scheme is conservative and fully-computable (e.g. the method is fully-discrete and there are no open-ended questions regarding the solvability of the sub-problems). One key ingredient of our method is that the parabolic sub-step is reformulated in terms of the velocity and the internal energy in a way that makes the method conservative, invariant domain preserving, and second-order accurate (see Section 5).

The remainder of the paper is organized as follows. We recall the compressible Navier–Stokes model and introduce the notation in Section 2. The overall principle of the method is summarized in Section 3.3. As usual, the devil is in the details: we discuss technical aspects of the hyperbolic substep and the parabolic substep in Section 4 and Section 5, respectively. The key results of the two sections are Theorems 4.2 and 5.5. We discuss the full method in Section 6. The main statement summarizing the results of the paper is Theorem 6.1. The method is illustrated numerically in Section 7. Some conclusions and open problems are reported in Section 8.

2. The compressible Navier–Stokes equation

In this section we define the notation and recall the Navier–Stokes equations.

2.1. Notation

The fluid occupies a bounded, polyhedral domain D in $\mathbb{R}^d$. The space dimension d is either 2 or 3 for simplicity. The dependent variable is $\boldsymbol{u} := (\rho, \boldsymbol{m}, E)^\intercal \in \mathbb{R}^{d+2}$, where ρ is the density, $\boldsymbol{m}$ the momentum, 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.

Given some Lipschitz flux $\mathbb{f} : \mathbb{R}^{d+2} \to \mathbb{R}^{(d+2)\times d}$, $\mathbb{f}(\boldsymbol{u}(\boldsymbol{x}))$ is a matrix with entries $\mathbb{f}_{ij}(\boldsymbol{u}(\boldsymbol{x}))$, $1 \le i \le d+2$, $1 \le j \le d$ and $\nabla\cdot\mathbb{f}(\boldsymbol{u}(\boldsymbol{x}))$ is a column vector with entries $(\nabla\cdot\mathbb{f}(\boldsymbol{u}))_i = \sum_{1\le j\le d}\partial_{x_j}\mathbb{f}_{ij}(\boldsymbol{u}(\boldsymbol{x}))$. For any $\boldsymbol{n} = (n_1 \dots, n_d)^\intercal \in \mathbb{R}^d$, we denote by $\mathbb{f}(\boldsymbol{u})\boldsymbol{n}$ the column vector with entries $\sum_{1\le l\le d}\mathbb{f}_{il}(\boldsymbol{u})n_l$, where $i \in \{1{:}d+2\}$. Given two integers $m \le n$, the symbol $\{m{:}n\}$ represents the set of integers $\{m, m+1, \dots, n\}$. Given two second-order tensors $\mathbb{s}$ and $\mathbb{e}$ in $\mathbb{R}^{d\times d}$, we denote the full tensor contraction operation by $\mathbb{s}{:}\mathbb{e} := \sum_{i,j\in\{1:d\}}\mathbb{s}_{ij}\mathbb{e}_{ij}$. As usual $\boldsymbol{a}\cdot\boldsymbol{b} := \sum_{i\in\{1:d\}}a_ib_i$ denotes the Euclidean inner-product in $\mathbb{R}^d$, and $\boldsymbol{a}\otimes\boldsymbol{b}$ is the second-order tensor with entries $(a_ib_j)_{i,j\in\{1:d\}}$. For any smooth vector field $\boldsymbol{a} : D \mapsto \mathbb{R}^d$, $\nabla\boldsymbol{a}$ is the second-order tensor with entries $(\partial_j a_i)_{i,j\in\{1:d\}}$. The Euclidean norm in $\mathbb{R}^d$ and the Frobenius norm in $\mathbb{R}^{d\times d}$ are denoted by $\|\cdot\|_{\ell^2}$.

2.2. Model description

Given some initial time t_0 with initial data $\boldsymbol{u}_0 := (\rho_0, \boldsymbol{m}_0, E_0)$, we look for $\boldsymbol{u}(t) := (\rho, \boldsymbol{m}, E)(t)$ solving the compressible Navier–Stokes system in some weak sense:

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

$$\partial_t\boldsymbol{m} + \nabla\cdot\big(\boldsymbol{v}\otimes\boldsymbol{m} + p(\boldsymbol{u})\mathbb{I} - \mathbb{s}(\boldsymbol{v})\big) = \boldsymbol{f}, \tag{2.1b}$$

$$\partial_t E + \nabla\cdot\big(\boldsymbol{v}(E + p(\boldsymbol{u})) - \mathbb{s}(\boldsymbol{v})\boldsymbol{v} + \boldsymbol{k}(\boldsymbol{u})\big) = \boldsymbol{f}\cdot\boldsymbol{v}, \tag{2.1c}$$

where $p(\boldsymbol{u})$ is the pressure, $\mathbb{I} \in \mathbb{R}^{d\times d}$ is the identity matrix, $\boldsymbol{f}$ is a prescribed external force, $\mathbb{s}(\boldsymbol{v})$ is the viscous stress tensor and $\boldsymbol{k}(\boldsymbol{u})$ is the heat-flux. We assume that the fluid is Newtonian and that the heat-flux follows Fourier's law, that is to say:

$$\mathbb{s}(\boldsymbol{v}) := 2\mu\mathbb{e}(\boldsymbol{v}) + (\lambda - \tfrac{2}{3}\mu)\nabla\cdot\boldsymbol{v}\mathbb{I}, \qquad \mathbb{e}(\boldsymbol{v}) := \nabla^{\mathrm{s}}\boldsymbol{v} := \tfrac{1}{2}\big(\nabla\boldsymbol{v} + (\nabla\boldsymbol{v})^\intercal\big),$$

$$\boldsymbol{k}(\boldsymbol{u}) := -c_v^{-1}\kappa\nabla e.$$

The constants $\mu > 0$ and $\lambda \geq 0$ are the shear and the bulk viscosities, respectively. The constant κ is the thermal conductivity and c_v is the heat capacity at constant volume. We will assume throughout that the coefficient $c_v^{-1}\kappa$ is constant and does not depend on the state $\boldsymbol{u}(t)$.

For the sake of completeness we recall the following standard result regarding the viscous stress tensor $\mathbb{s}(\boldsymbol{v})$.

Lemma 2.1. *Let $k := \max(0, \frac{d}{3}(1 - \frac{3\lambda}{2\mu})) \in [0, 1)$. Then the following holds true for all smooth vector fields $\boldsymbol{v}$ in $\mathbb{R}^d$:*

$$\mathbb{s}(\boldsymbol{v}):\nabla\boldsymbol{v} \geq 2\mu(1-k)\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2. \tag{2.2}$$

Proof. We have $\mathbb{s}(\boldsymbol{v}):\nabla\boldsymbol{v} = 2\mu\nabla^{\mathrm{s}}\boldsymbol{v}:\nabla^{\mathrm{s}}\boldsymbol{v} + (\lambda - \frac{2}{3}\mu)(\nabla\cdot\boldsymbol{v})^2$ and

$$\nabla^{\mathrm{s}}\boldsymbol{v}:\nabla^{\mathrm{s}}\boldsymbol{v} = \sum_{i,j\in\{1:d\}} |\mathbb{e}(\boldsymbol{v})_{ij}|^2 \geq \sum_{i\in\{1:d\}} |\mathbb{e}(\boldsymbol{v})_{ii}|^2 = \sum_{i\in\{1:d\}} |\partial_i \boldsymbol{v}_i|^2 \geq \tfrac{1}{d}(\nabla\cdot\boldsymbol{v})^2.$$

The result follows readily. □

We assume that the pressure $p(\boldsymbol{u})$ is derived from a complete equation of state. That is to say, introducing the specific volume $v := \rho^{-1}$, there exists a specific entropy $\sigma(v, e)$ where $\sigma : \mathbb{R}^+\times\mathbb{R}^+ \to \mathbb{R}$ is concave. We assume that the differential of $\sigma(v, e)$ is consistent with the Gibbs identity $T\,\mathrm{d}\sigma = \mathrm{d}e + p\,\mathrm{d}v$; therefore, setting $s(\rho, e) := \sigma(v, e)$, we have $T^{-1} := \frac{\partial s}{\partial e}$, $p := -\rho^2 T\frac{\partial s}{\partial \rho}$, see Menikoff and Plohr [7], Harten et al. [8] for more details.

The *admissible set* of (2.1) is

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

This is to say, we expect any reasonable solution $\boldsymbol{u}(t)$ of (2.1) to stay in $\mathcal{A}$. Following the terminology of Chueh et al. [9] we say that $\mathcal{A}$ is an invariant domain of (2.1). Important properties we want to maintain at the discrete level are thus the positivity of the density $\rho \geq 0$ and the positivity of the specific internal energy $e(\boldsymbol{u}) = \rho^{-1}E - \frac{1}{2}\|\boldsymbol{v}\|_{\ell^2}^2$.

That the pressure $p(\boldsymbol{u})$ is defined by a complete equation of state is essential for the splitting technique that we are going to used later. We insist again that the source term $\boldsymbol{f}$ is assumed to be prescribed. If $\boldsymbol{f}$ were to depend on the density (which would be the case for gravity in a star) or on the temperature (which would be the case of the gray-radiation equations), then the handling of the source term would have to be modified accordingly and this would entail additional difficulties. This type of problem is out of the scope of the present paper.

We conclude the section by briefly commenting on boundary conditions for system (2.1). For the sake of simplicity and to avoid analytical technicalities we assume that the no-slip and the thermally insulating boundary conditions are enforced on the entire boundary ∂D:

$$\boldsymbol{v}_{|\partial D} = \boldsymbol{0}, \qquad \boldsymbol{k}(\boldsymbol{u})\cdot\boldsymbol{n}_{|\partial D} = 0. \tag{2.4}$$

Notice that (2.4) closes the system (2.1), i. e., no further boundary condition has to be enforced. We refer the reader to [4, §3.5], as well as Sections 4 and 5 for additional details. In principle it is possible to enforce numerous other boundary conditions. A careful analysis of all of these alternative boundary conditions is beyond the scope of the present paper.

3. Strang splitting and stability properties of the hyperbolic and parabolic limits

We will separate the parabolic part and the hyperbolic part of the compressible Navier–Stokes system (2.1) by using Strang's splitting. To this end, we first identify a hyperbolic (Section 3.1) and a parabolic (Section 3.2) limit, then define the corresponding continuous solution operators S_1 and S_2, and finally identify associated stability properties. Both operators are then combined to form a solution operator for (2.1); see Section 3.3. We make no claim of originality about the operator splitting technique. The idea is not new and has been applied in the context of the compressible Navier–Stokes equation by Demkowicz et al. [6] among others. The novel contribution of the present work is the following:

(i) The construction of discrete solution operators $S_{1,h}$ and $S_{2,h}$ that when sequentially compounded yield conservation, preservation of the invariant domain properties of the continuous operators (stated Assumptions 3.1 and 3.2 in Sections 3.1 and 3.2), and satisfaction of a discrete energy balance.

(ii) Specific choice of transformation of variables at the intermediate step making the analysis and an efficient implementation possible.

3.1. Hyperbolic limit

The first asymptotic limit of (2.1) that we discuss is the vanishing viscosity limit, i.e., $\mu, \lambda \to 0$, with vanishing external forces $\boldsymbol{f}$. In this case the governing equations for $\boldsymbol{u}(t)$ reduce to

$$\partial_t \rho + \nabla\cdot(\boldsymbol{v}\rho) = 0, \tag{3.1a}$$
$$\partial_t \boldsymbol{m} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{m} + p(\boldsymbol{u})\mathbb{I}) = \boldsymbol{0}, \tag{3.1b}$$
$$\partial_t E + \nabla\cdot(\boldsymbol{v}(E + p(\boldsymbol{u}))) = 0, \tag{3.1c}$$
$$\boldsymbol{v}\cdot\boldsymbol{n}_{|\partial D} = 0. \tag{3.1d}$$

Here, in the vanishing viscosity limit, the no-slip boundary condition (2.4) is replaced by the slip condition (3.1d). We assume in the following that there exists some Banach space $\mathcal{B}_1$ with sufficient smoothness so that, provided $\boldsymbol{u}_0 \in \mathcal{B}_1 \cap \mathcal{A}$, some reasonable notion of entropy/viscosity solution of (3.1) can be established for some time interval (t_0, t^*). Giving a precise definition of the functional-space $\mathcal{B}_1$ is beyond the scope of this manuscript and somewhat irrelevant for our purpose. The reader is referred to Lions [10], Feireisl [11] for further insights on this very difficult question. Here, by slight abuse of notation $\mathcal{B}_1 \cap \mathcal{A}$ shall mean $\{\boldsymbol{v} \in \mathcal{B}_1 | \boldsymbol{v}(\boldsymbol{x}) \in \mathcal{A} \text{ for a.e. } \boldsymbol{x} \in D\}$. Let $S_1(\cdot, t_0)$ denote the solution map to (3.1); that is, $S_1(t, t_0)(\boldsymbol{u}_0) = \boldsymbol{u}(t)$ for a.e. $t \in (t_0, t^*)$. We introduce a stability notion for the solution map $S_1(\cdot, t_0)$:

Assumption 3.1 (*Stable Hyperbolic Solution Operator*)**.** Let $\boldsymbol{u}_0 \in \mathcal{B}_1 \cap \mathcal{A}$. Recalling that s denotes the specific entropy, we set $s_{\min} := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} s(\rho_0(\boldsymbol{x}), e(\boldsymbol{u}_0(\boldsymbol{x})))$ and introduce the set:

$$\mathcal{C}(\boldsymbol{u}_0) = \left\{\boldsymbol{u} = (\rho, \boldsymbol{m}, E) | \rho > 0,\ e > 0,\ s(e, \rho) \geq s_{\min}\right\}. \tag{3.2}$$

We make the following assumptions:

(i) The set $\mathcal{C}(\boldsymbol{u}_0)$ is invariant under $S_1(., t_0)$ for all $\boldsymbol{u}_0 \in \mathcal{A} \cap \mathcal{B}_1$, i.e., we have $S_1(t, t_0)(\boldsymbol{u}_0)(\boldsymbol{x}) \in \mathcal{C}(\boldsymbol{u}_0)$ for a.e. $\boldsymbol{x} \in D$ and a.e. $t \in (t_0, t^*)$. We say $\mathcal{C}(\boldsymbol{u}_0)$ is an invariant domain of (3.1).

(ii) There exists a family of entropy pairs $(\eta, \boldsymbol{q})$ (for instance a subset of generalized entropies, cf. Harten et al. [8]) such that the following inequality holds in the distribution sense in $D\times(t_0, t^*)$:

$$\partial_t \eta(S_1(t, t_0)(\boldsymbol{u}_0)) + \nabla\cdot(\boldsymbol{q}(S_1(t, t_0)(\boldsymbol{u}_0))) \leq 0.$$

3.2. Parabolic limit

The second asymptotic regime of interest in this manuscript is the diffusive or parabolic regime. The limit is formally obtained by assuming dominant diffusive terms and dominant external forces in (2.1). Then, the governing equations for $\boldsymbol{u}(\boldsymbol{x}, t)$ reduce to

$$\partial_t \rho = 0, \tag{3.3a}$$
$$\partial_t \boldsymbol{m} - \nabla\cdot(\mathbb{s}(\boldsymbol{v})) = \boldsymbol{f}, \tag{3.3b}$$
$$\partial_t E + \nabla\cdot(\boldsymbol{k}(\boldsymbol{u}) - \mathbb{s}(\boldsymbol{v})\boldsymbol{v}) = \boldsymbol{f}\cdot\boldsymbol{v}, \tag{3.3c}$$
$$\boldsymbol{v}_{|\partial D} = \boldsymbol{0}, \qquad \boldsymbol{k}(\boldsymbol{u})\cdot\boldsymbol{n}_{|\partial D} = 0. \tag{3.3d}$$

Since (3.3a) implies $\rho(\boldsymbol{x}, t) = \rho_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$, (3.3b) is equivalent to $\rho\partial_t\boldsymbol{v} - \nabla\cdot(\mathbb{s}(\boldsymbol{v})) = \boldsymbol{f}$. Taking the dot product of (3.3b) and $\boldsymbol{v}$ and subtracting the result from (3.3c) gives $\partial_t(E - \frac{1}{2}\rho\boldsymbol{v}^2) + \nabla\cdot\boldsymbol{k}(\boldsymbol{u}) - \mathbb{s}(\boldsymbol{v}):\nabla\boldsymbol{v} = 0$. Consequently, (3.3) is equivalent to solving

$$\rho_0\partial_t\boldsymbol{v} - \nabla\cdot(\mathbb{s}(\boldsymbol{v})) = \boldsymbol{f}, \qquad \boldsymbol{v}_{|\partial D} = \boldsymbol{0}, \tag{3.4a}$$
$$\rho_0\partial_t e - c_v^{-1}\kappa\Delta e = \mathbb{s}(\boldsymbol{v}):\mathbb{e}(\boldsymbol{v}), \qquad \partial_n e = 0, \tag{3.4b}$$
$$E := \rho_0 e + \tfrac{1}{2}\rho_0\boldsymbol{v}^2. \tag{3.4c}$$

Notice that $\partial_t \int_D E \,\mathrm{d}x = \int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x$; i.e., the variation of the total energy is equal to the power of the external sources. Existence and uniqueness of (3.4) can be established via standard parabolic solution theory, Gilbarg and Trudinger [12]. For the sake of argument we will simply assume that there exists two Banach spaces $\mathcal{B}_2$ and $\mathcal{B}_3$ such that the above problem is well-posed for all $\boldsymbol{u}_0 \in \mathcal{B}_2$ and all $\boldsymbol{f} \in \mathcal{B}_3$. Similarly to the hyperbolic case, we introduce the solution map $S_2(t, t_0)(\boldsymbol{u}_0, \boldsymbol{f}) = \boldsymbol{u}(t)$ to (3.3). Although the following assumption could easily be formulated rigorously in form of a theorem by specifying $\mathcal{B}_2$ and $\mathcal{B}_3$, we prefer to make it an assumption to stay general and avoid distracting technicalities.

Assumption 3.2 (*Stable Parabolic Solution Operator*)**.** Let $\boldsymbol{u}_0 \in \mathcal{A} \cap \mathcal{B}_2$ and $\boldsymbol{f} \in \mathcal{B}_3$. We define $e_{\min} = \operatorname{ess\,inf}_{\boldsymbol{x}\in D} e(\boldsymbol{u}_0(\boldsymbol{x}))$ and set

$$\mathcal{D}(\boldsymbol{u}_0) := \big\{\boldsymbol{u} = (\rho, \boldsymbol{m}, E) | \rho > 0,\ \ e \geq e_{\min}\big\}. \tag{3.5}$$

By possibly making t^* smaller we assume that:

(i) The set $\mathcal{D}(\boldsymbol{u}_0)$ is invariant under $S_2(., t_0)$ for all $\boldsymbol{u}_0 \in \mathcal{A} \cap \mathcal{B}_2$ and all $\boldsymbol{f} \in \mathcal{B}_3$, i.e., $S_2(t, t_0)(\boldsymbol{u}_0, \boldsymbol{f})(\boldsymbol{x}) \in \mathcal{D}(\boldsymbol{u}_0)$ for a.e. $\boldsymbol{x} \in D$ and a.e. $t \in (t_0, t^*)$. We say $\mathcal{D}(\boldsymbol{u}_0)$ is an invariant domain for (3.3).

(ii) The functional setting defining $S_2(t, t_0)$ is smooth enough such that

$$\int_D E(t)\,\mathrm{d}x = \int_D E(t_0)\,\mathrm{d}x + \int_{t_0}^{t}\int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x. \tag{3.6}$$

Our goal in the remainder of the paper is to construct a space and time approximation that is formally second-order accurate and complies in some reasonable sense with the stability properties stated in Assumption 3.1 and in Assumption 3.2.

Remark 3.3 (*Vacuum*)**.** In this paper we assume that no vacuum forms. It has been established in Hoff and Serre [13, Thm. 2] that the compressible Navier–Stokes equation may lose continuous dependency with respect to the initial data when vacuum occurs. It is shown therein that one can construct initial data in one dimension such that continuous dependency is actually lost. □

Remark 3.4 (L^p *Estimates*)**.** Using $\rho > 0$ and the entropy $\eta(\boldsymbol{u}) = \rho$ in Assumption 3.1 we infer the estimate $\|\rho\|_{L^\infty(t_0,t^*;L^1(D))} \leq \|\rho_0\|_{L^\infty(t_0,t^*;L^1(D))}$. Using $\rho > 0, e > 0$, (3.6) implies $\|\rho e\|_{L^\infty(t_0,t^*;L^1(D))} + \frac{1}{2}\|\rho \boldsymbol{v}^2\|_{L^\infty(t_0,t^*;L^1(D))} = \|\rho_0 e_0\|_{L^1(D)} + \frac{1}{2}\|\rho_0 \boldsymbol{v}_0^2\|_{L^1(D)} + \int_{t_0}^{t}\int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x$. □

3.3. Stability of Strang splitting

We propose to approximate (2.1) in time by using Strang's operator splitting. To be able to do that without going too much into the functional analysis details, we add one more assumption which can always be shown to hold true if $\boldsymbol{u}_0$ is smooth enough and t^* is small enough.

Assumption 3.5 (*Smoothness Compatibility*)**.** The following holds true for a.e. $t \in (t_0, t^*)$:

(i) For all $\boldsymbol{u}_0 \in \mathcal{B}_1 \cap \mathcal{A}$, $S_1(t, t_0)(\boldsymbol{u}_0) \in \mathcal{B}_2$.

(ii) For all $\boldsymbol{u}_0 \in \mathcal{B}_2 \cap \mathcal{A}$ and all $\boldsymbol{f} \in \mathcal{B}_3$, $S_2(t, t_0)(\boldsymbol{u}_0, \boldsymbol{f}) \in \mathcal{B}_1$.

Let $\tau \in (0, t^* - t_0]$ be some time step and let $\boldsymbol{u}_0 \in \mathcal{B}_1 \cap \mathcal{A}$ be some admissible initial data at time t_0. The version of Strang's splitting technique we consider in this paper consists of approximating the solution to (2.1) at $t := t_0 + \tau$ as follows:

$$S_1(t_0 + \tau, t_0 + \tfrac{1}{2}\tau) \circ S_2(t_0 + \tau, t_0) \circ (S_1(t_0 + \tfrac{1}{2}\tau, t_0)(\boldsymbol{u}_0), \boldsymbol{f}). \tag{3.7}$$

The above operations are well-posed by virtue of Assumption 3.5. The following result is elementary but is essential since it is the template for the approximation technique that we propose.

Lemma 3.6. *The following holds true for all $\boldsymbol{u}_0 \in \mathcal{B}_1 \cap \mathcal{A}$, all $\boldsymbol{f} \in \mathcal{B}_3$, all $\tau \in (0, t^* - t_0]$, and a.e. $\boldsymbol{x} \in D$:*

$$S_1(t_0 + \tau, t_0 + \tfrac{1}{2}\tau) \circ S_2(t_0 + \tau, t_0) \circ (S_1(t_0 + \tfrac{1}{2}\tau, t_0)(\boldsymbol{u}_0), \boldsymbol{f})(\boldsymbol{x}) \in \mathcal{A}.$$

Proof. By Assumption 3.1(i) and of Assumption 3.5(i) we have $S_1(t_0+\frac{1}{2}\tau, t_0)(\boldsymbol{u}_0) \in \mathcal{B}_2 \cap \mathcal{C}(\boldsymbol{u}_0) \subset \mathcal{B}_2 \cap \mathcal{A}$. Similarly, by Assumptions 3.2(i) and 3.5(ii) it follows that $S_2(t_0+\tau, t_0) \circ (S_1(t_0+\frac{1}{2}\tau, t_0)(\boldsymbol{u}_0), \boldsymbol{f}) \in \mathcal{B}_1 \cap \mathcal{D}(\boldsymbol{u}_0) \subset \mathcal{B}_1 \cap \mathcal{A}$. Finally, the result follows by repeating the first argument. □

We now discuss the space and time approximation of the evolution operators S_1 and S_2. The two key difficulties to overcome are to ensure that $\mathcal{C}(\boldsymbol{u}_0)$ remains invariant under the fully discrete version of S_1, and $\mathcal{D}(\boldsymbol{u}_0)$ remains invariant under the fully discrete version of S_2. We describe the discretization of the hyperbolic step (3.1) in Section 4, then we describe the discretization of the parabolic step (3.3) in Section 5.

4. Explicit hyperbolic step

In this section we describe the discrete setting that is used to approximate (3.1). The reader who is familiar with the theory developed in Guermond et al. [4,5] is invited to skip this section and move on to Section 5.

4.1. Discrete setting for the space approximation

For the explicit hyperbolic step we use the exact same setting as described in [4,5]. The method is discretization agnostic and can be implemented with finite volumes, discontinuous finite elements, and continuous finite elements. To avoid technicalities when approximating the parabolic problem, we are going to restrict the presentation to continuous finite elements. We assume to have at hand a sequence of shape-regular meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$, where $\mathcal{H}$ is the index set of the sequence. One may think of h as being the typical mesh-size. Given some mesh $\mathcal{T}_h$, we denote by $P(\mathcal{T}_h)$ a scalar-valued finite element space with basis functions $\{\varphi_i\}_{i\in\mathcal{V}}$. We assume that $P(\mathcal{T}_h) \subset C^0(\overline{D};\mathbb{R})$. We restrict ourselves to continuous Lagrange finite elements for the sake of simplicity and we assume that $\varphi_i \geq 0$ for all $i \in \mathcal{V}$. We denote by $\mathcal{V}^\partial$ the set of the degrees of freedom that are located on the boundary ∂D. The set $\mathcal{V}^\circ$ is composed of all the interior degrees of freedom. We introduce the vector-valued approximation space $\boldsymbol{P}(\mathcal{T}_h) := (P(\mathcal{T}_h))^{d+2}$. We set

$$m_{ij} = \int_D \varphi_i\varphi_j \,\mathrm{d}x, \quad \boldsymbol{c}_{ij} = \int_D \varphi_i \nabla\varphi_j \,\mathrm{d}x, \quad \boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}, \quad m_i = \int_D \varphi_i \,\mathrm{d}x.$$

The definitions of the coefficients m_{ij}, $\boldsymbol{c}_{ij}$ and m_i for the case of finite volumes and discontinuous finite element discretizations can be found in [5, §4].

4.2. Hyperbolic update

Let t_n be some time and $\boldsymbol{u}^n := \boldsymbol{u}(t_n)$. We now explain how we approximate the update $S_1(t_{n+1}, t_n)(\boldsymbol{u}^n)$. First, let $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{V}} \mathsf{U}_i^n\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a corresponding finite element approximation of $\boldsymbol{u}^n$. We assume that $\boldsymbol{u}_h^n$ is an admissible state, i.e.,

$$\mathsf{U}_i^n \in \mathcal{A}, \qquad \forall i \in \mathcal{V}.$$

Let τ be the current time step size and set $t_{n+1} := t_n + \tau$. Note that τ has to be chosen for each time step t_n subject to a suitable hyperbolic CFL condition; see (4.3)–(4.4) and Theorem 4.2. We now construct an approximation $\boldsymbol{u}_h^{n+1} := \sum_{i\in\mathcal{V}} \mathsf{U}^{n+1}\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ for the new time step t_{n+1} by combining a low-order approximation and a high-order approximation through a convex limiting technique described in [4,5].

The low order update is obtained as follows:

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

where $d_{ij}{}^{\mathrm{L},n}$ is defined by

$$d_{ij}^{L,n} := \max\big(\widehat{\lambda}_{\max}(\boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \widehat{\lambda}_{\max}(\boldsymbol{n}_{ji}, \mathsf{U}_j^n, \mathsf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}\big). \tag{4.1}$$

Here, $\widehat{\lambda}_{\max}(\boldsymbol{n}, \mathsf{U}_L, \mathsf{U}_R)$ is any upper bound on the maximum wave speed in the Riemann problem with left data U_i^n, right data U_j^n, and flux $\mathbb{f}(\boldsymbol{v})\boldsymbol{n}_{ij}$. One can use for instance the two rarefaction approximation discussed in Guermond

and Popov [14, Lem. 4.3] (see also Toro [15, Eq. (4.46)]) or any other guaranteed upper bound. For all $j \in \mathcal{I}(i)\backslash\{i\}$ we introduce the auxiliary states

$$\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_i^n + \mathsf{U}_j^n) - (\mathbb{f}(\mathsf{U}_j^n) - \mathbb{f}(\mathsf{U}_i^n))\frac{\boldsymbol{c}_{ij}}{2{d_{ij}}^{\mathrm{L},n}}. \tag{4.2}$$

The following statement is a key result on which the convex limiting strategy is based.

Lemma 4.1 (*Invariance of the Auxiliary States*)**.** *Let $\mathcal{U} \subset \mathcal{A}$ be any convex invariant domain for (3.1) such that $\mathsf{U}_i^n, \mathsf{U}_j^n \in \mathcal{U}$. Then the state $\overline{\mathsf{U}}_{ij}^n$ defined in (4.2) with ${d_{ij}}^{\mathrm{L},n}$ as defined in (4.1) belongs to $\mathcal{U}$.*

A possibly invariant-domain-violating and formally high-order solution, ${\boldsymbol{u}_h}^{\mathrm{H},n+1}$, is obtained by appropriately reducing the graph viscosity and replacing the lumped mass matrix by the full mass matrix (see, e.g., [4, §3.3-§3.4] and [5, §6]). The final high-order invariant-domain-preserving update $\boldsymbol{u}_h^{n+1}$ is obtained by applying convex limiting between the low-order solution ${\mathsf{U}_i}^{\mathrm{L},n+1}$ and the high-order solution ${\mathsf{U}_i}^{\mathrm{H},n+1}$ with relaxed bounds. The local bounds are computed using the auxiliary states (4.2) (see e.g., [4, §4] and [5, §7]). In the numerical illustrations reported at the end of the paper we limit the density from above and from below and the specific entropy from below. The relaxation technique for the bounds is explained in [4, §4.7] and [5, §7.6]. For further reference we introduce

$$\tau_0(\boldsymbol{u}_h^n) := \min_{i\in\mathcal{V}} \frac{m_i}{2|{d_{ii}}^{\mathrm{L},n}|}, \qquad \text{with} \qquad {d_{ii}}^{\mathrm{L},n} := -\sum_{j\in\mathcal{I}(i)\backslash\{i\}} {d_{ij}}^{\mathrm{L},n}. \tag{4.3}$$

The ratio $\tau/\tau_0(\boldsymbol{u}_h^n)$ is henceforth denoted CFL and called Courant–Friedrichs–Lewy number:

$$\mathrm{CFL} := \frac{\tau}{\tau_0(\boldsymbol{u}_h^n)}. \tag{4.4}$$

Let $S_{1h}(t_n+\tau, t_n) : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ denote the nonlinear operator defined by setting $S_{1h}(t_n+\tau, t_n)(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$. The key result regarding the hyperbolic update is the following.

Theorem 4.2 (*Invariance*)**.** *Let $\boldsymbol{u}_h^n \in \mathcal{A}$ and let $\mathcal{C}(\boldsymbol{u}_h^n)$ be as defined in (3.2).*

(i) *If no relaxation is applied on the entropy bounds, then $S_{1h}(t_n+\tau, t_n)(\boldsymbol{u}_h^n) \in \mathcal{C}(\boldsymbol{u}_h^n)$ for all $\tau \le \tau_0(\boldsymbol{u}_h^n)$. In other words, $\mathcal{C}(\boldsymbol{u}_h^n)$ is invariant under $S_{1h}(t_n+\tau, t_n)$ if* CFL ≤ 1.

(ii) *In case of relaxation of the entropy bounds in the convex limiter, there exists $c(h)$ with $\lim_{h\to 0} c(h) = 1$ and $s_{\min} \ge c(h)s_{\min}$ so that the same statement holds with the constraint $s(\rho, e) \ge s_{\min}$ in (3.2) replaced by $s(\rho, e) \ge c(h)s_{\min}$.*

(iii) *In both cases $\mathcal{A}$ is invariant under $S_{1h}(t_n+\tau, t_n)$ provided that $\tau \le \tau_0(\boldsymbol{u}_h^n)$.*

Remark 4.3 (*Second-Order in Time*)**.** In practice the method is made second-order accurate in time by using a strong stability preserving explicit Runge Kutta method. For instance it is sufficient to use SSPRK(2,2) (i.e., Heun's scheme) to achieve second-order accuracy in time. This is done as follows: one computes $\boldsymbol{w}_h^1 = S_{1h}(t_n+\tau, t_n)(\boldsymbol{u}_h^n)$ and $\boldsymbol{w}_h^2 = S_{1h}(t_n+2\tau, t_n+\tau)(\boldsymbol{w}_h^1)$ and one sets $\boldsymbol{u}_h^{n+1} = \frac{1}{2}\boldsymbol{u}_h^n + \frac{1}{2}\boldsymbol{w}_h^2$.

5. Implicit parabolic step

We now describe the discrete setting that is used to approximate the parabolic step (3.3). We use the same finite element setting that was introduced in Section 4.1.

5.1. Density and velocity update

Let again $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{V}} \mathsf{U}_i^n\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a finite element approximation of $\boldsymbol{u}^n$. We assume that $\boldsymbol{u}_h^n$ is an admissible state, i.e.,

$$\mathsf{U}_i^n \in \mathcal{A}, \qquad \forall i \in \mathcal{V}. \tag{5.1}$$

Let τ be the chosen hyperbolic time step size (see Section 4) for t_n. We now construct an approximation $\boldsymbol{u}_h^{n+1} = \sum_{i\in\mathcal{V}} \mathsf{U}_i^{n+1}\varphi_i$ of $S_2(t_n+\tau, t_n)(\boldsymbol{u}^n, \boldsymbol{f})$ as follows. Since the evolution equation for the density in (3.3) is $\partial_t \rho = 0$, the density is updated by setting

$$\varrho_i^{n+1} := \varrho_i^n, \qquad \forall i \in \mathcal{V}. \tag{5.2}$$

Next, the velocity $\boldsymbol{v}^n$ has to be updated. For this, we introduce the bilinear form associated with viscous dissipation,

$$a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}) : \mathbb{e}(\boldsymbol{w})\,\mathrm{d}x, \qquad \boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{H}_0^1(D) := H_0^1(D;\mathbb{R}^d). \tag{5.3}$$

Let $\{\boldsymbol{e}_k\}_{k\in\{1:d\}}$ be the canonical Cartesian basis of $\mathbb{R}^d$. For any $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$ we define the $d\times d$ matrix $\mathbb{B}_{ij} \in \mathbb{R}^{d\times d}$ by setting

$$(\mathbb{B}_{ij})_{kl} := a(\varphi_j \boldsymbol{e}_l, \varphi_i \boldsymbol{e}_k) := \int_D \mathbb{s}(\varphi_j \boldsymbol{e}_l) : \nabla^{\mathrm{s}}(\varphi_i \boldsymbol{e}_k)\,\mathrm{d}x, \qquad \forall k, l \in \{1:d\}. \tag{5.4}$$

Let $\boldsymbol{f}_h^{n+\frac12} := \sum_{j\in\mathcal{V}} \mathsf{F}_j^{n+\frac12}\varphi_j \in \boldsymbol{P}(\mathcal{T}_h)$ be an approximation of $\boldsymbol{f}(t_n + \frac12\tau)$ (at least second-order accurate in time and space). We use the Crank–Nicolson technique to compute $\boldsymbol{u}_h^{n+1}$. More precisely we solve for the unknown $\mathsf{V}^{n+\frac12}$ given by the following linear system:

$$\begin{cases} \varrho_i^n m_i \mathsf{V}^{n+\frac12} + \frac12\tau \sum_{j\in\mathcal{I}(i)} \mathbb{B}_{ij}\mathsf{V}^{n+\frac12} = m_i \mathsf{M}_i^n + \frac12\tau m_i \mathsf{F}_i^{n+\frac12}, & \forall i \in \mathcal{V}^\circ \\ \mathsf{V}_i^{n+\frac12} = \mathbf{0}, & \forall i \in \mathcal{V}^\partial, \end{cases} \tag{5.5a}$$

where $\mathsf{U}_i^n =: (\varrho_i^n, \boldsymbol{M}_i^n, E_i^n)$, and set

$$\mathsf{V}_i^{n+1} := 2\mathsf{V}^{n+\frac12} - \mathsf{V}_i^n, \qquad \mathsf{M}_i^{n+1} := \varrho_i^{n+1}\mathsf{V}_i^{n+1}, \qquad \forall i \in \mathcal{V}. \tag{5.5b}$$

We then introduce $\boldsymbol{v}_h^{n+\frac12} := \sum_{i\in\mathcal{V}} \mathsf{V}_i^{n+\frac12}\varphi_i$ and define

$$\mathsf{K}_i^{n+\frac12} := \frac{1}{m_i}\int_D \mathbb{s}(\boldsymbol{v}^{n+\frac12}) : \mathbb{e}(\boldsymbol{v}^{n+\frac12})\varphi_i\,\mathrm{d}x, \qquad \forall i \in \mathcal{V}. \tag{5.6}$$

Notice that $\sum_{i\in\mathcal{V}} m_i \mathsf{K}_i^{n+\frac12} = a(\boldsymbol{v}^{n+\frac12}, \boldsymbol{v}^{n+\frac12})$ owing to the partition of unity property. The main properties of the above definitions are summarized in the following result.

Lemma 5.1 (*Velocity Update*)**.** (i) *For every $i \in \mathcal{V}$ we have $\mathsf{K}_i^{n+\frac12} \geq 0$.*
(ii) *The following global energy balance holds true:*

$$\sum_{i\in\mathcal{V}} \tfrac12 m_i \varrho_i^n (\mathsf{V}_i^{n+1})^2 + \tau a(\boldsymbol{v}^{n+\frac12}, \boldsymbol{v}^{n+\frac12}) = \sum_{i\in\mathcal{V}} \tfrac12 m_i \varrho_i^n (\mathsf{V}_i^n)^2 + \sum_{i\in\mathcal{V}} \tau m_i \mathsf{F}_i^{n+\frac12}\cdot\mathsf{V}_i^{n+\frac12}. \tag{5.7}$$

Proof. (i) The inequality $\mathsf{K}_i^{n+\frac12} \geq 0$ is a consequence of (2.2) and $\varphi_i \geq 0$. (ii) We take the dot product of (5.5a) with $2\mathsf{V}_i^{n+\frac12}$ and recalling that $\mathsf{V}^{n+\frac12} = \frac12(\mathsf{V}_i^{n+1} + \mathsf{V}_i^n)$ we obtain for every $i \in \mathcal{V}^\circ$

$$\tfrac12 m_i \varrho_i^n (\mathsf{V}_i^{n+1})^2 + \tau a(\boldsymbol{v}^{n+\frac12}, \mathsf{V}_i^{n+\frac12}\varphi_i) = \tfrac12 m_i \varrho_i^n (\mathsf{V}_i^n)^2 + \tau m_i \mathsf{F}_i^{n+\frac12}\cdot\mathsf{V}_i^{n+\frac12}.$$

For every $i \in \mathcal{V}^\partial$ we have $\mathsf{V}_i^{n+\frac12} = \mathbf{0}$, which in turn implies that $\mathsf{V}_i^{n+1} = -\mathsf{V}_i^n$, i.e., $(\mathsf{V}_i^{n+1})^2 = (\mathsf{V}_i^n)^2$. Moreover, we have $a(\boldsymbol{v}^{n+\frac12}, \mathsf{V}_i^{n+\frac12}\varphi_i) = 0$ and $\mathsf{F}_i^{n+\frac12}\cdot\mathsf{V}_i^{n+\frac12} = 0$. Hence, for every $i \in \mathcal{V}^\partial$ we have

$$\tfrac12 m_i \varrho_i^n (\mathsf{V}_i^{n+1})^2 + \tau a(\boldsymbol{v}^{n+\frac12}, \mathsf{V}_i^{n+\frac12}\varphi_i) = \tfrac12 m_i \varrho_i^n (\mathsf{V}_i^n)^2 + \tau m_i \mathsf{F}_i^{n+\frac12}\cdot\mathsf{V}_i^{n+\frac12}.$$

Summing over $i \in \mathcal{V}$ and using the partition of unity property ($\sum_{i\in\mathcal{V}} \varphi_i = 1$) yields (5.7). □

Remark 5.2 (*Approximation Order*)**.** The update V_i^{n+1} constructed by (5.5) is formally second-order accurate in time and space since (5.5a) is a Crank–Nicolson time step.

5.2. Internal energy update (first-order)

The update of the internal energy entails some subtleties regarding the minimum principle when using the second-order Crank–Nicolson time stepping. Therefore, we first formulate the method with the backward Euler time stepping. The second-order extension is presented in Section 5.3. Let us introduce the bilinear form associated with the thermal diffusion

$$b(e, w) := c_v^{-1}\kappa \int_D \nabla e\cdot\nabla w \,\mathrm{d}x, \qquad \forall e, w \in H^1(D).$$

For any $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$ we set

$$\beta_{ij} := b(\varphi_j, \varphi_i). \tag{5.8}$$

Notice that the partition of unity property implies that $\beta_{ii} = -\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \beta_{ij}$. This implies in particular that for all $v_h := \sum_{j\in\mathcal{V}} \mathsf{V}_j\varphi_j \in P(\mathcal{T}_h)$ we have

$$b(v_h, \varphi_i) = \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \beta_{ij}(\mathsf{V}_j - \mathsf{V}_i). \tag{5.9}$$

This expression will be useful to prove the minimum principle on the internal energy. We further assume that

$$\beta_{ij} \le 0, \quad \forall i \ne j \in \mathcal{V}. \tag{5.10}$$

This condition is known to be satisfied for meshes composed of simplices in two and three space dimensions under the so-called acute angle condition, cf. e.g., Brandts et al. [16, §5.2], Xu and Zikatanov [17, Eq. (2.5)]. This is in particular true for Delaunay meshes. Although it can be done, it is not the purpose of this paper to relax this condition.

Recalling the viscous dissipation $\mathsf{K}_i^{n+\frac12}$ defined in (5.6), we now construct a low-order update of the internal energy $\mathsf{e}_i^{\mathrm{L},n+1}$ as follows. For all $i \in \mathcal{V}$ first set $\mathsf{e}_i^n := (\varrho_i^n)^{-1}E_i^n - \frac12\|\mathsf{V}_i^n\|_{\ell^2}^2$, then solve the linear system

$$m_i\varrho_i^n(\mathsf{e}_i^{\mathrm{L},n+1} - \mathsf{e}_i^n) + \tau\sum_{j\in\mathcal{I}(i)} \beta_{ij}\mathsf{e}_j^{\mathrm{L},n+1} = \tau m_i\mathsf{K}_i^{n+\frac12}, \qquad \forall i \in \mathcal{V}. \tag{5.11}$$

Recall that the boundary conditions (3.4b) together with the partition of unity property imply that

$$\sum_{i\in\mathcal{V}} m_i\varrho_i^n(\mathsf{e}_i^{\mathrm{L},n+1} - \mathsf{e}_i^n) = \tau\sum_{i\in\mathcal{V}} m_i\mathsf{K}_i^{n+\frac12} = \tau a(\boldsymbol{v}^{n+\frac12}, \boldsymbol{v}^{n+\frac12}). \tag{5.12}$$

This identity is used in the proof of Theorem 5.5.

Lemma 5.3 (*Minimum Principle*). *Let U^n be an admissible state. Then for all $\tau > 0$:*

$$\min_{j\in\mathcal{V}} \mathsf{e}_j^{\mathrm{L},n+1} \ge \min_{j\in\mathcal{V}}(\mathsf{e}_j^n + \tfrac{\tau}{\varrho_j^n}\mathsf{K}_j^{n+\frac12}) \ge \min_{j\in\mathcal{V}} \mathsf{e}_j^n \ge 0.$$

Proof. Recalling that $\sum_{j\in\mathcal{I}(i)} \beta_{ij} = 0$, we infer that

$$m_i\varrho_i^n(\mathsf{e}_i^{\mathrm{L},n+1} - \mathsf{e}_i^n) + \tau\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \beta_{ij}(\mathsf{e}_j^{\mathrm{L},n+1} - \mathsf{e}_i^{\mathrm{L},n+1}) = \tau m_i\mathsf{K}_i^{n+\frac12},$$

Let i be the index in $\mathcal{V}$ where $\mathsf{e}_i^{\mathrm{L},n+1}$ is minimal. Then $0 \ge \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \beta_{ij}(\mathsf{e}_j^{\mathrm{L},n+1} - \mathsf{e}_i^{\mathrm{L},n+1})$ because we have assumed that $\beta_{ij} \le 0$ for all $j \in \mathcal{I}(i)\backslash\{i\}$. Moreover, the definition of $\mathsf{K}_i^{n+\frac12}$ implies that $\mathsf{K}_i^{n+\frac12} \ge 0$ since we assumed $\varphi_i \ge 0$. All this implies that

$$m_i\varrho_i^n(\mathsf{e}_i^{\mathrm{L},n+1} - \mathsf{e}_i^n) \ge m_i\varrho_i^n(\mathsf{e}_i^{\mathrm{L},n+1} - \mathsf{e}_i^n) + \tau\sum_{j\in\mathcal{I}(i)\backslash\{i\}} \beta_{ij}(\mathsf{e}_j^{\mathrm{L},n+1} - \mathsf{e}_i^{\mathrm{L},n+1}) = \tau m_i\mathsf{K}_i^{n+\frac12} \ge 0.$$

In conclusion $\min_{j\in\mathcal{V}} \mathsf{e}_j^{\mathrm{L},n+1} =: \mathsf{e}_i^{\mathrm{L},n+1} \ge \mathsf{e}_i^n + \frac{\tau}{\varrho_i^n}\mathsf{K}_i^{n+\frac12} \ge \min_{j\in\mathcal{V}}\big(\mathsf{e}_j^n + \frac{\tau}{\varrho_j^n}\mathsf{K}_j^{n+\frac12}\big)$. □

5.3. Internal energy update (Second-order)

We now explain how to approximate the internal energy with a second-order Crank–Nicolson time stepping scheme. This is done by combining the low-order update and the second-order update using flux-corrected transport limiting (FCT); the reader is referred to e.g., Boris and Book [18], Zalesak [19], Kuzmin et al. [20].

We start by defining the high-order update of the internal energy, $\mathsf{e}_i{}^{\mathrm{H},n+1}$, as follows: We first compute $\mathsf{e}_i{}^{\mathrm{H},n+\frac{1}{2}}$ by solving

$$m_i\varrho_i^n(\mathsf{e}_i{}^{\mathrm{H},n+\frac{1}{2}}-\mathsf{e}_i^n)+\tfrac{1}{2}\tau\sum_{j\in\mathcal{I}(i)}\beta_{ij}\mathsf{e}_i{}^{\mathrm{H},n+\frac{1}{2}}=\tfrac{1}{2}\tau m_i\mathsf{K}_i^{n+\frac{1}{2}},\qquad\forall i\in\mathcal{V}.\tag{5.13}$$

and then set

$$\mathsf{e}_i{}^{\mathrm{H},n+1}=2\mathsf{e}_i{}^{\mathrm{H},n+\frac{1}{2}}-\mathsf{e}_i^n,\qquad\forall i\in\mathcal{V}.$$

In general, positivity properties for the Crank–Nicolson scheme can only be guaranteed under highly restrictive time-step size constraints. We do not assume that such time-step conditions are met. We just assume that the time-step size is dictated by the CFL constraints of the hyperbolic part. We thus resort to flux-corrected transport limiting, or alternatively convex limiting, to preserve positivity properties. Rewriting (5.13) by multiplying (5.13) by 2 and replacing $\mathsf{e}_i{}^{\mathrm{H},n+\frac{1}{2}}$ by $\frac{1}{2}(\mathsf{e}_i{}^{\mathrm{H},n+1}+\mathsf{e}_i^n)$ gives:

$$m_i\varrho_i^n(\mathsf{e}_i{}^{\mathrm{H},n+1}-\mathsf{e}_i^n)+\tfrac{1}{2}\tau\sum_{j\in\mathcal{I}(i)}\beta_{ij}(\mathsf{e}_j{}^{\mathrm{H},n+1}+\mathsf{e}_j^n)=\tau m_i\mathsf{K}_i^{n+\frac{1}{2}},\qquad\forall i\in\mathcal{V}.\tag{5.14}$$

We then take the difference between (5.11) and (5.14) to obtain

$$m_i\varrho_i^n(\mathsf{e}_i{}^{\mathrm{H},n+1}-\mathsf{e}_i{}^{\mathrm{L},n+1})=-\tfrac{1}{2}\tau\sum_{j\in\mathcal{I}(i)}\beta_{ij}(\mathsf{e}_j{}^{\mathrm{H},n+1}+\mathsf{e}_j^n-2\mathsf{e}_j{}^{\mathrm{L},n+1}).$$

Setting $A_{ij}:=-\frac{1}{2}\tau\beta_{ij}(\mathsf{e}_j{}^{\mathrm{H},n+1}-\mathsf{e}_i{}^{\mathrm{H},n+1}+\mathsf{e}_j^n-\mathsf{e}_i^n-2\mathsf{e}_j{}^{\mathrm{L},n+1}+2\mathsf{e}_i{}^{\mathrm{L},n+1})$, the above identity reads

$$m_i\varrho_i^n(\mathsf{e}_i{}^{\mathrm{H},n+1}-\mathsf{e}_i{}^{\mathrm{L},n+1})=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}A_{ij}.$$

Introducing $\mathsf{e}^{n,\min}:=\min_{j\in\mathcal{V}}\mathsf{e}_j{}^n$ we then define the FCT limiter coefficients as follows:

$$P_i^-:=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\min(A_{ij},0),\qquad Q_i^-:=m_i\varrho_i^n(\mathsf{e}^{n,\min}-\mathsf{e}_i{}^{\mathrm{L},n+1}),\tag{5.15a}$$

$$\ell_i^+=1,\qquad \ell_i^-:=\min\big(1,\tfrac{Q_i^-}{P_i^-}\big).\tag{5.15b}$$

Note that $P_i^-\le 0$ and $Q_i^-\le 0$ (owing to Lemma 5.3), therefore $\ell_i^-\ge 0$. By virtue of the definition of ℓ_i^- the inequality $\ell_i^-P_i^-\ge Q_i^-$ always holds true:

$$\ell_i^-P_i^-=\min\big(1,\tfrac{Q_i^-}{P_i^-}\big)P_i^-=-\min\big(1,\tfrac{Q_i^-}{P_i^-}\big)|P_i^-|=-\min(|P_i^-|,-Q_i^-)\ge Q_i^-\tag{5.16}$$

The high-order update of the internal energy is now defined by setting

$$m_i\varrho_i^n(\mathsf{e}_i{}^{n+1}-\mathsf{e}_i{}^{\mathrm{L},n+1})=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\ell_{ij}A_{ij},\qquad \ell_{ij}:=\begin{cases}\min(\ell_i^+,\ell_j^-), & \text{if } A_{ij}\ge 0,\\ \min(\ell_i^-,\ell_j^+), & \text{if } A_{ij}<0.\end{cases}\tag{5.17}$$

Lemma 5.4 (*Minimum Principle*). *The quantity* e^{n+1} *computed in* (5.17) *satisfies*

$$\min_{j\in\mathcal{V}}\mathsf{e}_j^{n+1}\ge\mathsf{e}^{n,\min}:=\min_{j\in\mathcal{V}}\mathsf{e}_j{}^n.\tag{5.18}$$

Proof. The above definitions imply

$$m_i\varrho_i^n(\mathsf{e}_i{}^{n+1}-\mathsf{e}_i{}^{\mathrm{L},n+1})\ge\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\ell_{ij}\min(A_{ij},0)\ge\ell_i^-\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\min(A_{ij},0)=\ell_i^-P_i^-\ge Q_i^-,$$

where we have used that $\ell_{ij} \le \ell_i^-$, the definition of P_i^-, and the inequality (5.16). This shows that the limiting enforces $m_i \varrho_i^n \mathsf{e}_i^{n+1} \ge m_i \varrho_i^n \mathsf{e}^{n,\min}$, i.e., $\mathsf{e}_i^{n+1} \ge \mathsf{e}^{n,\min}$. This in turn implies that $\min_{i\in\mathcal{V}} \mathsf{e}_i^{n+1} \ge \mathsf{e}^{n,\min} = \min_{j\in\mathcal{V}} \mathsf{e}_j{}^n$. □

5.4. Total energy update

Once the internal energy is updated according to (5.17), the total energy can be updated by setting

$$\mathsf{E}_i^{n+1} = \varrho_i^{n+1}\mathsf{e}_i^{n+1} + \tfrac{1}{2}\varrho_i^n \|\mathsf{V}_i^{n+1}\|_{\ell^2}^2, \qquad \forall i \in \mathcal{V}. \tag{5.19}$$

The main result of Section 5 is the following.

Theorem 5.5 (*Positivity and Conservation*)**.** *Let U^n be an admissible state. Let U^{n+1} be the state constructed by (5.2)–(5.5b)–(5.19), with the velocity update defined in (5.5) and the internal energy update defined in (5.17). Then, U^{n+1} is an admissible state, i.e., $\mathsf{U}_i^{n+1} \in \mathcal{A}$ for all $i \in \mathcal{V}$ and all τ, and the following holds for all $i \in \mathcal{V}$ and all τ:*

$$\varrho_i^{n+1} = \varrho_i^n > 0, \qquad \forall i \in \mathcal{V}, \tag{5.20a}$$

$$\min_{j\in\mathcal{V}} \mathsf{e}_j^{n+1} \ge \min_{j\in\mathcal{V}} \mathsf{e}_j^n > 0, \tag{5.20b}$$

$$\sum_{i\in\mathcal{V}} m_i \mathsf{E}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathsf{E}_i^n + \sum_{i\in\mathcal{V}} \tau m_i \mathsf{F}_i^{n+\frac{1}{2}} \cdot \mathsf{V}_i^{n+\frac{1}{2}}. \tag{5.20c}$$

Proof. (i) Since by assumption $\mathsf{U}_i^n \in \mathcal{A}$, we have $\varrho_i^n > 0$, whence $\varrho_i^{n+1} > 0$.
(ii) We have proved that $\min_{j\in\mathcal{V}} e_j^{n+1} \ge \min_{j\in\mathcal{V}} e_j^n \ge 0$ in Lemma 5.3.
(iii) We have established in (5.7) that

$$\sum_{i\in\mathcal{V}} \tfrac{1}{2} m_i \varrho_i^n (\mathsf{V}_i^{n+1})^2 + \tau a(\boldsymbol{v}^{n+\frac{1}{2}}, \boldsymbol{v}^{n+\frac{1}{2}}) = \sum_{i\in\mathcal{V}} \tfrac{1}{2} m_i \varrho_i^n (\mathsf{V}_i^n)^2 + \sum_{i\in\mathcal{V}} \tau m_i \mathsf{F}_i^{n+\frac{1}{2}} \cdot \mathsf{V}_i^{n+\frac{1}{2}}. \tag{5.21}$$

Recalling that $A_{ij} = -A_{ji}$ and $\ell_{ij} = \ell_{ji}$, we sum (5.17) over $i \in \mathcal{V}$ and obtain

$$\sum_{i\in\mathcal{V}} m_i \varrho_i^n \mathsf{e}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \varrho_i^n \mathsf{e}_i{}^{\mathrm{L},n}.$$

Invoking the identity (5.12) shows

$$\sum_{i\in\mathcal{V}} m_i \varrho_i^n \mathsf{e}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \varrho_i^n \mathsf{e}_i^n + \tau a(\boldsymbol{v}^{n+\frac{1}{2}}, \boldsymbol{v}^{n+\frac{1}{2}}). \tag{5.22}$$

Adding (5.21) and (5.22) gives (5.20c). □

We introduce a discrete nonlinear solution operator $S_{2h}(t_n + \tau, t_n) : \boldsymbol{P}(\mathcal{T}_h) \times \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ by setting $S_{2h}(t_n + \tau, t_n)(\boldsymbol{u}_h^n, \boldsymbol{f}_h^{n+\frac{1}{2}}) := \boldsymbol{u}_h^{n+1}$. Theorem 5.5 can then be rephrased as follows.

Corollary 5.6 (*Invariance*)**.** *Let $\boldsymbol{u}_h \in \boldsymbol{P}(\mathcal{T}_h) \cap \mathcal{A}$ and let $\boldsymbol{f}_h^{n+\frac{1}{2}} \in \boldsymbol{P}(\mathcal{T}_h)$. Then $\mathcal{D}(\boldsymbol{u}_h^n)$ is invariant under $S_{2h}(t_n + \tau, t_n)$ for all τ, i.e., $S_{2h}(t_n + \tau, t_n)(\boldsymbol{u}_h, \boldsymbol{f}_h^{n+\frac{1}{2}}) \in \mathcal{D}(\boldsymbol{u}_h^n) \subset \mathcal{A}$ for all $\tau > 0$.*

Remark 5.7 (*Definition of* $\mathsf{e}^{\min}$)**.** The definition of $\mathsf{e}^{\min}$ in (5.15a) can be slightly strengthened. The lower bound (5.18) holds for any number $\mathsf{e}^{\min}$ chosen in the interval $[\min_{j\in\mathcal{V}} \mathsf{e}_j^n, \min_{j\in\mathcal{V}} \mathsf{e}_j{}^{\mathrm{L},n}]$. However, selecting $\mathsf{e}^{\min}$ too close to $\min_{j\in\mathcal{V}} \mathsf{e}_j{}^{\mathrm{L},n}$ degenerates the accuracy order of the method to $\mathcal{O}(\tau)$ in the $L^\infty(D)$-norm. The numerical experiments reported in the paper are computed with $\mathsf{e}^{\min} := \min_{j\in\mathcal{V}} \mathsf{e}_j^n$.

Remark 5.8 (*Energy*)**.** Lemma 5.4 establishes that the minimum of the internal energy grows monotonically and Theorem 5.5 states that the temporal variation of the total energy is equal to the power of the sources. This implies in essence that a fully discrete counterpart of (3.6) holds true, which is exactly what one should expect.

6. Complete method

We now put all the pieces together and state the main result of the paper. Let $S_{1h}^{(2)}$ be a version of S_{1h} that is at least second-order accurate in time as discussed in Remark 4.3. Let $\boldsymbol{u}_h^n \in \boldsymbol{P}(\mathcal{T}_h)$ be an admissible state and let $\boldsymbol{f}_h^{n+\frac{1}{2}} \in \boldsymbol{P}(\mathcal{T}_h)$. Let us fix some number CFL > 0, which we call Courant–Friedrichs–Lewy number, and let $\tau_0(\boldsymbol{u}_h^n)$ be defined in (4.3). The time step τ is chosen by setting

$$\tau := \mathrm{CFL} \times \tau_0(\boldsymbol{u}_h^n). \tag{6.1}$$

The update $\boldsymbol{u}_h^{n+1} \in \boldsymbol{P}(\mathcal{T}_h)$ is computed as follows:

$$\boldsymbol{u}_h^{n+1} = S_{1h}^{(2)}(t_n+\tau, t_n+\tfrac{1}{2}\tau) \circ S_{2h}(t_n+\tau, t_n) \circ (S_{1h}^{(2)}(t_n+\tfrac{1}{2}\tau, t_n)(\boldsymbol{u}_h^n), \boldsymbol{f}_h^{n+\frac{1}{2}}). \tag{6.2}$$

Theorem 6.1 (*Invariance*)**.** *Let $\boldsymbol{u}_h^n \in \boldsymbol{P}(\mathcal{T}_h) \cap \mathcal{A}$ and $\boldsymbol{f}_h^{n+\frac{1}{2}} \in \boldsymbol{P}(\mathcal{T}_h)$. Then $\boldsymbol{u}_h^{n+1} \in \mathcal{A}$ provided* CFL *is small enough. Moreover, the mass is conserved $\sum_{i\in\mathcal{V}} m_i \varrho_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \varrho_i^n$ and, under the assumption that $\boldsymbol{f} \equiv \boldsymbol{0}$, the total energy is also conserved $\sum_{i\in\mathcal{V}} m_i \mathsf{E}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathsf{E}_i^n$.*

Proof. From Theorem 4.2 we infer that $S_{1h}^{(2)}(t_n+\frac{1}{2}\tau, t_n)(\boldsymbol{u}_h^n) \in \mathcal{A}$ if CFL is small enough. For example, for the SSPRK(2,2) and SSPRK(3,3) methods this holds with CFL $= 2$. From Corollary 5.6 we infer that $\boldsymbol{w}_h := S_{2h}(t_n+\tau, t_n)\big(S_{1h}^{(2)}(t_n+\frac{1}{2}\tau, t_n)(\boldsymbol{u}_h^n, \boldsymbol{f}_h^{n+\frac{1}{2}})\big) \in \mathcal{A}$ without any further restriction on τ. Using again Theorem 4.2 we infer that $S_{1h}^{(2)}(t_n+\tau, t_n+\frac{1}{2}\tau)(\boldsymbol{w}_h) \in \mathcal{A}$ provided $\frac{\tau}{2} \le \tau_0(\boldsymbol{w}_h)$, i.e., CFL $\le 2\tau_0(\boldsymbol{w}_h)/\tau_0(\boldsymbol{u}_h^n)$. □

Remark 6.2 (*CFL*)**.** Showing that Theorem 6.1 holds with a CFL number that is uniform with respect to the mesh size, i.e., $\tau_0(\boldsymbol{w}_h)/\tau_0(\boldsymbol{u}_h^n)$ can be bounded uniformly, would necessitate to prove some uniform bounds on $\boldsymbol{w}_h$. Except under very restrictive smallness assumptions on data, to the best of our knowledge this is a very challenging open problem that is well beyond the scope of the present paper.

7. Numerical illustration

We illustrate the approximation technique with a number of convergence tests and a computation of a shocktube benchmark problem.

7.1. Implementation details

All the tests reported below are done with the ideal gas equation of state, $s(\rho, e) = \log(e^{\frac{1}{\gamma-1}}\rho^{-1})$, with $\gamma = 1.4$. This in turn implies that $p = (\gamma-1)\rho e$, as well as $c_p = \frac{\gamma}{\gamma-1}$, and $c_v = \frac{1}{\gamma-1}$. We also assume that the ratio $\frac{\mu c_p}{\kappa} =: P_r$, called Prandtl number, is constant. Hence $c_v^{-1}\kappa = P_r^{-1}\frac{c_p}{c_v}\mu = \frac{\gamma}{P_r}\mu$. The bulk viscosity λ is set to 0.

All the computations are done with continuous $\mathbb{P}_1$ elements. The high-order method uses the entropy viscosity commutator described in [4, (3.15)–(3.16)] with the entropy ρs. Upper and lower bounds on the density are enforced by using the method described in [4, §4.4]. The relaxation of the bounds on the density is done by using the technique described in [4, §4.7]. The minimum principle on the specific entropy $\exp((\gamma-1)s) \ge \exp((\gamma-1)s^{\min})$ is enforced by proceeding as in [4, §4.6] with the constraint $\Psi(\mathsf{U}) := \rho e - \varrho^{\min}\rho^\gamma \ge 0$. The lower bound on the specific entropy for all $i \in \mathcal{V}$ is set with $\varrho_i^{\min} := \min_{j\in\mathcal{I}(i)} \rho_i^n e_i^n/(\rho_i^n)^\gamma$ and further relaxed by using [4, Eq. (4.14)]. The positivity of the internal energy is guaranteed by the minimum principle on the specific entropy, i.e., no limiting on the internal energy is done. High-performance implementations of the hyperbolic solver are available in form of open source software documented in Maier and Kronbichler [21], Maier and Tomas [22].

The demonstration code used here has not been parallelized. The linear system are solved by using the preconditioned CG version of PARDISO (phase = 23). The solution tolerance is set to 10^{-10} (parm(4) = 102). The reader is referred to Petra et al. [23].

7.2. 1D convergence tests

We estimate the convergence properties of the method on a smooth solution. We consider a one-dimensional viscous shockwave problem that has an exact solution which is described in Becker [24]. A partial English

Table 1
1D Viscous shockwave, $\mathbb{P}_1$ uniform meshes, Convergence tests, $t = 3$, CFL $= 0.4$.

I	$\delta_1(t)$	Rate	$\delta_2(t)$	Rate	$\delta_\infty(t)$	Rate
50	5.85E−02	–	3.11E−01	–	8.28E−03	–
100	2.50E−02	1.23	1.91E−01	0.71	2.82E−03	1.55
200	4.83E−03	2.37	3.27E−02	2.54	5.13E−04	2.46
400	1.07E−03	2.17	9.79E−03	1.74	9.32E−05	2.46
800	2.52E−04	2.09	2.29E−03	2.10	2.02E−05	2.21
1600	6.20E−05	2.02	5.76E−04	1.99	4.89E−06	2.05
3200	1.55E−05	2.00	1.46E−04	1.98	1.23E−06	1.99

translation of [24] and other exact solutions are found in Johnson [25]. The Navier–Stokes system (2.1) is solved over the real line with no source term, $\boldsymbol{f} = \boldsymbol{0}$.

One key assumption of [24] is that the Prandtl number $P_r := \frac{\mu c_P}{\kappa}$ is fixed and equal to $\frac{3}{4}$. Recall that μ is the shear viscosity and κ is the thermal conductivity. The bulk viscosity λ is set to 0.

We first construct a steady state solution. Let $\rho(x)$ be the density, $v(x)$ the velocity, and $e(x)$ the internal energy. Let v_0 be the velocity at infinity on the left ($v_0 := \lim_{x\to-\infty} v(x)$) and let v_1 be the velocity at infinity on the right ($v_1 := \lim_{x\to+\infty} v(x)$). We assume that $v_0 > v_1$. We define $v_{01} := \sqrt{v_0 v_1}$. Let ρ_0 be the density at infinity on the left. Since the solution is time-independent, the momentum is constant, say m_0. In the context of the above assumptions, it is shown in [24, Eq. (30.a)] (see also [25, Eq. (3.6)]) that the velocity profile $\mathbb{R} \ni x \mapsto v(x)$ is defined implicitly as the solution to the following equation:

$$x = \frac{2}{\gamma+1}\frac{\kappa}{m_0 c_v}\left\{\frac{v_0}{v_0-v_1}\log\Big(\frac{v_0-v(x)}{v_0-v_{01}}\Big) - \frac{v_1}{v_0-v_1}\log\Big(\frac{v(x)-v_1}{v_{01}-v_1}\Big)\right\}. \tag{7.1}$$

This equation is solved numerically to high accuracy by using a Newton technique. Notice that by convention, (7.1) implies that $v(0) = v_{01}$. Once $v(x)$ is known, the density and the internal energy at x are given by

$$\rho(x) = \frac{m_0}{v(x)}, \qquad e(x) = \frac{1}{2\gamma}\Big(\frac{\gamma+1}{\gamma-1}v_{01}^2 - v^2(x)\Big). \tag{7.2}$$

To obtain a time-dependent solution, which is computationally more challenging than solving a steady state solution, we construct a moving wave as follows. We first introduce the constant translation velocity v_∞ and we define

$$\boldsymbol{u}(x,t) := \begin{pmatrix} \rho(x - v_\infty t) \\ \rho(x - v_\infty t)(v_\infty + v(x - v_\infty t)) \\ \rho(x - v_\infty t)(e(x - v_\infty t) + \frac{1}{2}(v_\infty + v(x - v_\infty t))^2) \end{pmatrix}. \tag{7.3}$$

The field $\boldsymbol{u}$ solves (2.1) for any v_∞ since the Navier–Stokes equations are Galilean invariant. This solution is used for instance in Dumbser [26] for verification purposes.

We now compare the above solution to numerical simulations using the following parameters $\gamma = 1.4$, $\mu = 0.01$, $v_\infty = 0.2$, $v_0 = 1$, $\rho_0 = 1$. This gives $m_0 = 1$. Instead of enforcing v_1, we choose the pre-shock Mach number $M_0 = 3$, which then gives $v_1 = \frac{\gamma-1+2M_0^{-2}}{\gamma+1}$; see [25, Eq. (2.10)]. Notice that $\kappa = \frac{\mu c_p}{P_r}$ with $P_r = \frac{3}{4}$. We use the truncated domain $[-1, 1.5]$ (the larger the domain the higher the accuracy that can be reached on extremely fine grids). Inhomogeneous Dirichlet boundary conditions are enforced on all conserved quantities $\boldsymbol{u} = (\rho, \boldsymbol{m}, E)$ at the left and right boundary (see Section 2). The simulations are run until $t = 3$. The distance traveled by the shock is 0.6. For $q \in \{1, 2, \infty\}$, we compute a consolidated error indicator at the final time by adding the relative error in the L^q-norm of the density, the momentum, and the total energy as follows:

$$\delta_q(t) := \frac{\|\rho_h(t) - \rho(t)\|_{L^q(D)}}{\|\rho(t)\|_{L^q(D)}} + \frac{\|\boldsymbol{m}_h(t) - \boldsymbol{m}(t)\|_{L^q(D)}}{\|\boldsymbol{m}(t)\|_{L^q(D)}} + \frac{\|E_h(t) - E(t)\|_{L^q(D)}}{\|E(t)\|_{L^q(D)}}. \tag{7.4}$$

We show in Table 1 the results for 7 uniform grids. The coarsest grid has 50 grid points and the finest has 3200 grid points. The number of grid points is denoted by I. We observe second-order convergence in time and space in all the norms, as expected.

Table 2
2D Viscous shockwave, $\mathbb{P}_1$ nonuniform Delaunay meshes, $t = 3$, CFL $\in \{0.4, 0.9\}$.

CFL	I	$\delta_1(t)$	Rate	$\delta_2(t)$	Rate	$\delta_\infty(t)$	Rate
0.4	4458	8.99E−03	–	1.49E−02	–	1.20E−01	–
	17 589	1.35E−03	2.76	3.04E−03	2.31	3.23E−02	1.91
	34 886	5.19E−04	2.80	1.47E−03	2.13	1.44E−02	2.36
	69 781	2.45E−04	2.17	7.20E−04	2.05	7.93E−03	1.72
	139 127	1.04E−04	2.47	3.71E−04	1.93	3.27E−03	2.56
0.9	4458	6.99E−03	–	2.03E−02	–	1.58E−01	–
	17 589	9.51E−04	2.91	3.39E−03	2.61	3.61E−02	2.15
	34 886	3.98E−04	2.54	1.60E−03	2.20	1.55E−02	2.47
	69 781	1.79E−04	2.30	7.54E−04	2.17	8.23E−03	1.83
	139 127	8.17E−05	2.28	3.67E−04	2.09	3.28E−03	2.67

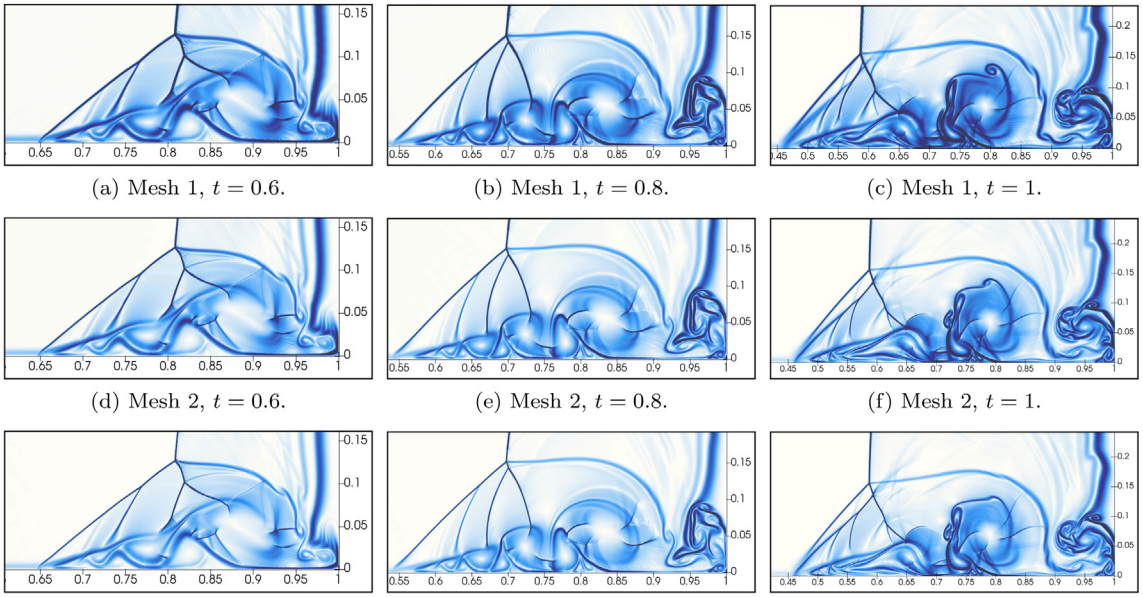
7.3. 2D convergence tests

We use again the exact shockwave solution described in Section 7.2 to verify the method in two-space dimensions. This test is also meant to verify that the method is genuinely second-order accurate on non-uniform meshes. Here we use nonuniform Delaunay triangulations. The convergence tests are done in the truncated domain $D = (-0.5, 1)\times(0, 1)$. In addition to inhomogeneous Dirichlet boundary conditions on the left and right sides we enforce periodic boundary conditions on $\{y = 0\}$ and $\{y = 1\}$. The length of the domain in the x-direction is slightly smaller than for the one-dimensional tests reported above. We do not expect to saturate the relative error indicators δ_1, δ_2 and δ_∞ due to boundary effects in this smaller computational domain since we restrict the meshsize not to be smaller than $1/425$. We use 5 meshes. These meshes are not nested to eliminate the risk of observing super-convergence effects. This makes having consistent convergence rates more difficult and therefore tests the robustness of the method. The meshsizes for these meshes are approximately $0.02, 0.01, 0.0707, 0.05, 0.003536$. The results are reported in Table 2 for the two CFL numbers 0.4 and 0.9. We observe that the method is second-order accurate both in time and space, for both CFL numbers, and in all error norms.

7.4. 2D shocktube test

As a final numerical test we simulate the interaction of a shock with a viscous boundary layer. The test case we consider has been introduced in the literature by Daru and Tenaud [27] and is further documented in Daru and Tenaud [28]. It is essentially a shocktube problem. The tube is the square cavity $D = (0, 1)^2$ with a diaphragm at $\{x = \frac{1}{2}\}$ separating it in two parts. The fluid is initially at rest. The state on the left-hand side of the diaphragm is $\rho_L = 120$, $v_L = 0$, $p_L = \rho_L/\gamma$. The right state is $\rho_R = 1.2$, $v_R = 0$, $p_R = \rho_R/\gamma$. We use the ideal gas equation of state $p = (\gamma - 1)\rho e$ with $\gamma = 1.4$. The bulk viscosity is set to 0. The Prandtl number is $Pr = 0.73$. The no-slip and the thermally insulating boundary conditions (2.4) are enforced throughout. The diaphragm is broken at $t = 0$. A shock, a contact and a rarefaction wave are created. The viscous shock and the contact move to the right. The rarefaction wave moves to the left. As the shock and the contact waves progress to the right they create thin viscous boundary layers on the top and the bottom walls of the tube. The shock hits the right wall at approximately $t \approx 0.2$ and is then reflected. The shock interacts with the contact discontinuity on its way back to the left. Complex interactions occur and the contact discontinuity stays stationary close to the right wall thereafter. The shock wave then continues its motion to the left and interacts with the viscous boundary layer which it created while moving to the right. This interaction is very strong and a lambda shock is formed as a result. We refer to [27, §6] and [28, §5&§6] for full descriptions of the various mechanisms at play in this problem.

The computations reported in this paper are done in the half domain $(0, 1)\times(0, \frac{1}{2})$. Symmetry with respect to the horizontal axis $\{y = \frac{1}{2}\}$ is obtained by enforcing the slip boundary condition instead of the no-slip boundary condition (2.4). This is achieved algebraically by simply replacing the homogeneous Dirichlet condition $\mathsf{V}_i^{n+\frac{1}{2}} = \mathbf{0}$ in (5.5) by $\boldsymbol{n}\cdot\mathsf{V}_i^{n+\frac{1}{2}} = \mathbf{0}$ at $\{y = \frac{1}{2}\}$. The weak bilinear form (5.3) then enforces the tangential trace of the normal viscous stress to be zero. In strong form these two conditions amount to enforcing the normal component of the



(a) Mesh 1, $t = 0.6$. (b) Mesh 1, $t = 0.8$. (c) Mesh 1, $t = 1$.

(d) Mesh 2, $t = 0.6$. (e) Mesh 2, $t = 0.8$. (f) Mesh 2, $t = 1$.

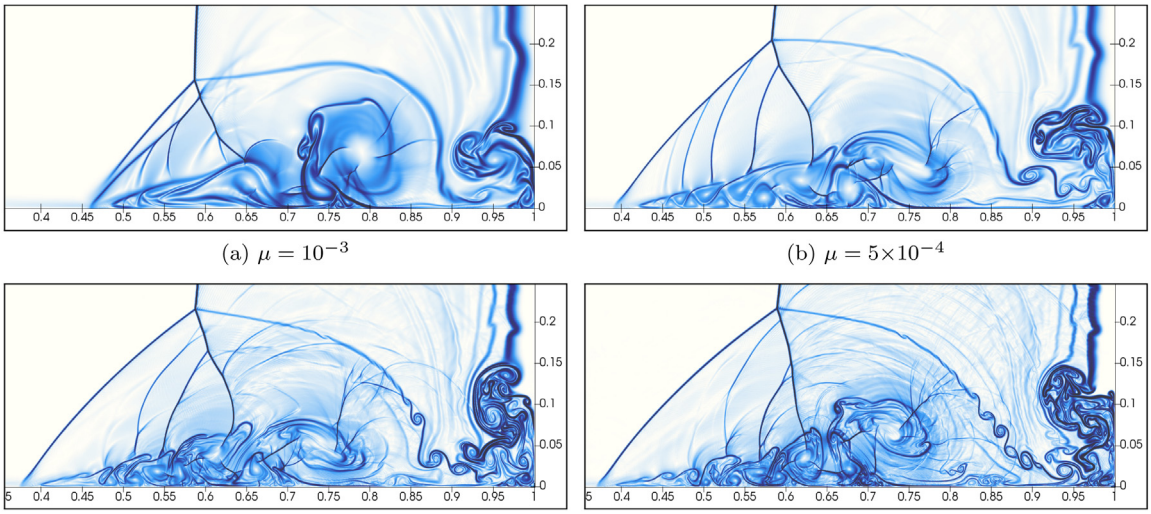
(g) Mesh 3, $t = 0.6$. (h) Mesh 3, $t = 0.8$. (i) Mesh 3, $t = 1$.

Fig. 1. 2D shocktube test. Density at $t \in \{0.6, 0.8, 1\}$ with $\mu = 10^{-3}$. Meshes with increasing refinement level: Mesh 1, 359388 grid point; Mesh 2, 684996 grid point; Mesh 3, 859765 grid points.

velocity to be zero and the normal derivative of the tangent component of the velocity to be zero at $\{y = \frac{1}{2}\}$. The CFL number used for these computations is 0.95 (see (4.4) and (6.1)). The computations are done with nonuniform meshes that are progressively refined. The meshes are highly nonuniform to concentrate the grid points in the right part of the cavity. In mesh 1 the meshsize is about 0.0007 on $\{0.3 \leq x \leq 1, y = 0\}$ and 0.0014 on $\{0.5 \leq x \leq 1, y = 0.5\}$ (359388 grid points). The meshsize in the second mesh is about 0.0005 on $\{0.3 \leq x \leq 1, y = 0\}$ and 0.001 on $\{0.5 \leq x \leq 1, y = 0.5\}$ (684996 grid points). For mesh 3 the meshsize is about 0.0004 on $\{0.3 \leq x \leq 1, y = 0\}$ and 0.001 on $\{0.5 \leq x \leq 1, y = 0.5\}$ (859765 grid points).

We start by demonstrating the behavior of the method under nonuniform mesh refinement. We show in Fig. 1 the gradient of the density field at $t \in \{0.6, 0.8, 1\}$ for the three meshes: Mesh 1 to Mesh 3. More precisely, denoting $g(\boldsymbol{x}) = \|\nabla \rho_h(\boldsymbol{x})\|_{\ell^2}$, $g_{\min} = \min_{\boldsymbol{x}\in D} g(\boldsymbol{x})$, $g_{\max} = \max_{\boldsymbol{x}\in D} g(\boldsymbol{x})$, we visualize the quantity $e^{-10\frac{g-g_{\min}}{g_{\max}-g_{\min}}}$ to amplify the contrast. We observe that the results at $t = 0.6$ and at $t = 0.8$ vary very little as the grids are refined. Some local changes are noticeable for the solution at $t = 1$, but the overall structure of the flow seems to be converging when the meshsize decreases. There is some disagreement in the literature on the solution at $t = 1$ for $\mu = 10^{-3}$. For instance various schemes are tested in Sjögreen and Yee [29] on meshes ranging from 1000×500 grid points to 4000×2000 grid points (in the half domain), but the results reported therein seem to depend on the scheme that is chosen. It is remarkable though that our results on the finest grid (Fig. 1(i)) are strikingly similar to those reported Fig. 8d in Daru and Tenaud [28] and Fig. 11l in Zhou et al. [30] (see also Fig. 5a in [28] and Fig. 6c in [30]); these three figures are almost Xerox copies of each other. But none of the results reported in [29] (and [31]) agree with the results shown in Fig. 1 (and Fig. 8d in [28] and Fig. 11l in [30]). In conclusion, it seems that our results agree very well with those reported in Daru and Tenaud [28] and Zhou et al. [30] but disagree with those reported in Sjögreen and Yee [29] (and Kotov et al. [31]), thereby shedding some doubts on the correctness of the computations in [29,31]. A detailed quantitative comparisons with [28] using extremely fine meshes is in preparation.

As a last numerical illustration we recompute the density field at $t = 1$ on Mesh 4 for four increasingly smaller viscosities $\mu \in \{10^{-3}, 5\times10^{-4}, 2\times10^{-4}, 10^{-4}\}$. The results are reported in Fig. 2. We observe that for decreasing viscosity the flow field develops increasingly more pronounced and smaller vortex structures. This confirms that the influence of the artificial graph viscosity of the hyperbolic step (see Section 7.2) is well below the viscous effects introduced by the physical viscosity μ and the thermal conductivity κ.



(a) $\mu = 10^{-3}$ (b) $\mu = 5\times10^{-4}$

(c) $\mu = 2\times10^{-4}$ (d) $\mu = 10^{-4}$

Fig. 2. 2D shocktube test, Mesh 3. Density at $t = 1$ for $\mu \in \{10^{-3}, 5\times10^{-4}, 2\times10^{-4}, 10^{-4}\}$.

8. Conclusions and outlook

A fully discrete second-order order accurate method for solving the compressible Navier–Stokes equations has been introduced. The novelty of this work lies in the guaranteed invariant domain preservation of the fully discrete method under the usual hyperbolic CFL condition. The method relies on the operator-splitting strategy in order to preserve invariant set stability properties. There is, in principle, no limitation for the accuracy in space. We also notice that the method exhibits quite robust behavior (in the eye-ball norm) for flows containing strong shock interactions with viscous layers. At this point in time, it is not yet clear how to develop a third-order accurate (in-time) invariant-domain-preserving scheme.

Declaration of competing interest

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

References

[1] D. Grapsas, R. Herbin, W. Kheriji, J.-C. Latché, An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations, SMAI J. Comput. Math. 2 (2016) 51–97.

[2] T. Gallouët, L. Gastaldo, R. Herbin, J.-C. Latché, An unconditionally stable pressure correction scheme for the compressible barotropic Navier-Stokes equations, M2AN Math. Model. Numer. Anal. 42 (2) (2008) 303–331.

[3] X. Zhang, On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations, J. Comput. Phys. 328 (2017) 301–343.

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

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

[6] L. Demkowicz, J.T. Oden, W. Rachowicz, A new finite element method for solving compressible Navier-Stokes equations based on an operator splitting method and h-p adaptivity, Comput. Methods Appl. Mech. Engrg. 84 (3) (1990) 275–326.

[7] R. Menikoff, B.J. Plohr, The Riemann problem for fluid flow of real materials, Rev. Modern Phys. 61 (1) (1989) 75–130.

[8] A. Harten, P.D. Lax, C.D. Levermore, W.J. Morokoff, Convex entropies and hyperbolicity for general Euler equations, SIAM J. Numer. Anal. 35 (6) (1998) 2117–2127, (electronic).

[9] K.N. Chueh, C.C. Conley, J.A. Smoller, Positively invariant regions for systems of nonlinear diffusion equations, Indiana Univ. Math. J. 26 (2) (1977) 373–392.

[10] P.-L. Lions, Mathematical Topics in Fluid Mechanics. Vol. 2, in: Oxford Lecture Series in Mathematics and its Applications, vol. 10, The Clarendon Press, Oxford University Press, New York, 1998, p. xiv+348, Compressible models, Oxford Science Publications.

[11] E. Feireisl, Dynamics of Viscous Compressible Fluids, in: Oxford Lecture Series in Mathematics and its Applications, vol. 26, Oxford University Press, Oxford, 2004, p. xii+212.
[12] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Springer, 2015.
[13] D. Hoff, D. Serre, The failure of continuous dependence on initial data for the Navier-Stokes equations of compressible flow, SIAM J. Appl. Math. 51 (4) (1991) 887–898.
[14] J.-L. Guermond, B. Popov, Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations, J. Comput. Phys. 321 (2016) 908–926.
[15] E.F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, third ed., Springer-Verlag, Berlin, 2009, p. xxiv+724, A practical introduction.
[16] J. Brandts, S. Korotov, M. Křížek, The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, Linear Algebra Appl. 429 (10) (2008) 2344–2357.
[17] J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comp. 68 (228) (1999) 1429–1446.
[18] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works, J. Comput. Phys. 135 (2) (1997) 170–186, With an introduction by Steven T. Zalesak, Commemoration of the 30th anniversary of J. Comput. Phys.; J. Comput. Phys. 11 (1) (1973) 38–69.
[19] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, J. Comput. Phys. 31 (3) (1979) 335–362.
[20] D. Kuzmin, R. Löhner, S. Turek, Flux–Corrected Transport, in: Scientific Computation, Springer, 2005, 3-540-23730-5.
[21] M. Maier, M. Kronbichler, Massively parallel 3D computation of the compressible Euler equations with an invariant-domain preserving second-order finite-element scheme, 2020.
[22] M. Maier, I. Tomas, The step-69 tutorial program: implementation of a graph-based scheme for Euler's equation of compressible gas dynamics. Deal.ii Library, URL https://www.dealii.org/developer/doxygen/deal.II/step_69.html.
[23] C.G. Petra, O. Schenk, M. Lubin, K. Gäertner, An augmented incomplete factorization approach for computing the schur complement in stochastic optimization, SIAM J. Sci. Comput. 36 (2) (2014) C139–C162.
[24] R. Becker, Stoßwelle und detonation, Z. Phys. 8 (1) (1922) 321–362.
[25] B.M. Johnson, Analytical shock solutions at large and small prandtl number, J. Fluid Mech. 726 (2013) R4, 12.
[26] M. Dumbser, Arbitrary high order $P_N P_M$ schemes on unstructured meshes for the compressible Navier-Stokes equations, Comput. Fluids 39 (1) (2010) 60–76.
[27] V. Daru, C. Tenaud, Evaluation of TVD high resolution schemes for unsteady viscous shocked flows, Comput. & Fluids 30 (1) (2001) 89–113.
[28] V. Daru, C. Tenaud, Numerical simulation of the viscous shock tube problem by using a high resolution monotonicity-preserving scheme, Comput. Fluids 38 (3) (2009) 664–676.
[29] B. Sjögreen, H. Yee, Grid convergence of high order methods for multiscale complex unsteady viscous compressible flows, J. Comput. Phys. 185 (1) (2003) 1–26.
[30] G. Zhou, K. Xu, F. Liu, Grid-converged solution and analysis of the unsteady viscous flow in a two-dimensional shock tube, Phys. Fluids 30 (1) (2018) 016102.
[31] M. Kotov, L. Ruleva, S. Solodovnikov, I. Kryukov, S. Surzhikov, Multiple flow regimes in a single hypersonic shock tube experiment, in: 30th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, 2014, http://dx.doi.org/10.2514/6.2014-2657, AIAA 2014-2657.