FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Numerická schémata pro rovnice vazkého stlačitelného proudění: Analýza a geometrie

RNDr. Bc. Radim Hošek

disertační práce
k získání akademického titulu **doktor (Ph.D.)**
v oboru **Aplikovaná matematika**

Školitel : prof. RNDr. Eduard Feireisl, DrSc.

Katedra matematiky

Plzeň, 2017

FACULTY
OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

Numerical Schemes for Equations of Viscous Compressible Flows: Analysis and Geometry

Radim Hošek

dissertation thesis
for taking academic degree **Doctor of Philosophy (Ph.D.)**
in specialization **Applied Mathematics**

Supervisor: Eduard Feireisl

Department of Mathematics

Pilsen, 2017

Deklarace, Declaration

Prohlašuji, že jsem tuto práci zpracoval samostatně. Je založena na vědeckých výsledcích, jejichž jsem autorem nebo spoluatorem, a na zdrojích, které jsou citovány a uvedeny v seznamu literatury. Při tvorbě práce byly zachovány postupy ve vědecké práci obvyklé.

Plzeň/Varšava, 18. únor 2017

I declare that the presented thesis is my original work. It is based on the scientific results which I have authored or co-authored and on cited sources which are contained in the list of references. Working on the thesis I have used solely the standard scientific methods.

Pilsen/Warsaw, 18th February, 2017

.....
Radim Hošek

Abstrakt

Tato disertační práce obsahuje komentovaný soubor vědeckých článků s výsledky v oblasti analýzy numerických schémat pro modely vazkých stlačitelných tekutin. Tyto jsou založeny především na průkopnickém výsledku Karpera, který dokázal konvergenci numerických řešení ke slabému řešení stlačitelných Navierových–Stokesových rovnic, jež popisují dynamiku tekutiny v takzvaném *barotropním režimu*. Tyto vědecké články jsou rozděleny do dvou skupin, první z nich obsahuje výsledky z numerické analýzy, ta druhá potom výsledky z oblasti geometrie sítí, na nichž jsou metody definovány.

První z těchto výsledků obsahuje návrh nové numerické metody pro barotropní tekutinu založenou na konečných diferencích a hlavní část jejího konvergenčního důkazu. Další dokazuje konvergenci zobecněné varianty Karperovy metody pro úplný systém, který zahrnuje i bilanci teploty. V posledním článku jsou odvozeny chybové odhady pro původní Karperovu metodu.

Výše zmíněné numerické metody mají speciální požadavky na geometrické vlastnosti využitých sítí. Druhá skupina výsledků obsahuje dva články týkající se existence právě takových tříd sítí, které splňují požadované vlastnosti. Tyto výsledky jsou doplněny článkem na téma simplexových sítí v obecné dimenzi. Motivací pro tento výsledek je společný jmenovatel dvou předchozích, a sice konstrukce publikovaná Sommervillem v roce 1923.

Klíčová slova

stlačitelné tekutiny, numerická schémata, barotropní tekutina, Navierovy–Stokesovy rovnice, Navierovy–Stokesovy–Fourierovy rovnice, Sommervillův čtyřstěn, třída sítí, dobře středovaná síť, zjemnění sítě, dláždění

Abstract

The dissertation thesis presents a commented collection of research articles with results in the analysis of numerical schemes for systems that model viscous compressible fluids. They are mainly based on a pioneering work of Karper, who proved a convergence of a numerical scheme to weak solutions of the compressible Navier–Stokes system, which represents a flow of a fluid in the so called *isentropic regime*. These research articles are split into two groups, first of them being results in the numerical analysis, the other one dealing with the underlying geometry.

First of the results contains a design of a new finite-difference numerical scheme for the isentropic flow and a major part of its convergence proof. The next one proves convergence of a generalization of a variant of Karper's method for the complete system including the balance of the temperature. In the last article the error estimates for the original Karper's method are shown.

The above mentioned numerical methods have particular geometrical requirements on the underlying meshes. The second group of results contains two articles on existence of such families of meshes satisfying required assumptions. These results are accompanied by an article on simplicial meshes for a general dimension. Its motivation comes from a common denominator of the two previous results, which is a construction introduced by Sommerville in 1923.

Keywords

compressible fluids, numerical scheme, barotropic fluid, Navier–Stokes system, Navier–Stokes–Fourier system, Sommerville tetrahedra, family of meshes, well-centered mesh, boundary-fitted mesh, mesh refinement, tessellations.

Zusammenfassung

Die vorliegende Dissertation befasst sich mit der Analyse der numerischen Schemen für Strömungsmodelle der viskosen komprimierbaren Fluiden und besteht aus einer kommentierten Sammlung von wissenschaftlichen Artikeln des Autors. Die Ergebnisse basieren vor allem auf der Pionierarbeit von Karper, der die Konvergenz eines numerischen Schemas gegen die schwache Lösung des komprimierbares Navier–Stokes–Systems für Fluidströmungen in dem sogenannten *barotropen Regime* bewies.

Die erhaltene Artikel werden in zwei Teile strukturiert, wobei der erste Abschnitt der numerischen Analyse gewidmet ist und der zweite sich mit der zugrundeliegenden Geometrie befasst. In dem ersten Beitrag konstruiert der Autor eine neue Finite–Differenzen–Schema für barotrope Fluidströmungen und großenteils zeigt er seine Konvergenz. Der zweite Artikel liefert den Beweis der Konvergenz von einer verallgemeinerten Variante von Karper-Verfahren für das Gesamtsystem, welches das Gleichgewicht der Temperatur einschließt. Am Ende erstes Teils der Dissertation werden die Fehlerabschätzungen für das ursprüngliche Karper-Verfahren in weiterem Artikel gezeigt.

Die oben genannten numerischen Schemen haben einige besondere Anforderungen an die unterliegenden Gittern. Zweiter Abschnitt der Dissertation fängt mit zwei Artikel an, in den die Existenz solcher Gitterfamilien, die die erforderlichen Eigenschaften erfüllen, bewiesen wird. In der letzten wissenschaftlichen Arbeit liefert der Autor ein Ergebnis für die simplizialen Gitter in einer allgemeinen Dimension. Die Motivation dafür stammt von dem gemeinsamen Nenner der zwei früheren Ergebnisse, das heißt die Konstruktion von Sommerville aus dem Jahr 1923.

Schlüsselwörter

komprimierbare Fluide, barotrope Fluide, numerische Schemen, Navier–Stokes–System, Navier–Stokes–Fourier–System, Sommerville–Tetraeder, Gitterfamilie, wohlzentrierte Gitter, rand-angepasste Gitter, Gitterverfeinerung, Parkettierung

Acknowledgements

Hereby I would like to express the thanks to my supervisor Eduard Feireisl for the creative freedom I was given, for all contacts to various great mathematical minds (himself included) and for the opportunity to learn about the mathematical culture in various parts of the world. Next thanks belong to all my superiors Šárka Nečasová, Gabriela Holubová, Pavel Drábek and Eduard Feireisl for their tolerance. This was a necessary ingredient to make my studies and scientific work straddled in both University of West Bohemia and Institute of Mathematics feasible. A great help in technical and administrative issues was provided by my classmate Jonáš Volek, who also deserves a special word of thanks. And I would like to thank Trygve Karper, whose style of scientific work was maybe the most appealing to me, for the chance to collaborate with him even though shortly.

My personal thanks go to my parents, who told me on my first school day that I will study for at least 13 years. Finally they turned into 23 years and all of them with their great support which I appreciate. A special apology belongs to my sister Barbora Šedivá, a first real scientist in our family, for not having spent much time with her in recent years. Last but not least, I would like to express my appreciation to my wife Tereza Hošková for creating a great home for us, a safe port from which I could set out for my cruises to the world of mathematics and where I love to return.

In Warsaw, 18th February 2017.

Radim Hošek

I would also like to acknowledge all the financial sources that helped me achieve my scientific results, even those beyond the scope of this thesis.

- European Union's Seventh Framework Program (FP7/2007-2013)/ERC Grant MATHEF 320078,
- Grant 13-00863S of the Grant Agency of the Czech Republic,
- Grant GA1300522S of the Grant Agency of the Czech Republic,
- Grant SGS-2016-003 of the University of West Bohemia.

Preface

The compressible Navier–Stokes–Fourier system is generally accepted as the complete macroscopic description of the flow of a compressible, viscous and heat conducting fluid, whose representative is the air, for instance. Description of its behavior is of general interest, not only for precise weather forecasting, but also in aerodynamics and other fields.

For nonlinear problems of this type the existence of smooth solutions is confined to a possibly very short time interval. The first results on existence of a global in time weak solution to this system in three spatial dimensions were developed in the 1990s for the simpler isentropic case, where the balance of internal energy is dropped. Moreover till these days the existence proof does not cover the physically most interesting case of a diatomic gas in three spatial dimensions.

The ambiguity between the society's high demand on using such system and the lack of rigorous mathematical knowledge is a major driving force in developing theoretical results and effective (and relevant) computational scheme to this complex problem. In Prague there is a long lasting tradition of mathematical modeling of fluid dynamics, connected with names like Babuška or Nečas, among others.

Till these days the so called Prague school keeps contributing to the cutting-edge research in this field. To support this claim we mention the ERC Advanced Grant, carried out at the Mathematical Institute of the Czech Academy of Sciences, awarded to Eduard Feireisl. I had the opportunity to join his team for the duration of my PhD studies.

The project is focused on the analysis of the *complete system* of compressible viscous fluids with one of its aims to bridge the rather separated analytic and numerical subcultures together. This also logically became the scope of my work in the past years. The research goal for the thesis has been stated rather generally, 'to contribute with original results to the field of mathematics of compressible, viscous and heat conducting fluids'.

The resulting dissertation thesis is compiled as a commented collection of various results achieved within the MATHEF grant activities, some of them co-authored with the team members and guests. The main results of these articles are pointed out and commented in the text, with the technical details being

omitted. An interested readers can find them in the scientific articles which are appended to the thesis in their original form. To be more specific, the following articles and results in those contained are included in the thesis:

- **[35]** R. Hošek and B. She. Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension. *Submitted to Journal of Numerical Mathematics. Preprint available at math.cas.cz*, 2017.
- **[16]** E. Feireisl, R. Hošek, and M. Michálek. A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains. *SIAM J. Numer. Anal.*, 54(5):3062–3082, 2016.
- **[14]** E. Feireisl, R. Hošek, D. Maltese, and A. Novotný. Error estimates for a numerical method for the compressible NavierStokes system on sufficiently smooth domains. *ESAIM: M2AN*, 51(1):279–319, 2017.
- **[30]** R. Hošek. Face-to-face partition of 3D space with identical well-centered tetrahedra. *Appl. Math.*, 60(6):637–651, 2015.
- **[32]** R. Hošek. Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains. *Appl. Math.*, 61(3):233–251, 2016.
- **[31]** R. Hošek. Construction and shape optimization of simplical meshes in d-dimensional space. *Submitted to Disc. Comp. Geom. Preprint available at ArXiv.org*, June 2016.

This all is preceded by a chapter with necessary introduction into the field. For better orientation of the reader, the reference numbers of the included scientific papers of the author are printed in bold.

The 'methodology' explanation, required by the respective regulation for a dissertation thesis, shrinks to a simple statement, that *the standard rigorous mathematical methods* will be used. The consistency is a big advantage of mathematics; using various methods can bring the researcher to various results, but never to contradictory ones.

Contents

Chapter 1

Introduction

1.1 Systems of Equations

The cornerstone of the flow of viscous compressible fluids is the (compressible) Navier–Stokes system,

$$\partial_t \varrho + \mathrm{div}_x\,(\varrho \mathbf{u}) = 0, \tag{1.1}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x p = \mathrm{div}_x \mathbb{S} + \varrho \mathbf{f}, \tag{1.2}$$

equations that represent the classical mechanics principles of conservation of mass and momentum. The state of the physical system is assumed to be described by two observable macroscopic quantities, ϱ representing the mass density, $\mathbf{u}$ the velocity. We skip the derivation of the system, one can find its brief derivation in [20], or a more detailed version in [24].

Usually the system is confined to a bounded domain $\Omega \subset \mathbb{R}^d$, we focus mainly on the physically relevant case $d = 3$. The behaviour of the velocity at the boundary should be specified. There is no general answer to that question, there are more acceptable possibilities, depending on the the fluid constitution and *flow regime*, see [43] for a detailed discussion. We will use the so called *no-slip* boundary condition, which reads

$$\mathbf{u}|_{\partial\Omega} = \mathbf{0}. \tag{1.3}$$

To determine the solution of an evolutionary equation, we must prescribe initial conditions. We set $\varrho(0,.) = \varrho_0 > 0$ and $(\varrho\mathbf{u})(0,.) = m_0$, where the initial functions have some minimal smoothness. To complete the system, we should supply the constitutive relations. We will be interested solely on Newtonian, i.e. *linearly constituted* fluids. In particular, the stress tensor $\mathbb{S}$ takes the form

$$\mathbb{S}(\nabla_x \mathbf{u}) = \mu\left(\nabla_x \mathbf{u} + \nabla_x^T \mathbf{u} - \frac{2}{d}\,\mathrm{div}_x \mathbf{u}\mathbb{I}\right) + \eta \mathrm{div}_x \mathbf{u}\,\mathbb{I}, \tag{1.4}$$

where the *shear and bulk viscosity* coefficients μ, η, that may depend on variables (ϱ, ϑ), are non-negative. Notice that the shear viscosity depends on the traceless part of the symmetric velocity gradient, the bulk viscosity only on the divergence of velocity. The details about the form of the stress tensor can be found in [7].

Also non-newtonian fluids are of the scientific interest (see [4]), including the so called *implicitly constituted* fluids, for which the mutual dependence of $\mathbb{S}$ and $\nabla_x \mathbf{u}$ cannot be expressed with a single-valued function, see works of the Prague group of Málek [1, 2] or Rajagopal [44].

For compressible flow the pressure is a known function of density, we assume it takes the form

$$p(\varrho) = a\varrho^\gamma, \tag{1.5}$$

which satisfies all necessary assumptions that one uses in the existence proof. For the details and possible relaxations of (1.5) see [20, Section 4.3]. This is a tremendous difference compared to the incompressible Navier–Stokes, where pressure is an unknown quantity, that is implicitly determined by the motion of the fluid as a Lagrange multiplier to the problem.

Last comment is aimed to the external force $\mathbf{f}$, whose effect is usually omitted in the analysis, i.e. it is assumed that $\mathbf{f} \equiv \mathbf{0}$, bearing in mind that including the external force does not bring any additional difficulties.

1.2 Existence Results for Compressible Navier–Stokes

The more famous incompressible Navier–Stokes system can be obtained from (1.1–1.2) after employing the incompressibility constraint. Upon the reasonable assumption of the constant initial distribution of the density it reads

$$\operatorname{div}_x \mathbf{u} = 0, \tag{1.6}$$

$$\partial_t \mathbf{u} + \operatorname{div}_x(\mathbf{u} \otimes \mathbf{u}) + \nabla_x p = \operatorname{div}_x \mathbb{S}. \tag{1.7}$$

These equations are known for almost 200 years, yet still the question of existence and quality of its solution is not satisfactorily answered. This issue is topical even now in the 21^{st} century, which is reflected in its incorporation among the seven Millennium Problems of the Clay Institute, see [11].

As the attempts at obtaining (long time) smooth solution even for smooth data (which are not only small perturbations of an equilibrium state) failed, Leray [41] in 1930s introduced the notion of weak solution, which replaces differential equations with a system of integral identities. The results of Leray underwent further improvements, e.g. by Hopf [29], Ladyzhenskaya [40], Caffarelli, Kohn & Nirenberg [3], to name at least a few.

In 1998 P. L. Lions [42] generalized Leray's theory also for the compressible case. His work represents the first results on existence for arbitrarily large date and/or time interval for the compressible flow. The supervisor of myself, Eduard Feireisl, later improved Lions' result, see [12].

In the 1980s Valli & Zajączkowski [47, 48] proved a local existence result for strong solutions, later improved by Cho, Choe & Kim [6]. Upon the assumptions of certain regularity of the data, the existence of strong solution to (1.1–1.2) is proved. However, the system is nonlinear, and one cannot exclude the development of singularities that lead to a breakdown of the solution. The life span of the classical solution might therefore be extremely short.

We find it important to emphasize that while the strong solution is unique, the notion of the weak solution is too benevolent which leaves the question of uniqueness opened. Even more disturbing are the results on the *compressible Euler equations*, a system that is supposed to describe a flow of a compressible inviscid fluid, i.e. system (1.1, 1.2, 1.4), where $\mu = \eta = 0$. The system has been proved to posses infinitely many weak solutions for certain initial data, see DeLellis & Székelyhidi [10], which satisfy additional admissibility criterion based on energy. Before this result, it was believed that a proper admissibility criterion could help to pick the *correct* solution. Chiodaroli, DeLellis & Kreml [5] later improved the result also for data that admit strong solution, which disproved the conjecture that the non-uniqueness is caused by insufficient smoothness of the initial data.

The previous paragraph illustrated the difficulties with the uniqueness of weak solutions to flow problems. Let us get back to the compressible Navier–Stokes system. Thanks to Feireisl, Jin & Novotný we have additional result at hand, the so called *weak-strong uniqueness*, see [17]. Roughly speaking, when the data are smooth enough to admit a strong solution, than any weak solution of the system must coincide with the strong one on its life span. The tool for proving this result is the *relative energy*, which is of an independent interest, because in general it can measure a distance of a weak solution to any couple of functions. These functions are the strong solution when proving the weak-strong uniqueness, but it can also be a solution of a different system, achieving the *singular limits* results, see [22]. A discrete version of the relative energy finds its use also in analysis of numerical schemes, which we will focus on later.

1.3 Closing the System Energetically: From Navier–Stokes to Navier–Stokes–Fourier

We start this section with a brief motivation. More detailed derivation of the following can be found e.g. in [20, Section 4.2]. To get the energy of the compressible Navier–Stokes system, we formally perform the following:

- Multiply the momentum equation by velocity $\mathbf{u}$ and integrate over domain.
- Multiply the continuity equation by $\frac{1}{2}|\mathbf{u}|^2$ and integrate over the domain.
- Multiply the continuity equation with a term $B'(\varrho)$, where $B \in C^1(\mathbb{R}^+)$ to get the *renormalized continuity equation*

$$\partial_t B(\varrho) + \mathrm{div}_x(B(\varrho)\mathbf{u}) + (B'(\varrho)\varrho - B(\varrho))\mathrm{div}_x\mathbf{u} = 0, \tag{1.8}$$

  and pick $B(x) = P(x) := x\int_1^x \frac{p(z)}{z^2}\,\mathrm{d}z$, the *pressure potential* so that

$$B'(\varrho)\varrho - B(\varrho) = p(\varrho).$$

Combination of these three ingredients together with some integration by parts and algebraic manipulation leads to the following form of the total energy equality,

$$\int_\Omega \partial_t \left(\frac{1}{2} \varrho |\mathbf{u}|^2 + P(\varrho) \right) \mathrm{d}x + \int_\Omega \mathbb{S}(\nabla_x \mathbf{u}) : \nabla_x \mathbf{u} \, \mathrm{d}x = 0. \tag{1.9}$$

The presence of the viscosity, represented by the positive *dissipation term* in (1.9) leads to conclusion that compressible Navier–Stokes system is an incomplete system from the point of view of energy. The part of energy that is dissipated through viscosity and turns into heat that is not captured in the system. For closing the system, we include the *internal energy balance*,

$$\partial_t(\varrho e) + \mathrm{div}_x(\varrho e \mathbf{u}) + \mathrm{div}_x \mathbf{q} = \mathbb{S} : \nabla_x \mathbf{u} - p \, \mathrm{div}_x \mathbf{u}. \tag{1.10}$$

The evolution of the internal energy is balanced by the internal energy flux $\mathbf{q}$ and the source terms of the mechanical origin. One could also include external heat sources, which we omit similarly as we omitted the external forces in the momentum equation.

The internal energy balance can be replaced by different balances. There is a thermodynamic reasoning for transforming internal energy balance (1.10) into a balance of temperature

$$\varrho c_V \left(\vartheta_t + \mathbf{u} \cdot \nabla_x \vartheta \right) + \mathrm{div}_x \mathbf{q} = \mathbb{S} : \nabla_x \mathbf{u} - \vartheta \frac{\partial p}{\partial \vartheta} \, \mathrm{div}_x \mathbf{u}, \tag{1.11}$$

or the balance of entropy

$$\partial_t(\varrho s) + \mathrm{div}_x(\varrho s \mathbf{u}) + \mathrm{div}_x \left(\frac{\mathbf{q}}{\vartheta} \right) = \frac{1}{\vartheta} \left(\mathbb{S} : \nabla_x \mathbf{u} - \frac{\mathbf{q} \cdot \mathbf{u}}{\vartheta} \right) =: \sigma, \tag{1.12}$$

see [12] for detailed derivation. The formulations (1.10, 1.11, 1.12) are equivalent from the point of view of classical solutions, however, not indeed in the weak formulation.

A constitutive relation that should be added is the *Fourier's Law*

$$\mathbf{q} = -\kappa \nabla_x \vartheta, \tag{1.13}$$

where the *heat conductivity* κ may depend[1] on (ϱ, ϑ).

The boundary condition that is usually supplied together with (1.10) is the following

$$\mathbf{q} \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad \text{i.e. } \nabla_x \vartheta \cdot \mathbf{n}|_{\partial\Omega} = 0,$$

representing the insulated domain where no flux of energy through the boundary is possible.

The pressure depends also on temperature, it is assumed to take the form

$$p(\varrho, \vartheta) = a_1 \varrho^\gamma + a_2 \varrho + \varrho \vartheta.$$

The definition of the weak solution for the Navier–Stokes–Fourier is not straightforward, for example the entropy equation is relaxed to an inequality, allowing a non-negative entropy production. See the proper definition of weak

[1] For the existence proof it is even necessary to assume a certain form of such dependance, the existence of solution is not known for the case of constant heat conductivity. On contrary, the constant viscosity coefficients rather ease the proof.

solutions in [12], the main claim is that the compressible Navier–Stokes–Fourier system under certain assumptions possesses a global in time weak solution. Moreover, for *sufficiently regular* data there exists a (local in time) unique strong solution, see [48]. Similarly to the isentropic case, a weak-strong uniqueness result was proved in [23], i.e. the classical solution is unique also within the class of weak solutions.

The precise formulation of the above results would cover many pages and goes beyond the scope of this thesis.

1.4 Convergent Numerical Schemes

The next topic is connected with numerical schemes to both compressible systems introduced above. The basic properties of numerical schemes are its *solvability* (which is not obvious for implicit nonlinear schemes that are used) and *stability*. Roughly speaking, this ensures that the scheme produces discrete solutions that do not breakdown. The connection of the scheme with the target system is represented by *consistency*. Usually an exact solution is plugged into the scheme and shown to fulfill the scheme up to a remainder that vanishes with vanishing discretization parameter. For schemes to the compressible Navier–Stokes(–Fourier) system, the classical solution is not at hand for longer time intervals. Hence for consistency the numerical solution is plugged into the weak formulation, showing that the remainder converges to zero for vanishing discretization parameter.

The above steps would be sufficient to guarantee convergence of schemes for linear problems, but they do not guarantee that the numerical quantities converge to the solution in our case indeed.

In this sense, the result of Karper [37] (see also Karlsen & Karper [36]) is a pioneering work, being the first scheme for which a convergence of its solution to a weak solution[2] of the target compressible Navier–Stokes system in three spatial dimensions is proved. The result is based on the machinery developed by Lions [42] for proving the existence of the weak solutions. The Lions' proof uses solutions of an approximative problem and compactness arguments to show the convergence. Loosely speaking, these solutions of an approximative problem are replaced by the solutions of a numerical scheme in Karper [37] or its recent user-friendly summary in Feireisl, Karper & Pokorný [20, Part II].

The Karper's scheme is a combined finite-volume/finite-element scheme which includes the upwind technique for convective terms. It uses a tetrahedral mesh of a polyhedral domain and approximates density with piecewise constants and velocity with Crouzeix–Raviart elements, piecewise affine functions with a jump across the boundary of an element with a zero mean, see [9]. These elements are non-conforming, since the finite-dimensional space of numerical functions is not a subspace of the target one, as the velocity is assumed to be in the Sobolev space $W^{1,2}$, which has a well defined traces and thus does not allow jumps.

The scheme was later extended also to the case when the target domain is smooth in [18]. One of the motivations for that was the prospect of deriving error estimates of the method, by virtue of the relative energy functional. However, this relative energy functional can measure the distance to a smooth solution only, which is not known to exist in polyhedral domains.

[2] To be precise, it is a weak convergence of a subsequence of the numerical solutions.

Another extension by Feireisl, Karper & Novotný accustomed the scheme to the complete Navier–Stokes–Fourier system in [19]. They managed to prove the convergence of the numerical solutions (up to a subsequence) to a weak solution.

1.5 Sommerville Tetrahedra

For results in the geometric part of this thesis we include a single topic, the *Sommerville tetrahedra*. In 1923, D.M.Y. Sommerville introduced several families of tetrahedral tiling of the three-dimensional space, see [45]. We will present here the basic construction.

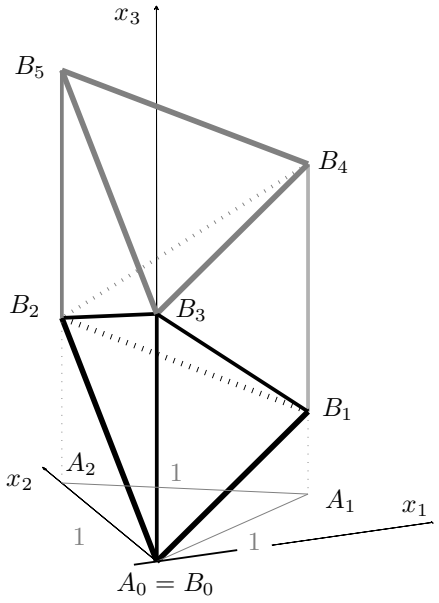


Figure 1.1: Illustration of the Sommerville's construction.

We take a unit equilateral triangle $A_0A_1A_2$, a positive parameter p and create points $B_0, \dots, B_5$ in the following way: $B_0 = A_0$, the first two coordinates of B_1 coincide with those of A_1, the third being p; B_2 is 'above' A_2 in the height $2p$; B_3 'above' A_0 in the height $3p$ and so on. We get three identical tetrahedra $\mathrm{co}\{B_0, B_1, B_2, B_3\}$, $\mathrm{co}\{B_1, B_2, B_3, B_4\}$ and $\mathrm{co}\{B_2, B_3, B_4, B_5\}$, that build a skew prism, see the sketch in Figure 1.1. It is easy to see that creating the points

$$B_z = [A_{i(z)}, zp],\ z \in \mathbb{Z}, \quad \text{where} \quad i(z) \equiv z \mod 3,$$

we can define the set of tetrahedra

$$\big\{\mathrm{co}\{B_z, B_{z+1}, B_{z+2}, B_{z+3}\}; z \in \mathbb{Z}\big\},$$

that tile the whole infinite triangular prism. Repeating the construction above every equilateral triangle in the xy-plane gives a tiling of the three-dimensional space.

Chapter 2

Contributions of the Author in Analysis

2.1 Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension. [35]

As the Karper's numerical scheme to compressible Navier–Stokes system introduced in [37] did not get a warm acceptance within the numerical community, which was driven mainly by the lack of its effective implementation, there was an effort to develop a similar result for a simpler numerical scheme.

The main idea suggested by Karper himself was to use finite differences. The different nature of the density and velocity, representing the state and the flux, respectively, leads naturally to the use of a *staggered grid*. Our space domain, a box, is equidistantly divided into small cubes, centers of which represent the primary grid, whereas the centers of the faces of these cubes represent the points of the dual grid. The density is defined on the primary grid, the velocity on the dual one, while only s-th component is defined at points that are centers of faces, whose normal vector is $\mathbf{e}_s$.

Central differences are used for discretizing the spatial derivatives, the backward Euler method replaces the time derivatives, making the scheme implicit which contributes to its stability. The upwind technique for the convective terms contributes to numerical diffusion and a parabolic regularization is used for the continuity equation.

We point out the main points of the paper.

- The preservation of the positivity of density is proved. The main ingredient is a discrete version of renormalized continuity equation (1.8), for which we need a version of the standard Taylor's Theorem generalized for functions with jumps in their highest order derivatives. We proved such result in a spin-off paper [33], which we did not include in this thesis.
- Existence of the solution to the nonlinear implicit scheme is proved through a fixed point argument.

- Energy estimates ensuring stability as well as some compactness results were derived.

- A version of discrete Sobolev inequality was proved, following the framework developed by Coudière et al. [8].

- A consistency formulation for both continuity and momentum equations is derived. The constraint for the *adiabatic exponent* γ in (1.5) is $\gamma > 3/2$ for the three-dimensional case.

In the meantime, the group of Gallouët has been independently working on the same topic, having produced a number of papers [25, 26, 27].

The theoretical treatment of the scheme, having been almost solely developed by the author is supplemented by a numerical experiments conducted by Bangwei She from Institute of Mathematics, CAS.

The entire original research paper **[35]**, that has been submitted to *Journal of Numerical Mathematics*, is attached as Appendix A.

2.2 A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains. [16]

The numerical method for the compressible Navier–Stokes–Fourier in [19] is designed for a target system in the same polyhedral domain as the numerical scheme. Similarly to the isentropic regime, the polyhedral domain is known to admit weak solutions only. However, we would like to have a chance to compare the numerical solution with (at least local-in-time) strong solution. Following the transition from [37] to [18] for isentropic case, we generalize the result in [19] for the complete system to a smoother domain. The compressible Navier–Stokes–Fourier system is assumed to take place in a C^1 bounded domain, while the numerical domain depends on the discretization parameter, containing the target domain in its interior, but not exceeding its boundary by more that an h-margin, where h is the discretization parameter.

The structure of the proof itself follows the one introduced in [19], with adjustments to the new situation in the domain geometry. Loosely speaking, the main difficulties addressed in the paper are to show that the contribution of the method arising from the difference of the domains is vanishing with $h \to 0$.

The contribution of the author was partial, mainly in conducting some of the estimates, accommodating those from [19] to the new setting. Moreover, due to requirements of the editor on the article's length, most of these estimates are not explicitly included in the final article.

The entire original research paper **[16]**, that was published in *SIAM Journal of Numerical Analysis*, is attached as Appendix B.

2.3 Error estimates for a numerical method for the compressible Navier–Stokes system on sufficiently smooth domains. [14]

In this result an error estimate for a solution of Karper-type scheme to a classical solution of compressible Navier–Stokes system is proved. It is motivated by a work of Gallouet et al. [28], where a similar result is achieved for polyhedral domains. Our improvement goes in two directions; firstly we assume that the fluid is confined to a smooth domain (of the class C^3), while the scheme is designed in a *boundary-fitted* polyhedral domain. The main reason for this separation of the domains is the lack of existence result for a classical solution in polyhedral domains, hence it is not excluded that the result in [28] might be void. The second direction of improvement is that the target solution is not assumed to be classical a priori, but its sufficient regularity is a consequence of sufficient regularity of the data and boundedness of the density. In particular, a weak solution exists, and

- there also exists a local in time strong solution (due to [6]),
- the weak solution coincides with a strong solution on its life span (due to weak-strong uniqueness result from [17]),
- the classical solution is global thanks to the bounded density (blow-up criteria in [46]).

The main tool in the article is a discrete version of a relative energy inequality, which is accustomed to measure a distance of a numerical solution of the scheme to a couple of regular functions. This is performed in three steps. First, a discrete energy ineguality is derived. Then it is accommodated for comparing with a projected smooth function and finally the (suitably extended) classical solution of the compressible Navier–Stokes is taken as these regular functions. All these steps lead to obtaining an inequality for which se can use an argument of the Gronwall type.

The result of **[14]** was also reformulated as an *unconditional* convergence and error estimates result in our paper [15]. The unconditionality resides in having no additional requirements on the classical solution, however, the crucial assumption is boundedness of the sequence of numerical densities. For the similarity of [15] to **[14]** it is not included in this thesis.

The contribution of the author resides mainly in the particular estimates that were conducted for the proof.

The entire original research paper **[14]** that was published in *ESAIM: Mathematical Modelling and Numerical Analysis*, is attached as Appendix C.

Chapter 3

Contributions of the Author in Geometry

The works dealing with numerical schemes mentioned in the previous chapter required certain results on the underlying geometry. The lack of those results lead to production of a series of papers, where the existence results of families of meshes with given properties were shown.

All the three results commented in this chapter had been summarized in the author's overview article [34], which was accepted for publishing in conference proceedings from *PANM 18* and is not included in the thesis.

3.1 Face-to-face partition of 3D space with identical well-centered tetrahedra. [30]

The first result is motivated by the our joint paper [**16**] with Feireisl & Michálek commented in section 2.2. The problem in [**16**] assumes a family of polyhedral domains $\{\Omega_h\}_{h\to 0}$ approximating $\Omega \in C^1$ in the following sense

$$\Omega \subset \Omega_h \subset \{x \in \mathbb{R}^3, \mathrm{dist}[x, \Omega] \leq h\}. \tag{3.1}$$

Every Ω_h is expected to admit a *face-to-face* polyhedral mesh $\mathcal{T}_h$, whose every element possesses a *special point* in its interior: a segment connecting these special points of two neighbouring elements is perpendicular to the common face of those elements.

Moreover, the family of meshes $\{\mathcal{T}_h\}_{h\to 0}$ is supposed to be regular, i.e. a regularity ratio of the elements is supposed to have a positivity lower bound independent of h. The standard regularity ratio compares a radius of the largest ball contained in an element and its diameter.

The easiest choice is to take tetrahedral elements, then for this special point one can take the center of an inscribed sphere. Also different equivalent regularity ratios can be used for tetrahedra.

In [**30**] we use the Sommerville's construction presented in section 1.5 to create a face-to-face tetrahedral mesh of a three-dimensional space and determine

the range of parameters p for which the tetrahedra contain the centers of their inspheres in their interiors (the so called *well-centeredness property*, introduced by VanderZee in [49]). Moreover, a shape-optimal value for p is determined. It is shown that for such case all the elements are identical. For this shape-optimality result an alternative regularity criterion is used, but we would get the same result also for using the criterion commented above [34, Theorem 2].

Scaling the result, we can create a face-to-face tessellation of the whole three-dimensional space using a single type of tetrahedra, whose size does not exceed h. Then, the numerical domain Ω_h is a union of those elements, whose intersection with Ω is not void.

This result is an own work of the author and the entire original research paper **[30]**, that was published in *Applications of Mathematics*, is attached as Appendix D.

3.2 Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains. [32]

The second geometrical result answers in an affirmative way the question of existence of family of *boundary-fitted* meshes for the numerical scheme investigated in **[14]**, see also section 2.3. Notice that in general, the numerical domain Ω_h is not contained in Ω.

After every refinement of a boundary-fitted mesh new vertices on the boundary of the computational domain are created, which must be shifted to the boundary to recover the boundary-fitted property. The fundamental question is, whether we can ensure that after infinitely many refinement steps, the elements do not degenerate.

Such result was available in 2D, see [38]. This result could not be generalized straightforwardly to a higher dimension, since the one-dimensionality of the boundary is a crucial ingredient of the proof.

Again, the Sommerville tetrahedra play a central role in the proof, since the *shape-optimal* one from the previous section can be decomposed into eight identical tetrahedra, similar to the original one, see Křížek [39]. As a consequence, any tetrahedron can be decomposed to eight tetrahedra of a half diameter, with certain regularity ratio being preserved. We work with a special regularity criterion based on the similarity of a tetrahedron to *the Sommerville's tetrahedron*. The whole proof is conducted in the terms of this *Sommerville regularity criterion*.

We derive assumptions on the initial mesh under which a certain minimal regularity of the elements during the refinements and shifts is guaranteed.

The result, containing rather laborious and technical proof, is an own work of the author and the entire original research paper **[32]**, that was published in *Applications of Mathematics*, is attached as Appendix E.

3.3 Construction and shape optimization of simplical meshes in d-dimensional space. [31]

The last of these geometric results is motivated by the Sommerville's construction itself, which can be interpreted as a method which creates a tessellation of d-dimensional space of of $(d-1)$-dimensional one. The idea is to take a simplex of a tessellation of the $(d-1)$-dimensional space and create the infinite prism made of d-dimensional simplices above it, we recall the sketch in Figure 1.1. This induction step can be easily supplemented by an initial step, which is an equidistant tiling of a line. A graph-theory tool of vertex coloring is used to ensure the face-to-face property of the tessellations. An interesting property is the equivolumetricity of the tessellation.

Similarly to the original Sommerville's construction, each of these induction steps is determined up to a positive parameter. Therefore, a simplical tiling of d-dimensional space is determined up to a d-dimensional vector of parameters. In the second part of **[31]** we determine a shape optimal value of this vector of parameters.

After a series of observations the optimization problem reduces to a constrained optimization problem for which the necessary *Karush-Kuhn-Tucker conditions* hold. Unfortunately we are not able to show convexity of our problem, for which KKT conditions are also sufficient. Instead we show existence of the maximizer and also existence of a unique vector satisfying these conditions which then necessarily must be the maximizer.

An interesting observation is that the result of the optimization would be the same, if one optimizes at every level of the construction, which is a simple one-dimensional optimization. In other words, the shape optimal tessellation of a space cannot be created from a sub-optimal tessellation of its hyperplane.

This result is an own work of the author and the entire original research paper **[31]**, that has been submitted to *Discrete and Computational Geometry*, is attached as Appendix F.

Chapter 4

Conclusion

To conclude, the goal of my PhD studies to contribute to the theory of viscous compressible heat conducting fluid was to a certain extent fulfilled. My main contributions to the field were itemized in the previous two chapters. It requires a certain time gap to be able to evaluate the merit of this contribution. By the date of this thesis submission the sets of citations and self-citations of the author's work were coinciding.

The number of open problems is increasing in general in the entire mathematics, not excluding my subfield. We introduce a non-exhaustive list of the tips for possible continuation of the work introduced in this thesis.

- The numerical solutions of the finite difference scheme from **[35]** should be shown to converge to a weak solution of the compressible Navier–Stokes system, in the spirit of Karper [37].

- An extension of the finite-difference scheme from **[35]** for a smooth target domain admitting classical solution would be of interest, compare with the result [18] for the Karper's scheme. The idea is to use some kind of a penalty method to supress the flow outside the physical domain.

- The numerical solutions could be shown to convergence to a *dissipative measure-valued solution*, a concept introduced in [13], according to which such solution coincides with a strong solution emanating from the same initial data on its life-span (a measure-valued–strong uniqueness result). The Karper's scheme has been recently shown to converge through this dissipative measure-valued solution to a classical solution of the compressible Navier–Stokes system in [21] for the whole range of physically relevant adiabatic exponents $\gamma \in (1, 2)$.

- The optimality result on the d-dimensional tessellations can be generalized also to a wider class of these tessellations. In the step from tessellation of a hyperplane to a tessellation of a d-dimensional space a general vector can be used instead of the more restrictive $p_d \mathbf{e}_d$, which is the case considered in **[31]**.

- The tessellations from [**31**] can be proved to possess another interesting properties. For $d = 4$ we should be able to prove, that the tessellation is using only simplex of a single type and its reflections, moreover, it builds a well-centered mesh. Our ultimate goal to prove the well-centered result for a general d is not easy to be obtained, since the simplices are no longer identical for $d \geq 5$ and all we control is the regularity ratio of the *worst* simplex in the tessellation, which is not enough for well-centeredness. It can be shown that there exist two simplices with the same regularity, one of them being well-centered, while the other one not.

This list is far from being complete, but it suggests the directions of the work in the nearest future. It is not excluded, that some of the above listed items will be solved by the time of this thesis' defense.

Bibliography

[1] M. Bulíček, P. Gwiazda, J. Málek, and A. Świerczewska-Gwiazda. On steady flows of incompressible fluids with implicit power-law-like rheology. *Advances in calculus of variations*, 2(2):109–136, 2009.

[2] M. Bulíček, P. Gwiazda, J. Málek, and A. Świerczewska-Gwiazda. On unsteady flows of implicitly constituted incompressible fluids. *SIAM Journal on Mathematical Analysis*, 44(4):2756–2801, 2012.

[3] L. Caffarelli, R. Kohn, and L. Nirenberg. Partial regularity of suitable weak solutions of the Navier-Stokes equations. *Commun. Pure Appl. Math.*, 35:771–831, 1982.

[4] R. Chhabra and J. Richardson. *Non-Newtonian Flow and Applied Rheology: Engineering Applications*. Elsevier Science, 2011.

[5] E. Chiodaroli, C. De Lellis, and O. Kreml. Global ill-posedness of the isentropic system of gas dynamics. *Communications on Pure and Applied Mathematics*, 68(7):1157–1190, 2015.

[6] Y. Cho, H. J. Choe, and H. Kim. Unique solvability of the initial boundary value problems for compressible viscous fluids. *Journal de mathématiques pures et appliquées*, 83(2):243–275, 2004.

[7] A. Chorin and J. Marsden. An introduction to mathematical fluid mechanics. *Spring Verlang, New York*, 1979.

[8] Y. Coudière, T. Gallouët, and R. Herbin. Discrete Sobolev inequalities and Lp error estimates for finite volume solutions of convection diffusion equations. *ESAIM: M2AN*, 35:767–778, 2001.

[9] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations i. *Revue française d'automatique, informatique, recherche opérationnelle. Mathématique*, 7(3):33–75, 1973.

[10] C. de Lellis and L. Székelyhidi. On admissibility criteria for weak solutions of the Euler equations. *Archive for Rational Mechanics and Analysis*, 195(1):225, 2010.

[11] C. L. Fefferman. Existence and smoothness of the Navier-Stokes equation. In *The millennium prize problems*, pages 57–67. Providence, RI: American Mathematical Society (AMS); Cambridge, MA: Clay Mathematics Institute, 2006.

[12] E. Feireisl. *Dynamics of viscous compressible fluids.* Oxford: Oxford University Press, 2004.

[13] E. Feireisl, P. Gwiazda, A. Świerczewska-Gwiazda, and E. Wiedemann. Dissipative measure-valued solutions to the compressible Navier–Stokes system. *Calculus of Variations and Partial Differential Equations*, 55(6):141, 2016.

[14] E. Feireisl, R. Hošek, D. Maltese, and A. Novotný. Error estimates for a numerical method for the compressible Navier-Stokes system on sufficiently smooth domains. *ESAIM: M2AN*, 51(1):279–319, 2017.

[15] E. Feireisl, R. Hošek, D. Maltese, and A. Novotný. Unconditional convergence and error estimates for bounded numerical solutions of the barotropic Navier–Stokes system. *To appear in Numerical Methods for Partial Differential Equations. Preprint available at math.cas.cz*, 2017.

[16] E. Feireisl, R. Hošek, and M. Michálek. A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains. *SIAM J. Numer. Anal.*, 54(5):3062–3082, 2016.

[17] E. Feireisl, B. J. Jin, and A. Novotný. Relative entropies, suitable weak solutions, and weak-strong uniqueness for the compressible Navier–Stokes system. *Journal of Mathematical Fluid Mechanics*, 14(4):717–730, 2012.

[18] E. Feireisl, T. Karper, and M. Michálek. Convergence of a numerical method for the compressible Navier–Stokes system on general domains. *Numer. Math.*, 134(4):667–704, 2016.

[19] E. Feireisl, T. Karper, and A. Novotný. A convergent numerical method for the Navier–Stokes–Fourier system. *IMA J. Numer. Anal.*, 36(4):1477, 2015.

[20] E. Feireisl, T. G. Karper, and M. Pokorný. *Mathematical theory of compressible viscous fluids: Analysis and numerics.* Birkhäuser, 2016.

[21] E. Feireisl and M. Lukáčová-Medviďová. Convergence of a mixed finite element-finite volume scheme for the isentropic Navier-Stokes system via dissipative measure-valued solutions. *Preprint available at math.cas.cz*, 2016.

[22] E. Feireisl and A. Novotný. *Singular limits in thermodynamics of viscous fluids.* Basel: Birkhäuser, 2009.

[23] E. Feireisl and A. Novotný. Weak–strong uniqueness property for the full Navier–Stokes–Fourier system. *Archive for Rational Mechanics and Analysis*, 204(2):683–706, 2012.

[24] G. Gallavotti. *Foundations of fluid dynamics.* Springer Science & Business Media, 2013.

[25] T. Gallouët, R. Herbin, J.-C. Latché, and D. Maltese. Convergence of the MAC scheme for the compressible stationary Navier–Stokes equations. *ArXiv e-prints*, July 2016.

[26] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Implicit MAC scheme for compressible Navier–Stokes equations: unconditional error estimates. *Preprint*, 2016.

[27] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Convergence of the marker-and-cell scheme for the semi-stationary compressible Stokes problem. *Mathematics and Computers in Simulation*, 2016. available on line.

[28] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Error estimates for a numerical approximation to the compressible barotropic Navier–Stokes equations. *IMA J. Numer. Anal.*, 36(2):543–592, 2016.

[29] E. Hopf. Über die Anfangswertaufabe für die hydrodynamischen Grundgleichungen. *Math. Nachr.*, 4:213–231, 1951.

[30] R. Hošek. Face-to-face partition of 3D space with identical well-centered tetrahedra. *Appl. Math.*, 60(6):637–651, 2015.

[31] R. Hošek. Construction and shape optimization of simplical meshes in d-dimensional space. *Submitted to Disc. Comp. Geom. Preprint available at ArXiv.org*, June 2016.

[32] R. Hošek. Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains. *Appl. Math.*, 61(3):233–251, 2016.

[33] R. Hošek. Expressing the remainder of Taylor polynomial when the function lacks smoothness. *Submitted to Elemente der Mathematik. Preprint available at math.cas.cz*, 2017.

[34] R. Hošek. The role of Sommerville tetrahedra in numerical mathematics. In *Programy a algoritmy numerické matematiky 18*, 2017.

[35] R. Hošek and B. She. Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension. *Submitted to Journal of Numerical Mathematics*, 2017.

[36] K. H. Karlsen and T. K. Karper. A convergent mixed method for the Stokes approximation of viscous compressible flow. *IMA Journal of Numerical Analysis*, 32(3):725, 2012.

[37] T. K. Karper. A convergent FEM-DG method for the compressible Navier-Stokes equations. *Numer. Math.*, 125(3):441–510, 2013.

[38] S. Korotov, M. Křížek, and P. Neittaanmäki. On the existence of strongly regular families of triangulations for domains with a piecewise smooth boundary. *Appl. Math., Praha*, 44(1):33–42, 1999.

[39] M. Křížek. An equilibrium finite element method in three-dimensional elasticity. *Apl. Mat.*, 27:46–75, 1982.

[40] O. Ladyzhenskaya. The mathematical theory of viscous incompressible flow. New York - London - Paris: Gordon and Breach Science Publishers. XVIII, 224 p. (1969)., 1969.

[41] J. Leray. Sur le mouvement d'un liquide visqueux emplissant l'espace. *Acta Math.*, 63:193–248, 1934.

[42] P.-L. Lions. *Mathematical topics in fluid mechanics. Vol. 2: Compressible models.* Oxford: Clarendon Press, 1998.

[43] N. V. Priezjev and S. M. Troian. Influence of periodic wall roughness on the slip behaviour at liquid/solid interfaces: molecular-scale simulations versus continuum predictions. *Journal of Fluid Mechanics*, 554:25–46, 2006.

[44] K. R. Rajagopal. On implicit constitutive theories for fluids. *Journal of Fluid Mechanics*, 550:243–249, 003 2006.

[45] D. M. Y. Sommerville. Space-filling tetrahedra in Euclidean space. *Proc. Edinburgh Math. Soc.*, 41:49–57, 1923.

[46] Y. Sun, C. Wang, and Z. Zhang. A Beale–Kato–Majda blow-up criterion for the 3-d compressible Navier–Stokes equations. *Journal de mathématiques pures et appliquées*, 95(1):36–47, 2011.

[47] A. Valli. An existence theorem for compressible viscous fluids. *Ann. Mat. Pura Appl. (4)*, 130:197–213, 1982.

[48] A. Valli and W. M. Zajączkowski. Navier–Stokes equations for compressible fluids: global existence and qualitative properties of the solutions in the general case. *Communications in mathematical physics*, 103(2):259–296, 1986.

[49] E. VanderZee, A. N. Hirani, D. Guoy, and E. A. Ramos. Well-centered triangulation. *SIAM J. Sci. Comput.*, 31(6):4497–4523, 2010.

List of publications of the author

Papers published in a reviewed journal

- **[14]** E. Feireisl, R. H., D. Maltese, and A. Novotný. Error estimates for a numerical method for the compressible Navier–Stokes system on sufficiently smooth domains. *ESAIM: M2AN*, 51(1):279–319, 2017.
- **[16]** E. Feireisl, R. H., and M. Michálek. A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains. *SIAM J. Numer. Anal.*, 54(5):3062–3082, 2016.
- **[32]** R. H. Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains. *Appl. Math.*, 61(3):233–251, 2016.
- **[30]** R. H. Face-to-face partition of 3D space with identical well-centered tetrahedra. *Appl. Math.*, 60(6):637–651, 2015.
- P. Drábek and R. H. Properties of solution diagrams for bistable equations. *Electron. J. Differ. Equ.*, 2015(156):1–19, 2015.

Papers accepted for publication

- [15] E. Feireisl, R. H., D. Maltese, and A. Novotný. Unconditional convergence and error estimates for bounded numerical solutions of the barotropic navier-stokes system. *To appear in NMPDE*, 2017.
- [34] R. H. The role of Sommerville tetrahedra in numerical mathematics. In *Programy a algoritmy numerick matematiky 18*, 2017.
- R. H. Walks in path graph on four vertices and Fibonacci sequence. *To appear in Miskolc Mathematical Notes*, 2017.

Submitted papers

- **[35]** R. H. and B. She. Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension. *Submitted to Journal of Numerical Mathematics*, 2017.
- [33] R. H. Expressing the remainder of Taylor polynomial when the function lacks smoothness. *Submitted to Elemente der Mathematik*, 2017.
- **[31]** R. H. Construction and shape optimization of simplical meshes in d-dimensional space. *Submitted to Disc. Comp. Geom. Preprint available at ArXiv.org*, June 2016.

Appendix A

R. H., B. She: Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension.

Stability and consistency of a finite difference scheme for compressible viscous isentropic flow in multi-dimension

Radim Hošek, Bangwei She

Institute of Mathematics, Czech Academy of Sciences*

January 26, 2017

Abstract

Motivated by the work of Karper [27], we propose a numerical scheme to compressible Navier–Stokes system in multi-spatial dimensions, based on finite differences. The backward Euler method is applied for the time discretization, while a staggered grid, with continuity and momentum equations on different grids, is used in space. The existence of a solution to the implicit nonlinear scheme, strictly positivity of the numerical density, stability and consistency of the method are proved. The theoretical part is complemented by computational results that are performed in two spatial dimensions.

Key words: compressible Navier-Stokes, finite difference method, positivity preserving, energy stability, consistency

1 Introduction

The compressible Navier–Stokes system as a set of balance laws for mass and momentum, describes the flow of isentropic viscous gas, where the thermal effects are neglected. Let ϱ, $\mathbf{u}$ be the density and velocity field, the governing equations read

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{1}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x p(\varrho) = \mathrm{div}_x \mathbb{S} + \mathbf{f}. \tag{2}$$

Unlike the incompressible case, pressure in here is a function of density, assumed as

$$p(\varrho) = a\varrho^\gamma, \quad a > 0, \gamma > 1, \tag{3}$$

where the important features of the pressure are its convexity and asymptotic behaviour. Discussions about weakening this assumption can be found in [10]. For the consistency formulation, we need $\gamma > \frac{3}{2}$ for three-dimensional flow, which covers the case of a monatomic gas.

For the sake of easing the computation, the viscous stress tensor is assumed to take the form $\mathbb{S} = \mu \nabla_x \mathbf{u}$, $\mu > 0$ is the viscosity coefficient, and $\mathrm{div}_x \mathbb{S} = \mu \Delta_x \mathbf{u}$. We also omit the external forces, i.e. we set $\mathbf{f} \equiv 0$, bearing in mind that including them would not bring any insurmountable difficulties.

The system is complemented with initial conditions

$$\varrho|_{t=0} = \varrho_0 > 0, \qquad \mathbf{u}|_{t=0} = \mathbf{u}_0, \tag{4}$$

and homogeneous Dirichlet boundary condition for velocity

$$\mathbf{u}|_{\partial\Omega} = 0, \tag{5}$$

where $\Omega \subset \mathbb{R}^d$ is assumed to be a bounded Lipschitz domain, for space dimension $d = 2$ or 3. The time interval is $[0, T]$, without any assumptions on its size. More over, we expect the regularity $\varrho_0 \in L^\gamma(\Omega)$, $\mathbf{u}_0 \in W_0^{1,2}(\Omega)$.

The existence of strong solutions to (1–5) for *sufficiently smooth* initial data was proved in [37], however only for a possibly small time interval $[0, T^\star)$. Therefore, it was welcome, when the unconditional existence of *weak solution* was proved by Lions [32] and further developed in [18]. However, the existence result still requires $\gamma > \frac{3}{2}$, which does not cover the case of a diatomic gas. There are results on full system describing compressible flow, i.e. considering also the balance law of energy. Numerical schemes can be found in the framework of finite difference, finite

*The research of the authors leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC Grant Agreement 320078. The Institute of Mathematics of the Academy of Sciences of the Czech Republic is supported by RVO:67985840.

volume, finite element, discontinuous Galerkin, gas kinetic BGK or mixtures of them. Representative examples are [31, 24, 3, 30, 6, 29, 35, 36, 38]. While considering the isentropic case, L^2-stable scheme has been studied in [1, 19], where upwind and pressure projection are the main technique. All speed asymptotic-preserving scheme can be found in [25] especially for low Mach number limit. Error estimates for the isentropic case was studied in [12, 23, 33]. The convergence of the compressible Navier-Stokes to its incompressible limit was numerically measured by a relative entropy at low Mach regime in [17]. Recently, Gallouët et al. proposed a MAC scheme similar to ours, for which they prove convergence results for (semi)stationary flows [20, 22], and error estimates for compressible Navier–Stokes [21].

Concerning the convergence of the numerical methods, to our best knowledge there is only one result in [27], where the scheme is based on finite element combined with discontinuous Galerkin method and uses also *upwind flux*. For linear problems, stability and consistency is enough to ensure convergence.In [27], Karper mimicked the proof of existence of weak solution for compressible Navier–Stokes system by Lions [32] and then showed for vanishing discretization parameter the convergence of the numerical solution, up to a subsequence, to a weak solution. This work had been further extended for smooth domains using non-fitted mesh [15] and to a heat conducting case [13, 14].

The scheme in [27] did not obtain a grateful acceptance, being labeled as *too academic* within the computational community. Therefore, an effort to prove convergence of a simpler numerical scheme motivated our result. In [28], Karper suggests a finite difference scheme for one dimensional compressible Navier–Stokes and shows its convergence. Moreover, it is suggested there to extend the result to multi-dimension, which we bring in this paper. Our result can be viewed as a starting point for two possible directions. One of them is continuation in the spirit of [27] in order to prove convergence of the (subsequence of) numerical solution to a weak solution. The other direction could be proving a convergence to measure-valued solution, which, in a suitable setting, coincides with a strong solution on its (possibly short) life span, see [11, 16].

In this paper we present the theoretical results of stability and consistency followed by numerical experiments. The paper is organized as follows. We explain the detailed scheme in Section 2. Then comes the proofs of positivity preserving of density, existence of the solution at any time level, energy stability and derivation of uniform estimates in Section 3, the consistency formulation in Section 4 and finally, numerical tests of the method in Section 5.

2 The numerical method

2.1 Time discretization

We discretize the time step equidistantly using Δt ($T = N_t \Delta t$) and define function only at these time instants $f^k := f(k\Delta t)$. The time derivative is approximated by the backward Euler method,

$$(\partial_h^t f)^n := \frac{f^n - f^{n-1}}{\Delta t}, \quad n = 1, 2, \ldots, N_t.$$

2.2 Spatial grids

2.2.1 Primary and dual grids

For convenience, the domain in our problem is set as $Q_T = I \times \Omega = [0, T] \times (0, L_x)^d$. A staggered grid is used in our spatial discretization. The domain Ω is uniformly discretized with mesh size $h = L_x/N_x$, i.e. $\overline{\Omega} := \bigcup \overline{Q_K}$ where the element Q_K is given by

$$Q_K = ((i-1)h; ih) \times ((j-1)h; jh) \times ((k-1)h; kh), \forall i, j, k \in \{1, \ldots, N_x\},$$

for example in three dimensions. The primary grid $\mathcal{T}$ is built by the centers K of these elements. Boundary of each element Q_K is created by faces F_σ, whose centers σ build the secondary grid $\mathcal{E}$, cf. Figure 1 which depicts the simpler two-dimensional case. Points $\sigma \in \mathcal{E}$ belonging to $\partial\Omega$ form $\mathcal{E}_{\text{ext}}$, while $\mathcal{E}_{\text{int}} = \mathcal{E} \setminus \mathcal{E}_{\text{ext}}$. We denote $\mathcal{E}(K)$ as the set of points that are at the center of the faces of element Q_K,

$$\mathcal{E}(K) := \left\{ \sigma = K \pm \frac{h}{2}\mathbf{e}_s, K \in \mathcal{T}, s = 1, \ldots, d \right\},$$

where $\mathbf{e}_s$ is a unit basis vector in one of the space directions (i.e. either $\mathbf{e}_1, \mathbf{e}_2$ or $\mathbf{e}_3$). Note that σ is linked with the direction of its normal vector $\mathbf{e}_s$, we denote it also as

$$\sigma_{,s\pm} = K \pm \frac{h}{2}\mathbf{e}_s.$$

On the other hand, any $\sigma \in \mathcal{E}_{\text{int}}$ adjacent to the elements K and $L \in \mathcal{N}(K)$ of the primary mesh, where $\mathcal{N}(K)$ is the collection neighbouring elements of K, we write $\sigma = K|L$ if $L = K + h\mathbf{e}_s$ for some $s = 1, \ldots, d$.

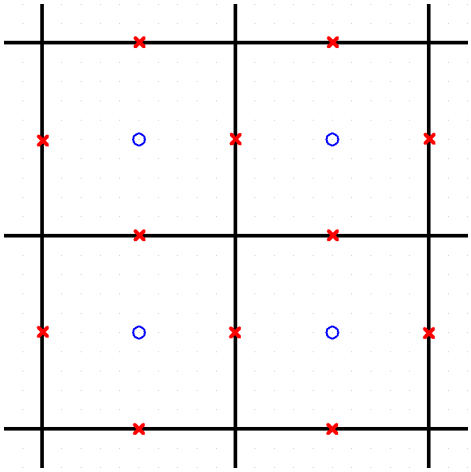


Figure 1: Space discretization: Blue circles o and red crosses × are the points of primary mesh and dual mesh, respectively.

2.2.2 Transferring quantities between grids

For any quantity f_h defined on the primary mesh $\mathcal{T}$ we denote its value at K as f_K. It can be interpolated to the dual one for $\sigma = K|L \in \mathcal{E}_{\text{int}}$ with

$$\{f\}_\sigma = \frac{1}{2}(f_K + f_L).$$

Mainly vector quantities are defined on dual grid. We define only the s-th component g^s_σ of vector quantity $\mathbf{g}_h$ on each face $\sigma \in \mathcal{E}$, if $\mathbf{e}_s$ is the normal vector to the face F_σ. Then the projection to the primary grid reads

$$\bar{\mathbf{g}}_K = \frac{1}{2}\sum_{s=1}^{d}\left(g^s_{\sigma,s+} + g^s_{\sigma,s-}\right)\mathbf{e}_s. \tag{6}$$

2.2.3 Extending discrete quantities

We will compute numerical solutions using decreasing discretization parameter and investigate the weak limit of the numerical solutions, considering these being L^p functions. For this purpose we interpret discrete quantities defined in primary mesh $\mathcal{T}$ as piecewise constant functions with respect to this mesh, defined by

$$f_h(\mathbf{x}) = f_K, \text{ for } \mathbf{x} \in Q_K.$$

We denote the space of piecewise constant functions with respect to the grid $\mathcal{T}$ by

$$X(\mathcal{T}) = \{f \in L^\infty(\Omega); f|_K \equiv f_K \in \mathbb{R}\}.$$

Discrete vector quantity defined component-wise on dual mesh (g^s_σ) can be also identified with piecewise constants, which is

$$g^s_h(\mathbf{x} - \frac{h}{2}\mathbf{e}_s) = g^s_{\sigma,s-}, \text{ for } \mathbf{x} \in Q_K. \tag{7}$$

for all $\sigma \in \mathcal{E}$. Note that the s-th component of $\mathbf{g}$ is constant in the neighbourhood Q_σ of σ, which is the center of the face F_σ. The space of such functions is denoted by $X(\mathcal{E})^d$, we also define

$$X(\mathcal{E}_{\text{int}})^d = \left\{\mathbf{g} \in X(\mathcal{E})^d; \mathbf{g}|_{\mathcal{E}_{\text{ext}}} = \mathbf{0}\right\}.$$

To indicate the mesh-dependence of these functions, here and hereafter we equip them with subscript h. This subscript will be omitted any time where values at particular points of the mesh are considered. Then, for all $f_h \in X(\mathcal{T})$, $\mathbf{g}_h \in X(\mathcal{E}_{\text{int}})^d$ we have

$$h^d \sum_{K\in\mathcal{T}} f_K = \int_\Omega f_h \,\mathrm{d}x, \qquad h^d \sum_{\sigma\in\mathcal{E}_{\text{int}}} g^s_\sigma \mathbf{e}_s = \int_\Omega \mathbf{g}_h \,\mathrm{d}x. \tag{8}$$

Besides the piecewise constant extension $\mathbf{g}_h \in X(\mathcal{E}_{\text{int}})^d$, we will need also an extension to the space of functions with piecewise constant first order derivatives, i.e. piecewise linears with respect to primary cells,

$$\widehat{\mathbf{g}}(\mathbf{x}) = \sum_{s=1}^{d} \left((g^s_{\sigma,s+} - g^s_{\sigma,s-}) \left(\frac{x_s}{h} - \left\lfloor \frac{x_s}{h} \right\rfloor \right) + g^s_{\sigma,s-} \right) \mathbf{e}_s, \qquad \text{for } \mathbf{x} = (x_1, \dots, x_d) \in Q_K,$$

where $\lfloor \cdot \rfloor$ is the floor rounding operator. The values of discrete quantities outside Ω that we use, will be extrapolated according to the boundary conditions, see Section 2.5.

2.2.4 Projection of continuous quantities to the grid

We will also need to project smooth quantities to our grids. We define the projection operators $\Pi^P : L^1(\Omega) \to X(\mathcal{T})$ and $\Pi^D : W^{1,1}_0(\Omega; \mathbb{R}^d) \to X(\mathcal{E}_{\text{int}})^d$ with

$$(\Pi^P \phi)_K = \frac{1}{h^d} \int_{Q_K} \phi \, \mathrm{d}x, \qquad (\Pi^D \mathbf{v})_\sigma = \frac{\mathbf{e}_s}{h^{d-1}} \int_{F_\sigma} v^s \, \mathrm{d}S_x.$$

Note that $\mathbf{v} \in (W^{1,1}(\Omega))^d$ is the minimal requirement so that $\mathbf{v}$ has bounded traces and the projection Π^D is well defined. The zero trace at $\partial\Omega$ guarantees that $\Pi^D \mathbf{v}|_\sigma = \mathbf{0}$ for $\sigma \in \mathcal{E}_{\text{ext}}$.

The projection to the primary grid satisfies

$$\sum_{K \in \mathcal{T}} (\Pi^P \phi)_K = \int_\Omega \phi \, \mathrm{d}x, \tag{9}$$

and using Taylor expansion and (7), one can derive the following estimates,

$$\|\Pi^P \phi - \phi\|_{L^p(\Omega)} \le h \|\nabla \phi\|_{L^p(\Omega)}, \qquad \|\Pi^D \mathbf{v} - \mathbf{v}\|_{L^p(\Omega;\mathbb{R}^d)} \le h \|\nabla \mathbf{v}\|_{L^p(\Omega)}. \tag{10}$$

2.3 Standard Difference Operators

2.3.1 Definitions

In this paper we use two basic difference operators

$$(\partial^s_h f)_\sigma = \frac{f_L - f_K}{h}, \qquad \text{for } f_h \in X(\mathcal{T}), \tag{11}$$

$$(\partial^s_h g^s)_K = \frac{g^s_{\sigma,s+} - g^s_{\sigma,s-}}{h}, \qquad \text{for } \mathbf{g}_h \in X(\mathcal{E}_{\text{int}})^d. \tag{12}$$

A property worth noticing is that the discrete derivatives and therefore also all first order differential operators can be viewed as mappings between the grids. The mixed derivative is defined as

$$(\partial^r_h g^s)_{K + \frac{h}{2}\mathbf{e}_s \pm \frac{h}{2}\mathbf{e}_r} = \mp \frac{g^s_{K+\frac{h}{2}\mathbf{e}_s} - g^s_{K+\frac{h}{2}\mathbf{e}_s \pm h\mathbf{e}_r}}{h}, \qquad \text{for } \mathbf{g}_h \in X(\mathcal{E}_{\text{int}})^d \text{ and every } K \in \mathcal{T}. \tag{13}$$

Notice that (13) can cover (12) if $r = s$ and $K + \frac{h}{2}\mathbf{e}_s \pm \frac{h}{2}\mathbf{e}_r \in \Omega$.
We can naturally define the discrete divergence operator with

$$(\mathrm{div}_h \mathbf{g})_K = \sum_{s=1}^{d} (\partial^s_h g^s)_K,$$

and Laplace operators by

$$(\Delta_h f)_K = (\mathrm{div}_h \partial^s f)_K = \frac{1}{h^2} \sum_{L \in \mathcal{N}(K)} (f_L - f_K), \qquad (\Delta_h g^s)_\sigma = \frac{1}{h^2} \sum_{r=1}^{d} (g^s_{\sigma - \mathbf{e}_r} - 2g^s_\sigma + g^s_{\sigma + \mathbf{e}_r}).$$

2.3.2 Calculus for the discrete operators

From the definition of differential operators one deduces the following two properties that are a discrete counterpart of the integration by parts.

Lemma 2.1. *Let $f_h \in X(\mathcal{T}), \mathbf{g}_h \in X(\mathcal{E}_{\text{int}})^d, \mathbf{v}_h \in X(\mathcal{E}_{\text{int}})^d$. Then*

$$\sum_{K \in \mathcal{T}} (\mathrm{div}_h \mathbf{g})_K f_K = - \sum_{\sigma \in \mathcal{E}_{\text{int}}} g^s_\sigma (\partial^s_h f)_\sigma. \tag{14}$$

$$- \sum_{\sigma \in \mathcal{E}_{\text{int}}} (\Delta_h v^s)_\sigma g^s_\sigma = \sum_{K \in \mathcal{T}} \sum_{s=1}^{d} \left((\partial^s_h g^s)_K (\partial^s_h v^s)_K + \frac{1}{2} \sum_{\substack{r=1 \\ r \neq s}}^{d} \sum_{i=1}^{2} (\partial^r_h g^s)_{K + \frac{h}{2}\mathbf{e}_s + (-1)^i \frac{h}{2}\mathbf{e}_r} (\partial^r_h v^s)_{K + \frac{h}{2}\mathbf{e}_s + (-1)^i \frac{h}{2}\mathbf{e}_r} \right). \tag{15}$$

The proof can be found in the Appendix A. Taking $\mathbf{v}_h = \mathbf{g}_h$ in (15) we obtain

$$-\sum_{\sigma\in\mathcal{E}_{\rm int}}(\Delta_h g^s)_\sigma g^s_\sigma = \sum_{K\in\mathcal{T}}\sum_{s=1}^{d}\left(|\partial_h^s g^s|^2_K + \frac{1}{2}\sum_{\substack{r=1\\ r\neq s}}^{d}\left(|\partial_h^r g^s|^2_{K+\frac{h}{2}\mathbf{e}_s+\frac{h}{2}\mathbf{e}_r} + |\partial_h^r g^s|^2_{K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_r}\right)\right) =: \sum_{K\in\mathcal{T}}\sum_{s=1}^{d}\sum_{r=1}^{d}|\widetilde{\partial_h^r g^s}|^2_K. \tag{16}$$

2.3.3 Inverse estimates

Inverse estimate is a typical powerful tool for obtaining compactness result for a sequence of numerical solutions. We introduce its analogue for our finite difference setting in the following two lemmas.

Lemma 2.2. *Let $f_h \in X(\mathcal{T})$ and $\mathbf{g}_h \in X(\mathcal{E}_{\rm int})^d$. Then we have*

$$\|\partial_h^s f\|_{L^p(\Omega)} \le c(p)h^{-1}\|f\|_{L^p(\Omega)}, \qquad \|{\rm div}_h\mathbf{g}\|_{L^p(\Omega)} \le c(p)h^{-1}\|\mathbf{g}\|_{L^p(\Omega)},$$

with a positive constant $c(p)$, independent of h.

Proof. We observe, by virtue of the generalized triangle inequality, that

$$h^d\sum_{\sigma\in\mathcal{E}_{\rm int}}|(\partial_h^s f)_\sigma|^p = h^{d-p}\sum_{\sigma\in\mathcal{E}_{\rm int}}|f_L - f_K|^p \le c(p)h^{d-p}\sum_{K\in\mathcal{T}}|f_K|^p,$$
$$h^d\sum_{K\in\mathcal{T}}|({\rm div}_h\mathbf{g})_K|^p = h^{d-p}\sum_{K\in\mathcal{T}}|g^s_{\sigma,s+} - g^s_{\sigma,s-}|^p \le c(p)h^{d-p}\sum_{\sigma\in\mathcal{E}_{\rm int}}|g^s_\sigma|^p.$$

Using (8) concludes the proof. □

Lemma 2.3. *Let $p > q \ge 1$ and $f_h \in X(\mathcal{T}), \mathbf{g}_h \in X(\mathcal{E}_{\rm int})^d$. Then we have the estimate*

$$\|f\|_{L^p(\Omega)} \le c(p,q)h^{d(\frac{1}{p}-\frac{1}{q})}\|f\|_{L^q(\Omega)}, \qquad \|\mathbf{g}\|_{L^p(\Omega)} \le c(p,q)h^{d(\frac{1}{p}-\frac{1}{q})}\|\mathbf{g}\|_{L^q(\Omega)}.$$

Proof. We show the proof for $f \in X(\mathcal{T})$ only, leaving the other part to the kind reader. By definition, $\|f\|^p_{L^p(K)} = h^d|f_K|^p$, which implies $\|f\|_{L^p(K)} = h^{d(\frac{1}{p}-\frac{1}{q})}\|f\|_{L^q(K)}$. Then from the inequality

$$\sqrt[m]{S^m+1} \le S+1, \quad S\ge 0, m\ge 1, \qquad \text{setting } S = \frac{a^q}{b^q}, m = \frac{p}{q},$$

we deduce $\sqrt[p]{A^p+B^p} \le \sqrt[q]{A^q+B^q}$ and using induction also $\sqrt[p]{\sum_i a_i^p} \le \sqrt[q]{\sum_i a_i^q}$, which implies

$$\|f\|_{L^p(\Omega)} = \sqrt[p]{\sum_{K\in\mathcal{T}}\|f\|^p_{L^p(K)}} \le c(p,q)h^{d(\frac{1}{p}-\frac{1}{q})}\sqrt[p]{\sum_{K\in\mathcal{T}}\|f\|^p_{L^q(K)}} \le c(p,q)h^{d(\frac{1}{p}-\frac{1}{q})}\sqrt[q]{\sum_{K\in\mathcal{T}}\|f\|^q_{L^q(K)}} = c(p,q)h^{d(\frac{1}{p}-\frac{1}{q})}\|f\|_{L^q(\Omega)}.$$

□

Remark 1. *Analogously one would show that for any quantity f that is piecewise constant in time with respect to Δt-equidistant discretization of $[0,T]$ and any $p > q \ge 1$ it holds that*

$$\|f\|_{L^p(0,T)} \lesssim (\Delta t)^{(\frac{1}{p}-\frac{1}{q})}\|f\|_{L^q(0,T)}, \tag{17}$$

2.4 Upwind discretization and upwind derivative

The 'upwinding' or 'upstreaming' is a method vastly used in finite volume schemes for discretizing flow quantities. For its locally conservative properties (see [9, Section 1.1]), it appears useful in wider set of methods. First, we set $f^+ = \max\{0, f\}$, $f^- = \min\{0, f\}$. Then, we can write $f = f^+ + f^-$, $f^+ = \frac{1}{2}(f+|f|)$ and $f^- = \frac{1}{2}(f-|f|)$.

Let $\mathbf{u}_h \in X(\mathcal{E}_{\rm int})^d$ and $\sigma = K|L, L = K + h\mathbf{e}_s, s = 1,\dots,d$. Then we define the upwind flux of the quantity $f \in X(\mathcal{T})$ with respect to velocity $\mathbf{u}$ by

$${\rm Up}[f,\mathbf{u}]_\sigma = f_K(u^s_\sigma)^+ + f_L(u^s_\sigma)^-,$$

and the *upwind discrete derivative* and the upwind divergence with

$$\partial_s^{\rm Up}[f,\mathbf{u}]_K = \frac{{\rm Up}[f,\mathbf{u}]_{\sigma,s+} - {\rm Up}[f,\mathbf{u}]_{\sigma,s-}}{h}, \qquad {\rm div}_{\rm Up}[f,\mathbf{u}]_K = \sum_{s=1}^{d}\partial_s^{\rm Up}[f,\mathbf{u}]_K.$$

The following lemma is then a simple corollary of Lemma 2.1.

Lemma 2.4. *Let $f_h \in X(\mathcal{T}), \mathbf{v}_h = [v^1, \cdots, v^d] \in X(\mathcal{E}_{\rm int})^d$, then* $\sum_{K\in\mathcal{T}} \mathrm{div}_{\rm Up}[f,\mathbf{v}]_K = 0.$

The next lemma shows the difference between upwinding and averaging. It can be obtained by direct calculation.

Lemma 2.5. *Let $f \in X(\mathcal{T}), \mathbf{v} = [v^1, v^2, v^3] \in X(\mathcal{E}_{\rm int})^d$. Then,*

$$\mathrm{Up}[f,\mathbf{v}]_\sigma = \{f\}_\sigma (v^s)_\sigma - \frac{h}{2}|v^s_\sigma|(\partial^s_h f)_\sigma.$$

2.5 The Method

We introduce the following implicit scheme,

$$\partial^t_h \varrho^n_K + \mathrm{div}_{\rm Up}[\varrho^n, \mathbf{u}^n]_K - h^\alpha(\Delta_h \varrho^n)_K = 0, \tag{18}$$

$$\partial^t_h(\{\varrho\bar{u}\}_\sigma)^n + \{\mathrm{div}_{\rm Up}[\varrho^n\bar{\mathbf{u}}^n, \mathbf{u}^n]\}_\sigma + \big(\partial^s_h p(\varrho^n)\big)_\sigma \mathbf{e}_s - \mu(\Delta_h \mathbf{u}^n)_\sigma - h^\alpha \sum_{r=1}^d \big\{\partial^r_h\big(\{\bar{\mathbf{u}}^n\}\partial^r_h\varrho^n\big)\big\}_\sigma = 0, \tag{19}$$

for all $K \in \mathcal{T}, \sigma \in \mathcal{E}_{\rm int}$ and $n = \{1, \ldots, N_t\}$, with initial values

$$\varrho^0_K = \Pi^P \varrho_0, \qquad \bar{\mathbf{u}}^0_K = \Pi^P \mathbf{u}_0. \tag{20}$$

and boundary conditions

$$\mathbf{u}^n_\sigma = 0, \qquad (\mathbf{n}\cdot\nabla_h\rho^n)_\sigma = 0, \qquad \text{for } \sigma \in \mathcal{E}_{\rm ext}, n = 0, \ldots N_t, \tag{21}$$

To be more specific, the boundary conditions are implemented as $\rho_{\sigma-\frac{h}{2}\mathbf{e}_s} = \rho_{\sigma+\frac{h}{2}\mathbf{e}_s}$ and $\mathbf{u}_{\sigma+\frac{h}{2}\mathbf{e}_r-\frac{h}{2}\mathbf{e}_s} = -\mathbf{u}_{\sigma+\frac{h}{2}\mathbf{e}_r+\frac{h}{2}\mathbf{e}_s}$ for any $\sigma \in \mathcal{E}_{\rm ext}$ and $r \neq s$.

The way of projecting the initial velocity is motivated by the fact that nothing like (9) holds true for Π^D and also that we do not need the initial velocity on the faces $\sigma \in \mathcal{E}$.

Remark 2. *There is no boundary condition for density on the continuous level. However we need to equip the scheme with the no flux boundary condition for the density due to the additional artificial diffusion term in the scheme, which regularizes the continuity equation.*

3 Existence, stability and energy estimates

We start with showing the stability of the numerical method and deriving energy estimates. Prior to that we introduce two auxiliary results.

3.1 Renormalized continuity equation

Under certain regularity assumptions, density and velocity that satisfy continuity equation are known to satisfy its *renormalized* form (see DiPerna, Lions [5] or [10, Proposition 4.2]). Here we introduce its discrete counterpart.

Lemma 3.1. *Let $(\varrho_h, \mathbf{u}_h)$ satisfy the discrete continuity equation (18). Then for any $B \in C^2(\mathbb{R})$, $(\varrho_h, \mathbf{u}_h)$ satisfy the discrete renormalized equation,*

$$h^d \sum_{K\in\mathcal{T}} \Big(\partial^t_h B(\varrho^n_K) + (B'(\varrho^n_K)\varrho^n_K - B(\varrho^n_K))(\mathrm{div}_h \mathbf{u}^n)_K + \mathcal{P}_K\Big) = 0, \tag{22}$$

where

$$\mathcal{P}_K = \Delta t \frac{B''(\varrho^\eta_K)}{2}|\varrho^n_K|^2 + \frac{1}{2}\sum_{s=1}^d \Big((h^\alpha + h u^{s\ +}_{\sigma,s-})B''(\varrho^{n,\star}_{\sigma,s-})|(\partial^s_h\varrho)_{\sigma,s-}|^2 + (h^\alpha - h u^{s\ -}_{\sigma,s+})B''(\varrho^{n,\star}_{\sigma,s+})|(\partial^s_h\varrho)_{\sigma,s+}|^2\Big). \tag{23}$$

The intermediate values $\varrho^\eta_K, \varrho^{n,\star}_{\sigma,s\pm}$ are from the Lagrangian remainders of Taylor expansions.

Proof. We multiply (18) with $B'(\varrho^n_K)$ and handle the uprising terms.

Step 1. Using the Taylor expansion for the discrete time derivative of $B(\varrho^n_K)$ we get

$$\partial^t_h B(\varrho^n_K) = \frac{B(\varrho^n_K) - B(\varrho^{n-1}_K)}{\Delta t} = B'(\varrho^n_K)\partial^t_h\varrho^n_K - \frac{\Delta t}{2}B''(\varrho^\eta_K)|\partial^t_h\varrho^n_K|^2,$$

i.e. the time derivative term yields the first terms in both (22) and (23).

Step 2. We omit the time index n which is constant along the whole rest of the proof.

As $\mathrm{div}_{\mathrm{Up}}[\varrho,\mathbf{u}]_K = \sum_{s=1}^{d}(\partial_h^s \mathrm{Up}[\varrho,\mathbf{u}])_K$, we will prove it for one component only, leaving the summation over s as the very last step of the proof. Using the notation $\sigma, s- = J|K$ and $\sigma, s+ = K|L$, we can write

$$\begin{aligned} B'(\varrho_K)(\partial_h^s \mathrm{Up}[\varrho,\mathbf{u}])_K &= \frac{B'(\varrho_K)}{h}\Big(\varrho_K u^{s\,+}_{\sigma,s+} + \varrho_L u^{s\,-}_{\sigma,s+} - \varrho_J u^{s\,+}_{\sigma,s-} - \varrho_K u^{s\,-}_{\sigma,s-}\Big) \\ &= \frac{B'(\varrho_K)}{h}\Big(\varrho_K(u^s_{\sigma,s+} - u^s_{\sigma,s-}) + u^{s\,-}_{\sigma,s+}(\varrho_L - \varrho_K) + u^{s\,+}_{\sigma,s-}(\varrho_K - \varrho_J)\Big). \end{aligned} \tag{24}$$

Taylor expansion gives

$$\begin{aligned} B(\varrho_L) - B(\varrho_K) &= B'(\varrho_K)(\varrho_L - \varrho_K) + \frac{1}{2}B''(\varrho^\star_{\sigma,s+})(\varrho_L - \varrho_K)^2 \\ B(\varrho_K) - B(\varrho_J) &= B'(\varrho_K)(\varrho_K - \varrho_J) - \frac{1}{2}B''(\varrho^\star_{\sigma,s-})(\varrho_K - \varrho_J)^2, \end{aligned} \tag{25}$$

which, having used the definition of discrete derivative, yields

$$\begin{aligned} \frac{1}{h}B'(\varrho_K)(\varrho_L - \varrho_K) &= (\partial_h^s B(\varrho))_{\sigma,s+} - \frac{1}{2h}B''(\varrho^\star_{\sigma,s+})(\varrho_L - \varrho_K)^2, \\ \frac{1}{h}B'(\varrho_K)(\varrho_K - \varrho_J) &= (\partial_h^s B(\varrho))_{\sigma,s-} + \frac{1}{2h}B''(\varrho^\star_{\sigma,s-})(\varrho_K - \varrho_J)^2. \end{aligned} \tag{26}$$

Substitution from (26) into (24) yields

$$\begin{aligned} B'(\varrho_K)\partial_h^s \mathrm{Up}[\varrho,\mathbf{u}]_K &= B'(\varrho_K)\varrho_K(\partial_h^s u^s)_K + u^{s\,-}_{\sigma,s+}(\partial_h^s B(\varrho))_{\sigma,s+} + u^{s\,+}_{\sigma,s-}(\partial_h^s B(\varrho))_{\sigma,s-} \\ &- \frac{1}{2h}u^{s\,-}_{\sigma,s+}B''(\varrho^\star_{\sigma,s+})(\varrho_L - \varrho_K)^2 + \frac{1}{2h}u^{s\,+}_{\sigma,s-}B''(\varrho^\star_{\sigma,s-})(\varrho_K - \varrho_J)^2. \end{aligned} \tag{27}$$

The last two terms are a contribution to $\mathcal{P}$, while the first three are rewritten as

$$\begin{aligned} &B'(\varrho_K)\varrho_K(\partial_h^s u^s)_K + u^{s\,-}_{\sigma,s+}(\partial_h^s B(\varrho))_{\sigma,s+} + u^{s\,+}_{\sigma,s-}(\partial_h^s B(\varrho))_{\sigma,s-} \\ &\quad= \big(B'(\varrho_K)\varrho_K - B(\varrho_K)\big)(\partial_h^s u^s)_K \\ &\qquad + \frac{B(\varrho_K)}{h}\left(\underbrace{u^s_{\sigma,s+} - u^{s\,-}_{\sigma,s+}}_{u^{s\,+}_{\sigma,s+}} + \underbrace{u^{s\,+}_{\sigma,s-} - u^s_{\sigma,s-}}_{-u^{s\,-}_{\sigma,s-}}\right) + \frac{B(\varrho_L)}{h}u^{s\,-}_{\sigma,s+} - \frac{B(\varrho_J)}{h}u^{s\,+}_{\sigma,s-} \\ &\quad= \big(B'(\varrho_K)\varrho_K - B(\varrho_K)\big)(\partial_h^s u^s)_K + \partial_h^{\mathrm{Up}}[B(\varrho),\mathbf{u}]_K. \end{aligned} \tag{28}$$

Let us substitute (28) to (27), sum over s and over $K \in \mathcal{T}$. Thanks to Lemma 2.4 , we obtain (22).

Step 3. To conclude the proof we show that the artificial diffusion term will contribute to (23) only. By virtue of (25), we get

$$\begin{aligned} -h^\alpha B'(\varrho_K)(\Delta_h \varrho)_K &= -h^{\alpha-2}B'(\varrho_K)\left((\varrho_L - \varrho_K) - (\varrho_K - \varrho_J)\right) \\ &= -h^\alpha(\Delta_h B(\varrho))_K + \frac{1}{2}h^{\alpha-2}B''(\varrho^\star_{\sigma,s+})(\varrho_L - \varrho_K)^2 + \frac{1}{2}h^{\alpha-2}B''(\varrho^\star_{\sigma,s-})(\varrho_K - \varrho_J)^2. \end{aligned} \tag{29}$$

Summing (29) over s and over $K \in \mathcal{T}$, the first term on the right-hand side vanishes due to Neumann boundary condition of the density, while the other two terms contribute to the pollution term (23). □

Note that $\mathcal{P}_\mathcal{K} \geq 0$ provided B is convex.

Remark 3. *One can weaken the assumptions on B in Lemma 3.1 and allow jumps of its second derivatives, paying the price that all $B''(\xi), \xi \in (a,b)$ in (23) are replaced by some $B_2(\xi) \in \mathrm{co}\{B''_-(z), B''_+(z)\}$, which are the one-sided second derivatives of B at ξ. Anyway, $P_K \geq 0$ as long as B is convex. The proof of such assertion remains the same as in Lemma 3.1, with one exception. Instead of the standard Taylor's Theorem one just uses its generalized version, see [26].*

3.2 Positivity of density

We show that the discrete density is positive. Motivated by Karper [27], we present a complete proof of the following lemma. The lemma plays a role of an induction step, where the initial step is $0 < \varrho_K^0 = h^{-d}\int_K \varrho_0 \,\mathrm{d}x$ for all $K \in \mathcal{T}$, since $\varrho_0 > 0$ by assumption.

Lemma 3.2. *Suppose that $\varrho_h^n \in X(\mathcal{T})$ and $\mathbf{u}_h^n \in X(\mathcal{E}_{\mathrm{int}})^d$ satisfy (18), where $\varrho_h^{n-1} > 0$ in Ω_h. Then*

$$\varrho_h^n > 0, \text{ in } \Omega_h.$$

Proof. The proof is stated in two steps, and the first being its nonnegativity. We use the renormalized continuity equation (22) with the one-parametric family of functions

$$B_\eta(z) = \begin{cases} (-z)^\eta & \text{for } z < 0, \\ 0 & \text{for } z \geq 0, \end{cases}$$

for $\eta > 1$. Notice that every B_η satisfies the weakened assumptions of Lemma 3.1 in the sense of Remark 3, i.e. $B_\eta \in C^1(\mathbb{R})$ and B_η'' is a continuous function, with an exception in the form of a jump discontinuity at 0, but since B_η is convex, we have $P_K > 0$. Moreover, $\eta \to 1^+$ yields $B_\eta(z) \to B(z) = \max\{-z, 0\}$ and

$$B_\eta'(z)z - B_\eta(z) = (\eta - 1)(-z)^\eta \to 0, \qquad \text{as } \eta \to 1^+, \quad \text{for } z < 0, \tag{30}$$

while for $z \geq 0$ the convergence is satisfied trivially. Since by assumption $\varrho_K^0 > 0$, it remains to show the induction step. Then (22) together with $P_K > 0$ and $B_\eta(\varrho_K^{n-1}) = 0$ for all $K \in \mathcal{T}$ (since we assume $\varrho_K^{n-1} > 0$) yields

$$\sum_{K\in\mathcal{T}} B_\eta(\varrho_K^n) \leq -\Delta t \sum_{K\in\mathcal{T}} \left(B_\eta'(\varrho_K^n)\varrho_K^n - B_\eta(\varrho_K^n)\right)(\mathrm{div}_h \mathbf{u}^n)_K. \tag{31}$$

Sending $\eta \to 1^+$ in (31), one gets by virtue of (30) that

$$\sum_{K\in\mathcal{T}} \max\{-\varrho_K^n, 0\} \leq 0,$$

from which we conclude $\varrho_K^n \geq 0$ for any $K \in \mathcal{T}$.

Next we show that the density is strictly positive. Choose $K \in \mathcal{T}$ such that $\varrho_K^n \leq \varrho_L^n$ for all $L \in \mathcal{T}$. Then we have

$$\begin{aligned} \varrho_K^n - \varrho_K^{n-1} &= -\Delta t \mathrm{div}_{\mathrm{Up}}[\varrho^n, \mathbf{u}^n]_K + \Delta t h^\alpha (\Delta_h \varrho^n) \\ &\geq -\frac{\Delta t}{h}\sum_{s=1}^d \left(\varrho_K^n u_{\sigma_{s,+}}^s - \varrho_K^n u_{\sigma_{s,-}}^s + (\varrho_{K+h\mathbf{e}_s}^n - \varrho_K^n)u_{\sigma_{s,+}}^{s\,-} + (\varrho_K^n - \varrho_{K-h\mathbf{e}_s}^n)u_{\sigma_{s,-}}^{s\,+}\right) \\ &\geq -\Delta t \varrho_K^n (\mathrm{div}_h \mathbf{u}^n)_K \geq -\Delta t \varrho_K^n |(\mathrm{div}_h \mathbf{u}^n)_K|, \end{aligned} \tag{32}$$

where we have used the minimality of ϱ_K^n to estimate the last term on the first row and last two terms on the second row from below with 0. Then, from (32) we get

$$\varrho_L^n \geq \varrho_K^n \geq \frac{1}{1 + \Delta t |(\mathrm{div}_h \mathbf{u}^n)_K|}\varrho_K^{n-1} > 0, \qquad \text{for any } L \in \mathcal{T},$$

which concludes the proof. □

3.3 Energy estimates

For the upcoming energy estimates we will need to handle the convective term, where we use the following identity.

Lemma 3.3. *For the convective term from (19), the following identity holds,*

$$h^d \sum_{K\in\mathcal{T}} \mathrm{div}_{\mathrm{Up}}[\varrho^n \bar{\mathbf{u}}^n, \mathbf{u}^n]_K \cdot \bar{\mathbf{u}}_K = -h^d \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \mathrm{Up}[\varrho^n, \mathbf{u}^n]_\sigma \left(\partial_h^s \frac{|\bar{\mathbf{u}}^n|^2}{2}\right)_\sigma + \mathcal{N}, \tag{33}$$

where $\mathcal{N}$, the numerical diffusion term reads

$$\mathcal{N} = \frac{h^{d+1}}{4} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |\mathrm{Up}[\varrho^n, \mathbf{u}^n]_\sigma||(\partial_h^s \bar{\mathbf{u}}^n)_\sigma|^2.$$

Proof. We omit the time index n for the sake of brevity. Applying Lemma 2.1, the left hand side $\mathcal{L}$ of (33) equals

$$\mathcal{L} = -h^d \sum_{\sigma\in\mathcal{E}_{\rm int}} \mathrm{Up}[\varrho\bar{\mathbf{u}},\mathbf{u}]_\sigma \cdot (\partial_h^s \bar{\mathbf{u}})_\sigma := h^{d-1} \sum_{\sigma\in\mathcal{E}_{\rm int}} \mathcal{L}_\sigma.$$

Considering $\sigma = K|L$, we can write

$$\begin{aligned}
\mathcal{L}_\sigma &= -\big(\varrho_K \bar{\mathbf{u}}_K u_\sigma^{s\,+} + \varrho_L \bar{\mathbf{u}}_L u_\sigma^{s\,-}\big)\cdot(\bar{\mathbf{u}}_L - \bar{\mathbf{u}}_K)\\
&= \varrho_K u_\sigma^{s\,+}\left(\frac{|\bar{\mathbf{u}}|_K^2}{2} + \frac{|\bar{\mathbf{u}}|_K^2}{2} - \bar{\mathbf{u}}_K\cdot\bar{\mathbf{u}}_L + \frac{|\bar{\mathbf{u}}_L|^2}{2} - \frac{|\bar{\mathbf{u}}_L|^2}{2}\right) + \varrho_L u_\sigma^{s\,-}\left(\frac{|\bar{\mathbf{u}}_K|^2}{2} - \frac{|\bar{\mathbf{u}}_K|^2}{2} + \bar{\mathbf{u}}_K\cdot\bar{\mathbf{u}}_L - \frac{|\bar{\mathbf{u}}_L|^2}{2} - \frac{|\bar{\mathbf{u}}_L|^2}{2}\right)\\
&= (\varrho_K u_\sigma^{s\,+} + \varrho_L u_\sigma^{s\,-})\left(\frac{|\bar{\mathbf{u}}_K|^2}{2} - \frac{|\bar{\mathbf{u}}_L|^2}{2}\right) + (\varrho_K u_\sigma^{s\,+} - \varrho_L u_\sigma^{s\,-})\left|\frac{\bar{\mathbf{u}}_K - \bar{\mathbf{u}}_L}{2}\right|^2\\
&= -h\mathrm{Up}[\varrho,\mathbf{u}]_\sigma\left(\partial_h^s\frac{|\bar{\mathbf{u}}|^2}{2}\right)_\sigma + \frac{h^2}{4}|\mathrm{Up}[\varrho,\mathbf{u}]_\sigma||(\partial_h^s\bar{\mathbf{u}})_\sigma|^2.
\end{aligned}$$

Summation over σ concludes the proof. □

Now we can deduce the following energy estimates on the numerical solution.

Theorem 3.4. *Let $(\varrho_h,\mathbf{u}_h)$ be the numerical solution obtained through the scheme (18–20). For any time step $m = 1,\dots,N_t$ the following stability estimate holds,*

$$h^d\sum_{K\in\mathcal{T}}\left(\varrho_K^m\frac{|\bar{\mathbf{u}}_K^m|^2}{2} + \frac{1}{\gamma-1}p(\varrho_K^m)\right) + \Delta t h^d\mu\sum_{n=1}^m\sum_{K\in\mathcal{T}}\sum_{r=1}^d\sum_{s=1}^d|\widetilde{\partial_h^r u^{s,n}}|_K^2 + \sum_{j=1}^4 N_j^m \le h^d\sum_{K\in\mathcal{T}}\left(\varrho_K^0\frac{|\bar{\mathbf{u}}_K^0|^2}{2} + \frac{1}{\gamma-1}p(\varrho_K^0)\right), \tag{34}$$

where

$$\begin{aligned}
N_1^m &= \Delta t h^d\sum_{n=1}^m\sum_{K\in\mathcal{T}}\sum_{s=1}^d\frac{1}{2}\Big((h^\alpha + h^2(u_{\sigma,s-}^{s,n})^+)p''(\varrho_{\sigma,s-}^{n,\star})|(\partial_h^s\varrho^n)_{\sigma,s-}|^2 + (h^\alpha - h^2(u_{\sigma,s+}^{s,n})^-)p''(\varrho_{\sigma,s+}^{n,\star})|(\partial_h^s\varrho^n)_{\sigma,s+}|^2\Big),\\
N_2^m &= (\Delta t)^2h^d\sum_{n=1}^m\sum_{K\in\mathcal{T}}\frac{p''(\varrho_K^\eta)}{2}|(\partial_h^t\varrho_K)^n|^2,\\
N_3^m &= (\Delta t)^2h^d\sum_{n=1}^m\sum_{K\in\mathcal{T}}\frac{\varrho_K^{n-1}}{2}|(\partial_h^t\bar{\mathbf{u}}_K)^n|^2,\\
N_4^m &= \Delta t h^{d+1}\frac{1}{4}\sum_{n=1}^m\sum_{\sigma\in\mathcal{E}_{\rm int}}|\mathrm{Up}[\varrho^n,\mathbf{u}^n]_\sigma||(\partial_h^s\bar{\mathbf{u}}^n)_\sigma|^2.
\end{aligned}$$

Proof. We take the scalar product of the discrete momentum equation (19) and $h^d(u^s)_\sigma^n\mathbf{e}_s$, sum over $\sigma\in\mathcal{E}_{\rm int}$ and handle term by term.

Time difference term. We use the notation $\sigma = K|L$ and the definition of projection to primary grid (6) to get

$$\frac{h^d}{2}\sum_{\sigma\in\mathcal{E}_{\rm int}}\partial_h^t(\varrho_K\bar{\mathbf{u}}_K + \varrho_L\bar{\mathbf{u}}_L)^n u_\sigma^{s,n}\cdot\mathbf{e}_s = h^d\sum_{K\in\mathcal{T}}\partial_h^t(\varrho_K\bar{\mathbf{u}}_K)^n\cdot\bar{\mathbf{u}}_K^n. \tag{35}$$

Convective term. Using the projection into primary grid (6), Lemma 3.3, summation by parts (14) and the continuity equation (18), we can write

$$\begin{aligned}
h^d\sum_{\sigma\in\mathcal{E}_{\rm int}}&\frac{\mathrm{div}_{\rm Up}[\varrho^n\bar{\mathbf{u}}^n,\mathbf{u}^n]_K + \mathrm{div}_{\rm Up}[\varrho^n\bar{\mathbf{u}}^n,\mathbf{u}^n]_L}{2}\cdot(u^s)_\sigma^n\mathbf{e}_s = h^d\sum_{K\in\mathcal{T}}\mathrm{div}_{\rm Up}[\varrho^n\bar{\mathbf{u}}^n,\mathbf{u}^n]_K\cdot\bar{\mathbf{u}}_K^n\\
&= -h^d\sum_{\sigma\in\mathcal{E}_{\rm int}}\mathrm{Up}[\varrho^n,\mathbf{u}^n]_\sigma\left(\partial_h^s\frac{|\bar{\mathbf{u}}^n|^2}{2}\right)_\sigma + \mathcal{N} = h^d\sum_{K\in\mathcal{T}}(\mathrm{div}_{\rm Up}[\varrho^n,\mathbf{u}^n])_K\frac{|\bar{\mathbf{u}}_K^n|^2}{2} + \mathcal{N}\\
&= -h^d\sum_{K\in\mathcal{T}}(\partial_h^t\varrho_K)^n\frac{|\bar{\mathbf{u}}_K^n|^2}{2} + h^{d+\alpha}\sum_{K\in\mathcal{T}}(\Delta_h\varrho^n)_K\frac{|\bar{\mathbf{u}}_K^n|^2}{2} + \mathcal{N}.
\end{aligned} \tag{36}$$

Pressure term. Using (14), one gets

$$h^d\sum_{\sigma\in\mathcal{E}_{\rm int}}(\partial_h^s p(\varrho^n))_\sigma\mathbf{e}_s\cdot(u^s)_\sigma^n = -h^d\sum_{K\in\mathcal{T}}p(\varrho_K^n)(\mathrm{div}_h\mathbf{u}^n)_K.$$

Then, we apply Lemma 3.1 with $B(z) = \frac{1}{\gamma-1}p(z)$ to deduce

$$-h^d \sum_{K\in\mathcal{T}} p(\varrho_K^n)(\mathrm{div}_h \mathbf{u}^n)_K = \frac{h^d}{\gamma-1}\sum_{K\in\mathcal{T}}(\partial_h^t p(\varrho_K))^n + h^d \sum_{K\in\mathcal{T}}(\mathcal{P}_K)^n. \tag{37}$$

Viscosity term. Direct application of (16) gives

$$-h^d\mu \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} (\Delta_h \mathbf{u}^n)_\sigma \cdot (u^s)_\sigma^n \mathbf{e}_s = \mu h^d \sum_{K\in\mathcal{T}}\sum_{s=1}^{d}\sum_{r=1}^{d} |\widetilde{\partial_h^s u^s}|_K^2. \tag{38}$$

Additional term. Using (6) and summation by parts (14), we can write

$$-h^{d+\alpha}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\sum_{r=1}^{d}\left\{\partial_h^r\left(\{\bar{\mathbf{u}}^n\}\partial_h^r\varrho^n\right)\right\}_\sigma\cdot(u^s)_\sigma^n\mathbf{e}_s = -h^{d+\alpha}\sum_{K\in\mathcal{T}}\sum_{r=1}^{d}\partial_h^r\left(\{\bar{\mathbf{u}}^n\}\partial_h^r\varrho^n\right)_K\cdot\bar{\mathbf{u}}_K^n = h^{d+\alpha}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\{\bar{\mathbf{u}}^n\}_\sigma(\partial_h^s\varrho^n)_\sigma\cdot(\partial_h^s\bar{\mathbf{u}}^n)_\sigma,$$

Then, employing

$$\{\bar{\mathbf{u}}^n\}_\sigma\cdot(\partial_h^s\bar{\mathbf{u}}^n)_\sigma = \frac{1}{2h}(\bar{\mathbf{u}}_L^n+\bar{\mathbf{u}}_K^n)\cdot(\bar{\mathbf{u}}_L^n-\bar{\mathbf{u}}_K^n) = \frac{|\bar{\mathbf{u}}_L^n|^2-|\bar{\mathbf{u}}_K^n|^2}{2h} = \left(\partial_h^s\frac{|\bar{\mathbf{u}}^n|^2}{2}\right)_\sigma,$$

to the chain of equalities above and using (14) with the no-flux boundary condition for density (21), we obtain

$$-h^{d+\alpha}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\sum_{r=1}^{d}\left\{\partial_h^r\left(\{\bar{\mathbf{u}}^n\}\partial_h^r\varrho^n\right)\right\}_\sigma\cdot(u^s)_\sigma^n\mathbf{e}_s = h^{d+\alpha}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\left(\partial_h^s\frac{|\bar{\mathbf{u}}^n|^2}{2}\right)_\sigma(\partial_h^s\varrho^n)_\sigma = -h^{d+\alpha}\sum_{K\in\mathcal{T}}(\Delta_h\varrho^n)_K\frac{|\bar{\mathbf{u}}_K^n|^2}{2}. \tag{39}$$

Final step. We observe the identity

$$\partial_h^t(\varrho_K\bar{\mathbf{u}}_K)^n\bar{\mathbf{u}}_K^n - \partial_h^t\varrho_K^n\left(\frac{|\bar{\mathbf{u}}_K^n|^2}{2}\right) = \partial_h^t\left(\varrho_K\frac{|\bar{\mathbf{u}}_K|^2}{2}\right)^n + \varrho_K^{n-1}\frac{|\bar{\mathbf{u}}_K^n-\bar{\mathbf{u}}_K^{n-1}|^2}{2}. \tag{40}$$

Finally, we collect the right-hand sides of (35–39), employ (40), multiply by Δt and sum over time to obtain the desired result. Notice that the artificial diffusion terms get canceled out. □

3.4 Existence of the numerical solution

As the numerical scheme (18-19) is implicit and nonlinear, the existence of its solution (i.e. of the quantities in the next step) is not a priori known. We prove it in the upcoming section using Schaeffer's fixed point theorem, see e.g. [8, Theorem 9.2.4]. Note that nothing about the uniqueness of the solution is claimed.

Theorem 3.5 (Schaeffer's fixed point theorem)**.** *Let $\mathcal{S} : Z \to Z$ be a continuous mapping defined on a finite-dimensional space Z and let the set*

$$\{z \in Z, z = \kappa\mathcal{S}(z), \kappa\in[0,1]\},$$

be nonempty and bounded. Then there exists $z \in Z$ such that

$$z = \mathcal{S}(z).$$

Before stating the existence theorem, we prove an auxiliary lemma concerning the viscosity term.

Lemma 3.6. *Let $\mathcal{L} : X(\mathcal{E}_{\mathrm{int}})^d \to X(\mathcal{E}_{\mathrm{int}})^d$ be a linear mapping given by*

$$(\mathcal{L}(\mathbf{v}))_\sigma := (\Delta_h\mathbf{v})_\sigma. \tag{41}$$

Then its inverse operator $\mathcal{L}^{-1}$ is bounded with constant depending on the discretization parameter h.

Proof. Note that for fixed h $X(\mathcal{E}_{\mathrm{int}})^d$ is a finite-dimensional space and thus all norms are equivalent. Therefore, we aim at proving

$$\|\mathcal{L}(\mathbf{v})\|_\infty \geq c(h) > 0, \qquad \text{for all } \mathbf{v} = (v^1, v^2, v^3) \in X(\mathcal{E}_{\mathrm{int}})^d, \|\mathbf{v}\|_\infty = 1. \tag{42}$$

From $\|\mathbf{v}\|_\infty = 1$ we have that $|v_\sigma^s| = 1$ for some $\sigma \in \mathcal{E}_{\mathrm{int}}$. Without loss of generality we may assume that $v_\sigma^s = -1$. And as $v_{\sigma'}^s = 0$ when $\sigma' \in \mathcal{E}_{\mathrm{ext}}$, there exist $K_1, K_2 \in \mathcal{T}$ such that

$$(\partial^s v^s)_{K_2} \geq \frac{h}{N_x} = \frac{h^2}{L_x}, \qquad \text{and} \quad (\partial^s v^s)_{K_1} \leq -\frac{h^2}{L_x},$$

where K_1, K_2 differ only in the s-component and $hN_x = L_x$. Therefore, using the same argument as before, there exists $\widetilde{\sigma} \in \mathcal{E}_{\rm int}$ such that

$$\|(\Delta_h \mathbf{v})\|_\infty \geq |(\partial_h^s \partial_h^s \mathbf{v})_{\widetilde{\sigma}}| \geq \frac{(\partial^s v^s)_{K_2} - (\partial^s v^s)_{K_1}}{L_x} \geq \frac{2h^2}{L_x^2},$$

which is (42) with $c(h) = \frac{2h^2}{L_x^2}$. □

Theorem 3.7. *Let $p(\varrho) = a\varrho^\gamma$ and $\varrho_h^{n-1} \in X(\mathcal{T})$, $\mathbf{u}_h^{n-1} \in X(\mathcal{E}_{\rm int})^d$ be given; $\varrho_K^{n-1} > 0$ for all $K \in \mathcal{T}$. Then the numerical scheme (18-19) admits a solution*

$$\varrho_h^n \in X(\mathcal{T}), \varrho_K^n > 0 \text{ for all } K \in \mathcal{T}, \mathbf{u}_h^n \in X(\mathcal{E}_{\rm int})^d.$$

Moreover, is satisfies the discrete conservation of mass

$$\sum_{K\in\mathcal{T}} \varrho_K^n = \sum_{K\in\mathcal{T}} \varrho_K^{n-1}. \tag{43}$$

Proof. We show the existence in two steps. We treat the continuity equation first.

Step 1. We claim, that for ϱ_h^{n-1} given, the continuity scheme (18) provides a unique solution depending continuously on the parameter $\mathbf{u}_h^n \in X(\mathcal{E}_{\rm int})^d$.

In fact, for all $K \in \mathcal{T}$ (18) builds a system of N_e linear equations with N_e unknowns, where N_e denotes the number of points in the primary mesh, where ϱ_K^{n-1} represents the (known) right-hand side and $\mathbf{u}_h^n$ is a parameter.

The associated homogeneous problem

$$\varrho_K^n + \Delta t \mathrm{div}_{\rm Up}[\varrho^n, \mathbf{u}^n]_K - \Delta t\, h^\alpha (\Delta_h \varrho^n)_K = 0, \tag{44}$$

admits a unique solution and hence the trivial one. It is easy to verify that $\varrho_h^n \equiv 0$ indeed solves (44). To show uniqueness one uses the same procedure as in the proof of Lemma 3.2 to get

$$\sum_{K\in\mathcal{T}} \max\{-\varrho_K^n, 0\} \leq 0,$$

and hence $\varrho_K^n = 0$ for all $K \in \mathcal{T}$.

Therefore, for given ϱ_h^{n-1}, the continuity scheme (18) supplies us with a unique solution $\varrho_h^n = \varrho_h^n(\mathbf{u}_h^n)$, where the mapping

$$\mathbf{u}_h^n \mapsto \varrho_h^n(\mathbf{u}_h^n),$$

is continuous in $X(\mathcal{E}_{\rm int})^d$. Moreover, Lemma 3.2 gives $\varrho_h^n > 0$. The discrete conservation of mass (43) can be obtained by simple summation of the discrete continuity equation (18) over all $K \in \mathcal{T}$.

Step 2. We rewrite the momentum scheme (19) as follows,

$$\mu(\Delta_h \mathbf{u}^n)_\sigma \mathbf{e}_r = \kappa \mathcal{F}_\sigma(\mathbf{u}_h^n), \qquad \sigma \in \mathcal{E}_{\rm int}, \tag{45}$$

where

$$\mathcal{F}_\sigma(\mathbf{u}_h^n) := -\frac{\{\varrho^n[\mathbf{u}_h^n]\bar{\mathbf{u}}^n\}_\sigma - \{\varrho^{n-1}\bar{\mathbf{u}}^{n-1}\}_\sigma}{\Delta t} - \{\mathrm{div}_{\rm Up}[\varrho^n[\mathbf{u}_h^n]\bar{\mathbf{u}}^n, \mathbf{u}^n]\}_\sigma - (\partial_h^s p(\varrho^n[\mathbf{u}_h^n]))_\sigma + h^\alpha \sum_{r=1}^d \{\partial_h^r(\{\bar{\mathbf{u}}^n\}\partial_h^r \varrho^n[\mathbf{u}_h^n])\}_\sigma.$$

Note that $\bar{\mathbf{u}}^{n-1}$ was determined in the previous step and $\bar{\mathbf{u}}^0$ is given by the initial conditions (20). We define $\mathcal{F} := (\mathcal{F}_\sigma)_{\{\sigma\in\mathcal{E}_{\rm int}\}}$ together with $\mathbf{u}_{\sigma'}^n = 0$ for $\sigma' \in \mathcal{E}_{\rm ext}$.

We are searching for $\mathbf{u}_h^n$ being a fixed point of the mapping $\mathcal{F} \circ \mathcal{L}^{-1}$, with $\mathcal{L}$ defined by (41). We verify the assumptions of the Schaeffer's fixed point theorem (Theorem 3.5). As $\mathcal{F}$ is clearly continuous and $\mathcal{L}^{-1}$ is linear and bounded, their composition is continuous in the finite dimensional space $X(\mathcal{E}_{\rm int})^d$. Any possible solution $\mathbf{u}_{h,\kappa}^n$ of (45) is indeed a solution of the momentum scheme with the diffusion constant enlarged to $\frac{\mu}{\kappa}$, i.e. the energy estimate (34), with μ replaced by μ/κ, implies that $\mathcal{F}(\mathbf{u}_{h,\kappa}^n)$ is bounded in $X(\mathcal{E}_{\rm int})^d$, independently of κ. The boundedness of $\mathcal{L}^{-1}$ further implies that also

$$\left\{\mathbf{u}_{h,\kappa}^n \in X(\mathcal{E}_{\rm int})^d, \quad \mathbf{u}_{h,\kappa}^n = \kappa \mathcal{F} \circ \mathcal{L}^{-1}(\mathbf{u}_{h,\kappa}^n)\right\}, \tag{46}$$

is bounded independently of κ. Note that the set in (46) is nonempty, as zero obviously solves (45) with $\kappa = 0$.

□

3.5 Uniform bounds

The convergence proof requires some compactness results which are usually gained through uniform bounds of approximate quantities. In sequel, the notation $A \lesssim B$ means $A \leq cB$, where $c > 0$ is a constant that does not depend on the discretization parameter h. $A \approx B$ means $A \lesssim B$ and $B \lesssim A$.

The energy estimate (34) allows us to establish the following uniform bounds.

Proposition 3.8. *Let $(\varrho_h, \mathbf{u}_h)$ be a numerical solution obtained through the scheme (18)–(20) and let the total initial energy D be defined by*

$$D = \int_\Omega \frac{1}{2}\varrho_0 \mathbf{u}_0^2 + \frac{1}{\gamma - 1} p(\varrho_0)\,\mathrm{d}x. \tag{47}$$

Then

$$\|\varrho_h\|_{L^\infty(L^\gamma(\Omega))} \lesssim D, \qquad \|p(\varrho_h)\|_{L^\infty(L^1(\Omega))} \lesssim D, \qquad \|\sqrt{\varrho_h}\bar{\mathbf{u}}_h\|_{L^\infty(L^2(\Omega))} \lesssim D. \tag{48}$$

Note that the constant D depends solely on the initial data $(\varrho_0, \mathbf{u}_0)$.

Proof. Due to the convexity of $p(z)$, all the terms in the left hand side of (34) are non-negative, hence every single one can be estimated by the right-hand side of (34).

Further, it is the definition of initial conditions to the numerical scheme (20), property (9) and the inequality $\|\varrho_0\|_{L^1(\Omega)} \leq \|\varrho_0\|_{L^\gamma(\Omega)}$ that guarantee

$$\frac{h^d}{\gamma - 1} \sum_{K \in \mathcal{T}} p(\varrho_K^0) \leq \frac{1}{\gamma - 1} \int_\Omega p(\varrho_0)\,\mathrm{d}x.$$

Then we apply the Jensen's inequality on each cell twice to get also

$$h^d \sum_K \varrho_K^0 |\bar{\mathbf{u}}_K^0|^2 \leq \sum_{K \in \mathcal{T}} \varrho_K \left(\int_{Q_K} |\mathbf{u}_0|^2\,\mathrm{d}x \right) \leq \sum_{K \in \mathcal{T}} h^{-d} \int_{Q_K} \int_{Q_K} \varrho_0 |\mathbf{u}_0|^2\,\mathrm{d}x\,\mathrm{d}y = \sum_{K \in \mathcal{T}} \int_{Q_K} \varrho_0 |\mathbf{u}_0|^2\,\mathrm{d}x = \int_\Omega \varrho_0 |\mathbf{u}_0|^2\,\mathrm{d}x.$$

□

The reader can observe a slight abuse of notation, concerning the Bochner spaces. Since we have not defined the extension of our discrete quantities to integrable functions in time, keep in mind that the equiintegrability of some v_h in a Bochner space $L^q(0, T; X)$ should be understood as

$$\left(\Delta t \sum_{n=1}^{N_t} (\|v^n\|_X)^q \right)^{\frac{1}{q}} \leq c,$$

which corresponds to the standard Bochner norm for the piecewise constant extension in time.

Using Hölder inequality one deduces from (48) also

$$\|\varrho_h \bar{\mathbf{u}}_h\|_{L^\infty\left(L^{\frac{2\gamma}{\gamma+1}}(\Omega;\mathbb{R}^d)\right)} \lesssim D. \tag{49}$$

3.6 Discrete inequality of the Sobolev type and velocity estimates

Similarly to the continuous case, we would like to obtain information about better (equi)integrability of $\mathbf{u}_h$. For this purpose we introduce a version of the discrete Sobolev embedding theorem. Both the claim and its proof are inspired by an analogous assertion from [2, Lemma 1]. Prior to that, we introduce the following auxiliary algebraic inequality.

Lemma 3.9. *For any $a, b \in \mathbb{R}$ and any $p > 2$ the following inequality holds,*

$$\left| |a|^{p-1}a - |b|^{p-1}b \right| \leq \frac{p}{2} \left(|a|^{p-1} + |b|^{p-1} \right) |a - b|. \tag{50}$$

Proof. Without loss of generality we can assume that $a \geq b$, then it holds, that

$$|a|^{p-1}a - |b|^{p-1}b \geq 0. \tag{51}$$

It can be shown through discussing the signs of a and b:

1. Let $a \geq 0, b \geq 0$. Then the left-hand side of (51) equals $a^p - b^p \geq b^{p-1}(a - b) \geq 0$.
2. Let $a \geq 0, b < 0$. Then $a^p - |b|^{p-1}b \geq 0$.

3. Let $a<0, b<0$. Then $(-b)|b|^{p-1}-(-a)|a|^{p-1}\geq |b|^{p-1}(a-b)\geq 0$.

Therefore, it remains to show that $|a|^{p-1}a-|b|^{p-1}b\leq \frac{p}{2}(|a|^{p-1}+|b|^{p-1})(a-b)$. We will use Taylor expansion of the function $f(x):=|x|^{p-1}x$, notice that $f'(x)=p|x|^{p-1}$ and $f''(x)=p(p-1)|x|^{p-3}x$ is increasing.

Then

$$\begin{aligned} |x|^{p-1}x &= |a|^{p-1}a+p|a|^{p-1}(x-a)+\frac{1}{2}f''(\zeta_a)(x-a)^2,\\ |x|^{p-1}x &= |b|^{p-1}b+p|b|^{p-1}(x-b)+\frac{1}{2}f''(\zeta_b)(x-b)^2. \end{aligned} \tag{52}$$

We take $x=\frac{1}{2}(a+b)$ and subtract the equations in (52) to obtain

$$|a|^{p-1}a-|b|^{p-1}b=p(|a|^{p-1}+|b|^{p-1})\frac{a-b}{2}+(f''(\zeta_b)-f''(\zeta_a))\frac{(a-b)^2}{4}. \tag{53}$$

As $\zeta_a\geq\zeta_b$ and f'' is increasing, the last term on the right-hand side of (53) is negative which, together with (51), recovers (50). □

Now we can prove the Sobolev-type inequality for discrete quantities.

Proposition 3.10. *Let $\mathbf{w}=(w^1,\cdots,w^d)\in X(\mathcal{E}_{\rm int})^d$, then the following inequality holds*

$$h^d\left(\sum_{\sigma\in\mathcal{E}_{\rm int}}|w^s_\sigma|^q\right)^{\frac{2}{q}}\lesssim h^d\sum_{r=1}^d\sum_{s=1}^d\sum_{K\in\mathcal{T}}|\widetilde{\partial^r_h w^s}|^2=:\|\widetilde{\nabla_h\mathbf{w}_h}\|_2^2 \qquad \text{for } \begin{cases} d=2, & q\in[1,\infty),\\ d=3, & q\in[1,6]. \end{cases}$$

Proof. We start with $d=3$, $\mathbf{v}\in X(\mathcal{E}_{\rm int})^d$, $\mathbf{v}|_{\partial\Omega}=0$, whose relation to $\mathbf{w}$ will be specified later. Any component v^s of $\mathbf{v}$ can be expressed, by virtue of the definition (13), as

$$v^s_\sigma(x)=h\sum_{K\in\mathcal{T}}(\partial^r v^s)_K\chi_K^{r,s}(x), \tag{54}$$

for any $r=1,\dots,d$, where the characteristic function $\chi_K^{r,s}$ equals one at $x\in\sigma$, for which K participates on creating the value v_σ and zero otherwise. In particular, if $r=s$, we define

$$\chi_K^{s,s}(x)=\begin{cases}1 & \text{if } x\in Q_\sigma:(\sigma-K)\cdot\mathbf{e}_s\geq 0\wedge(\sigma-K)\cdot\mathbf{e}_p=0,\forall p\in\{1,\dots,d\}\setminus\{s\},\\ 0 & \text{otherwise,}\end{cases} \tag{55}$$

and for $r\neq s$

$$\chi_K^{r,s}(x)=\begin{cases}1 & \text{if } x\in Q_\sigma:(\sigma-K)\cdot\mathbf{e}_s=\frac{1}{2}\wedge(\sigma-K)\cdot\mathbf{e}_r\geq 0\wedge(\sigma-K)\cdot\mathbf{e}_p=0,p\in\{1,\dots,d\}\setminus\{r,s\},\\ 0 & \text{otherwise.}\end{cases} \tag{56}$$

We comment on the definitions (55–56), that since every x belongs to three distinct cubes Q_σ, we pick always the one, whose face F_σ has the normal vector $\mathbf{e}_s$, where s is indicated by the second item at the upper index of $\chi_K^{\cdot,s}$ and was fixed at the beginning of the proof.

Integrating (54) over Ω and estimating the characteristic functions $\chi_K^{s,s}$ from above yields

$$\int_\Omega|v^s_\sigma|\,\mathrm{d}x\leq h\sum_{K\in\mathcal{T}}|\partial^s v^s|_K h^{d-1}\leq h^d\sum_{K\in\mathcal{T}}|\partial^s v^s|_K. \tag{57}$$

Further, denoting $\dot K=K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_{r_1}$ and $\ddot K=K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_{r_2}$, we can express

$$|v^s_\sigma(x)|^2=\sum_{K\in\mathcal{T}}(\partial^{r_1}v_s)_{\dot K}h\chi_K^{r_1,s}(x)\sum_{K\in\mathcal{T}}(\partial^{r_2}v_s)_{\ddot K}h\chi_K^{r_2,s}(x)\leq\left(\sum_{K\in\mathcal{T}}|(\partial^{r_1}v_s)_{\dot K}|h\overline{\chi}_K^{r_1,s}(x)\right)\left(\sum_{K\in\mathcal{T}}|(\partial^{r_2}v_s)_{\ddot K}|h\overline{\chi}_K^{r_2,s}(x)\right)$$

with three mutually distinct indices r_1,r_2,s, where

$$\overline{\chi}_K^{r_i,s}(x)=\begin{cases}1 & \text{if } x\in Q_\sigma:(\sigma-K)\cdot\mathbf{e}_s=\frac{h}{2}\wedge(\sigma-K)\cdot\mathbf{e}_p=0,p\in\{1,\dots,d\}\setminus\{r_i,s\},\\ 0 & \text{otherwise,}\end{cases} \tag{58}$$

which is a dominating function to $\chi_K^{r_i,s}$, independent of r_i. In particular, $\overline{\chi}_K^{r_1,s}(x),\overline{\chi}_K^{r_2,s}(x)$ depend only on (x_{r_2},x_s), (x_{r_1},x_s), respectively. Thus we can compute

$$\begin{aligned}
&\int_{\mathbb{R}}\int_{\mathbb{R}}(v_\sigma^s)^2\,\mathrm{d}x_{r_1}\,\mathrm{d}x_{r_2}\\
&\quad\le \sum_{K\in\mathcal{T}} h|(\partial_h^{r_1}v^s)_{\dot K}|\int_{\mathbb{R}}\overline{\chi}_K^{r_1,s}(x_{r_2},x_s)\,\mathrm{d}x_{r_2}\sum_{K\in\mathcal{T}} h|(\partial_h^{r_2}v^s)_{\ddot K}|\int_{\mathbb{R}}\overline{\chi}_K^{r_2,s}(x_{r_1},x_s)\,\mathrm{d}x_{r_1}\\
&\quad= h^4\sum_{K\in\mathcal{T}}|(\partial_h^{r_1}v^s)_{\dot K}|\sum_{K\in\mathcal{T}}|(\partial_h^{r_2}v^s)_{\ddot K}|(\mathbf{1}_{K-\frac{h}{2}\mathbf{e}_s}(x_s))^2,
\end{aligned}\tag{59}$$

where we used the fact that after integrating with respect to x_{r_1}, x_{r_2}, the functions $\overline{\chi}_K^{r_1,s}(x), \overline{\chi}_K^{r_2,s}(x)$ leave their projections to the line x_s, which are in both cases equal to $\mathbf{1}_{K-\frac{h}{2}\mathbf{e}_s}(x_s)$. Integrating (59) over the remaining variable x_s, we get

$$\int_{\mathbb{R}^d}|v_\sigma^s|^2\,\mathrm{d}x=\int_\Omega|v_\sigma^s|^2\le h^6\sum_{K\in\mathcal{T}}|(\partial_h^{r_1}v^s)_{\dot K}||(\partial_h^{r_1}v^s)_{\ddot K}|.\tag{60}$$

Having all the ingredients, we enter the main part of the proof. We start with the standard interpolation inequality and substitute from (57) and (60) to obtain

$$\|v_h^s\|_{\frac{3}{2}}^{\frac{3}{2}}\le\|v_h^s\|_1^{\frac{1}{2}}\|v_h^s\|_2\le h^{\frac{9}{2}}\left(\sum_{K\in\mathcal{T}}|\partial_h^s v^s|_K\sum_{K\in\mathcal{T}}|\partial_h^{r_1}v^s|_{\dot K}\sum_{K\in\mathcal{T}}|\partial_h^{r_2}v^s|_{\ddot K}\right)^{\frac{1}{2}}.\tag{61}$$

Using the AG-inequality $ABC\le\frac{1}{3^3}(A+B+C)^3$, (61) becomes

$$\|v_h^s\|_{\frac{3}{2}}^{\frac{3}{2}}\le 3^{-\frac{3}{2}}h^{\frac{9}{2}}\left(\sum_{K\in\mathcal{T}}|\partial_h^s v^s|_K+\sum_{K\in\mathcal{T}}|\partial_h^{r_1}v^s|_{\dot K}+\sum_{K\in\mathcal{T}}|\partial_h^{r_2}v^s|_{\ddot K}\right)^{\frac{3}{2}}\lesssim h^{\frac{9}{2}}\left(\sum_{K\in\mathcal{T}}\sum_{r=1}^{d}|\partial_h^r v^s|_K\right)^{\frac{3}{2}}.\tag{62}$$

Now we set $v_h^s=|w_h^s|^3w_h^s$ and apply Lemma 3.9 to (62), to get

$$\|w_h^s\|_6^6\le\left(\frac{2}{3}h^d\sum_{r=1}^{d}\sum_{K\in\mathcal{T}}|\widetilde{\partial_h^r w^s}|_K\{|w^s|^3\}_K^{\star r}\right)^{\frac{3}{2}}.\tag{63}$$

where $\{v^s\}_K^{\star r}$ is rather unusual interpolation. In particular,

$$\{v^s\}_K^{\star r}:=\begin{cases}\frac{1}{2}\left(v_{K+\frac{h}{2}\mathbf{e}_s}+v_{K+\frac{h}{2}\mathbf{e}_s-\mathbf{e}_r}\right)+\frac{1}{2}\left(v_{K+\frac{h}{2}\mathbf{e}_s}+v_{K+\frac{h}{2}\mathbf{e}_s+\mathbf{e}_r}\right) & \text{for } r\neq s,\\ v_{K+\frac{h}{2}\mathbf{e}_s}+v_{K-\frac{h}{2}\mathbf{e}_s} & \text{for } r=s.\end{cases}$$

However, all we care about is its estimate $\sum_{K\in\mathcal{T}}\{|v^s|\}_K^\star\le 2\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|v^s|$. With that and Cauchy-Schwarz inequality, (63) remains

$$\|w_h^s\|_6^6\lesssim\left\|\sum_{r=1}^{d}|\widetilde{\partial_h^r w^s}|\right\|_2^{\frac{3}{2}}\left\||w_h^s|^3\right\|_2^{\frac{3}{2}},$$

i.e., after summation over all components

$$\|w_h^s\|_6^{6-\frac{9}{2}}\lesssim\|\widetilde{\nabla_h w_h^s}\|_2^{\frac{3}{2}}.$$

The proof for $d=2$ follows the same step and is a bit simpler. We have

$$|v_\sigma^s(x)|^2\le\left(\sum_{K\in\mathcal{T}}(\partial_h^r v^s)_{\dot K}h\chi_K^{r,s}(x)\right)\left(\sum_{K\in\mathcal{T}}(\partial_h^s v^s)_K h\chi_K^{s,s}(x)\right),\tag{64}$$

where $r=s\in\{1,2\}, r\neq s$ and $\dot K=K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_r$. We recall the definition of $\overline{\chi_K^{r,s}}$ (58) and introduce $\overline{\chi}_K^{s,s}$, a dominating function to $\chi_K^{s,s}$, with

$$\overline{\chi}_K^{s,s}(x)=\begin{cases}1 & \text{if } x\in Q_\sigma:(\sigma-K)\cdot\mathbf{e}_p=0, p\in\{1,\dots,d\}\setminus\{r_i,s\},\\ 0 & \text{otherwise.}\end{cases}$$

Similarly as before, $\overline{\chi}_K^{r,s}(x)=\overline{\chi}_K^{r,s}(x_s)$ and $\overline{\chi}_K^{s,s}(x)=\overline{\chi}_K^{s,s}(x_r)$. Therefore, the integration of (64) yields

$$\int_\Omega|v^s(x)|^2\,\mathrm{d}x\le h^4\sum_{K\in\mathcal{T}}(\partial_h^r v^s)_{\dot K}\sum_{K\in\mathcal{T}}(\partial_h^s v^s)_K.\tag{65}$$

Then we set $\mathbf{v} = |\mathbf{w}|^{\lambda-1}\mathbf{w}$, with $\mathbf{w} \in X(\mathcal{E}_{\text{int}})^2$ and $\lambda > 2$. Substituting into (65) and applying Lemma 3.9 one gets

$$\|w^s\|_{2\lambda}^{\lambda} \lesssim \left(h^2 \sum_{K\in\mathcal{T}} \{|w^s|^{\lambda-1}\}\|_{\dot{K}}^{\star r}(\partial_h^r w^s)_{\dot{K}}\right)^{\frac{1}{2}} \left(h^2 \sum_{K\in\mathcal{T}} \{|w^s|^{\lambda-1}\}\|_K^{\star s}(\partial_h^s w^s)_K\right)^{\frac{1}{2}} \lesssim \||w^s|^{\lambda-1}\|_p \|\partial_h^r w^s\|_{p'}^{\frac{1}{2}} \|\partial_h^s w^s\|_{p'}^{\frac{1}{2}}, \quad (66)$$

where we applied Hölder's inequality in the last step. Now we fix p with $2\lambda = p(\lambda-1)$ (and therefore $p'(\lambda+1) = 2\lambda$) and divide both sides of (66) with the norm of w^s and apply the Young inequality to get

$$\|w^s\|_{2\lambda} \lesssim \sum_{r=1}^{2} \|\partial_h^r w^s\|_{p'}. \quad (67)$$

The final step is the chain of inequalities build on (67) and standard Lebesgue embeddings

$$\|w^s\|_q \lesssim \|w^s\|_{2\lambda} \lesssim \|\widetilde{\nabla_h w^s}\|_{\frac{2\lambda}{\lambda+1}} \lesssim \|\widetilde{\nabla_h w^s}\|_2,$$

as $p' = \frac{2\lambda}{\lambda+1} < 2$ for any admissible λ.

□

Remark 4. *To prove discrete Sobolev inequality we use the cross derivatives of the velocity, which are, in the finite difference scheme, employed in a rather awkward way. It is interesting that in the three-dimensional case, thanks to the interpolation (61), we do not need to use all 3×3 derivatives, but only 3×2, as we could alternatively use the same derivative twice in the inequality in (61).*

Due to the positivity of the density we can deduce from (34) that

$$\|\widetilde{\nabla_h \mathbf{u}_h}\|_{L^2(L^2(\Omega))} \lesssim D. \quad (68)$$

and using Proposition 3.10 we get also that

$$\|\mathbf{u}_h\|_{L^2(L^q(\Omega))} \lesssim D, \qquad \|\bar{\mathbf{u}}_h\|_{L^2(L^q(\Omega))} \lesssim D, \quad (69)$$

with $q \in [1, 6]$ for $d = 3$ and $q \in [1, \infty)$ for $d = 2$.

4 Consistency of the numerical method

One step towards the convergence to a weak solution is the consistency of numerical solutions, i.e. verifying that the numerical solution satisfies the weak formulation of the problem up to a residual term $\mathcal{R}(\varrho_h, \mathbf{u}_h)$ which satisfies

$$\mathcal{R}(\varrho_h, \mathbf{u}_h) \to 0, \quad \text{as } h \to 0.$$

In this section we formulate the results both for $d = 2, d = 3$. The difference in these cases occurs only in the inverse estimates and the discrete Sobolev inequality (Proposition 3.10) and its consequences, mainly the velocity integrability (69).

We want to emphasize that our result on consistency is not the only possibility. Our goal was to enable as large set of admissible values for γ as possible. Stronger assumptions on the integrability properties of test functions is the price to pay.

4.1 Preliminary material for proving consistency

First, we show some useful estimates on projections and artificial diffusion terms in order to shorten the proofs of consistency. First let us recall the estimates (10).

Lemma 4.1. *Let $\phi \in W^{1,p}(\Omega)$. Then*

$$\|\partial_h \Pi^P \phi\|_{L^p(\Omega)} \lesssim \|\nabla\phi\|_{L^p(\Omega)}, \quad \|\partial_h \Pi^P \Pi^D \mathbf{v}\|_{L^p(\Omega)} \lesssim \|\nabla \mathbf{v}\|_{L^p(\Omega)}, \quad (70)$$

$$\|\Pi^P \Pi^D \mathbf{v} - \mathbf{v}\|_{L^p(\Omega)} \lesssim h\|\nabla \mathbf{v}\|_{L^p(\Omega)}, \quad (71)$$

$$\|\Pi^P \nabla_h \Pi^P \Pi^D \mathbf{v} - \nabla \mathbf{v}\|_{L^p(\Omega)} \lesssim h\|\nabla_x^2 \mathbf{v}\|_{L^p(\Omega)}. \quad (72)$$

Proof. Estimates (70) are the direct consequences of the mean value theorem, with its double application in the latter case,

$$|(\partial^s \Pi^P \phi)_\sigma| = h^{-1}|\phi(\xi_L) - \phi(\xi_K)| \lesssim |\nabla_h \phi|, \quad \text{with some } \xi_K \in Q_K, \xi_L \in Q_L,$$
$$|(\partial^s \Pi^P \Pi^D \mathbf{v})_\sigma| = h^{-1}|\mathbf{v}(\widetilde{\xi_L}) - \mathbf{v}(\widetilde{\xi_K})| \lesssim |\nabla_h \mathbf{v}|, \quad \text{with some } \widetilde{\xi_K} \in Q_K, \widetilde{\xi_L} \in Q_L.$$

Similarly, to get (71) we can write

$$|(\Pi^P \Pi^D \mathbf{v} - \mathbf{v}| \leq |\Pi^P \Pi^D \mathbf{v} - \Pi^D \mathbf{v}| + |\Pi^D \mathbf{v} - \mathbf{v}| \lesssim h|\nabla_x \widehat{\Pi^D \mathbf{v}}| + h|\nabla_x \mathbf{v}| \lesssim h|\nabla_x \mathbf{v}|.$$

To prove (72) we show using Taylor expansion that

$$|(\Pi^P \partial_h^s \Pi^P \Pi^D v^r)(x) - \partial^s v^r(x)| \lesssim h|\nabla_x^2 v^r|, \tag{73}$$

where $x \in K$. Let us denote $L = K + h\mathbf{e}_s$, $J = K - h\mathbf{e}_s$, then

$$(\Pi^P \partial_h^s \Pi^P \Pi^D v^r)(x) = \frac{1}{2h}\left((\Pi^P \Pi^D v^r)_L - (\Pi^P \Pi^D v^r)_J\right). \tag{74}$$

Expressing the Taylor expansion of v^r at each cell K gives

$$v^r(x) = v^r(x_K) + \nabla_x v^r(x_K)(x - x_K) + \frac{1}{2}(x - x_K)^T \nabla_x^2 v^r(\xi(x))(x - x_K),$$

where x_K is its center. Further, as the affine function with zero mean belong to the kernel of the combined projection $Pi^{PD} := \Pi^P \Pi^D$, we have

$$(\Pi^P \Pi^D v^r)_K = v^r(x_K) + \frac{1}{4h^2}\left(\int_{F_{\sigma,r+}} \nabla_x^2 v^r(\xi(x))(x - x_K)^2 \,\mathrm{d}S_x + \int_{F_{\sigma,r-}} \nabla_x^2 v^r(\xi(x))(x - x_K)^2 \,\mathrm{d}S_x\right). \tag{75}$$

Combining (74) and (75), we can write

$$\left|\frac{1}{2h}\left((\Pi^P \Pi^D v^r)_L - (\Pi^P \Pi^D v^r)_J\right) - \partial^s v^r(x)\right| \lesssim \left|\frac{1}{2h}\left(v^r(x_L) - v^r(x_J)\right) - \partial^s v^r(x)\right| + h|\nabla^2 v^r|. \tag{76}$$

Further we use the Mean Value Theorem to express

$$\partial^s v^r(x) = \partial^s v^r(x_K) + \nabla_x \partial^s v^r(\xi_K)(x - x_K), \tag{77}$$

for $x \in K$. The combination of (74, 76, 77) finally yields (73), which proves (72).

□

We introduce the following lemma that will simplify the treatment of the artificial viscosity term.

Lemma 4.2. *Let ϱ_h be obtained through the scheme (18–19) with $\gamma > 1$. Then it holds that*

$$h^\alpha \|\partial_h^s \varrho_h\|_{L^2(0,T,L^2(\Omega))} \lesssim h^\beta c(D),$$

with $\beta = \frac{\alpha}{2} + \min\{0, d(\frac{1}{4} - \frac{1}{\gamma})\}$ and D is defined by (47).

Proof. First let $\gamma \geq 2$. We use the renormalized equation (22) with $B(z) = z^2$. Thanks to the fact that $\mathcal{P}_K \geq 0$, we obtain

$$h^\alpha \int_0^T \int_\Omega (\partial_h \varrho_h)^2 \leq \int_\Omega \varrho_0^2 \,\mathrm{d}x - \int_\Omega \varrho^2(T) \,\mathrm{d}x + \int_0^T \int_\Omega |\varrho_h|^2 |\mathrm{div}_h \mathbf{u}_h| \,\mathrm{d}x \lesssim D^2 + \|\varrho_h\|^2_{L^\infty(0,T;L^4(\Omega))} \|\mathrm{div}_h \mathbf{u}_h\|_{L^2(0,T;L^2(\Omega))}, \tag{78}$$

where we used the Hölder inequality and energy estimate (48). Applying the inverse estimate to the latter term in (78), one gets

$$\|\varrho_h\|^2_{L^4(\Omega)} \lesssim h^{\min\left\{0, 2d\left(\frac{1}{4} - \frac{1}{\gamma}\right)\right\}} \|\varrho_h\|^2_{L^\gamma} \lesssim D^2 h^{\min\{0, \frac{d(\gamma-4)}{2\gamma}\}}. \tag{79}$$

Combining (78)–(79)) together with the energy estimates (48) and (68), one gets

$$h^\alpha \|\partial_h^s \varrho_h\|_{L^2(L^2)} = h^{\frac{\alpha}{2}} \|h^{\frac{\alpha}{2}} \partial_h^s \varrho_h\|_{L^2(0,T,L^2(\Omega))} \leq h^{\alpha/2} D^{1/2} + h^{\frac{\alpha}{2} + \min\{0, d(\frac{1}{4} - \frac{1}{\gamma})\}} D^{\frac{3}{2}}.$$

For $\gamma \in (1, 2)$, one just uses one more inverse estimate to get $\|\varrho_h\|_\gamma \lesssim h^{d(\frac{1}{2} - \frac{1}{\gamma})} \|\varrho_h\|_\gamma$, but this term will be dominated by $h^\beta c(D)$ for low values of h, anyway. □

Let us write out explicitly the assumptions on α and γ that ensure $\beta > 0$ in Lemma 4.2.

$$\beta > 0 \quad \text{if we have} \quad d = 2: \begin{cases} \gamma \in (1,4), & \alpha > \frac{4}{\gamma} - 1, \\ \gamma \geq 4, & \alpha > 0, \end{cases} \quad \text{or} \quad d = 3: \begin{cases} \gamma \in (1,4), & \alpha > \frac{6}{\gamma} - \frac{3}{2}, \\ \gamma \geq 4, & \alpha > 0. \end{cases} \tag{80}$$

The two following lemmas find their use in the proof of consistency of the momentum scheme.

Lemma 4.3. *For any $f_h \in X(\mathcal{T})$, $\mathbf{g}_h \in X(\mathcal{E}_{\mathrm{int}})^d$, $\mathbf{v} \in W^{2,q}(\Omega)$ we have*

$$\int_\Omega f_h \mathrm{div}_x \mathbf{v} \,\mathrm{d}x = \int_\Omega f_h \mathrm{div}_h(\Pi_h^D \mathbf{v}) \,\mathrm{d}x, \tag{81}$$

Proof. The proof of both identities is based on the Divergence theorem and decomposition of the domain Ω to cells Q_K, where f_h and $(\nabla_h \mathbf{g})$ are constant. The chain of equalities

$$\begin{aligned} \int_\Omega f_h \mathrm{div}_x \mathbf{v} \,\mathrm{d}x &= \sum_{K\in\mathcal{T}} f_K \int_{Q_K} \mathrm{div}_x \mathbf{v} \,\mathrm{d}x = \sum_{K\in\mathcal{T}} f_K \int_{\partial Q_K} \mathbf{v}\cdot\mathbf{n} \,\mathrm{d}S_x \\ &= h^2 \sum_{K\in\mathcal{T}} f_K \sum_{s=1}^d \Big((\Pi^D \mathbf{v})_{\sigma,s+} - (\Pi^D \mathbf{v})_{\sigma,s-} \Big) = h^d \sum_{K\in\mathcal{T}} f_K (\mathrm{div}_h \Pi_h^D \mathbf{v})_K = \int_\Omega f_h \mathrm{div}_h(\Pi_h^D \mathbf{v}) \,\mathrm{d}x, \end{aligned}$$

recovers (81). □

Next, let us define the extension for $(\partial_h^r g^s)_{K+\frac{h}{2}\mathbf{e}_s \pm \frac{h}{2}\mathbf{e}_r}$ for $r \neq s$ and $\mathbf{g} \in X(\mathcal{E}_{\mathrm{int}})^d$ to be piecewise constant in its neighbourhood. In particular we define

$$(q^{r,s})(x) = (q^{r,s})_{K+\frac{h}{2}\mathbf{e}_s \pm \frac{h}{2}\mathbf{e}_r}, \qquad \text{when } x - \frac{h}{2}\mathbf{e}_s \mp \frac{h}{2}\mathbf{e}_r \in Q_K \wedge x \in \Omega, \tag{82}$$

with

$$q^{r,s} = \partial_h^r g_h^s \qquad \text{or} \qquad q^{r,s} = \partial_h^r g_h^s \partial_h^r v_h^s, \tag{83}$$

where $\mathbf{g} \in X(\mathcal{E}_{\mathrm{int}})^d$ and $\mathbf{v} \in X(\mathcal{E})^d$.

As a consequence of (82) we have also

$$h^d \sum_{K\in\mathcal{T}} \left(\frac{1}{2}(q^{r,s})_{K+\frac{h}{2}\mathbf{e}_s+\frac{h}{2}\mathbf{e}_r} + \frac{1}{2}(q^{r,s})_{K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_r} \right) = \int_\Omega (q^{r,s}) \,\mathrm{d}x,$$

and thus also

$$h^d \sum_{K\in\mathcal{T}} \sum_{s=1}^d \left((q^{s,s})_K + \sum_{\substack{r=1 \\ r\neq s}}^d \left(\frac{1}{2}(q^{r,s})_{K+\frac{h}{2}\mathbf{e}_s+\frac{h}{2}\mathbf{e}_r} + \frac{1}{2}(q^{r,s})_{K+\frac{h}{2}\mathbf{e}_s-\frac{h}{2}\mathbf{e}_r} \right) \right) = \int_\Omega \sum_{r=1}^d \sum_{s=1}^d q^{r,s} \,\mathrm{d}x, \tag{84}$$

where $q^{r,s}$ satisfies (83). The core of the argument is that all nonzero $q_K^{r,s}$ are covered twice with one-half, beside the border ones, whose intersection with Ω is of the size $h^d/2$.

The extension (82) might be viewed as another mesh, and that is the reason why Gallouet at al. define it at the beginning in [20]. We prefer to state it here at the only place where we use it. Notice that for $r = s$ we have $\partial_h^s g^s \in X(\mathcal{T})$, for which the extension is defined in Section 2.2.3.

Lemma 4.4. *Let $\mathbf{g} \in X(\mathcal{E}_{\mathrm{int}})^d$ and $\mathbf{v} \in W_0^{1,1}(\Omega)$. Then it holds that*

$$\begin{aligned} &\sum_{K\in\mathcal{T}} \sum_{s=1}^d \left((\partial_h^s g^s)_K (\partial_h^s \Pi^D \mathbf{v})_K + \frac{1}{2} \sum_{\substack{r=1 \\ r\neq s}}^d \sum_{i=1}^2 (\partial_h^r g^s)_{K+\frac{h}{2}\mathbf{e}_s+(-1)^i\frac{h}{2}\mathbf{e}_r} (\partial_h^r \Pi^D \mathbf{v})_{K+\frac{h}{2}\mathbf{e}_s+(-1)^i\frac{h}{2}\mathbf{e}_r} \right) \\ &\quad = \int_\Omega \sum_{s=1}^d \sum_{r=1}^d \widetilde{\partial_h^r g_h^s \partial_x^r v^s}(x) \,\mathrm{d}x + R = \int_\Omega \widetilde{\nabla_h \mathbf{g}_h} : \nabla_x \mathbf{v} \,\mathrm{d}x + R, \end{aligned} \tag{85}$$

where $|R| \leq h \|\widetilde{\nabla_h \mathbf{g}_h}\|_2 \|\nabla_x^2 \mathbf{v}\|_2$.

Proof. Let $\dot{K} := K + \frac{h}{2}\mathbf{e}_s + (-1)^i \frac{h}{2}\mathbf{e}_r$ for $i \in \{1,2\}$, no matter whether $r \neq s$ or not. If we extend $\mathbf{v}$ with zero outside Ω, we can express

$$(\partial_h^r(\Pi^D \mathbf{v})^s)_{\dot{K}} = \frac{1}{h}\left[\frac{1}{h^{d-1}}\int_{F_{\sigma+}} v^s \,\mathrm{d}S_x - \frac{1}{h^{d-1}}\int_{F_{\sigma-}} v^s \,\mathrm{d}S_x\right], \tag{86}$$

where $\sigma\pm := \dot{K} \pm \frac{h}{2}\mathbf{e}_r$.

Similarly as in the proof of Lemma 4.1, we use Taylor theorem to express

$$v^s(x) = v^s(\sigma) + \nabla_x v^s(\sigma)(x-\sigma) + \frac{1}{2}\nabla_x^2 v^s(x-\sigma), \tag{87}$$

for $x \in F_\sigma$ Substituting (87) into (86) yields

$$(\partial^r(\Pi^D \mathbf{v})^s)_{\dot{K}} \leq \frac{1}{h}(v^s(\sigma+) - v^s(\sigma-)) + h(|\nabla_x^2 v^s(\sigma+)| + |\nabla_x^2 v^s(\sigma-)|),$$

as the affine function with zero mean belongs to the kernel of the projection Π^D. Then we use the mean value theorem twice to get for $x \in Q_{\dot{K}}$ (we apologize for an abuse of notation)

$$\begin{aligned}
&(\partial^r(\Pi^D \mathbf{v})^s)_{\dot{K}} - \partial_x^r v^s(x) \\
&\leq \frac{1}{h}(v^s(\sigma+) - v^s(\sigma-)) - \partial_x^r v^s(x) + h(|\nabla_x^2 v^s(\sigma+)| + |\nabla_x^2 v^s(\sigma-)|) \\
&= \partial_x^r v^s(\xi) - \partial_x^r v^s(x) + h(|\nabla_x^2 v^s(\sigma+)| + |\nabla_x^2 v^s(\sigma-)|) \\
&\leq h\sqrt{2}\nabla_x \partial_x^r v^s(\xi') + + h(|\nabla_x^2 v^s(\sigma+)| + |\nabla_x^2 v^s(\sigma-|).
\end{aligned}$$

Now we have for any $\dot{K}$

$$(\partial_h^r g^s)_{\dot{K}}(\partial^r(\Pi^D \mathbf{v})^s)_{\dot{K}} \leq (\partial_h^r g^s)_{\dot{K}} h^{-d}\int_{Q_{\dot{K}}} \partial_x^r v^s \,\mathrm{d}x + (\partial_h^r g^s)_{\dot{K}} h^{1-d}\int_{Q_{\dot{K}}} |\nabla_x^2 v^s(x)| \,\mathrm{d}x. \tag{88}$$

Finally we apply (84) to (88) and Cauchy-Schwarz inequality to obtain (85). □

4.2 Consistency of the continuity scheme

The weak formulation of the continuity method reads as follows.

Theorem 4.5 (Consistency formulation for the continuity)**.** *Let $\varrho_h, \hat{\mathbf{u}}_h$ be piecewise constant and piecewise affine representations, respectively in space and piecewise constant in time, of the solution to the numerical scheme (18–19), with the following parameters: $\gamma > \frac{2d}{d+2}, \alpha > \max\left\{\frac{d(4-\gamma)}{2\gamma}, 0\right\}$, i.e.*

$$\begin{aligned}
d = 2: &\qquad \gamma > 1, \qquad \alpha > \max\left\{\frac{4}{\gamma} - 1, 0\right\}, \\
d = 3: &\qquad \gamma > \frac{6}{5}, \qquad \alpha > \max\left\{\frac{6}{\gamma} - \frac{3}{2}, 0\right\}.
\end{aligned} \tag{89}$$

Then for any $\phi \in C^2(\Omega)$ it holds that

$$\int_\Omega \partial_h^t \varrho_h^n \phi \,\mathrm{d}x - \int_\Omega \varrho_h^n \hat{\mathbf{u}}_h^n \cdot \nabla_x \phi \,\mathrm{d}x = h^{\theta_1}\langle \mathbf{r}_h, \nabla_x \phi\rangle + h^{\theta_2}\langle \mathbb{Q}_h, \nabla_x^2 \phi\rangle,$$

where $\theta_1, \theta_2 > 0$ and $\|\mathbf{r}_h\|_{L^1(0,T,L^{p'}(\Omega))} \lesssim 1$, $\|\mathbb{Q}_h\|_{L^1(0,T;L^{q'}(\Omega))} \lesssim 1$ for $p' = \frac{p}{p-1}$ and $q' = \frac{q}{q-1}$ satisfying

$$d = 2: \quad \begin{cases} p \geq 2 \\ q > \frac{2\gamma}{3\gamma - 2} \\ q \geq 1 \end{cases} \qquad \text{or} \qquad d = 3: \quad \begin{cases} p \geq 2 \\ p > \frac{6\gamma}{5\gamma - 6} \\ q > \frac{6\gamma}{7\gamma - 6} \\ q \geq 1 \end{cases}.$$

Proof. We multiply (18) with $h^d(\Pi^P \phi)_K$ and sum over $K \in \mathcal{T}$. Then we handle the product term by term as following.

Time derivative. We use (9) to get

$$h^d \sum_{K\in\mathcal{T}} (\partial_h^t \varrho_K)^n (\Pi^P \phi) = \sum_{K\in\mathcal{T}} (\partial_h^t \varrho_K)^n \int_K \phi(x) \,\mathrm{d}x = \int_\Omega (\partial_h^t \varrho_h)^n \phi \,\mathrm{d}x.$$

Convective term. Using the definition of the projection and standard integration by parts we get

$$\begin{aligned}
&h^d \sum_{K\in\mathcal{T}} \mathrm{div}_{\mathrm{Up}}[\varrho^n, \mathbf{u}^n]_K (\Pi^P \phi)_K \\
&= \int_\Omega \mathrm{div}_{\mathrm{Up}}[\varrho_h^n, \mathbf{u}_h^n]\phi \,\mathrm{d}x \\
&= -\sum_{s=1}^d \int_\Omega \mathrm{Up}[\varrho_h^n, \mathbf{u}_h^n] \frac{\phi(. + \frac{h}{2}\mathbf{e}_s) - \phi(. - \frac{h}{2}\mathbf{e}_s)}{h} \,\mathrm{d}x \\
&= -\sum_{s=1}^d \int_\Omega \{\varrho_h^n\} u_h^{s,n} \partial_h^s \phi \,\mathrm{d}x + \sum_{s=1}^d \int_\Omega \frac{h}{2} |u^{s,n}| (\partial_h^s \varrho_h^n) \partial_h^s \phi \,\mathrm{d}x =: I_1 + R_1,
\end{aligned}$$

where the equality on the last row follows from the application of Lemma 2.5. Further

$$\begin{aligned}
I_1 &= -\int_\Omega \{\varrho_h^n\} \mathbf{u}_h^n \cdot \nabla_h \phi \,\mathrm{d}x = -\int_\Omega \varrho_h^n \mathbf{u}_h^n \cdot \nabla_h \phi \,\mathrm{d}x - \int_\Omega \left(\frac{\varrho_h^n(x + \frac{h}{2}\mathbf{e}_s) - \varrho_h^n(x)}{2} - \frac{\varrho_h^n(x) - \varrho_h^n(x - \frac{h}{2}\mathbf{e}_s)}{2} \right) \mathbf{u}_h^n \cdot \nabla_h \phi \,\mathrm{d}x \\
&=: I_2 + R_2.
\end{aligned}$$

Then, using standard integration by parts together with $\mathbf{u}_h|_{\partial\Omega} = 0$, the identities

$$\partial_h^s v_h^s|_K = \partial_h^s \hat{v}_h^s|_K \equiv \partial^s \hat{v}_h^s|_K,$$

for any $\mathbf{v} = (v^1, v^2, v^3) \in X(\mathcal{E}_{\mathrm{int}})^d$ and $\mathrm{div}_x \hat{\mathbf{u}}$ being constant on each cell, we get

$$\begin{aligned}
I_2 &= \int_\Omega \mathrm{div}_h(\varrho_h^n \mathbf{u}_h^n)\phi \,\mathrm{d}x = \sum_{K\in\mathcal{T}} \varrho_K^n \int_K \mathrm{div}_h \mathbf{u}_h^n \phi = \sum_{K\in\mathcal{T}} \varrho_K^n \int_{Q_K} \phi \mathrm{div}_x \hat{\mathbf{u}}_h^n \,\mathrm{d}x \\
&= \sum_{K\in\mathcal{T}} \int_{Q_K} \phi \mathrm{div}_x(\varrho_h^n \hat{\mathbf{u}}_h^n) = -\int_\Omega \varrho_h^n \hat{\mathbf{u}}_h^n \cdot \nabla_x \phi \,\mathrm{d}x.
\end{aligned}$$

We need to show that the residual terms R_1, R_2 contribute to $\mathbf{r}_h, \mathbb{Q}_h$. To see that, we perform summation by parts to R_1, R_2 to obtain

$$|R_1| + |R_2| \lesssim h \left| \int_\Omega \partial_h(\mathbf{u}_h^n \nabla_x \phi) \varrho_h^n \,\mathrm{d}x \right| \leq h \int_\Omega |\nabla_h \mathbf{u}_h^n| |\nabla_x \phi| |\varrho_h^n| \,\mathrm{d}x + h \int_\Omega |\mathbf{u}_h^n| |\partial_h \nabla_x \phi| |\varrho_h^n| \,\mathrm{d}x =: R_1' + R_2'.$$

Using Hölder inequality with exponents p_1, p_2, p, where $\frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p} = 1$, and using inverse estimates we can estimate

$$|R_1'| = h \int_\Omega |\nabla_h \mathbf{u}_h^n| |\nabla_x \phi| |\varrho_h^n| \,\mathrm{d}x \lesssim h \|\nabla_h \mathbf{u}_h^n\|_{p_1} \|\varrho_h^n\|_{p_2} \|\nabla_x \phi\|_p, \lesssim h^{\theta_1} \|\nabla_h \mathbf{u}_h^n\|_2 \|\varrho_h^n\|_\gamma \|\nabla_x \phi\|_p, \tag{90}$$

where $\theta_1 > 0$ as long as $p > \frac{2d\gamma}{\gamma(2+d)-2d}$, which implies the restriction on γ such that $\gamma > \frac{2d}{2+d}$, see also Remark 5.

Similarly we deduce

$$|R_2'| \lesssim h^\theta \|\mathbf{u}_h^n\|_{q_1} \|\varrho_h^n\|_\gamma \|\nabla_x^2 \phi\|_q, \tag{91}$$

where $\theta > 0$ if $q \geq 1$ and $q > \frac{dq_1\gamma}{(q_1+dq_1-d)\gamma - dq_1}, \gamma > \frac{dq_1}{q_1+dq_1-d}, q_1 \geq 1$. More specifically, the lower bounds read

$$d = 3: \ q > \frac{6\gamma}{7\gamma - 6} \text{ with } q_1 = 6, \quad \text{or} \quad d = 2: \ q = q(q_1) > \frac{2\gamma}{3\gamma - 2} \text{ with } q_1 \text{ arbitrarily large.}$$

We recall also the basic constraint $\gamma \geq 1$ which is crucial for stability of the method.

Then, summing over time one gets

$$\Delta t \sum_{n=1}^{N_t} (|R_1| + |R_2|) \lesssim h^{\theta_1} c(D) \|\nabla_x \phi\|_p + h^\theta c(D) \|\nabla_x^2 \phi\|_q,$$

after using the energy estimates (48), (69) and (68).

Artificial viscosity term. We perform integration by parts (14) to get

$$h^{d+\alpha}\sum_{K\in\mathcal{T}}(\Delta_h\varrho^n)_K(\Pi^P\phi)_K = h^{d+\alpha}\sum_{\sigma\in\mathcal{E}_{\rm int}}(\nabla_h\varrho^n)_\sigma(\partial_h^s\Pi^P\phi)_\sigma,$$

which can be further estimated using Hölder inequality to obtain

$$h^{d+\alpha}\sum_{K\in\mathcal{T}}(\Delta_h\varrho^n)_K(\Pi^P\phi)_K \le h^\alpha\left(\int_\Omega(\partial_h\varrho_h^n)^2\,\mathrm{d}x\right)^{\frac12}\left(\int_\Omega(\partial_h\Pi^P\phi)^2\,\mathrm{d}x\right)^{\frac12} \lesssim h^\alpha\|\partial_h\varrho_h^n\|_2\|\nabla\phi\|_2, \tag{92}$$

where we used Lemma 4.1 in the last inequality. Then the summation over time and Lemma 4.2 supply the estimate $h^\beta c(D)$ as well as the lower bound on α, see (80). Moreover, $p\ge 2$ is required.

The existence of $\mathbf{r}_h, \mathbb{Q}_h$ with properties stated in the Theorem is a consequence of appropriate boundedness of terms on the right-hand sides of (90), (91), (92), the Riesz representation theorem and $\theta_2 = \min\{\theta,\beta\}$. □

Remark 5. *In the above computation, we can formally apply the inverse estimate to smooth functions as well. For instance in (90), since* $\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p}=1$, *we have*

$$0<\theta_1 = 1+d\left(\frac{1}{p_1}-\frac12\right)+d\left(\frac{1}{p_2}-\frac{1}{\gamma}\right) = 1+d\left(1-\frac12-\frac{1}{\gamma}-\frac{1}{p}\right) = d\left(\frac{\gamma(2+d)-2d}{2d\gamma}-\frac{1}{p}\right),$$

which indicates

$$p>\frac{2d\gamma}{\gamma(2+d)-2d},\quad \gamma>\frac{2d}{2+d}.$$

4.3 Consistency of the momentum scheme

Theorem 4.6 (Consistency formulation for the momentum)**.** *Let* $(\varrho_h^n,\mathbf{u}_h^n)$ *be piecewise constant representations of the solution to numerical scheme (18–19) with* $\Delta t\approx h$ *and the following parameters*

$$\gamma>\frac{d}{2}, \alpha>\max\left\{\frac{d(4-\gamma)}{2\gamma},0\right\}. \tag{93}$$

Then for any $\mathbf{v}\in C^2(\Omega)^3$, *it holds that*

$$\begin{aligned}&\int_\Omega\partial_h^t(\varrho_h\bar{\mathbf{u}}_h)^n\cdot\mathbf{v}\,\mathrm{d}x-\int_\Omega\varrho_h^n\bar{\mathbf{u}}_h^n\otimes\bar{\mathbf{u}}_h^n:\nabla_x\mathbf{v}\,\mathrm{d}x-\int_\Omega p(\varrho_h^n)\mathrm{div}_x\mathbf{v}\,\mathrm{d}x+\mu\int_\Omega(\nabla_h\mathbf{u}_h^n):\nabla_x\mathbf{v}\,\mathrm{d}x\\ &\quad=h^{\theta_1}\langle\mathbf{r_h},\nabla_x\mathbf{v}\rangle+h^{\theta_2}\langle\mathbb{Q}_h,\nabla_x^2\mathbf{v}\rangle,\end{aligned} \tag{94}$$

with $\|\mathbf{r_h}\|_{L^1(0,T;L^{p'}(\Omega)}\lesssim 1$ *and* $\|\mathbb{Q}_h\|_{L^1(0,T;L^{q'}(\Omega)}\lesssim 1$, *where* $p'=\frac{p}{p-1}$ *and* $q'=\frac{q}{q-1}$ *which satisfy:*

$$d=2:\quad\begin{cases}p\ge 3,\\ p>\frac{2\gamma}{\gamma-1},\\ q>\frac{2\gamma}{\gamma-1},\end{cases}\qquad or\qquad d=3:\quad\begin{cases}p>\frac{6\gamma}{2\gamma-3},\\ q>\frac{6\gamma}{4\gamma-3}.\end{cases} \tag{95}$$

Proof. We multiply momentum scheme (19) by $h^d\Pi^D\mathbf{v}$ and handle term by term. We would like to point out, that the values of exponents θ_i may vary throughout the proof. To find the proper values of θ_i for (94) should be obtained as the minima of $\theta_i, i=1,2$ throughout their occurrences in the proof.

Time difference term. Using the transition between grids (6) one gets

$$h^d\sum_{\sigma\in\mathcal{E}_{\rm int}}\partial_h^t\{\varrho\bar{\mathbf{u}}\}_\sigma^n\cdot\Pi^D\mathbf{v} = h^d\partial_h^t\left(\sum_{K\in\mathcal{T}}(\varrho\bar{\mathbf{u}})_K\cdot(\Pi^P\Pi^D\mathbf{v})_K\right)^n = \int_\Omega\partial_h^t(\varrho_h\bar{\mathbf{u}}_h)^n\cdot v + R_1+R_2,$$

where

$$\begin{aligned}R_1 &= h^d\sum_{K\in\mathcal{T}}\sqrt{\varrho_K^{n-1}}\sqrt{\varrho_K^{n-1}}\frac{\bar{\mathbf{u}}_K^n-\bar{\mathbf{u}}_K^{n-1}}{\Delta t}\int_{Q_K}(\Pi^P\Pi^D\mathbf{v}-\mathbf{v})\,\mathrm{d}x\\ &\le\|\varrho_h^{n-1}\|_\gamma^{\frac12}h\|\nabla_x\mathbf{v}\|_{\frac{2\gamma}{\gamma-1}}\left((\Delta t)\int_\Omega\varrho_h^{n-1}\left(\frac{\bar{\mathbf{u}}_h^n-\bar{\mathbf{u}}^{n-1}}{\Delta t}\right)^2\mathrm{d}x\right)^{\frac12}(\Delta t)^{-\frac12}\\ &=: h^{\theta_1}\|\varrho_h^{n-1}\|_\gamma^{\frac12}\|\nabla_x\mathbf{v}\|_{\frac{2\gamma}{\gamma-1}}\mathbf{U}_h^n,\quad \theta_1=\frac12.\end{aligned}$$

By virtue of (34) we have $\Delta t \sum_{n=1}^{N_t} (\mathbf{U}_h^n)^2 \lesssim c(D)$, which implies, together with (48), that $\Delta t \sum_{n=1}^{N_t} |R_1| \lesssim h^{\frac{1}{2}} \|\nabla_x \mathbf{v}\|_{\frac{2\gamma}{\gamma-1}} c(D)$. The other residual term reads

$$R_2 = h^d \sum_{K\in\mathcal{T}} \bar{\mathbf{u}}_K^n \frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} (\Pi^P \Pi^D \mathbf{v} - \mathbf{v}). \tag{96}$$

From energy estimates (34) we have

$$\Delta t^2 h^d \sum_{n=1}^{N_t} \sum_{K\in\mathcal{T}} (\varrho_K^\eta)^{\gamma-2} \left(\frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} \right)^2 \lesssim c(D). \tag{97}$$

Using the properties of Legendre remainder points of strictly convex functions, formulated e.g. in [7, Lemma 2.1], (97) implies also

$$(\Delta t)^\gamma h^d \sum_{n=1}^{N_t} \sum_{K\in\mathcal{T}} \left(\frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} \right)^\gamma \lesssim c(D,\gamma). \tag{98}$$

Thus, applying Hölder inequality and estimate (71) to (96), one gets

$$|R_2| \le \|\mathbf{u}^n\|_6 h \|\nabla_x \mathbf{v}\|_{\frac{6\gamma}{5\gamma-6}} (\Delta t)^{-\frac{\gamma-1}{\gamma}} \left((\Delta t)^{\gamma-1} h^d \sum_{K\in\mathcal{T}} \left(\frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} \right)^\gamma \right)^{\frac{1}{\gamma}} \le h^{\frac{1}{\gamma}} \|\mathbf{u}_h^n\|_6 \|\nabla_x \mathbf{v}\|_{\frac{6\gamma}{5\gamma-6}} H_h^n,$$

where (98) yields $\Delta t \sum_{n=1}^{N_t} (H_h^n)^\gamma \lesssim c(D)$. This together with (69) implies

$$\Delta t \sum_{n=1}^{N_t} |R_2| \lesssim h^{\frac{1}{\gamma}} c(D,\gamma) \|\nabla_x \mathbf{v}\|_{\frac{6\gamma}{5\gamma-6}}, \quad \text{for } d = 3,$$

For $d = 2$ we can estimate analogously

$$|R_2| \le h^{\frac{1}{\gamma}} \|\mathbf{u}^n\|_{p_1} \|\nabla_x \mathbf{v}\|_p H_h^n, \qquad \text{with } \frac{p_1\gamma}{(p_1-1)\gamma - p_1},$$

and thus we get a lower bound $p = p(p_1) > \frac{\gamma}{\gamma-1}$, as p_1 can be arbitrarily large.

In both choices of d we have $\theta_2 = \frac{1}{\gamma}$. It is possible, but not effective to lower the integrability exponent of $\nabla_x \mathbf{v}$ by inverse estimates, since this constraint on integrability is not active.

Notice that we used the relation $\Delta t \approx h$ in this part of the proof.

Convective term. We use the transition between grids, summation by parts (14) and Lemma 2.5 to obtain

$$\begin{aligned}
&\Delta t h^d \sum_{n=1}^{N_t} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \{\mathrm{div}_{\mathrm{Up}}[\varrho^n \bar{\mathbf{u}}^n, \mathbf{u}^n]\}_\sigma \cdot (\Pi^D \mathbf{v})_\sigma = -\Delta t h^d \sum_{n=1}^{N_t} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \mathrm{Up}[\varrho^n \bar{\mathbf{u}}^n, \mathbf{u}^n] \cdot (\partial_h^s \Pi^P \Pi^D \mathbf{v})_\sigma \\
&= -\Delta t h^d \sum_{n=1}^{N_t} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} u_\sigma^{s,n} \{\varrho_h^n \bar{\mathbf{u}}_h^n\}_\sigma \cdot (\partial_h^s \Pi^P \Pi^D \mathbf{v})_\sigma + h^{d+1} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |u_\sigma^{s,n}| \partial_h^s (\varrho^n \bar{\mathbf{u}}^n)_\sigma \cdot (\partial_h^s \Pi^P \Pi^D \mathbf{v})_\sigma \\
&= -h^d \Delta t \sum_{n=1}^{N_t} \sum_{K\in\mathcal{T}} \varrho^n \bar{\mathbf{u}}^n \otimes \bar{\mathbf{u}}^n : \{\nabla_h \Pi^P \Pi^D \mathbf{v}\}_K + R_3 = -\Delta t \sum_{n=1}^{N_t} \sum_{K\in\mathcal{T}} \int_{Q_K} \varrho_K^n \bar{\mathbf{u}}_K^n \otimes \bar{\mathbf{u}}_K^n : \nabla_x \mathbf{v} + R_3 + R_4.
\end{aligned}$$

We need to estimate the residual terms. Before starting that, we perform summation by parts (14) to R_3 and we split the discrete derivative of the product,

$$\begin{aligned}
R_3 &= -h^{d+1} \sum_{K\in\mathcal{T}} (\varrho^n \bar{\mathbf{u}}^n)_K \cdot \sum_{s=1}^d (\partial_h^s (|u^{s,n}| \partial_h^s \Pi^P \Pi^D \mathbf{v}))_K \\
&= -h^{d+1} \sum_{K\in\mathcal{T}} (\varrho^n \bar{\mathbf{u}}^n)_K \cdot \sum_{s=1}^d (\partial_h^s u^{s,n})_K (\partial_h^s \Pi^P \Pi^D \mathbf{v})_{\sigma,s-} - h^{d+1} \sum_{K\in\mathcal{T}} (\varrho^n \bar{\mathbf{u}}^n)_K \cdot \sum_{s=1}^d (\partial_h^s \partial_h^s \Pi^P \Pi^D \mathbf{v})_K u_{\sigma,s+}^{s,n} \\
&=: R_{3,1} + R_{3,2}.
\end{aligned}$$

Then we use Hölder inequality using $p, p_1, p_2 : \frac{1}{p} + \frac{1}{p_1} + \frac{1}{p_2} = 1$, the inequality

$$|\partial_h^s |g|| \le |\partial_h^s g|,$$

relation (70) and inverse estimate (Lemma 2.3) twice to get

$$|R_{3,1}| \le h^{d+1} \sum_{K\in\mathcal{T}} |\varrho_K^n \bar{\mathbf{u}}_K^n| \sum_{s=1}^{d} |(\partial_h^s u^{s,n})_K||(\partial_h^s \Pi^P \Pi^D \mathbf{v})_{\sigma,s-}| \le h\|\nabla_h \mathbf{u}\|_{p_1} \|\varrho_h^n \bar{\mathbf{u}}_h^n\|_{p_2} \|\partial_h \Pi^P \Pi^D \mathbf{v}\|_p$$
$$\lesssim h^{\theta_1} \|\nabla_h \mathbf{u}_h^n\|_2 \|\varrho_h^n \bar{\mathbf{u}}_h^n\|_{\frac{2\gamma}{\gamma+1}} \|\nabla_x \mathbf{v}\|_p,$$

where the exponent θ_1 remains positive as long as $p > \frac{2d\gamma}{2\gamma - d}, \gamma > \frac{d}{2}$, i.e.,

$$d = 2: \ p > \frac{2\gamma}{\gamma - 1}, \gamma > 1, \quad \text{or} \quad d = 3: \ p > \frac{6\gamma}{2\gamma - 3}, \gamma > \frac{3}{2}.$$

Similarly, for $R_{3,2}$ we use the same tools and Mean Value Theorem to obtain

$$|R_{3,2}| \le h^{d+1} \sum_{K\in\mathcal{T}} |\varrho_K^n \bar{\mathbf{u}}_K^n| \sum_{s=1}^{d} |u_{\sigma,s+}^{s,n}||(\partial_h^s \partial_h^s \Pi^P \Pi^D \mathbf{v})_K| \lesssim h\|\mathbf{u}_h^n\|_{q_1} \|\varrho_h^n \bar{\mathbf{u}}_h^n\|_{q_2} \|\partial_h |\nabla_x \mathbf{v}|\|_q$$
$$\lesssim h^{\theta_2} \|\mathbf{u}_h^n\|_{q_1} \|\varrho_h^n \bar{\mathbf{u}}_h^n\|_{\frac{2\gamma}{\gamma+1}} \|\nabla_x^2 \mathbf{v}\|_q,$$

where θ_2 is positive as long as

$$d = 2: \ q = q(q_1) > \frac{2\gamma}{2\gamma - 1}, \quad \text{or} \quad d = 3: \ q > \frac{6\gamma}{4\gamma - 3},$$

since q_1 can be arbitrarily large for $d = 2$ and $q_1 = 6$ for $d = 3$.

Applying summation over time, uniform estimates (49, 69) and the assumptions on test function $\mathbf{v}$ one gets that $\Delta t \sum_{n=1}^{N_t} |R_{3,1}| + |R_{3,2}| = c(D) \left(h^{\theta_1} \|\nabla_x \mathbf{v}\|_p + h^{\theta_2} \|\nabla_x^2 \mathbf{v}\|_q \right)$.

Pressure term. By virtue of summation by parts (14) and Lemma 4.3 we write the following chain of equalities:

$$h^d \sum_{\sigma\in\mathcal{E}_{\rm int}} (\partial_h^s p(\varrho^n))_\sigma \mathbf{e}_s \cdot (\Pi_h^D \mathbf{v})_\sigma = -h^d \sum_{K\in\mathcal{T}} p(\varrho_K^n) \mathrm{div}_h (\Pi_h^D \mathbf{v})_K = -\int_\Omega p(\varrho_h^n) \mathrm{div}_x \mathbf{v} \, \mathrm{d}x.$$

Viscosity term. We apply summation by parts (Lemma 2.1) and Lemma 4.4 to get

$$h^d \Delta t \sum_{\sigma\in\mathcal{E}_{\rm int}} (\Delta_h \mathbf{u}_h^n)_\sigma \cdot (\Pi^D \mathbf{v})_\sigma^s = h^d \Delta t \int_\Omega \widetilde{\nabla_h \mathbf{u}_h^n} : \nabla_x \mathbf{v} + R_4,$$

where $\Delta t \sum_{n=1}^{N_t} |R_4| \lesssim h\|\widetilde{\nabla_h \mathbf{u}_h}\|_{2,2} \|\nabla_x^2 \mathbf{v}\|_2 \lesssim C(D) h^\theta \|\nabla_x^2 \mathbf{v}\|_q$, with $\theta_2 = 1 + \min\left\{0, d(\frac{1}{2} - \frac{1}{q})\right\}$, thus $\theta_2 > 0$ as long as $q > \frac{2d}{2+d}$, i.e.,

$$d = 2: \ q > 1, \quad \text{or} \quad d = 3: \ q > \frac{6}{5}.$$

Artificial viscosity term. Finally we treat the last term using summation by parts and transition between grids to get

$$R_5 := h^{d+\alpha} \sum_{\sigma\in\mathcal{E}_{\rm int}} \sum_{r=1}^{d} \{\partial_h^r (\{\bar{\mathbf{u}}^n\} \partial_h^r \varrho^n)\}_\sigma \cdot (\Pi^D \mathbf{v})_\sigma = h^{d+\alpha} \sum_{K\in\mathcal{T}} \sum_{r=1}^{d} \sum_{s=1}^{d} \partial_h^r (\{\bar{u}^{s,n}\} \partial_h^r \varrho^n)_K (\Pi^P \Pi^D v^s)_K$$
$$= h^{d+\alpha} \sum_{\sigma\in\mathcal{E}_{\rm int}} \sum_{r=1}^{d} \{\bar{u}^{r,n}\}_\sigma (\partial_h^s \varrho^n)_\sigma (\partial_h^s \Pi^P \Pi^D v^r)_\sigma,$$

where in the last inequality we interchanged the role of r and s in order to get the standard summation over σ, which is associated with s. Applying the Hölder inequality we get

$$|R_5| \le h^\alpha \|\bar{\mathbf{u}}_h^n\|_6 \|\partial_h^s \varrho_h^n\|_2 \|\nabla_x \mathbf{v}\|_3.$$

Summation over time together with applying the uniform bounds gives

$$\Delta t \sum_{n=1}^{N_t} |R_5| \le h^\alpha \|\mathbf{u}_h\|_{2,6} \|\nabla_x \mathbf{v}\|_3 \|\partial_h^s \varrho_h\|_{2,2} \le h^{\theta_1} c(D) \|\nabla \mathbf{v}\|_3,$$

where $\theta_1 = \beta > 0$ is ensured by the assumptions on the lower bounds of α, see (80). Moreover, $p \ge 3$ is required. □

5 Numerical experiments

In this section we perform two numerical experiments for the scheme in two dimensional space, one with Dirichlet boundary condition and the other is periodic type. Our computational domain is always $\Omega = [0,1]^2$, and some constants are chosen as $\mu = 0.01,\ a = 1.0, \gamma = 1.4.\ \alpha = 1.86$ is chosen to satisfy the restriction (80).

Implementation – fix point iteration for the implicit scheme We solve the implicit nonlinear scheme by fix-point iteration. Given the data $\left(\varrho_h^n, \mathbf{u}_h^n\right)$ at time t^n, let $\left(\varrho_h^{n,0}, \mathbf{u}_h^{n,0}\right) = \left(\varrho_h^n, \mathbf{u}_h^n\right)$, then for $\ell = 0, 1, \cdots$, we linearize the nonlinear system and solve

$$\frac{\varrho_K^{n,\ell+1} - \varrho_K^{n,0}}{\Delta t} + \mathrm{div}_{\mathrm{Up}}[\varrho^{n,\ell}, \mathbf{u}^{n,\ell}]_K - h^\alpha(\Delta_h \varrho^{n,\ell})_K = 0,$$

$$\frac{\{\varrho^{n,\ell+1}\bar{\mathbf{u}}^{n,\ell+1}\}_\sigma - \{\varrho^{n,0}\bar{\mathbf{u}}^{n,0}\}_\sigma}{\Delta t} + \{\mathrm{div}_{\mathrm{Up}}[\varrho^{n,\ell}\bar{\mathbf{u}}^{n,\ell}, \mathbf{u}^{n,\ell}]\}_\sigma + \left(\partial_h^s p(\varrho^{n,\ell})\right)_\sigma e_s$$
$$- \mu(\Delta_h \mathbf{u}^{n,\ell+1})_\sigma - h^\alpha \sum_{r=1}^{d} \{\partial_h^r(\hat{u}^{n,\ell}\partial_h^r \varrho^{n,\ell+1})\}_\sigma = 0,$$

until $\|w_h^{n,\ell} - w_h^{n,\ell+1}\| < \xi\|w_h^{n,\ell}\|$, for $w_h \in \{\varrho_h, \mathbf{u}_h\}$, where ξ is a very small positive parameter, e.g. $\xi = 1.0e-6$. Then the solution at next time step t^{n+1} is obtain by $w_h^{n+1} = w_h^{n,\ell+1}$. As we solve the above iterative steps explicitly, a CFL condition is required for preserving the stability $\Delta t = \mathrm{CFL}\frac{h_{\min}}{|\mathbf{u}|_{\max}}$ with $\mathrm{CFL} = 0.6$.

5.1 Cavity flow

In this experiment we simulate the two dimensional cavity flow supplied with Dirichlet data $\mathbf{u} = (16x^2(1-x)^2, 0)^T$ on the top boundary, and zero otherwise. Starting with the initial values $\mathbf{u} = \mathbf{0}$, and $\varrho = 1$ we show in Figure 2 the evolution of the contour mapping for density and velocity components till time $T = 1$ with mesh parameter $h = 1/128$. In order to present the Experimental Order of Convergence(EOC), we calculate the errors in relative norms for different mesh sizes till $t = 0.1$ while the reference solution is computed at the fine mesh $h = 1/512$. From Table 1 we observe first order convergence.

Table 1: Convergence results of cavity flow

h	$\|e_{\nabla\mathbf{u}}\|_{l^2(L^2)}$	EOC	$\|e_{\mathbf{u}}\|_{l^2(L^2)}$	EOC	$\|e_\varrho\|_{l^1(L^1)}$	EOC	$\|e_\varrho\|_{l^\infty(L^\gamma)}$	EOC
1/32	9.22e-03	–	2.84e-01	–	6.08e-05	–	1.79e-03	–
1/64	4.46e-03	1.05	1.37e-01	1.05	2.79e-05	1.12	9.15e-04	0.97
1/128	2.06e-03	1.11	7.14e-02	0.94	1.45e-05	0.95	4.79e-04	0.93
1/256	9.03e-04	1.19	3.09e-02	1.21	5.98e-06	1.27	2.11e-04	1.18

5.2 Gresho-vortex

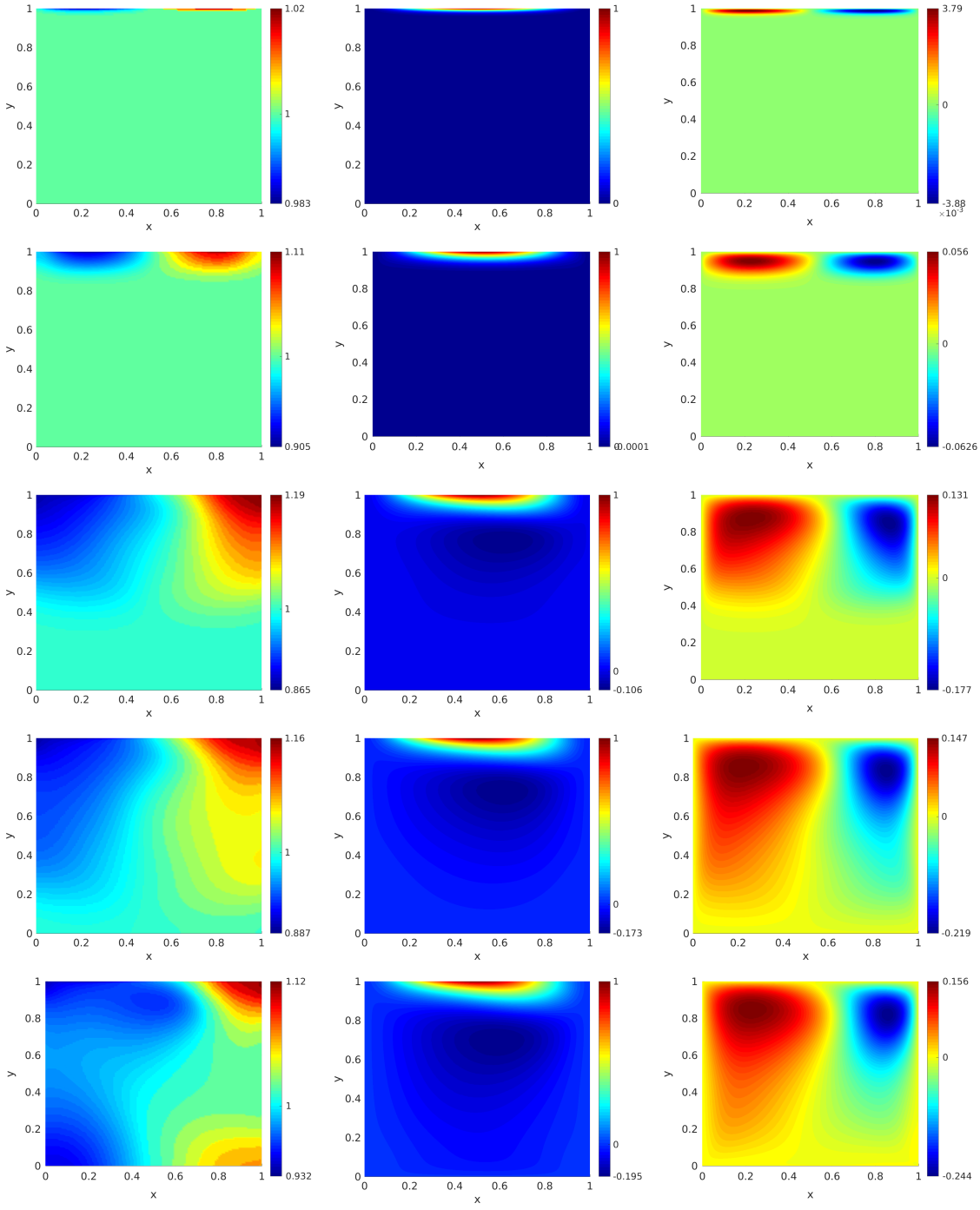
This experiment is an example of rotating vortex, that has been studied in [4, 17, 34] and reference therein for the isentropic flow. Initially, a vortex of radius $R = 0.2$ is prescribed at location $(x_0, y_0) = (0.5, 0.5)$ with the velocity field given by

$$\begin{cases} u_1(0,x,y) = u_r(r) * (y - 0.5)/r, \\ u_2(0,x,y) = u_r(r) * (0.5 - x)/r. \end{cases}$$

where $r = \sqrt{(x-0.5)^2 + (y-0.5)^2}$ and the radial velocity of the vortex u_r is

$$u_r(r) = \sqrt{\gamma}\begin{cases} 2r/R & \text{if } 0 \le r < R/2, \\ 2(1 - r/R) & \text{if } R/2 \le r < R, \\ 0 & \text{if } r \ge R. \end{cases}$$

By setting the periodic boundary condition, we show in Figure 3 the evolution of the flow with mesh parameter $h = 1/128$, from which we see obvious diffusion effects. Analogous to the settings of the previous cavity test, EOC Table 2 indicates similarly first order convergence in the related norms.

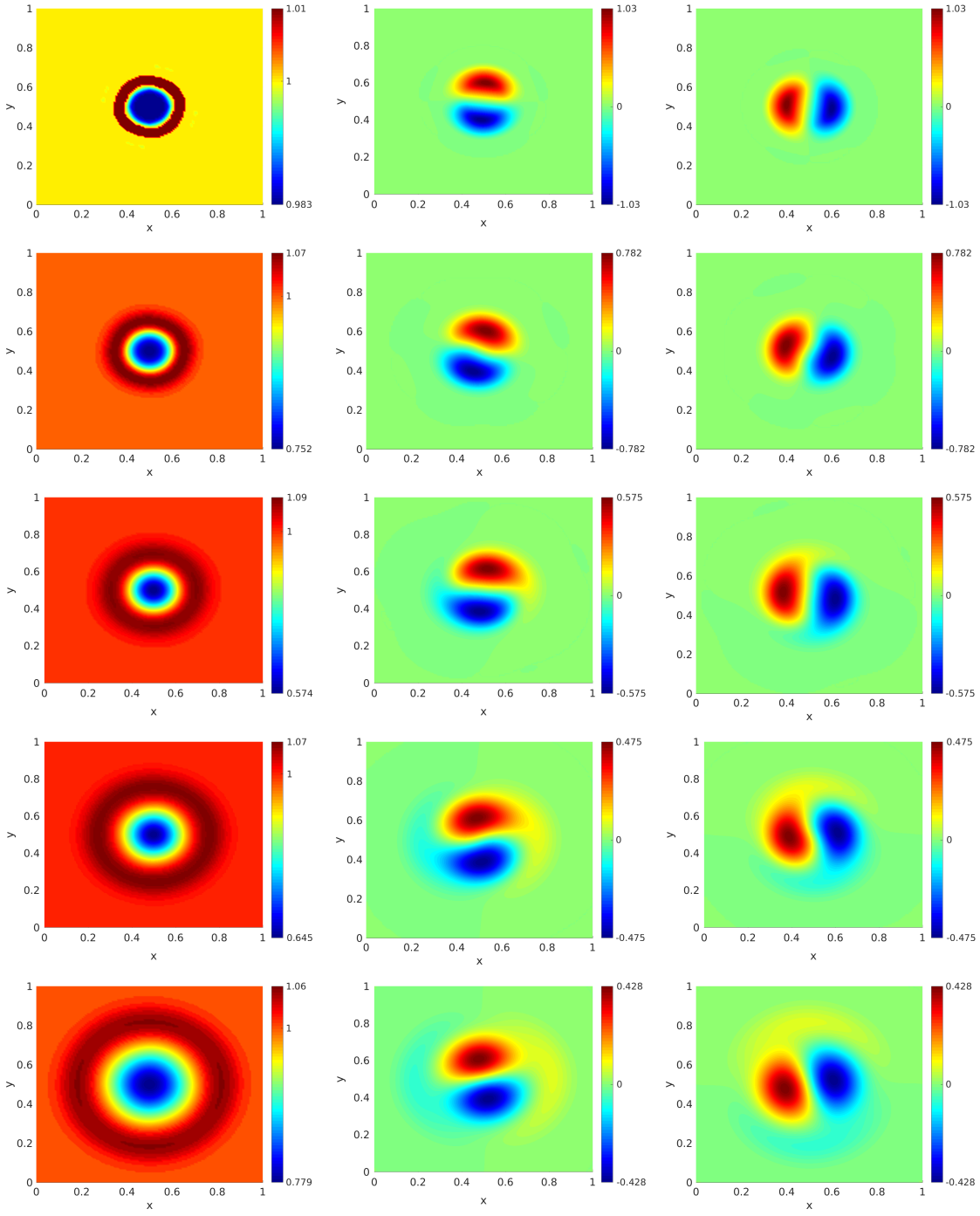


(a) density ϱ (b) velocity component u_1 (c) velocity component u_2

Figure 2: Time evolution of cavity flow, from top to bottom are $t = 0.01, 0.1, 0.5, 0.75, 1$, from left to right are densities and velocity components

Table 2: Convergence results of Gresho vortex test

h	$\|e_{\nabla \mathbf{u}}\|_{l^2(L^2)}$	EOC	$\|e_{\mathbf{u}}\|_{l^2(L^2)}$	EOC	$\|e_{\varrho}\|_{l^1(L^1)}$	EOC	$\|e_{\varrho}\|_{l^\infty(L^\gamma)}$	EOC
1/32	1.10e-02	–	3.74e-01	–	4.40e-04	–	1.35e-02	–
1/64	5.57e-03	0.98	1.88e-01	1.00	2.22e-04	0.99	6.72e-03	1.00
1/128	2.69e-03	1.05	8.71e-02	1.11	1.02e-04	1.12	3.10e-03	1.12
1/256	1.15e-03	1.22	3.37e-02	1.37	3.86e-05	1.40	1.16e-03	1.42



(a) density ϱ (b) velocity component u_1 (c) velocity component u_2

Figure 3: Time evolution of gresho vortex, from top to bottom are $t = 0.01, 0.05, 0.1, 0.15, 0.2$, from left to right are densities and velocity components

A Appendix: Proof of Lemma 2.1

Before proving Lemma 2.1, let us introduce a simplified version in one dimension.

Lemma A.1. *Let the computational domain Ω degenerate to one dimensional interval $I = [a, b]$ and be equally divided into M intervals of the size $h = \frac{b-a}{M}$. Assume that the functions f, ψ, v are discretized at the interval centres, while g, ϕ are located at the division points. Moreover, we assume homogeneous Dirichlet boundary conditions for g, v, e.g.*

$$g_{1/2} = 0, \quad g_{M+1/2} = 0, \quad v_0 = -v_1, \quad v_{M+1} = -v_M.$$

Then the following equalities hold

$$\sum_{i=1}^{M} f_i \frac{g_{i+1/2} - g_{i-1/2}}{h} = -\sum_{i=1}^{M-1} \frac{f_{i+1} - f_i}{h} g_{i+1/2} \tag{99a}$$

$$\sum_{i=1}^{M-1} \frac{\phi_{i+3/2} - 2\phi_{i+1/2} + \phi_{i-1/2}}{h^2} g_{i+1/2} = -\sum_{i=1}^{M} \frac{\phi_{i+1/2} - \phi_{i-1/2}}{h} \frac{g_{i+1/2} - g_{i-1/2}}{h}. \tag{99b}$$

$$-\sum_{i=1}^{M} \frac{\psi_{i+1} - 2\psi_i + \psi_{i-1}}{h^2} v_i = \frac{1}{2}\sum_{i=1}^{M} \frac{\psi_{i+1} - \psi_i}{h} \frac{v_{i+1} - v_i}{h} + \frac{1}{2}\sum_{i=1}^{M} \frac{\psi_i - \psi_{i-1}}{h} \frac{v_i - v_{i-1}}{h} \tag{99c}$$

Proof. By using the boundary conditions $g_{1/2} = g_{M+1/2} = 0$ we directly obtain (99a)

$$\begin{aligned}
&\sum_{i=1}^{M} f_i \frac{g_{i+1/2} - g_{i-1/2}}{h} = \frac{1}{h}\Big(\sum_{i=1}^{M} f_i g_{i+1/2} - \sum_{i=1}^{M} f_i g_{i-1/2}\Big) = \frac{1}{h}\Big(\sum_{i=1}^{M} f_i g_{i+1/2} - \sum_{j=0}^{M-1} f_{j+1} g_{j+1/2}\Big) \\
&= \frac{1}{h}\Big(\sum_{i=1}^{M-1} f_i g_{i+1/2} - \sum_{j=1}^{M-1} f_{j+1} g_{j+1/2} + f_M g_{M+1/2} - f_1 g_{1/2}\Big) = -\sum_{i=1}^{M-1} \frac{f_{i+1} - f_i}{h} g_{i+1/2},
\end{aligned}$$

and (99b)

$$\begin{aligned}
&\sum_{i=1}^{M-1} \frac{\phi_{i+3/2} - 2\phi_{i+1/2} + \phi_{i-1/2}}{h^2} g_{i+1/2} = \frac{1}{h^2}\left(\sum_{j=2}^{M} (\phi_{j+1/2} - \phi_{j-1/2}) g_{j-1/2} - \sum_{i=1}^{M-1} (\phi_{i+1/2} - \phi_{i-1/2}) g_{i+1/2}\right) \\
&= -\sum_{i=1}^{M} \frac{\phi_{i+1/2} - \phi_{i-1/2}}{h} \frac{g_{i+1/2} - g_{i-1/2}}{h} - (\phi_{3/2} - \phi_{1/2}) g_{1/2} - (\phi_{M+1/2} - \phi_{M-1/2}) g_{M+1/2} \\
&= -\sum_{i=1}^{M} \frac{\phi_{i+1/2} - \phi_{i-1/2}}{h} \frac{g_{i+1/2} - g_{i-1/2}}{h}.
\end{aligned}$$

Applying the Dirichlet boundary condition for v we can show (99c)

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{M} \frac{\psi_{i+1} - \psi_i}{h} \frac{v_{i+1} - v_i}{h} + \frac{1}{2}\sum_{i=1}^{M} \frac{\psi_i - \psi_{i-1}}{h} \frac{v_i - v_{i-1}}{h} \\
&= \frac{1}{2h^2}\Big(\sum_{j=2}^{M+1} (\psi_j - \psi_{j-1}) v_j - \sum_{i=1}^{M} (\psi_{i+1} - \psi_i) v_i + \sum_{i=1}^{M} (\psi_i - \psi_{i-1}) v_i - \sum_{j=0}^{M-1} (\psi_{j+1} - \psi_j) v_j\Big) \\
&= \frac{1}{2h^2}\Big(-2\sum_{i=1}^{M} (\psi_{i+1} - 2\psi_i + \psi_{i-1}) v_i + (\psi_{M+1} - \psi_M)(v_{M+1} + v_M) - (\psi_1 - \psi_0)(v_1 + v_0)\Big) \\
&= -\frac{1}{h^2}\sum_{i=1}^{M} (\psi_{i+1} - 2\psi_i + \psi_{i-1}) v_i.
\end{aligned}$$

□

Lemma 2.1 is to show for $f \in X(\mathcal{T}), \mathbf{g} \in X(\mathcal{E}_{\rm int})^d$ the following equalities.

$$\sum_{K\in\mathcal{T}} (\mathrm{div}_h \mathbf{g})_K f_K = -\sum_{\sigma\in\mathcal{E}_{\rm int}} g^s_\sigma (\partial^s_h f)_\sigma.$$

$$-\sum_{\sigma\in\mathcal{E}_{\rm int}} (\Delta_h v^s)_\sigma g^s_\sigma = \sum_{K\in\mathcal{T}} \left((\partial^s_h g^s)_K (\partial^s_h v^s)_K + \frac{1}{2}\sum_{r=1, r\neq s}^{d}\sum_{i=1}^{2} (\partial^r_h g^s)_{K+\frac{h}{2}\mathbf{e}_s+(-1)^i\frac{h}{2}\mathbf{e}_r} (\partial^r_h v^s)_{K+\frac{h}{2}\mathbf{e}_s+(-1)^i\frac{h}{2}\mathbf{e}_r}\right).$$

Proof. It is obvious to obtain the first equality by using (99a) for $s = 1, \ldots, d$ and summing them up. The second equality can be done with same strategy by applying (99b) for the first term on the right hand side and (99c) for the latter term on the right hand side. Summing them up concludes the proof. □

References

[1] G. Ansanay-Alex, F. Babik, J. C. Latché, and D. Vola. An L2-stable approximation of the Navier—Stokes convection operator for low-order non-conforming finite elements. *Int. J. Numer. Meth. Fluids*, 66(5):555–580, 2011.

[2] Y. Coudière, T. Gallouët, and R. Herbin. Discrete Sobolev inequalities and Lp error estimates for finite volume solutions of convection diffusion equations. *ESAIM: M2AN*, 35:767–778, 7 2001.

[3] P. I. Crumpton, J. A. Mackenzie, and K. W. Morton. Cell vertex algorithms for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 109(1):1–15, 1993.

[4] P. Degond and M. Tang. All speed scheme for the low mach number limit of the isentropic Euler equations. *Commun. Comput. Phys.*, 10:1–31, 7 2011.

[5] R. J. DiPerna and P. L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones mathematicae*, 98(3):511–547, 1989.

[6] V. Dolejší and M. Feistauer. *Discontinuous Galerkin method*, volume 48 of *Springer Series in Computational Mathematics*. Springer, Cham, 2015. Analysis and applications to compressible flow.

[7] P. Drábek and R. Hošek. Properties of solution diagrams for bistable equations. *Electron. J. Differ. Equ.*, 2015(156):1–19, 2015.

[8] L. C. Evans. *Partial differential equations.* Providence, RI: American Mathematical Society, 1998.

[9] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of numerical analysis. Vol. 7: Solution of equations in $\mathbb{R}^n$ (Part 3). Techniques of scientific computing (Part 3)*, pages 713–1020. Amsterdam: North-Holland/ Elsevier, 2000.

[10] E. Feireisl. *Dynamics of viscous compressible fluids.* Oxford: Oxford University Press, 2004.

[11] E. Feireisl, P. Gwiazda, A. Świerczewska-Gwiazda, and E. Wiedemann. Dissipative measure-valued solutions to the compressible Navier–Stokes system. *Calculus of Variations and Partial Differential Equations*, 55(6):141, 2016.

[12] E. Feireisl, R. Hošek, D. Maltese, and A. Novotný. Error estimates for a numerical method for the compressible Navier–Stokes system on sufficiently smooth domains. *ESAIM: M2AN*, 51(1):279–319, 2017.

[13] E. Feireisl, R. Hošek, and M. Michálek. A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains. *SIAM J. Numer. Anal.*, 54(5):3062–3082, 2016.

[14] E. Feireisl, T. Karper, and M. Michálek. Convergence of a numerical method for the compressible Navier–Stokes system on general domains. *Numer. Math.*, 134(4):667–704, 2016.

[15] E. Feireisl, T. Karper, and A. Novotný. A convergent numerical method for the Navier–Stokes–Fourier system. *IMA J. Numer. Anal.*, 36(4):1477, 2015.

[16] E. Feireisl and M. Lukáčová-Medvid'ová. Convergence of a mixed finite element finite volume scheme for the isentropic Navier–Stokes system via dissipative measure-valued solutions. *preprint available at ArXiv*, Aug. 2016.

[17] E. Feireisl, M. Lukáčová-Medvid'ová, Š. Nečasová, A. Novotný, and B. She. Error estimate for a numerical approximation to the compressible barotropic Navier-Stokes equations. Preprint, 2016.

[18] E. Feireisl, A. Novotný, and H. Petzeltová. On the existence of globally defined weak solutions to the Navier-Stokes equations. *J. Math. Fluid Mech.*, 3(4):358–392, 2001.

[19] T. Gallouët, L. Gastaldo, R. Herbin, and J.-C. Latché. An unconditionally stable pressure correction scheme for the compressible barotropic Navier-Stokes equations. *ESAIM: M2AN*, 42:303–331, 3 2008.

[20] T. Gallouët, R. Herbin, J.-C. Latché, and D. Maltese. Convergence of the MAC scheme for the compressible stationary Navier-Stokes equations. *ArXiv e-prints*, July 2016.

[21] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Implicit MAC scheme for compressible Navier-Stokes equations: unconditional error estimates. *Preprint*, 2016.

[22] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Convergence of the marker-and-cell scheme for the semi-stationary compressible Stokes problem. *Mathematics and Computers in Simulation*, 2016. available on line.

[23] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Error estimates for a numerical approximation to the compressible barotropic Navier–Stokes equations. *IMA J. Numer. Anal.*, 36(2):543–592, 2016.

[24] F. Grasso and C. Meola. Euler and Navier-Stokes equations for compressible flows: finite-volume methods. In *Handbook of computational fluid mechanics*, pages 159–282. Academic Press, San Diego, CA, 1996.

[25] J. Haack, S. Jin, and J. Liu. An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. *Commun. Comput. Phys.*, 12:955–980, 10 2012.

[26] R. Hošek. Expressing the remainder of Taylor polynomial when the function lacks smoothness. *Preprint*, 2017.

[27] T. K. Karper. A convergent FEM-DG method for the compressible Navier-Stokes equations. *Numer. Math.*, 125(3):441–510, 2013.

[28] T. K. Karper. Convergent finite differences for 1D viscous isentropic flow in Eulerian coordinates. *Discrete Contin. Dyn. Syst., Ser. S*, 7(5):993–1023, 2014.

[29] C. M. Klaij, J. J. W. van der Vegt, and H. van der Ven. Space-time discontinuous Galerkin method for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 217(2):589–611, 2006.

[30] M. Kouhi and E. Oñate. An implicit stabilized finite element method for the compressible Navier-Stokes equations using finite calculus. *Comput. Mech.*, 56(1):113–129, 2015.

[31] M. Kupiainen and B. Sjögreen. A Cartesian embedded boundary method for the compressible Navier-Stokes equations. *J. Sci. Comput.*, 41(1):94–117, 2009.

[32] P.-L. Lions. *Mathematical topics in fluid mechanics. Vol. 2: Compressible models.* Oxford: Clarendon Press, 1998.

[33] B. Liu. The analysis of a finite element method with streamline diffusion for the compressible Navier-Stokes equations. *SIAM J. Numer. Anal.*, 38(1):1–16 (electronic), 2000.

[34] S. Noelle, G. Bispen, K. R. Arun, M. Lukáčová-Medvid'ová, and C.-D. Munz. A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM J. Sci. Comput.*, 36(6):B989–B1024, 2014.

[35] J. S. Park and C. Kim. Higher-order multi-dimensional limiting strategy for discontinuous Galerkin methods in compressible inviscid and viscous flows. *Comput. & Fluids*, 96:377–396, 2014.

[36] F. Renac, S. Gérald, C. Marmignon, and F. Coquel. Fast time implicit-explicit discontinuous Galerkin method for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 251:272–291, 2013.

[37] A. Valli. An existence theorem for compressible viscous fluids. *Ann. Mat. Pura Appl. (4)*, 130:197–213, 1982.

[38] K. Xu, C. Kim, L. Martinelli, and A. Jameson. BGK-based schemes for the simulation of compressible flow. *Int. J. Comput. Fluid Dyn.*, 7(3):213–235, 1996.

Appendix B

E. Feireisl, R. H., M. Michálek: A convergent numerical method for the full Navier–Stokes–Fourier system in smooth physical domains.

SIAM J. NUMER. ANAL.
Vol. 54, No. 5, pp. 3062–3082

A CONVERGENT NUMERICAL METHOD FOR THE FULL NAVIER–STOKES–FOURIER SYSTEM IN SMOOTH PHYSICAL DOMAINS*

EDUARD FEIREISL†, RADIM HOŠEK†, AND MARTIN MICHÁLEK†

Abstract. We propose a mixed finite volume–finite element numerical method for solving the full Navier–Stokes–Fourier system describing the motion of a compressible, viscous, and heat conducting fluid. The physical domain occupied by the fluid has a smooth boundary and it is approximated by a family of polyhedral numerical domains. Convergence and stability of the numerical scheme is studied. The numerical solutions are shown to converge, up to a subsequence, to a weak solution of the problem posed on the limit domain.

Key words. Navier–Stokes–Fourier system, finite element method, finite volume method, stability, general domain

AMS subject classifications. 65M80, 65M08, 35Q30

DOI. 10.1137/15M1011809

1. Introduction.

Numerical methods based on finite element approximations use a mesh over the physical domain Ω. If the boundary of the latter is curved, meshes built up by means of *polygonal* elements can only approximate the kinematic boundary $\partial\Omega$. On the other hand, rigorous *error* estimates of the numerical methods usually require the exact solution of the problem to be smooth. Smooth solutions, however, can exist only on *regular* physical domains. It is therefore of interest to study the convergence of a numerical scheme in the situation when a family of numerical polyhedral domains Ω_h approaches, in a certain sense, the limit physical domain Ω. To avoid technicalities and since we are primarily interested in *smooth* solutions of the problem, only bounded domains with a sufficiently smooth boundary $\partial\Omega \in C^1$ will be considered although the principal results of this paper can be easily extended to less regular geometries, say $\partial\Omega$ Lipschitz.

1.1. Navier–Stokes–Fourier system.

The motion of a compressible, viscous, and heat conducting fluid in the framework of continuum mechanics is characterized by three basic macroscopic (observable) quantities: the mass density $\varrho = \varrho(t,x)$, the absolute temperature $\vartheta = \vartheta(t,x)$, and the velocity field $\mathbf{u} = \mathbf{u}(t,x)$, depending on the time $t \in (0,T)$ and the reference (Eulerian) spatial position $x \in \Omega$. The time evolution of the fluid is governed by the *Navier–Stokes–Fourier system* of equations for Newtonian fluids, see e.g. Gallavotti [9]:

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{1.1}$$

*Received by the editors March 9, 2015; accepted for publication (in revised form) July 22, 2016; published electronically October 6, 2016.

http://www.siam.org/journals/sinum/54-5/M101180.html

Funding: The work of the first and second authors leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC Grant Agreement 320078. The Institute of Mathematics of the Academy of Sciences of the Czech Republic is supported by RVO:67985840. The work of the third author was supported by the grant SVV-2016-260335.

†Mathematical Institute ASCR, 11567 Prahal, Czech Republic (feireisl@math.cas.cz, hosek@math.cas.cz, michalek@math.cas.cz).

$$\partial_t(\varrho\mathbf{u}) + \mathrm{div}_x(\varrho\mathbf{u}\otimes\mathbf{u}) + \nabla_x p(\varrho,\vartheta) = \mu\Delta\mathbf{u} + (\mu+\lambda)\nabla_x \mathrm{div}_x\mathbf{u}, \tag{1.2}$$

$$c_v\left[\partial_t(\varrho\vartheta) + \mathrm{div}_x(\varrho\vartheta\mathbf{u})\right] - \mathrm{div}_x(\kappa(\vartheta)\nabla_x\vartheta) = \mu|\nabla_x\mathbf{u}|^2 + \lambda|\mathrm{div}_x\mathbf{u}|^2 - \vartheta\frac{\partial p(\varrho,\vartheta)}{\partial\vartheta}\mathrm{div}_x\mathbf{u}, \tag{1.3}$$

with the pressure

$$p(\varrho,\vartheta) = a_1\varrho^\gamma + a_2\varrho + \varrho\vartheta,\ a_1, a_2 > 0, \tag{1.4}$$

$\mu, c_v > 0$, $\lambda \geq -\frac{2}{3}\mu$, and $\gamma > 3$. The heat conductivity $\kappa = \kappa(\vartheta)$ is continuously differentiable, satisfying

$$\underline{\kappa}(1+\vartheta^2) \leq \kappa(\vartheta) \leq \overline{\kappa}(1+\vartheta^2),\ \underline{\kappa} > 0. \tag{1.5}$$

Then we denote the primitive function $K(\vartheta) = \int_0^\vartheta \kappa(z)\,\mathrm{d}z$, i.e., $\kappa(\vartheta)\nabla_x\vartheta = \nabla_x K(\vartheta)$. For the sake of simplicity, the effect of external mechanical and heat sources is omitted in (1.2) and (1.3), respectively. The specific form of the constitutive relations is inspired by similar assumptions introduced in [6]. In particular, the problem (1.1)–(1.5), supplemented with suitable boundary conditions, admits a global-in-time weak solution for any finite energy initial data; see [6, Chapter 7, Theorem 7.1]. In the context of the existence theory developed in [6], the assumption $\gamma > 3$ is optimal.

The system of equations (1.1)–(1.3) is supplemented with the *no-slip* and *no-flux* boundary conditions

$$\mathbf{u}|_{\partial\Omega} = 0, \qquad -\kappa(\vartheta)\nabla_x\vartheta\cdot\mathbf{n}|_{\partial\Omega} = 0; \tag{1.6}$$

the initial state of the fluid is given by

$$\varrho(0,\cdot) = \varrho_0 > 0,\ \vartheta(0,\cdot) = \vartheta_0 > 0,\ \mathbf{u}(0,\cdot) = \mathbf{u}_0. \tag{1.7}$$

1.2. Numerical analysis. We propose a modification of the numerical method for the Navier–Stokes–Fourier system developed in [8] adapted to the physical domain with a smooth boundary, where the target domain Ω is approximated by a family of polyhedral (numerical) domains $\{\Omega_h\}_{h>0}$. A similar problem has been treated in [7] in the context of *barotropic fluids*, where the original numerical method of Karlsen and Karper [12], and Karper [13] has been adapted to the smooth domain setting. In contrast with [7], the presence of the heat equation (1.3), together with the Neumann-type boundary condition $(1.6)_2$, creates new difficulties addressed in the present paper.

Motivated by Karper [13], we use a mixed finite element finite volume method, where the convective terms are approximated by the standard upwind operator, while the diffusive term in the momentum equation is handled by means of the discontinuous Galerkin method based on the nonconformal finite elements of Crouzeix–Raviart-type. Accordingly, we consider an *unfitted tetrahedral mesh* generating a family of numerical domains $\{\Omega_h\}_{h>0}$ such that (see section 2.2.1 for details)

$$\Omega \subset \overline{\Omega} \subset \Omega_h \subset \mathcal{U}_h[\Omega] \equiv \left\{ x \in R^3 \;\middle|\; \mathrm{dist}[x,\Omega] < h \right\}. \tag{1.8}$$

Since the diffusion coefficient in the heat equation (1.3) is nonlinear, it seems more convenient to use the finite volume scheme for the discretization of the heat flux as well. In order to prove stability and, more importantly, *consistency* of the resulting numerical method, the underlying mesh should be shape regular in the sense of

Eymard, Gallouët, and Herbin [3] and satisfy (1.8) at the same time. Examples of tetrahedral meshes complying with this stipulation were constructed in [11]. As a byproduct of our analysis, the theory developed here probably includes a treatment of variational crimes for the convection-diffusion equation with Neumann boundary conditions for finite volumes that can be of independent interest.

The paper is organized as follows. In section 2, we introduce the concept of a *weak solution* to the Navier–Stokes–Fourier system, together with the necessary numerical framework including the basic notation and properties of the underlying function spaces. In section 3, we define the numerical method and state our main result concerning convergence towards a weak solution of the Navier–Stokes–Fourier system. Having exhausted the preliminary material, we report certain relations and estimates already obtained in [8]. Section 4 deals with numerical analogues of the renormalized version of the continuity and thermal energy balance as well as a discrete version of the total energy balance. Section 5 addresses the issue of stability of the scheme, recalling the uniform bounds necessary for the limit passage. The material in these two sections is presented without proofs, with the references to the relevant parts of [8]. Section 6 is devoted to the problem of consistency and convergence of the scheme mimicking certain steps of the existence theory developed in [6, Chapter 7]. We conclude the paper in section 7 by showing *unconditional convergence* of the scheme on condition that the numerical solutions remain bounded independently of the step parameter h.

2. Preliminaries, weak solutions, numerical framework. In this section, we collect the preliminary material concerning solvability of the Navier–Stokes–Fourier system and the apparatus of numerical analysis used in the paper.

2.1. Weak solutions. We use the concept of *weak formulation* of the problem (1.1)–(1.7) introduced in [6, Chapter 4].

DEFINITION 2.1. *A triple of functions* $[\varrho, \vartheta, \mathbf{u}]$ *is a* weak solution *to the problem* (1.1)–(1.7) *in the space-time cylinder* $(0,T)\times\Omega$ *if*

$$\varrho \in L^\infty(0,T;L^\gamma(\Omega)),\ \vartheta \in L^2(0,T;L^6(\Omega)),\ \mathbf{u} \in L^2(0,T;W_0^{1,2}(\Omega;R^3)), \tag{2.1}$$

$$\varrho\mathbf{u} \in L^\infty(0,T;L^{\frac{2\gamma}{\gamma+1}}(\Omega;R^3)),\ \varrho\vartheta \in L^\infty(0,T;L^1(\Omega)), \tag{2.2}$$

$$\varrho \geq 0,\ \vartheta > 0 \ \textit{a.a. in}\ (0,T)\times\Omega, \tag{2.3}$$

$$\int_0^T \int_\Omega \Big[\varrho\partial_t\varphi + \varrho\mathbf{u}\cdot\nabla_x\varphi\Big]\ \mathrm{d}x\ \mathrm{d}t = -\int_\Omega \varrho_0\varphi(0,\cdot)\ \mathrm{d}x \tag{2.4}$$

for any $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$*;*

$$\begin{aligned}\int_0^T \int_\Omega \Big[\varrho\mathbf{u}\cdot\partial_t\varphi + \varrho\mathbf{u}\otimes\mathbf{u} : \nabla_x\varphi + p(\varrho,\vartheta)\mathrm{div}_x\varphi\Big]\ \mathrm{d}x\ \mathrm{d}t \\ = \int_0^T \int_\Omega \Big[\mu\nabla_x\mathbf{u} : \nabla_x\varphi + \lambda\mathrm{div}_x\mathbf{u}\ \mathrm{div}_x\varphi\Big]\ \mathrm{d}x\ \mathrm{d}t \\ - \int_\Omega \varrho_0\mathbf{u}_0\cdot\varphi(0,\cdot)\ \mathrm{d}x,\ \lambda = \frac{1}{3}\mu + \eta > 0,\end{aligned} \tag{2.5}$$

for any $\varphi \in C_c^\infty([0,T)\times\Omega;R^3)$*;*

$$\begin{aligned}(2.6)\qquad &\int_0^T\int_\Omega \Big[c_v\Big(\varrho\vartheta\partial_t\varphi+\varrho\vartheta\mathbf{u}\cdot\nabla_x\varphi\Big)-\overline{K(\vartheta)}\Delta\varphi\Big]\ \mathrm{d}x\ \mathrm{d}t\\ &+\int_0^T\int_\Omega\Big[\mu|\nabla_x\mathbf{u}|^2+\lambda|\mathrm{div}_x\mathbf{u}|^2\Big]\varphi\ \mathrm{d}x\ \mathrm{d}t\\ &-\int_0^T\int_\Omega\varrho\vartheta\mathrm{div}_x\mathbf{u}\varphi\ \mathrm{d}x\ \mathrm{d}t\le\int_\Omega c_v\varrho_0\vartheta_0\varphi(0,\cdot)\ \mathrm{d}x\end{aligned}$$

for any $\varphi\in C_c^\infty([0,T)\times\overline{\Omega})$*,* $\varphi\ge 0$*,* $\nabla_x\varphi\cdot\mathbf{n}|_{\partial\Omega}=0$*, where*

$$(2.7)\qquad \varrho\overline{K(\vartheta)}=\varrho K(\vartheta);$$

the energy inequality

$$\begin{aligned}(2.8)\quad &\int_\Omega\left[\frac{1}{2}\varrho|\mathbf{u}|^2+c_v\varrho\vartheta+\frac{a}{\gamma-1}\varrho^\gamma+b\varrho\log(\varrho)\right](\tau,\cdot)\ \mathrm{d}x\\ &\le\int_\Omega\left[\frac{1}{2}\varrho_0|\mathbf{u}_0|^2+c_v\varrho_0\vartheta_0+\frac{a}{\gamma-1}\varrho_0^\gamma+b\varrho_0\log(\varrho_0)\right]\ \mathrm{d}x\ \textit{for a.a. } \tau\in(0,T).\end{aligned}$$

The *existence* of weak solutions to the Navier–Stokes–Fourier system on an arbitrary time interval $(0,T)$ was proved in [6, Chapter 7, Theorem 7.1]. The interested reader may consult [6] for a thorough discussion concerning the inequalities in (2.6), (2.8) as well as the interpretation of (2.7). Further properties of weak solutions and, in particular, the problem of weak-strong uniqueness and conditional regularity are discussed in section 7.

2.2. Mesh, finite elements. In what follows, we make systematic use of the following notation:

$$a\lesssim b \text{ if } a\le cb,\ c>0 \text{ a constant},\ a\approx b \text{ if } a\lesssim b \text{ and } b\lesssim a.$$

Here, "constant" means a generic quantity independent of the size of the mesh and the time step used in the numerical scheme.

2.2.1. Mesh. Our numerical scheme is constructed over a family of *polyhedral* domains Ω_h approximating Ω in the sense specified in (1.8). Furthermore, we suppose that each Ω_h admits a *conformal shape regular tetrahedral mesh* consisting of a set of compact elements $E\in E_h$, a set of faces $\Gamma\in\Gamma_h$, along with the associated normals $\mathbf{n}$, and a family of control points $x_E\in\mathrm{int}[E]$, enjoying the following property (cf. Eymard, Gallouët, and Herbin [3, Chapter 3]):

1. The intersection $E\cap F$ of two elements $E,F\in E_h$, $E\ne F$, is either empty or have a common face, edge, or vertex.
2. For any $E\in E_h$, $\mathrm{diam}[E]\approx h$, $r[E]\approx h$, where r denotes the radius of the largest sphere contained in E.
3. If E and F are two neighboring elements sharing a common face Γ, then the segment $[x_E,x_F]$ is perpendicular to Γ. We denote $d_\Gamma=|x_E-x_F|>0$.

Remark 2.1. If the mesh is well-centered (cf. VanderZee et al. [18]), the point x_E can be taken as the center of the circumsphere of the element E. A well-centered mesh satisfying (1.8) for a given domain Ω was constructed in [11].

Remark 2.2. Since our method is based on finite elements of first order, the expected rate of convergence should be the same even if the polygonal approximation of the physical domain is replaced by more sophisticated "curved" elements; cf. Lenoir [14].

Each face $\Gamma \in \Gamma_h$ is associated with a normal vector $\mathbf{n}$. We shall write Γ_E whenever a face $\Gamma_E \subset \partial E$ is considered as a part of the boundary of the element E. In such a case, the normal vector to Γ_E is always the *outer* normal vector with respect to E. Moreover, for any function g continuous on each element E, we set

(2.9)
$$g^{\rm out}|_\Gamma = \lim_{\delta\to 0+} g(\cdot+\delta\mathbf{n}),\ g^{\rm in}|_\Gamma = \lim_{\delta\to 0+} g(\cdot-\delta\mathbf{n}),\ [[g]]_\Gamma = g^{\rm out}-g^{\rm in}, \{g\}_\Gamma = \frac{g^{\rm out}+g^{\rm in}}{2}.$$

For $\Gamma_E \subset \partial E$ we simply write g for $g^{\rm in}$. We also omit the subscript Γ if no confusion arises. Finally, we distinguish two families of faces,

$$\Gamma_{h,\rm ext} = \left\{ \Gamma \in \Gamma_h \ \Big|\ \Gamma \subset \partial\Omega_h \right\},\ \Gamma_{h,\rm int} = \Gamma_h \setminus \Gamma_{h,\rm ext}.$$

2.2.2. Piecewise linear finite elements. We start by introducing the space of piecewise constant functions

$$Q_h(\Omega_h) = \left\{ v \in L^2(\Omega_h) \ \Big|\ v|_E = a_E \in R \text{ for any } E \in E_h \right\},$$

along with the associated projection

$$\Pi_h^Q : L^1(\Omega_h) \to Q_h(\Omega_h),\ \Pi_h^Q[v] \equiv \hat{v},\ \Pi_h^Q[v]|_E = \frac{1}{|E|}\int_E v\ {\rm d}x \text{ for any } E \in E_h.$$

From standard Poincaré's inequality we get

$$\left\| v - \Pi_h^Q[v] \right\|_{L^q(\Omega_h)} \stackrel{<}{\sim} h \|\nabla_x v\|_{L^q(\Omega_h;R^3)},\ \text{for any } v \in W^{1,q}(\Omega_h),\ 1 \le q \le \infty, \tag{2.10}$$

(2.11)
$$\left\| v - \frac{1}{|\Gamma_E|}\int_{\Gamma_E} v\ {\rm dS}_x \right\|_{L^q(E)} + h^{1/q} \left\| v - \frac{1}{|\Gamma_E|}\int_{\Gamma_E} v\ {\rm dS}_x \right\|_{L^q(\Gamma_E)} \stackrel{<}{\sim} h \|\nabla_x v\||_{L^q(E)}$$

for any $\Gamma_E \subset \partial E$ and $1 \le q < \infty$. The same estimate holds also for $q = \infty$ with obvious modifications.

In order to establish the consistency of the numerical approximation of the heat flux term in (1.3), we shall need another projection

$$\Pi_h^B : C(\overline{\Omega}_h) \to Q_h(\Omega_h),\ \Pi_v^B[v]|_E = v(x_E),\ E \in E_h.$$

Obviously,

$$\|v - \Pi_h^B[v]\|_{L^\infty(\Omega_h)} \stackrel{<}{\sim} h \|\nabla_x v\|_{L^\infty(\Omega_h;R^3)} \text{ for any Lipschitz } v. \tag{2.12}$$

Next, we introduce the *Crouzeix–Raviart finite element spaces*

(2.13)
$$V_h(\Omega_h) = \left\{ v \in L^2(\Omega_h) \ \Big|\ v|_E = \text{affine},\ E \in E_h,\ \int_\Gamma [[v]]\ {\rm dS}_x = 0 \text{ for any } \Gamma \in \Gamma_{h,\rm int} \right\},$$

$$V_{h,0}(\Omega_h) = \left\{ v \in V_h \ \Big|\ \int_\Gamma v \ \mathrm{dS}_x = 0 \text{ for any } \Gamma \in \Gamma_{h,\mathrm{ext}} \right\}, \tag{2.14}$$

and the projection

$$\Pi_h^V : W^{1,q}(\Omega_h) \to V_h(\Omega_h), \ \int_\Gamma \Pi_h^V[v] \ \mathrm{dS}_x = \int_\Gamma v \ \mathrm{dS}_x \text{ for any } \Gamma \in \Gamma_h.$$

For a differential operator D, we denote

$$D_h v|_E = D(v|_E) \text{ for any } v \text{ differentiable on each element } E \in E_h.$$

It is easy to check that

$$\int_{\Omega_h} \mathrm{div}_h \Pi_h^V[\mathbf{u}] \ w \ \mathrm{d}x = \int_{\Omega_h} \mathrm{div}_h \mathbf{u} \ w \ \mathrm{d}x \text{ for any } w \in Q_h(\Omega_h), \tag{2.15}$$

$$\int_{\Omega_h} \nabla_h v \cdot \nabla_h \Pi_h^V[\varphi] \ \mathrm{d}x = \int_{\Omega_h} \nabla_h v \cdot \nabla_x \varphi \ \mathrm{d}x \text{ if } v \in V_{h,0}(\Omega_h), \ \varphi \in W_0^{1,2}(\Omega_h); \tag{2.16}$$

see Karper [13, Lemma 2.11]. Moreover, as a direct consequence of the shape regularity of the mesh, we record the error estimates

$$\left\| v - \Pi_h^V[v] \right\|_{L^q(\Omega_h)} + h \left\| \nabla_h \left(v - \Pi_h^V[v] \right) \right\|_{L^q(\Omega_h; R^3)} \stackrel{<}{\sim} h^m \left\| \nabla^m v \right\|_{L^q(\Omega_h; R^{3^m})}, \tag{2.17}$$

$m = 1, 2, \ 1 < q < \infty$, for any $v \in W^{m,q}(\Omega_h)$; see Karper [13, Lemma 2.7].

2.2.3. Upwind. We introduce the standard *upwind* operator $\mathrm{Up}[r, \mathbf{u}]$ defined on a face Γ as

$$\mathrm{Up}[r, \mathbf{u}] = r^{\mathrm{in}} [\tilde{\mathbf{u}} \cdot \mathbf{n}]^+ + r^{\mathrm{out}} [\tilde{\mathbf{u}} \cdot \mathbf{n}]^-, \tag{2.18}$$

where we have denoted $[c]^+ = \max\{c, 0\}$, $[c]^- = \min\{c, 0\}$, $\tilde{v} = \frac{1}{|\Gamma|} \int_\Gamma v \ \mathrm{dS}_x$. Such a definition makes sense as soon as $r \in Q_h(\Omega_h)$, $\mathbf{u} \in V_h(\Omega_h; R^3)$, and $\Gamma \in \Gamma_{h,\mathrm{int}}$.

After a bit tedious but straightforward manipulation carried over in full detail in [8, section 2.4, formula (2.17)], we deduce the formula

$$\begin{aligned} \int_{\Omega_h} r\mathbf{u} \cdot \nabla_x \phi \ \mathrm{d}x = & \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}[r, \mathbf{u}] [[F]] \ \mathrm{dS}_x \\ & + \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} (F - \phi) \ [[r]] \ [\tilde{\mathbf{u}} \cdot \mathbf{n}]^- \ \mathrm{dS}_x \\ & + \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \phi r (\mathbf{u} - \tilde{\mathbf{u}}) \cdot \mathbf{n} \ \mathrm{dS}_x \\ & + \int_{\Omega_h} (F - \phi) r \mathrm{div}_h \mathbf{u} \ \mathrm{d}x \end{aligned} \tag{2.19}$$

for any $r, F \in Q_h(\Omega_h)$, $\mathbf{u} \in V_{h,0}(\Omega_h; R^3)$, $\phi \in C^1(\overline{\Omega}_h)$.

Finally, we recall Jensen's inequality in the form

$$\sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} |\tilde{v}|^q \ \mathrm{dS}_x \stackrel{<}{\sim} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} |v|^q \ \mathrm{dS}_x, \ 1 \le q < \infty, \tag{2.20}$$

for any $v \in C(\overline{E})$, $E \in E_h$, together with

$$\sum_{\Gamma \in \Gamma_h} \int_\Gamma |v - \tilde{v}|^2 \ \mathrm{dS}_x \lesssim h \int_{\Omega_h} |\nabla_h v|^2 \ \mathrm{d}x \text{ for any } v \in V_{h,0}(\Omega_h; R^3) \tag{2.21}$$

that follows directly from Poincaré's inequality (2.11).

2.2.4. $L^p - L^q$ estimates and traces. Since the mesh is shape regular, we can derive the following estimates by a scaling argument. First, we have

$$\|v\|^q_{L^q(\partial E)} \lesssim \frac{1}{h} \left(\|v\|^q_{L^q(E)} + h^q \|\nabla_x v\|^q_{L^q(E;R^3)} \right), \tag{2.22}$$

$1 \leq q < \infty$, for any $v \in C^1(E)$, whence

$$\|w\|^q_{L^q(\partial E)} \lesssim \frac{1}{h} \|w\|^q_{L^q(E)} \text{ for any } 1 \leq q < \infty,\ w \in P_m, \tag{2.23}$$

where P_m denotes the space of polynomials of order m.

Similarly, we obtain

$$\|w\|_{L^p(E)} \lesssim h^{3\left(\frac{1}{p} - \frac{1}{q}\right)} \|w\|_{L^q(E)},\ 1 \leq q < p \leq \infty,\ w \in P_m, \tag{2.24}$$

and therefore

$$\|w\|_{L^p(\Omega_h)} \leq c h^{3\left(\frac{1}{p} - \frac{1}{q}\right)} \|w\|_{L^q(\Omega_h)}, \tag{2.25}$$

$1 \leq q < p \leq \infty$, for any $w|_E \in P_m(E)$, $E \in E_h$. There is an analogue of (2.24) and (2.25) for piecewise smooth functions of the time variable $t \in (0,T)$ for the discretization of order Δt. Specifically, we derive

$$\|w\|_{L^p(0,T)} \lesssim (\Delta t)^{\left(\frac{1}{p} - \frac{1}{q}\right)} \|w\|_{L^q(0,T)},\ 1 \leq q < p \leq \infty \tag{2.26}$$

for any w that is constant on any time segment $[j\Delta t, (j+1)\Delta t]$ contained in $[0,T]$.

2.2.5. Discrete Sobolev spaces. For $v \in Q_h(\Omega_h)$, let

$$\|v\|^2_{H^1_{Q_h}(\Omega_h)} = \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{[[v]]^2}{h} \ \mathrm{dS}_x$$

be a discrete analogue of the Sobolev gradient seminorm. Similarly, we introduce

$$\|v\|^2_{H^1_{V_h}(\Omega_h)} = \int_\Omega \left(|\nabla_h v|^2 \right) \ \mathrm{d}x \text{ for } v \in V_h(\Omega_h).$$

Recall that

$$\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma [[v]]^2 \ \mathrm{dS}_x \lesssim h \|v\|^2_{H^1_{V_h}(\Omega_h)} \text{ for any } v \in V_h(\Omega_h); \tag{2.27}$$

see Gallouët, Herbin, and Latché [10, Lemma 2.2].

We report a discrete analogue of the standard Sobolev embedding relations:

$$\|v\|_{L^6(\Omega_h)} \lesssim \left(\|v\|_{H^1_{Q_h}(\Omega_h)} + \|v\|_{L^2(\Omega_h)} \right),\ v \in Q_h(\Omega_h) \tag{2.28}$$

(see Chenais-Hillairet and Droniou [1, Lemma 6.1]), and

$$\|v\|_{L^6(\Omega_h)} \stackrel{<}{\sim} \|v\|_{H^1_{V_h}(\Omega_h)},\ v \in V_{h,0}(\Omega_h) \tag{2.29}$$

(see Gallouet, Herbin, and Latché [10, Lemma 3.2]).

Finally, let $[v]_\delta = v * \omega_\delta$ denote the spatial regularization by a convolution with a family of smooth kernels, specifically $\omega \in C_c^\infty(\{x \in R^3 \mid |x| < 1\})$ satisfying

$$\omega_\delta(y) = \frac{1}{\delta^3}\omega\left(\frac{y}{\delta}\right),\ \omega \geq 0,\ \omega(y) = \omega(|y|),\ \int_{R^3} \omega(y)\ \mathrm{d}y = 1.$$

We have

$$\int_{\{x\in\Omega_h \ \mid\ \mathrm{dist}[x,\partial\Omega_h]>\delta\}} |\nabla_x [v]_\delta|^2\ \mathrm{d}x \stackrel{<}{\sim} \frac{h}{\delta}\|v\|^2_{H^1_{Q_h}(\Omega_h)} \text{ for any } v \in Q_h(\Omega_h) \tag{2.30}$$

and

$$\int_{\Omega_h} |\nabla_x [v]_\delta|^2\ \mathrm{d}x \stackrel{<}{\sim} \frac{h}{\delta}\|v\|^2_{H^1_{V_h}(\Omega_h)} \text{ for any } v \in V_{h,0}(\Omega_h) \tag{2.31}$$

provided $0 < \delta \leq h$ (see Christiansen, Munthe-Kaas, and Owren [2, Proposition 5.67]). Note that the functions from $V_{h,0}$ can be extended to be zero outside Ω_h so that the regularization is well-defined.

3. Numerical scheme, main result. The numerical scheme is formally the same as in [8], the only difference is that the numerical domains Ω_h depend on the discretization step h. For this reason, it is convenient for the initial data ϱ_0, ϑ_0, $\mathbf{u}_0$ to be defined on the whole space R^3, $\mathbf{u}_0$ vanishing outside Ω.

We set

$$\varrho_h^0 = \Pi_h^Q[\varrho_0] \in Q_h(\Omega_h),\ \vartheta_h^0 = \Pi_h^Q[\vartheta_0] \in Q_h(\Omega_h),\ \mathbf{u}_h^0 = \Pi_h^V[\mathbf{u}_0] \in V_{h,0}(\Omega_h; R^3). \tag{3.1}$$

We fix the time step $\Delta t \approx h$ and introduce the discrete time derivative

$$D_t b_h^k = \frac{b_h^k - b_h^{k-1}}{\Delta t}.$$

The numerical solutions $[\varrho_h^k, \vartheta_h^k, \mathbf{u}_h^k]_{h>0}$, $k = 1, 2, \ldots,$

$$\varrho_h^k,\ \vartheta_h^k \in Q_h(\Omega_h),\ \mathbf{u}_h^k \in V_{h,0}(\Omega_h; R^3)$$

are defined successively by means of the numerical method:

$$\int_{\Omega_h} D_t \varrho_h^k \phi\ \mathrm{d}x - \sum_{\Gamma\in\Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}[\varrho_h^k, \mathbf{u}_h^k]\ [[\phi]]\ \mathrm{dS}_x + h^\alpha \sum_{\Gamma\in\Gamma_{h,\mathrm{int}}} \int_\Gamma [[\varrho_h^k]][[\phi]]\ \mathrm{dS}_x = 0 \tag{3.2}$$

for all $\phi \in Q_h(\Omega_h)$, with a parameter $0 < \alpha < 1$;

$$
\begin{aligned}
(3.3)\qquad \int_{\Omega_h} D_t(\varrho_h^k \widehat{\mathbf{u}}_h^k) \cdot \phi \,\mathrm{d}x - \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}[\varrho_h^k \widehat{\mathbf{u}}_h^k, \mathbf{u}_h^k] \cdot [[\widehat{\phi}]] \,\mathrm{dS}_x \\
+ \int_{\Omega_h} \left[\mu \nabla_h \mathbf{u}_h^k : \nabla_h \phi + \lambda \mathrm{div}_h \mathbf{u}_h^k \mathrm{div}_h \phi \right] \,\mathrm{d}x \\
- \int_{\Omega_h} p(\varrho_h^k, \vartheta_h^k) \mathrm{div}_h \phi \,\mathrm{d}x \\
+ h^\alpha \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma [[\varrho_h^k]] \left\{ \widehat{u_h^k} \right\} \cdot [[\widehat{\phi}]] \,\mathrm{dS}_x = 0
\end{aligned}
$$

for any $\phi \in V_{h,0}(\Omega_h; R^3)$;

$$
\begin{aligned}
(3.4)\qquad c_v \int_{\Omega_h} D_t(\varrho_h^k \vartheta_h^k) \phi \,\mathrm{d}x - c_v \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}[\varrho_h^k \vartheta_h^k, \mathbf{u}_h^k] \, [[\phi]] \,\mathrm{dS}_x \\
+ \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} [[K(\vartheta_h^k)]] \, [[\phi]] \,\mathrm{dS}_x \\
= \int_{\Omega_h} \left[\mu |\nabla_h \mathbf{u}_h^k|^2 + \lambda |\mathrm{div}_h \mathbf{u}_h^k|^2 \right] \phi \,\mathrm{d}x - \int_{\Omega_h} \varrho_h^k \vartheta_h^k \mathrm{div}_h \mathbf{u}_h^k \phi \,\mathrm{d}x
\end{aligned}
$$

for any $\phi \in Q_h(\Omega_h)$.

Remark 3.1. The terms proportional to h^α are needed for technical reasons explained in detail in [8, section 7.3]. They represent numerical counterparts of the artificial viscosity regularization used in [6, Chapter 7] and were introduced by Eymard et al. [4] to prove convergence of the momentum scheme (3.3).

Before stating our main result, it is convenient to extend the numerical solution to be defined for *any* $t \in R$. To this end, we set

$$
\varrho_h(t, \cdot) = \varrho_h^0, \ \vartheta_h(t, \cdot) = \vartheta_h^0, \ \mathbf{u}_h(t, \cdot) = \mathbf{u}_h^0 \text{ for } t \leq 0,
$$

$$
\varrho_h(t, \cdot) = \varrho_h^k, \ \vartheta_h(t, \cdot) = \vartheta_h^k, \ \mathbf{u}_h(t, \cdot) = \mathbf{u}_h^k \text{ for } t \in [k\Delta t, (k+1)\Delta t), \ k = 1, 2, \ldots,
$$

and, accordingly, the discrete time derivative of a quantity v_h is

$$
D_t v_h(t, \cdot) = \frac{v_h(t) - v_h(t - \Delta t)}{\Delta t}, \ t > 0.
$$

The main result of the present paper reads as follows.

THEOREM 3.1. *Let $\Omega \subset R^3$ be a bounded domain of class C^1 approximated by a family of polyhedral domains $\{\Omega_h\}_{h>0}$ in the sense specified in* (1.8)*, where each Ω_h admits a tetrahedral mesh satisfying the hypotheses introduced in section* 2.2.1*. Suppose that $\mu > 0$, $\lambda = \mu/3 + \eta > 0$, and that the pressure $p = p(\varrho, \vartheta)$ and the heat conductivity coefficient $\kappa = \kappa(\vartheta)$ comply with* (1.4), (1.5)*. Let $[\varrho_h, \vartheta_h, \mathbf{u}_h]_{h>0}$ be a family of numerical solutions resulting from the scheme* (3.1)–(3.4) *with*

$$
\Delta t \approx h
$$

such that $\varrho_h > 0$, $\vartheta_h > 0$ for all $h > 0$.

Then, at least for a suitable subsequence,

$$\varrho_h \to \varrho \text{ weakly-(*) in } L^\infty(0,T;L^\gamma(\Omega)) \text{ and strongly in } L^1((0,T)\times\Omega),$$

$$\vartheta_h \to \vartheta \text{ weakly in } L^2(0,T;L^6(\Omega)),$$

$\mathbf{u}_h \to \mathbf{u}$ *weakly in* $L^2(0,T;L^6(\Omega;R^3))$, $\nabla_h \mathbf{u}_h \to \nabla_x \mathbf{u}$ *weakly in* $L^2((0,T)\times\Omega;R^{3\times 3})$, *where* $[\varrho,\vartheta,\mathbf{u}]$ *is a weak solution of the Navier–Stokes–Fourier system* (1.1)–(1.7) *in* $(0,T)\times\Omega$ *in the sense of Definition* 2.1.

The existence of the numerical solutions $[\varrho_h,\vartheta_h,\mathbf{u}_h]$ was shown in [8, section 8.1]. The rest of the paper is basically devoted to the proof of Theorem 3.1. As some steps are essentially the same as in [8] we omit technicalities and focus only on the necessary modifications to accommodate the variable numerical domains.

4. Renormalization. The proof of convergence of the numerical method (3.1)–(3.4) mimics the principal steps of the existence theory developed in [6] based, among other things, on suitable *renormalization* of both the equation of continuity (1.1) and the heat equation (1.3). At the level of numerical solutions, we can deduce the following (see [8, sections 4.1, 4.2]):

1. *Renormalized continuity scheme.*

$$\int_{\Omega_h} D_t b(\varrho_h^k)\phi \,\mathrm{d}x - \sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma \mathrm{Up}[b(\varrho_h^k),\mathbf{u}_h^k]\,[[\phi]]\,\mathrm{dS}_x \tag{4.1}$$

$$+\int_{\Omega_h}\phi\,\left(b'(\varrho_h^k)\varrho_h^k - b(\varrho_h^k)\right)\mathrm{div}_h\mathbf{u}_h^k\,\mathrm{d}x = -\int_{\Omega_h}\frac{\Delta t}{2}b''(\xi_{\varrho,h}^k)\left(\frac{\varrho_h^k-\varrho_h^{k-1}}{\Delta t}\right)^2\phi\,\mathrm{d}x$$

$$-h^\alpha\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\phi\,b''(\overline{\eta}_{\varrho,h}^k)[[\varrho_h^k]]^2\,\mathrm{dS}_x - \frac{1}{2}\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\phi\,b''(\eta_{\varrho,h}^k)[[\varrho_h^k]]^2|\tilde{\mathbf{u}}_h^k\cdot\mathbf{n}|\,\mathrm{dS}_x$$

for any $\phi\in Q_h(\Omega_h)$, $b\in C^2(0,\infty)$, where $\xi_{\varrho,h}^k\in\mathrm{co}\{\varrho_h^{k-1},\varrho_h^k\}$ on each element $E\in E_h$ and $\eta_{\varrho,h}^k,\overline{\eta}_{\varrho,h}^k\in\mathrm{co}\{\varrho_h^k,(\varrho_h^k)^{\mathrm{out}}\}$ on each face $\Gamma\in\Gamma_{h,\mathrm{int}}$, where $\mathrm{co}\{A,B\}=[\inf\{A,B\},\sup\{A,B\}]$.

2. *Renormalized thermal energy scheme.*

$$c_v\int_{\Omega_h} D_t\left(\varrho_h^k\chi(\vartheta_h^k)\right)\phi\,\mathrm{d}x - c_v\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\mathrm{Up}(\varrho_h^k\chi(\vartheta_h^k),\mathbf{u}_h^k)\,[[\phi]]\,\mathrm{dS}_x \tag{4.2}$$

$$+\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\frac{1}{d_\Gamma}[[K(\vartheta_h^k)]]\,[[\chi'(\vartheta_h^k)\phi]]\,\mathrm{dS}_x$$

$$=\int_{\Omega_h}\left(\mu|\nabla_h\mathbf{u}_h^k|^2+\lambda|\mathrm{div}_h\mathbf{u}_h^k|^2\right)\chi'(\vartheta_h^k)\phi\,\mathrm{d}x - \int_{\Omega_h}\chi'(\vartheta_h^k)\varrho_h^k\vartheta_h^k\mathrm{div}_h\mathbf{u}_h^k\phi\,\mathrm{d}x$$

$$-c_v\frac{\Delta t}{2}\int_{\Omega_h}\chi''(\xi_{\vartheta,h}^k)\varrho_h^{k-1}\left(\frac{\vartheta_h^k-\vartheta_h^{k-1}}{\Delta t}\right)^2\phi\,\mathrm{d}x$$

$$+\frac{c_v}{2}\sum_{E\in E_h}\sum_{\Gamma_E\subset\partial E}\int_{\Gamma_E}\phi\chi''(\eta^k_{\vartheta,h})[[\vartheta^k_h]]^2(\varrho^k_h)^{\rm out}[\tilde{\mathbf{u}}^k_h\cdot\mathbf{n}]^-\ {\rm dS}_x$$

$$-h^\alpha c_v\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma[[\varrho^k_h]]\ [[\left(\chi(\vartheta^k_h)-\chi'(\vartheta^k_h)\vartheta^k_k\right)\phi]]\ {\rm dS}_x$$

for any $\phi\in Q_h(\Omega_h)$, $\chi\in C^2(0,\infty)$, with $\xi^k_{\vartheta,h}\in{\rm co}\{\vartheta^{k-1}_h,\vartheta^k_h\}$ and $\eta^k_{\vartheta,h}\in{\rm co}\{\vartheta^k_h,(\vartheta^k_h)^{\rm out}\}$.

Finally, exactly as in [8, section 4.3] we may use (4.1), (4.2), and the momentum scheme (3.3) to deduce the following:

- *Total energy balance.*

$$\begin{aligned}(4.3)\qquad &D_t\int_{\Omega_h}\left[\frac{1}{2}\varrho^k_h|\widehat{\mathbf{u}}^k_h|^2+c_v\varrho^k_h\vartheta^k_h+\frac{a}{\gamma-1}\left(\varrho^k_h\right)^\gamma+b\varrho^k_h\log(\varrho^k_h)\right]\ {\rm d}x\\&+\frac{\Delta t}{2}\int_{\Omega_h}\left(A\left|\frac{\varrho^k_h-\varrho^{k-1}_h}{\Delta t}\right|^2+\varrho^{k-1}_h\left|\frac{\widehat{\mathbf{u}}^k_h-\widehat{\mathbf{u}}^{k-1}_h}{\Delta t}\right|^2\right)\ {\rm d}x\\&-\sum_{E\in E_h}\sum_{\Gamma_E\subset\partial E}\int_{\Gamma_E}(\varrho^k_h)^{\rm out}[\tilde{\mathbf{u}}^k_h\cdot\mathbf{n}]^-\frac{\left|\widehat{\mathbf{u}}^k_h-\left(\widehat{\mathbf{u}}^k_h\right)^{\rm out}\right|^2}{2}\ {\rm dS}_x\\&+\frac{A}{2}\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma\left(h^\alpha+|\tilde{\mathbf{u}}^k_h\cdot\mathbf{n}|\right)[[\varrho^k_h]]^2\ {\rm dS}_x\le 0\\&\text{with } A=\min_{\varrho>0}\left\{a\gamma\varrho^{\gamma-2}+\frac{b}{\varrho}\right\}>0.\end{aligned}$$

5. Stability. Similarly to [8, section 5] we derive uniform bounds on the family of numerical solutions independent of the step h.

5.1. Mass conservation and energy bounds. Taking $\phi\equiv 1$ in the continuity scheme (3.2) we obtain

$$(5.1)\qquad \int_{\Omega_h}\varrho_h(t,\cdot)\ {\rm d}x=\int_{\Omega_h}\varrho^0_h\ {\rm d}x\approx\int_\Omega\varrho_0\ {\rm d}x\ \text{for any } h>0,$$

meaning the total mass is conserved by the scheme.

The total energy balance (4.3) gives rise to

$$\begin{aligned}(5.2)\qquad &\int_{\Omega_h}\left[\frac{1}{2}\varrho_h|\widehat{\mathbf{u}}_h|^2+c_v\varrho_h\vartheta_h+\frac{a}{\gamma-1}\left(\varrho_h\right)^\gamma+b\varrho_h\log(\varrho_h)\right](\tau,\cdot)\ {\rm d}x\\&\le\int_{\Omega_h}\left[\frac{1}{2}\varrho_h|\widehat{\mathbf{u}}_h|^2+c_v\varrho_h\vartheta_h+\frac{a}{\gamma-1}\left(\varrho_h\right)^\gamma+b\varrho_h\log(\varrho_h)\right](\tau,\cdot)\ {\rm d}x\\&\le\int_{\Omega_h}\left[\frac{1}{2}\varrho^0_h|\widehat{\mathbf{u}}^0_h|^2+c_v\varrho^0_h\vartheta^0_h+\frac{a}{\gamma-1}\left(\varrho^0_h\right)^\gamma+b\varrho^0_h\log(\varrho^0_h)\right]\ {\rm d}x\equiv E_{0,h},\ E_{0,h}\lesssim 1.\end{aligned}$$

In particular, we deduce the uniform bounds, independently of $h\to 0$:

$$(5.3)\qquad \sup_{\tau\in(0,T)}\|\sqrt{\varrho_h}\widehat{\mathbf{u}}_h(\tau,\cdot)\|_{L^2(\Omega_h)}\lesssim 1,$$

$$\sup_{\tau\in(0,T)}\|\varrho_h\vartheta_h(\tau,\cdot)\|_{L^1(\Omega_h)} \lesssim 1, \tag{5.4}$$

$$\sup_{\tau\in(0,T)}\|\varrho_h[\log\vartheta_h]^+(\tau,\cdot)\|_{L^1(\Omega_h)} \lesssim 1, \tag{5.5}$$

$$\sup_{\tau\in(0,T)}\|\varrho_h(\tau,\cdot)\|_{L^\gamma(\Omega_h)} \lesssim 1. \tag{5.6}$$

We also record the bounds on the numerical dissipation:

$$\sum_{k\geq 0}\int_{\Omega_h}\left[\left|\varrho_h^k-\varrho_h^{k-1}\right|^2+\varrho_h^{k-1}\left|\widehat{\mathbf{u}}_h^k-\widehat{\mathbf{u}}_h^{k-1}\right|^2\right]\,\mathrm{d}x \lesssim 1, \tag{5.7}$$

$$-\sum_{E\in E_h}\sum_{\Gamma_E\subset\partial E}\int_0^T\int_{\Gamma_E}(\varrho_h)^{\mathrm{out}}[\tilde{\mathbf{u}}_h\cdot\mathbf{n}]^-\left|\widehat{\mathbf{u}}_h-(\widehat{\mathbf{u}}_h)^{\mathrm{out}}\right|^2\,\mathrm{dS}_x\,\mathrm{d}t \lesssim 1, \tag{5.8}$$

and

$$\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_0^T\int_\Gamma\left(|\tilde{\mathbf{u}}_h\cdot\mathbf{n}|+h^\alpha\right)[[\varrho_h]]^2\,\mathrm{dS}_x\,\mathrm{d}t \lesssim 1. \tag{5.9}$$

5.2. Entropy bounds. The bounds resulting from the dissipation mechanism encoded in (3.3), (3.4) are obtained by taking $\chi=\log$, $\phi=1$ in the renormalized thermal energy scheme (4.2). Using the fact that

$$\int_{\Omega_h}\varrho_h^k\mathrm{div}_h\mathbf{u}_h^k\,\mathrm{d}x \leq -\int_{\Omega_h}D_t\Big(\varrho_h^k\log(\varrho_h^k)\Big)\,\mathrm{d}x \tag{5.10}$$

(cf. (4.1)), we arrive at

$$\begin{aligned}
&c_v\int_{\Omega_h}D_t\left(\varrho_h^k\log(\vartheta_h^k)\right)\,\mathrm{d}x-\int_{\Omega_h}D_t\left(\varrho_h^k\log(\varrho_h^k)\right)\,\mathrm{d}x \geq \\
&-\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\frac{1}{d_\Gamma}[[K(\vartheta_h^k)]]\,[[(\vartheta_h^k)^{-1}]]\,\mathrm{dS}_x+\int_{\Omega_h}\left(\mu|\nabla_h\mathbf{u}_h^k|^2+\lambda|\mathrm{div}_h\mathbf{u}_h^k|^2\right)\frac{1}{\vartheta_h^k}\,\mathrm{d}x \\
&+\frac{\Delta t}{2}c_v\int_{\Omega_h}(\xi_{\vartheta,h}^k)^{-2}\varrho_h^{k-1}\left(\frac{\vartheta_h^k-\vartheta_h^{k-1}}{\Delta t}\right)^2\,\mathrm{d}x \\
&-\frac{1}{2}c_v\sum_{E\in E_h}\sum_{\Gamma_E\subset\partial E}\int_{\Gamma_E}(\eta_{\vartheta,h}^k)^{-2}\left(\vartheta_h^k-\left(\vartheta_h^k\right)^{\mathrm{out}}\right)^2\left(\varrho_h^k\right)^{\mathrm{out}}[\tilde{\mathbf{u}}_h^k\cdot\mathbf{n}]^-\,\mathrm{dS}_x \\
&-h^\alpha c_v\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma[[\varrho_h^k]]\,[[\log(\vartheta_h^k)]]\,\mathrm{dS}_x,
\end{aligned} \tag{5.11}$$

where the parameters appearing in the numerical dissipation are the same as in (4.1), (4.2).

Now, exactly as in [8, section 5], inequality (5.11), together with the bounds already established, gives rise to the following estimates:

$$\sup_{\tau\in(0,T)}\|\varrho_h\log(\vartheta_h)(\tau,\cdot)\|_{L^1(\Omega_h)} \lesssim 1, \tag{5.12}$$

$$\int_0^T \int_{\Omega_h} \frac{1}{\vartheta_h} |\nabla_h \mathbf{u}_h|^2 \,\mathrm{d}x \,\mathrm{d}t \lesssim 1, \tag{5.13}$$

$$\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_0^T \int_\Gamma \frac{[[\vartheta_h^\beta]]^2}{d_\Gamma} \mathrm{dS}_x \,\mathrm{d}t \lesssim 1, \quad \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_0^T \int_\Gamma \frac{[[\log(\vartheta_h)]]^2}{d_\Gamma} \mathrm{dS}_x \,\mathrm{d}t \lesssim 1, \tag{5.14}$$

where $0 \leq \beta \leq 1$ and

$$\|\vartheta_h\|_{L^2(0,T;L^6(\Omega_h))} + \|\log(\vartheta_h)\|_{L^2(0,T;L^6(\Omega_h))} \lesssim 1. \tag{5.15}$$

We have also bounds on the numerical dissipation:

$$\sum_{k \geq 0} \int_{\Omega_h} (\xi_{\vartheta,h}^k)^{-2} \varrho_h^{k-1} \left(\vartheta_h^k - \vartheta_h^{k-1}\right)^2 \,\mathrm{d}x \lesssim 1, \ \xi_{\vartheta,h}^k \in \mathrm{co}\{\vartheta_h^{k-1}, \vartheta_h^k\}, \tag{5.16}$$

(5.17)

$$-\sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_0^T \int_{\Gamma_E} (\eta_{\vartheta,h})^{-2} [[\vartheta_h]]^2 \left(\varrho_h\right)^{\mathrm{out}} [\tilde{\mathbf{u}}_h \cdot \mathbf{n}]^- \,\mathrm{dS}_x \,\mathrm{d}t \lesssim 1, \ \eta_{\vartheta,h} \in \mathrm{co}\{\vartheta_h, \vartheta_h^{\mathrm{out}}\}.$$

5.3. Temperature estimates. Revisiting the thermal energy balance (4.2) for $\chi(\vartheta_h^k) = (\vartheta_h^k)^\beta$, $0 < \beta < 1$, and with the test function $\phi = 1$, we obtain

$$\begin{aligned} &-\beta \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} [[K(\vartheta_h)]] \, [[(\vartheta_h)^{\beta-1}]] \,\mathrm{dS}_x + \beta\mu \int_{\Omega_h} \vartheta_h^{\beta-1} |\nabla_h \mathbf{u}_h|^2 \,\mathrm{d}x \,\mathrm{d}t \\ &\quad + c_v \beta(1-\beta) \frac{\Delta t}{2} \sum_{k=1} \int_{\Omega_h} (\xi_{\vartheta,h}^k)^{\beta-2} \varrho_h^{k-1} \left(\frac{\vartheta_h^k - \vartheta_h^{k-1}}{\Delta t} \right)^2 \,\mathrm{d}x \\ &\quad + \frac{c_v}{2} \beta(1-\beta) \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_\Gamma (\eta_{\vartheta,h}^k)^{\beta-2} [[\vartheta_h^k]]^2 \left(\varrho_h^k\right)^{\mathrm{out}} [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \,\mathrm{dS}_x \\ &\lesssim c_v \int_{\Omega_h} D_t(\varrho_h^k (\vartheta_h^k)^\beta) \,\mathrm{d}x + \beta \int_{\Omega_h} \varrho_h^k (\vartheta_h^k)^\beta \mathrm{div}_h \mathbf{u}_h^k \,\mathrm{d}x + h^\alpha c_v (1-\beta) \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma [[\varrho_h^k]] \, [[(\vartheta_h^k)^\beta]] \,\mathrm{dS}_x. \end{aligned} \tag{5.18}$$

Arguing as in [8, section 5.3] we deduce from (5.18) the following estimates:

$$-\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \int_0^T \frac{1}{d_\Gamma} [[K(\vartheta_h)]] \, [[(\vartheta_h)^{\beta-1}]] \,\mathrm{dS}_x \lesssim 1 \text{ for all } 0 < \beta < 1, \tag{5.19}$$

$$\sum_{\Gamma \in \Gamma_h} \int_0^T \int_\Gamma \frac{\left[\left[\vartheta_h^{1+\frac{\beta}{2}}\right]\right]^2}{h} \,\mathrm{dS}_x \lesssim 1 \text{ for all } 0 \leq \beta < 1; \tag{5.20}$$

whence, in accordance with (2.28),

$$\|\vartheta_h\|_{L^p(0,T;L^q(\Omega_h))} \lesssim 1 \text{ for any } 1 \leq p < 3, \ 1 \leq q < 9. \tag{5.21}$$

Finally, returning to the thermal energy scheme (3.4) with $\phi = 1$, we may use the previous estimates to conclude

$$\int_0^T \int_{\Omega_h} |\nabla_h \mathbf{u}_h|^2 \,\mathrm{d}x \,\mathrm{d}t \lesssim 1, \tag{5.22}$$

and, in accordance with (2.29),

$$\|\mathbf{u}_h\|^2_{L^2(0,T;L^6(\Omega_h;R^3))} \lesssim 1. \tag{5.23}$$

6. Consistency and convergence. Our goal is to check that (i) the numerical method is *consistent* with the original weak formulation and (ii) the numerical solutions converge, modulo a suitable subsequence, to a weak solution of the problem as stated in Theorem 3.1.

6.1. Consistency. To begin, we claim that the proofs of consistency for the continuity scheme (3.2) and the momentum scheme (3.3) are exactly the same as in [8, sections 6.1, 6.2], where the upwind terms may be handled by means of formula (2.19).

6.1.1. Continuity and momentum scheme. Taking $\Pi_h^Q[\phi]$, $\phi \in C_c^\infty(R^3)$, as a test function in the continuity scheme (3.2) gives rise to

$$\int_{R^3} [D_t \varrho_h - \varrho_h \mathbf{u}_h \cdot \nabla_x \phi] \ \mathrm{d}x = \int_{R^3} R_h^1(t,\cdot) \cdot \nabla_x \phi \ \mathrm{d}x \tag{6.1}$$

for any $\phi \in C_c^\infty(R^3)$ provided ϱ_h, $\mathbf{u}_h$ were extended to be zero outside Ω_h. The remainder satisfies (see [7, section 6.1])

$$\left\| R_h^1 \right\|_{L^2(0,T;L^{\frac{6\gamma}{5\gamma-6}}(R^3;R^3))(R^3;R^3))} \stackrel{<}{\sim} h^\beta \text{ for some } \beta > 0. \tag{6.2}$$

The choice $\Pi_h^V[\phi]$, $\phi \in C_c^\infty(\Omega; R^3)$, as a test function in the momentum balance (3.3) gives rise to

$$\int_\Omega D_t(\varrho_h \widehat{\mathbf{u}}_h) \cdot \phi \ \mathrm{d}x - \int_\Omega (\varrho_h \widehat{\mathbf{u}}_h \otimes \mathbf{u}_h) : \nabla_x \phi \ \mathrm{d}x \tag{6.3}$$

$$+ \int_\Omega [\mu \nabla_h \mathbf{u}_h : \nabla_x \phi + \lambda \mathrm{div}_h \mathbf{u}_h \mathrm{div}_x \phi] \ \mathrm{d}x - \int_\Omega p(\varrho_h, \vartheta_h) \mathrm{div}_x \phi \ \mathrm{d}x = \int_\Omega \mathbb{R}_h^2 : \nabla_x \phi \ \mathrm{d}x$$

for any $\phi \in C_c^\infty(\Omega; R^3)$, where the remainder satisfies (see [7, section 6.2])

$$\left\| \mathbb{R}_h^2 \right\|_{L^1(0,T;L^{\frac{\gamma}{\gamma-1}}(\Omega;R^{3\times 3}))} \stackrel{<}{\sim} h^\beta \text{ for some } \beta > 0. \tag{6.4}$$

Since $\Omega \subset \Omega_h$ for any h and ϕ has compact support in Ω, all terms in (6.3) are well-defined.

6.1.2. Consistency for the thermal energy balance. Instead of working directly with the thermal energy scheme (3.4), we consider its renormalized variant (4.2). Motivated by [8, section 6.3], we take the nonlinearities χ belonging to the class

$$\chi \in W^{2,\infty}[0,\infty), \ \chi'(\vartheta) \geq 0, \ \chi''(\vartheta) \leq 0, \quad \chi(\vartheta) = \text{const for all } \vartheta > \vartheta_\chi. \tag{6.5}$$

We start by rewriting

$$\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} [[K(\vartheta_h^k)]] \ [[\chi'(\vartheta_h^k)\phi]] \ \mathrm{d}S_x \tag{6.6}$$

$$= \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} \{\phi\} \ [[K(\vartheta_h^k)]] \ [[\chi'(\vartheta_h^k)]] \mathrm{d}S_x + \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} \left\{\chi'\left(\vartheta_h^k\right)\right\} [[K(\vartheta_h^k)]] \ [[\phi]] \mathrm{d}S_x$$

for any $\phi \in Q_h(\Omega_h)$.

Next, take $\phi \in C^2(R^3)$ such that $\nabla_x \phi \cdot \mathbf{n} = 0$ on $\partial\Omega$, and use $\Pi_h^B[\phi]$ as a test function in the renormalized thermal energy scheme (4.2). In view of (6.6), we obtain

(6.7)
$$
\begin{aligned}
&c_v \int_{\Omega_h} D_t \left(\varrho_h^k \chi(\vartheta_h^k) \right) \Pi_h^B[\phi] \, \mathrm{d}x - c_v \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}(\varrho_h^k \chi(\vartheta_h^k), \mathbf{u}_h^k) \, [[\Pi_h^B[\phi] \,]] \, \mathrm{dS}_x \\
&\quad + \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} \left\{ \chi' \left(\vartheta_h^k \right) \right\} [[K(\vartheta_h^k)]] \, [[\Pi_h^B[\phi] \,]] \mathrm{dS}_x \\
&= \int_{\Omega_h} \left(\mu |\nabla_h \mathbf{u}_h^k|^2 + \lambda |\mathrm{div}_h \mathbf{u}_h^k|^2 \right) \chi'(\vartheta_h^k) \Pi_h^B[\phi] \, \mathrm{d}x - \int_{\Omega_h} \chi'(\vartheta_h^k) \vartheta_h^k \varrho_h^k \mathrm{div}_h \mathbf{u}_h^k \Pi_h^B[\phi] \, \mathrm{d}x \\
&\quad - h^\alpha c_v \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma [[\varrho_h^k]] \, [[\left(\chi(\vartheta_h^k) - \chi'(\vartheta_h^k) \vartheta_k^k \right) \Pi_h^B[\phi] \,]] \, \mathrm{dS}_x + \langle D_h, \phi \rangle \,,
\end{aligned}
$$

where

$$
\begin{aligned}
\langle D_h(t), \phi \rangle = &- \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{1}{d_\Gamma} \left\{ \Pi_h^B[\phi] \right\} [[K(\vartheta_h^k)]] \, [[\chi'(\vartheta_h^k)]] \mathrm{dS}_x \\
&- c_v \frac{\Delta t}{2} \int_{\Omega_h} \chi''(\xi_{\vartheta,h}^k) \varrho_h^{k-1} \left(\frac{\vartheta_h^k - \vartheta_h^{k-1}}{\Delta t} \right)^2 \Pi_h^B[\phi] \, \mathrm{d}x \\
&+ \frac{c_v}{2} \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \Pi_h^B[\phi] \chi''(\eta_{\vartheta,h}^k) [[\vartheta_h^k]]^2 \left(\varrho_h^k \right)^{\mathrm{out}} [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \, \mathrm{dS}_x .
\end{aligned}
$$

As χ satisfies (6.5), it is easy to check that $\langle D_h(t), \phi \rangle \geq 0$ whenever $\phi \geq 0$. Moreover, applying (6.7) with $\phi = 1$ we get

$$
\begin{aligned}
0 \leq \langle D_h(t), 1 \rangle \leq &h^\alpha c_v \sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma [[\varrho_h^k]] \, [[\left(\chi(\vartheta_h^k) - \chi'(\vartheta_h^k) \vartheta_h^k \right)]] \, \mathrm{dS}_x \\
&+ \int_{\Omega_h} \chi'(\vartheta_h^k) \vartheta_h^k \varrho_h^k \mathrm{div}_h \mathbf{u}_h^k \, \mathrm{d}x + c_v \int_{\Omega_h} D_t \left(\varrho_h^k \chi(\vartheta_h^k) \right) \, \mathrm{d}x,
\end{aligned}
$$

where the three integrals on the right-hand side are controlled by the estimates (5.4), (5.6), (5.9), (5.14), and (5.22). We may therefore conclude that

$$
0 \leq \langle D_h(t), \phi \rangle \stackrel{<}{\sim} R_h^3(t) \|\phi\|_{L^\infty(\Omega_h)}, \ \|R_h^3\|_{L^1(0,T)} \stackrel{<}{\sim} 1 \text{ whenever } \phi \geq 0. \tag{6.8}
$$

Note that (6.8) as well as other estimates derived in this section depend on the structural properties of the function χ postulated in (6.5).

Now, the discrete time derivative can be written as

$$
\int_{\Omega_h} D_t \left(\varrho_h^k \chi(\vartheta_h^k) \right) \Pi_h^B[\phi] \, \mathrm{d}x = \int_{\Omega_h} D_t \left(\varrho_h^k \chi(\vartheta_h^k) \right) \phi \, \mathrm{d}x
$$

$$
+ \int_{\Omega_h} \frac{\varrho_h^k - \varrho_h^{k-1}}{\Delta t} \chi(\vartheta_h^k) \left(\Pi_h^B[\phi] - \phi \right) \, \mathrm{d}x + \int_{\Omega_h} \varrho_h^{k-1} \frac{\chi(\vartheta_h^k) - \chi(\vartheta_h^{k-1})}{\Delta t} \left(\Pi_h^B[\phi] - \phi \right) \, \mathrm{d}x.
$$

As χ is bounded and $\Delta t \approx h$, we may use (2.12) to deduce

$$
\left| \int_{\Omega_h} \frac{\varrho_h^k - \varrho_h^{k-1}}{\Delta t} \chi(\vartheta_h^k) \left(\Pi_h^B[\phi] - \phi \right) \mathrm{d}x \right| \lesssim \left(\int_{\Omega_h} \frac{\left(\varrho_h^k - \varrho_h^{k-1} \right)^2}{\Delta t} \, \mathrm{d}x \right)^{1/2} \sqrt{h} \|\nabla_x \phi\|_{L^\infty(\Omega_h; R^3)},
$$

where the right-hand side is controlled by (5.7).

Similarly,

$$\left| \int_{\Omega_h} \sqrt{\varrho_h^{k-1}} \sqrt{\varrho_h^{k-1}} \frac{\chi(\vartheta_h^k) - \chi(\vartheta_h^{k-1})}{\Delta t} \left(\Pi_h^B[\phi] - \phi \right) \ \mathrm{d}x \right|$$

$$\lesssim \left(\Delta t \int_{\Omega_h} \varrho_h^{k-1} \left(\frac{\chi(\vartheta_h^k) - \chi(\vartheta_h^{k-1})}{\Delta t} \right)^2 \ \mathrm{d}x \right)^{1/2} \sqrt{h} \|\nabla_x \phi\|_{L^\infty(\Omega_h; R^3)} \|\varrho_h^{k-1}\|_{L^\gamma(\Omega_h)}^{1/2},$$

which can be bounded by means of (5.16). Indeed it is enough to check that

$$\chi(A) - \chi(B) \lesssim \frac{A-B}{A} \text{ whenever } A > B \geq 0$$

as long as χ belongs to the class (6.5).

Summing up the previous estimates, we may infer that

(6.9)
$$\left| \int_{\Omega_h} D_t \left(\varrho_h \chi(\vartheta_h) \right) \left(\Pi_h^B[\phi] - \phi \right) \ \mathrm{d}x \right| \lesssim \sqrt{h} \ R_h^4(t) \|\nabla_x \phi\|_{L^\infty(\Omega_h; R^3)}, \ \|R_h^4\|_{L^2(0,T)} \lesssim 1.$$

To handle the upwind term, we use formula (2.19) yielding

(6.10)
$$\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \mathrm{Up}[\varrho_h^k \chi(\vartheta_h^k), \mathbf{u}_h^k] \ [[\Pi_h^B[\phi] \ ]] \ \mathrm{d}S_x$$

$$= \int_{\Omega_h} \varrho_h^k \chi(\vartheta_h^k) \mathbf{u}_h^k \cdot \nabla_x \phi \ \mathrm{d}x - \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \left(\Pi_h^B[\phi] - \phi \right) [[\varrho_h^k \chi(\vartheta_h^k)]] [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \ \mathrm{d}S_x$$

$$+ \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \varrho_h^k \chi(\vartheta_h^k) \phi (\tilde{\mathbf{u}} - \mathbf{u}) \cdot \mathbf{n} \ \mathrm{d}S_x + \sum_{E \in E_h} \int_{E_h} \varrho_h^k \chi(\vartheta_h^k) \mathrm{div}_h \mathbf{u}_h^k \left(\phi - \Pi_h^B \phi \right) \ \mathrm{d}x.$$

We write

$$\sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \left(\Pi_h^B[\phi] - \phi \right) [[\varrho_h^k \chi(\vartheta_h^k)]] [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \ \mathrm{d}S_x =$$

$$\sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \left(\Pi_h^B[\phi] - \phi \right) \varrho_h^k [[\chi(\vartheta_h^k)]] [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \ \mathrm{d}S_x$$

$$+ \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \left(\Pi_h^B[\phi] - \phi \right) [[\varrho_h^k]] \chi((\vartheta_h^k)^{\mathrm{out}}) [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \ \mathrm{d}S_x,$$

where, by means of Hölder's and Jensen's inequalities, the error estimates (2.12), and the trace estimates (2.23),

$$\left| \sum_{E \in E_h} \sum_{\Gamma_E \subset \partial E} \int_{\Gamma_E} \left(\Pi_h^B[\phi] - \phi \right) (\varrho_h^k) [[\chi(\vartheta_h^k)]] [\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}]^- \ \mathrm{d}S_x \right|$$

$$\lesssim h^{3/2} \|\nabla_x \phi\|_{L^\infty(\Omega_h; R^3)} \left(\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma \frac{[[\chi(\vartheta_h^k)]]^2}{h} \ \mathrm{d}S_x \right)^{1/2} \left(\sum_{\Gamma \in \Gamma_{h,\mathrm{int}}} \int_\Gamma |\varrho_h^k|^2 |\tilde{\mathbf{u}}_h^k \cdot \mathbf{n}|^2 \ \mathrm{d}S_x \right)^{1/2}$$

$$\lesssim h^{3/2}\|\nabla_x\phi\|_{L^\infty(\Omega_h;R^3)}\left(\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\frac{[[\chi(\vartheta_h^k)]]^2}{h}\,\mathrm{dS}_x\right)^{1/2}\left(\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma|\varrho_h^k|^2|\mathbf{u}_h^k|^2\,\mathrm{dS}_x\right)^{1/2}$$

$$\lesssim h\|\nabla_x\phi\|_{L^\infty(\Omega_h;R^3)}\left(\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\frac{[[\chi(\vartheta_h^k)]]^2}{h}\,\mathrm{dS}_x\right)^{1/2}\left(\sum_{E\in E_h}\int_E|\varrho_h^k|^2|\mathbf{u}_h^k|^2\mathrm{d}x\right)^{1/2}.$$

Now, the relations (5.6), (5.14), and (5.23) may be used to control both integrals on the right-hand side in $L^2(0,T)$.

Furthermore, as χ is bounded, the integral

$$\sum_{E\in E_h}\sum_{\Gamma_E\subset\partial E}\int_{\Gamma_E}\left(\Pi_h^B[\phi]-\phi\right)[[\varrho_h^k]]\chi((\vartheta_h^k)^{\mathrm{out}})[\tilde{\mathbf{u}}_h^k\cdot\mathbf{n}]^-\,\mathrm{dS}_x$$

can be handled with the help of the energy estimate (5.9), (5.23), and (2.12).

Finally, we observe that the remaining two integrals on the right-hand side of (6.10) can be estimated by means of (2.11) and the available energy bounds (5.6), (5.22). Thus we conclude that

$$\left|\sum_{\Gamma\in\Gamma_h}\int_\Gamma\mathrm{Up}[\varrho_h^k\chi(\vartheta_h^k),\mathbf{u}_h^k]\,[[\Pi_h^B[\phi]\,]]\,\mathrm{dS}_x-\int_{\Omega_h}\varrho_h^k\chi(\vartheta_h^k)\mathbf{u}_h^k\cdot\nabla_x\phi\,\mathrm{d}x\right|\tag{6.11}$$

$$\lesssim h^{\frac{\gamma-2}{\gamma}}R_h^5(t)\|\nabla_x\phi\|_{L^\infty(\Omega_h;R^3)},\ \|R_h^5\|_{L^1(0,T)}\lesssim 1.$$

The most delicate part of the proof of consistency of the thermal energy scheme ((3.4)) is the heat flux term. We need the following auxiliary result.

LEMMA 6.1. *Let $\phi\in C^2(R^3)$ such that $\nabla_x\phi\cdot\mathbf{n}|_{\partial\Omega}=0$.*
Then

$$\left|\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma\frac{1}{d_\Gamma}[[v]]\,[[\,\Pi_h^B[\phi]\,]]\mathrm{dS}_x+\int_\Omega v\Delta\phi\,\mathrm{d}x\right|\lesssim\sqrt{h}\Big(\|v\|_{H^1_{Q_h}(\Omega_h)}+\|v\|_{L^\infty(\Omega_h)}\Big)\,\|\phi\|_{C^2}$$

for any $v\in Q_h(\Omega_h)$.

Proof:
First, by the Gauss–Green theorem,

$$\int_{\Omega_h}v\Delta\phi\,\mathrm{d}x=\sum_{E\in E_h}\int_E v\Delta\phi\,\mathrm{dx}=\sum_{E\in E_h}\int_{\partial E}v\nabla_x\phi\cdot\mathbf{n}\,\mathrm{dS}_x$$

$$=-\sum_{\Gamma\in\Gamma_{h,\mathrm{int}}}\int_\Gamma[[v]]\nabla_x\phi\cdot\mathbf{n}\,\mathrm{dS}_x+\int_{\partial\Omega_h}v\nabla_x\phi\cdot\mathbf{n}\,\mathrm{dS}_x,$$

where, furthermore,

$$\left|\int_{\Omega_h}v\Delta\phi\,\mathrm{d}x-\int_\Omega v\Delta\phi\,\mathrm{d}x\right|\le\left|\int_{\Omega_h\setminus\Omega}|v||\Delta\phi|\,\mathrm{d}x\right|\lesssim h\|v\|_{L^\infty(\Omega_h)}\|\phi\|_{C^2(R^3)}.\tag{6.12}$$

Next, going back to the definition of the projection Π_h^B, we get

$$\left|\nabla_x\phi\cdot\mathbf{n}-\frac{[[\Pi_h^B\phi]]}{d_\Gamma}\right|\lesssim h\|\phi\|_{C^2(\overline{\Omega})}\text{ on any face }\Gamma,$$

and, by Hölder's inequality,

(6.13)
$$\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma |\ [[v]]\ |\,{\rm dS}_x \leq \left(\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma \frac{[[v]]^2}{d_\Gamma}\ {\rm dS}_x\right)^{1/2}\left(\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma d_\Gamma {\rm dS}_x\right)^{1/2} \lesssim \|v\|_{H^1_{Q_h}(\Omega)}|\Omega|^{1/2}.$$

Thus it remains to control the integral $\int_{\partial\Omega_h} v\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x$. To this end, write

$$\int_{\Omega_h\setminus\Omega} v\Delta\phi\ {\rm d}x = \sum_{E\in E_h, E\not\subset\overline{\Omega}}\int_{E\setminus\Omega} v\Delta\phi\ {\rm d}x,$$

where the left-hand side is small in view of (6.12). Moreover, by the Gauss–Green theorem,

$$\sum_{E\in E_h, E\not\subset\overline{\Omega}}\int_{E\setminus\Omega} v\Delta\phi\ {\rm d}x = \int_{\partial\Omega_h} v\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x + \sum_{E\in E_h, E\not\subset\overline{\Omega}}\int_{\partial(E\setminus\Omega)\setminus\partial\Omega_h} v\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x.$$

Seeing that $\nabla_x\phi\cdot\mathbf{n}|_{\partial\Omega}=0$ we may infer that

$$\sum_{E\in E_h, E\not\subset\overline{\Omega}}\int_{\partial(E\setminus\Omega)\setminus\partial\Omega_h} v\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x = -\sum_{\Gamma\in\Gamma_{h,\rm int},\ \Gamma\subset\partial E,\ E\not\subset\overline{\Omega}}\int_{\Gamma\setminus\Omega}[[v]]\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x,$$

where, similarly to (6.13),

$$\left|\sum_{\Gamma\in\Gamma_{h,\rm int},\ \Gamma\subset\partial E,\ E\not\subset\overline{\Omega}}\int_{\Gamma\setminus\Omega}[[v]]\nabla_x\phi\cdot\mathbf{n}\ {\rm dS}_x\right|$$
$$\lesssim \|\phi\|_{C^1(R^3)}\left(\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma\frac{[[v]]^2}{d_\Gamma}\ {\rm dS}_x\right)^{1/2}\left(\sum_{\Gamma\in\Gamma_{h,\rm int},\ \Gamma\subset\partial E,\ E\not\subset\overline{\Omega}}\int_\Gamma d_\Gamma\ {\rm dS}_x\right)^{1/2}$$
$$\lesssim \|\phi\|_{C^1(R^3)}\|v\|_{H^1_{Q_h}(\Omega_h)}\left|\left\{x\in R^3\ \middle|\ {\rm dist}[x,\partial\Omega_h]<2h\right\}\right|^{1/2}\approx h^{1/2}\|\phi\|_{C^1}\|v\|_{H^1_{Q_h}(\Omega_h)}.$$

Now, we are ready to deal with the diffusion term

$$\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma\frac{1}{d_\Gamma}\left\{\chi'\left(\vartheta_h^k\right)\right\}[[K(\vartheta_h^k)]]\ [[\Pi_h^B[\phi]]]{\rm dS}_x.$$

Introducing a new function K_χ, $K'_\chi(\vartheta)=\chi'(\vartheta)K'(\vartheta)$, we rewrite the diffusive term with the help of the mean-value theorem as

$$\left\{\chi'\left(\vartheta_h^k\right)\right\}[[K(\vartheta_h^k)]] = [[K_\chi(\vartheta_h^k)]] + c_h^k(x)[[\vartheta_n^k]]^2,$$

where c_h^k is uniformly bounded. Consequently, we get

$$\sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma\frac{1}{d_\Gamma}\left\{\chi'\left(\vartheta_h^k\right)\right\}[[K(\vartheta_h^k)]]\ [[\Pi_h^B[\phi]]]{\rm dS}_x$$
$$= \sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma\frac{1}{d_\Gamma}[[K_\chi(\vartheta_h^k)]]\ [[\Pi_h^B[\phi]]]{\rm dS}_x + \sum_{\Gamma\in\Gamma_{h,\rm int}}\int_\Gamma c_h^k\frac{[[\vartheta_h^k]]^2}{d_\Gamma}[[\Pi_h^B[\phi]]]{\rm dS}_x.$$

Seeing that $|[[\Pi_h^B \phi]]| \leq h\|\nabla_x \phi\|_{L^\infty(\Omega_h;R^3)}$, we can estimate the last integral using the entropy bounds (5.14), while the first integral can be "replaced" by $\int_\Omega K_\chi(\vartheta_h^k)\Delta\phi \,\mathrm{d}x$ in view of Lemma 6.1.

Finally, observing that the remaining terms in (6.7) can be treated in a similar way, we sum up the previous estimates to obtain

$$\begin{aligned}(6.14)\quad &\int_{\Omega_h} D_t\left(\varrho_h^k \chi(\vartheta_h^k)\right)\phi \,\mathrm{d}x - \int_{\Omega_h} \varrho_h^k \chi(\vartheta_h^k)\mathbf{u}_h^k \cdot \nabla_x \phi \,\mathrm{d}x - \int_\Omega K_\chi(\vartheta_h^k)\Delta\phi \,\mathrm{d}x \\ &= \int_{\Omega_h} \left(\mu|\nabla_h \mathbf{u}_h^k|^2 + \lambda|\mathrm{div}_h \mathbf{u}_h^k|^2\right)\chi'(\vartheta_h^k)\phi \,\mathrm{d}x - \int_{\Omega_h} \chi'(\vartheta_h^k)\vartheta_h^k \varrho_h^k \mathrm{div}_h \mathbf{u}_h^k \phi \,\mathrm{d}x \\ &\quad + \langle D_h, \phi\rangle + h^\beta \left\langle R_h^6, \phi\right\rangle,\end{aligned}$$

for a certain $\beta > 0$, where

$$(6.15)\qquad \left|\left\langle R_h^6(t), \phi\right\rangle\right| \lesssim R_h^7(t)\|\phi\|_{C^2(R^3)},\ \|R_h^7\|_{L^1(0,T)} \lesssim 1.$$

Relation (6.14) holds for any test function $\phi \in C^2(R^2)$ such that $\nabla_x \phi \cdot \mathbf{n}|_{\partial\Omega} = 0$, and for any χ enjoying the properties stated in (6.5). The quantity D_h is a bounded measure satisfying (6.8).

We conclude by a simple observation that (6.14) gives rise to

$$\begin{aligned}(6.16)\quad &\int_{\Omega_h} D_t\left(\varrho_h^k \chi(\vartheta_h^k)\right)\phi \,\mathrm{d}x - \int_\Omega \varrho_h^k \chi(\vartheta_h^k)\mathbf{u}_h^k \cdot \nabla_x \phi \,\mathrm{d}x - \int_\Omega K_\chi(\vartheta_h^k)\Delta\phi \,\mathrm{d}x \\ &= \int_\Omega \left(\mu|\nabla_h \mathbf{u}_h^k|^2 + \lambda|\mathrm{div}_h \mathbf{u}_h^k|^2\right)\chi'(\vartheta_h^k)\phi \,\mathrm{d}x - \int_\Omega \chi'(\vartheta_h^k)\vartheta_h^k \varrho_h^k \mathrm{div}_h \mathbf{u}_h^k \phi \,\mathrm{d}x \\ &\quad + \langle D_h, \phi\rangle + h^\beta \left\langle R_h^6, \phi\right\rangle,\end{aligned}$$

where the integrals over the complements $\Omega_h \setminus \Omega$ were incorporated in D_h and R_h^6. As for the discrete time derivative, we claim that

$$\begin{aligned}(6.17)\quad &\int_0^T \psi(t) \int_{\Omega_h} D_t\left(\varrho_h \chi(\vartheta_h)\right)\phi \,\mathrm{d}x \,\mathrm{d}t \\ &= \psi(0)\int_{\Omega_h} \varrho_h^0 \chi(\vartheta_h^0)\phi \,\mathrm{d}x - \int_0^T \int_{\Omega_h} \left(\frac{\psi(t+\Delta t) - \psi(t)}{\Delta t}\right)\varrho_h \chi(\vartheta_h)\phi \,\mathrm{d}x\end{aligned}$$

for any $\psi \in C_c^\infty[0,T)$, where, by the mean-value theorem,

$$\left|\left(\frac{\psi(t+\Delta t) - \psi(t)}{\Delta t}\right) - \partial_t \psi\right| \lesssim \Delta t \sup_{s\in[0,T]} |\psi''(s)|.$$

Thus, with (6.17) in mind, we observe that (6.16) coincides with its analogue proved in [8, section 6.3, formula (6.25)].

6.2. Convergence. As observed above, the consistency formulation (6.3), (6.4), (6.16), and (6.17) is the same as in [8], whence the proof of convergence can be carried over by means of the arguments specified in [8, section 7]. We have proved Theorem 3.1.

7. Unconditional convergence. If the initial data $[\varrho_0, \vartheta_0, \mathbf{u}_0]$ are regular and the physical domain has a sufficiently smooth boundary, the Navier–Stokes–Fourier system is known to admit strong solutions, at least on a possibly short time interval. If

$$\varrho_0,\ \vartheta_0 \in W^{3,2}(\Omega),\ \varrho_0 > 0,\ \vartheta_0 > 0,\ \mathbf{u}_0 \in W^{3,2}(\Omega; R^3) \tag{7.1}$$

are the initial data satisfying the relevant *compatibility conditions*, and if Ω is of class $C^{2+\nu}$, then the problem (1.1)–(1.7) admits a (classical) solution

$$\varrho, \vartheta \in C([0, T_{\max}); W^{3,2}(\Omega)),\ \mathbf{u}_0 \in C([0, T_{\max}); W^{3,2}(\Omega; R^3)) \tag{7.2}$$

on a maximal time interval $[0, T_{\max})$; see Valli [15], [16], and Valli and Zajaczkowski [17].

On the other hand, as shown in [6, Chapter 7], the problem (1.1)–(1.7) endowed with the regular initial data (7.1) possesses a global-in-time weak solution in the sense of Definition 2.1. Weak and strong solutions emanating from the same initial data *should* coincide on their common existence time interval. As a matter of fact, the answer is not completely straightforward; however, the following result holds; see [5, Lemma 3.2].

PROPOSITION 7.1. *In addition to the hypotheses of Theorem* 3.1*, suppose that* $\Omega \subset R^3$ *is a bounded domain,* $\partial\Omega \in C^{2,\nu}$*, and that the initial data satisfy* (7.1)*. Let* $[\varrho, \vartheta, \mathbf{u}]$ *be a weak solution of the Navier–Stokes–Fourier system* (1.1)–(1.7) *enjoying extra regularity*

$$\varrho,\ \vartheta,\ \mathrm{div}_x \mathbf{u} \in L^\infty((0,T) \times \Omega),\ \mathbf{u} \in L^\infty((0,T) \times \Omega; R^3).$$

Then $[\varrho, \vartheta, \mathbf{u}]$ *coincides with the strong solution of the same problem as long as the latter exists.*

It turns out that the weak solutions possessing the regularity claimed in Proposition 7.1 are in fact strong. More specifically, we report the following assertion; see [5, Theorem 2.2].

PROPOSITION 7.2. *Under the hypotheses of Proposition* 7.1*, let* $[\varrho, \vartheta, \mathbf{u}]$ *be a weak solution of the Navier–Stokes–Fourier system, emanating from regular initial data satisfying* (7.1)*, and enjoying the extra regularity*

$$\varrho,\ \vartheta,\ \mathrm{div}_x \mathbf{u} \in L^\infty((0,T) \times \Omega),\ \mathbf{u} \in L^\infty((0,T) \times \Omega; R^3).$$

Then $[\varrho, \vartheta, \mathbf{u}]$ *is a strong (classical) solution of the problem in* $(0,T) \times \Omega$*.*

Combining the previous results with Theorem 3.1, we obtain the following statement concerning *unconditional convergence* of the numerical scheme (3.1)–(3.4).

THEOREM 7.1. *In addition to the hypotheses of Theorem* 3.1*, suppose that* $\Omega \subset R^3$ *is a bounded domain,* $\partial\Omega \in C^{2,\nu}$*, and the initial data satisfy* (7.1)*. Let* $[\varrho_h, \vartheta_h, \mathbf{u}_h]_{h>0}$ *be a family of numerical solutions constructed by means of the scheme* (3.1)–(3.4) *such that*

$$\varrho_h > 0,\ \vartheta_h > 0,\ \textit{and}\ \varrho_h,\ \vartheta_h,\ |\mathbf{u}_h|,\ |\mathrm{div}_h \mathbf{u}_h| \leq M$$

a.a. in $(0,T) \times \Omega$ *for a certain constant* M *independent of* h*.*

Then

$$\varrho_h \to \varrho \textit{ weakly-(*) in } L^\infty(0,T; L^\gamma(\Omega)) \textit{ and strongly in } L^1((0,T) \times \Omega),$$
$$\vartheta_h \to \vartheta \textit{ weakly in } L^2(0,T; L^6(\Omega)),$$

$\mathbf{u}_h \to \mathbf{u}$ *weakly in* $L^2(0,T;L^6(\Omega;R^3))$, $\nabla_h \mathbf{u}_h \to \nabla_x \mathbf{u}$ *weakly in* $L^2((0,T)\times\Omega;R^{3\times 3})$,

where $[\varrho,\vartheta,\mathbf{u}]$ *is the (unique) strong solution of the Navier–Stokes–Fourier system* (1.1)–(1.7) *in* $(0,T)\times\Omega$ *emanating from the initial data* (7.1).

REFERENCES

[1] C. Chainais-Hillairet and J. Droniou, *Finite-volume schemes for noncoercive elliptic problems with Neumann boundary conditions*, IMA J. Numer. Anal., 31 (2011), pp. 61–85.

[2] S. H. Christiansen, H. Z. Munthe-Kaas, and B. Owren, *Topics in structure-preserving discretization*, Acta Numer., 20 (2011), pp. 1–119.

[3] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods. In Handbook of numerical analysis, Vol.* VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[4] R. Eymard, T. Gallouët, R. Herbin, and J. C. Latché, *A convergent finite element-finite volume scheme for the compressible Stokes problem.* II. *The isentropic case*, Math. Comp., 79 (2010), pp. 649–675.

[5] E. Feireisl and Y. Sun, *Conditional regularity of very weak solutions to the Navier-Stokes-Fourier system*, in Recent Advances in Partial Differential Equations and Applications, Contemp. Math. 666, AMS, Providence, NY, 2016, pp. 179–199.

[6] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, 2004.

[7] E. Feireisl, T. Karper, and M. Michálek, *Convergence of a numerical method for the compressible Navier–Stokes system on general domains*, Numer. Math., http://dx.doi.org/10.1007/s00211-015-0786-6.

[8] E. Feireisl, T. Karper, and A. Novotný, *A convergent numerical method for the Navier-Stokes-Fourier system system*, IMA J. Numer. Math., http://dx.doi.org/10.1093/imanum/drv049.

[9] G. Gallavotti, *Foundations of Fluid Dynamics*, Springer-Verlag, New York, 2002.

[10] T. Gallouet, E. Herbin, and J.-C. Latché, *A convergent finite element-finite volume scheme for the compressible Stokes problem.* I. *The isothermal case*, Math. Comput., 78 (2009), pp. 1333–1352.

[11] R. Hošek, *Face-to-face partition of 3D space with identical well-centered tetrahedra*, Appl. Math., 60 (2015), pp 637–651.

[12] K. H. Karlsen and T. K. Karper, *A convergent mixed method for the Stokes approximation of viscous compressible flow*, IMA J. Numer. Anal., 32 (2012), pp. 725–764.

[13] T. K. Karper, *A convergent FEM-DG method for the compressible Navier-Stokes equations*, Numer. Math., 125 (2013), pp. 441–510.

[14] M. Lenoir, *Optimal isoparametric finite elements and error estimates for domains involving curved boundaries*, SIAM J. Numer. Anal., 23 (1986), pp. 562–580.

[15] A. Valli, *A correction to the paper: "An existence theorem for compressible viscous fluids"*, Ann. Mat. Pura Appl. (4), 132 (1983), pp. 399–400 1982.

[16] A. Valli, *An existence theorem for compressible viscous fluids*, Ann. Mat. Pura Appl. (4), 130 (1982), pp. 197–213.

[17] A. Valli and M. Zajaczkowski, *Navier-Stokes equations for compressible fluids: Global existence and qualitative properties of the solutions in the general case*, Comm. Math. Phys., 103 (1986), pp. 259–296.

[18] E. VanderZee, A. N. Hirani, D. Guoy, V. Zharnitsky, and E. A. Ramos, *Geometric and combinatorial properties of well-centered triangulations in three and higher dimensions*, Comput. Geom., 46 (2013), pp. 700–724.

Appendix C

E. Feireisl, R. H., D. Maltese, A. Novotný: Error estimates for a numerical method for the compressible Navier-Stokes system on sufficiently smooth domains.

ESAIM: M2AN 51 (2017) 279–319
DOI: 10.1051/m2an/2016022

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

ERROR ESTIMATES FOR A NUMERICAL METHOD FOR THE COMPRESSIBLE NAVIER–STOKES SYSTEM ON SUFFICIENTLY SMOOTH DOMAINS *, **, ***

Eduard Feireisl[1,2], Radim Hošek[1,2], David Maltese[1,2] and Antonín Novotný[1,2]

Abstract. We derive an *a priori* error estimate for the numerical solution obtained by time and space discretization by the finite volume/finite element method of the barotropic Navier–Stokes equations. The numerical solution on a convenient polyhedral domain approximating a sufficiently smooth bounded domain is compared with an exact solution of the barotropic Navier–Stokes equations with a bounded density. The result is unconditional in the sense that there are no assumed bounds on the numerical solution. It is obtained by the combination of discrete relative energy inequality derived in [T. Gallouët, R. Herbin, D. Maltese and A. Novotný, *IMA J. Numer. Anal.* **36** (2016) 543–592.] and several recent results in the theory of compressible Navier–Stokes equations concerning blow up criterion established in [Y. Sun, C. Wang and Z. Zhang, *J. Math. Pures Appl.* **95** (2011) 36–47] and weak strong uniqueness principle established in [E. Feireisl, B.J. Jin and A. Novotný, *J. Math. Fluid Mech.* **14** (2012) 717–730].

Mathematics Subject Classification. 35Q30, 65N12, 65N30, 76N10, 76N15, 76M10, 76M12.

Received August 26, 2015. Revised January 13, 2016. Accepted March 23, 2016.

Keywords and phrases. Navier–Stokes system, finite element numerical method, finite volume numerical method, error estimates.

* *The research of E. Feireisl leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC Grant Agreement 320078. The Institute of Mathematics of the Academy of Sciences of the Czech Republic is supported by RVO:67985840.*

** *The research of R. Hošek leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC Grant Agreement 320078. The Institute of Mathematics of the Academy of Sciences of the Czech Republic is supported by RVO:67985840.*

*** *The work of D. Maltese and A. Novotný has been supported by the MODTERCOM project within the APEX programme of the Provence–Alpes-Côte d'Azur region.*

[1] Institute of Mathematics of the Academy of Sciences of the Czech Republic, Žitná 25, 115 67 Praha 1, Czech Republic. feireisl@math.cas.cz

[2] Institut Mathématiques de Toulon, EA2134, University of Toulon, BP 20132, 839 57 La Garde, France.

Article published by EDP Sciences

1. Introduction

We consider the compressible Navier–Stokes equations in the barotropic regime in a space-time cylinder $Q_T = (0,T) \times \Omega$, where $T > 0$ is arbitrarily large and $\Omega \subset R^3$ is a bounded domain:

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{1.1}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x p(\varrho) = \mathrm{div}_x \mathbb{S}(\nabla_x \mathbf{u}). \tag{1.2}$$

In equations (1.1) and (1.2) $\varrho = \varrho(t,x) \geq 0$ and $\mathbf{u} = \mathbf{u}(t,x) \in R^3$, $t \in [0,T)$, $x \in \Omega$ are the unknown density and velocity fields, while $\mathbb{S}$ and p are the viscous stress and pressure characterizing the fluid *via* the constitutive relations

$$\mathbb{S}(\nabla_x \mathbf{u}) = \mu \left(\nabla_x \mathbf{u} + \nabla_x^t \mathbf{u} - \frac{2}{3} \mathrm{div}_x \mathbf{u} \mathbb{I} \right), \ \mu > 0, \tag{1.3}$$

$$p \in C^2(0,\infty) \cap C^1[0,\infty), \ p(0) = 0, \ p'(\varrho) > 0 \text{ for all } \varrho \geq 0, \ \lim_{\varrho \to \infty} \frac{p'(\varrho)}{\varrho^{\gamma - 1}} = p_\infty > 0, \tag{1.4}$$

where $\gamma \geq 1$.

The assumption $p'(0) > 0$ in (1.4) excludes the constitutive laws for pressure behaving as ϱ^γ as $\varrho \to 0^+$. The error estimates stated in Theorem 3.2 however still hold in the case $\lim_{\varrho \to 0^+} \frac{p'(\varrho)}{\varrho^{\gamma-1}} = 0$, in particular for the isentropic pressure laws $p(\varrho) = \varrho^\gamma$. The proof contains some additional technical difficulties, see also Remark 3.2.

Equations (1.1) and (1.2) are completed with the no-slip boundary conditions

$$\mathbf{u}|_{\partial\Omega} = 0, \tag{1.5}$$

and initial conditions

$$\varrho(0,\cdot) = \varrho_0, \ \mathbf{u}(0,\cdot) = \mathbf{u}_0, \ \varrho_0 > 0 \text{ in } \overline{\Omega}. \tag{1.6}$$

We notice that under assumption (1.3), we may write

$$\mathrm{div}_x \mathbb{S}(\nabla_x \mathbf{u}) = \mu \Delta \mathbf{u} + \frac{\mu}{3} \nabla_x \mathrm{div}_x \mathbf{u}. \tag{1.7}$$

The results on error estimates for numerical schemes for the compressible Navier–Stokes equations are in the mathematical literature on short supply. We refer the reader to papers of Liu [39, 40], Jovanović [28], Gallouet *et al.* [22].

In [22] the authors have developed a methodology of deriving unconditional error estimates for the numerical schemes to the compressible Navier–Stokes equations (1.1)–(1.6) and applied it to the numerical scheme (3.5)–(3.7) discretizing the system on polyhedral domains. They have obtained error estimates for the discrete solution with respect to a *classical solution* of the system on the same (polyhedral) domain. In spite of the fact that [22] provides the first and to the best of our knowledge so far the sole error estimate for discrete solutions of a finite volume/finite element approximation to a model of compressible fluids that does not need any assumed bounds on the numerical solution itself, it has two weak points: 1) The existence of classical solutions on at least a short time interval to the compressible Navier–Stokes equations is known for smooth C^3 domains (see [43] or [4]) but may not be in general true on the polyhedral domains. 2) The numerical solutions are compared with the classical exact solutions (as is usual in any previous existing mathematical literature). In this paper we address both points raised above and to a certain extent remove the limitations of the theory presented in [22].

More precisely, we generalize the result of Gallouet *et al.* ([22], Thm. 3.1) in two directions:

(1) The physical domain Ω filled by the fluid and the numerical domain Ω_h, $h > 0$ approximating the physical domain do not need to coincide.

(2) If the physical domain is sufficiently smooth (at least of class C^3) and the C^3-initial data satisfy natural compatibility conditions, we are able to obtain the unconditional error estimates with respect to any *weak exact solution with bounded density.*

As in [22], and in contrast with any other literature dealing with finite volume or mixed finite volume/finite element methods for compressible fluids [3, 10, 16–19, 23–25, 28, 29, 32–35, 44] and others) this result does not require any assumed bounds on the discrete solution: the sole bounds needed for the result are those provided by the numerical scheme. Moreover, in contrast with [22] and with all above mentioned papers, the exact solution is solely weak solution with bounded density. This seemingly weak hypothesis is compensated by the regularity and compatibility conditions imposed on initial data that make possible a (sophisticated) bootstrapping argument showing that weak solutions with bounded density are in fact strong solutions in the class investigated in [22].

These results are achieved by using the following tools:

(1) The technique introduced in [22] modified in order to accommodate non-zero velocity of the exact sample solution on the boundary of the numerical domain.
(2) Three fundamental recent results from the theory of compressible Navier–Stokes equations, namely
   - Local in time existence of strong solutions in class (2.11) and (2.12) by Cho *et al.* [4].
   - Weak strong uniqueness principle proved in [13] (see also [14]).
   - Blow up criterion for strong solutions in the class (2.11) and (2.12) by Sun *et al.* [41].

   The three above mentioned items allow to show that the weak solution with bounded density emanating from the sufficiently smooth initial data is in fact a strong solution defined on the large time interval $[0, T)$.
(3) Bootstrapping argument using recent results on maximal regularity for parabolic systems by Danchin [8], Denk *et al.* [5] and Krylov [36]. The last item allows to bootstrap the strong solution in the class Cho *et al.* [4] to the class needed for the error estimates in [22], provided a certain compatibility condition for the initial data is satisfied.

2. Preliminaries

2.1. Weak and strong solutions to the Navier–Stokes system

We introduce the notion of the weak solution to system (1.1)–(1.4):

Definition 2.1 (Weak solutions)**.** *Let* $\varrho_0 : \Omega \to [0, +\infty)$ *and* $\boldsymbol{u}_0 : \Omega \to \mathbb{R}^3$ *with finite energy* $E_0 = \int_\Omega (\frac{1}{2}\varrho_0 |\boldsymbol{u}_0|^2 + H(\varrho_0))\,\mathrm{d}x$ *and finite mass* $0 < M_0 = \int_\Omega \varrho_0\,\mathrm{d}x$*. We shall say that the pair* $(\varrho, \boldsymbol{u})$ *is a weak solution to the problem* (1.1)–(1.6) *emanating from the initial data* $(\varrho_0, \boldsymbol{u}_0)$ *if:*

(a) $\varrho \in C_{\mathrm{weak}}([0,T]; L^a(\Omega))$*, for a certain* $a > 1$*,* $\varrho \geq 0$ *a.e. in* $(0,T)$*, and* $\boldsymbol{u} \in L^2(0,T; W_0^{1,2}(\Omega; \mathbb{R}^3))$.
(b) *the continuity equation* (1.1) *is satisfied in the following weak sense*

$$\int_\Omega \varrho\varphi\,\mathrm{d}x\Big|_0^\tau = \int_0^\tau \int_\Omega (\varrho\partial_t\varphi + \varrho\boldsymbol{u}\cdot\nabla_x\varphi)\,\mathrm{d}x\,\mathrm{d}t,\ \forall\tau\in[0,T],\ \forall\varphi\in C_c^\infty([0,T]\times\overline{\Omega}). \tag{2.1}$$

(c) $\varrho\boldsymbol{u} \in C_{\mathrm{weak}}([0,T]; L^b(\Omega;\mathbb{R}^3))$*, for a certain* $b > 1$*, and the momentum equation* (1.2) *is satisfied in the weak sense,*

$$\begin{aligned}\int_\Omega \varrho\boldsymbol{u}\cdot\varphi\,\mathrm{d}x\Big|_0^\tau &= \int_0^\tau\int_\Omega (\varrho\boldsymbol{u}\cdot\partial_t\varphi + \varrho\boldsymbol{u}\otimes\boldsymbol{u} : \nabla\varphi + p(\varrho)\,\mathrm{div}\,\varphi)\,\mathrm{d}x\,\mathrm{d}t\\ &\quad - \int_0^\tau\int_\Omega (\mu\nabla\boldsymbol{u} : \nabla_x\varphi\,\mathrm{d}x\,\mathrm{d}t + (\mu+\lambda)\mathrm{div}\boldsymbol{u}\mathrm{div}\varphi)\,\mathrm{d}x\,\mathrm{d}t,\ \forall\tau\in[0,T],\ \forall\varphi\in C_c^\infty([0,T]\times\Omega;\mathbb{R}^3).\end{aligned} \tag{2.2}$$

(d) *The following energy inequality is satisfied*

$$\int_\Omega \left(\frac{1}{2}\varrho|\boldsymbol{u}|^2 + H(\varrho) \right) \mathrm{d}x \Big|_0^\tau + \int_0^\tau \int_\Omega \left(\mu|\nabla \boldsymbol{u}|^2 + (\mu+\lambda)|\operatorname{div} \boldsymbol{u}|^2 \right) \mathrm{d}x\,\mathrm{d}t \leq 0, \ \textit{for a.a. } \tau \in (0,T), \tag{2.3}$$

$$\textit{with } \ H(\varrho) = \varrho \int_1^\varrho \frac{p(z)}{z^2}\mathrm{d}z. \tag{2.4}$$

Here and hereafter the symbol $\int_\Omega g\,\mathrm{d}x\,|_0^\tau$ *is meant for* $\int_\Omega g(\tau,x)\,\mathrm{d}x - \int_\Omega g_0(x)\,\mathrm{d}x$.

In the above definition, we tacitly assume that all the integrals in the formulas (2.1)–(2.3) are defined and we recall that $C_{\rm weak}([0,T];L^a(\Omega))$ is the space of functions of $L^\infty([0,T];L^a(\Omega))$ which are continuous as functions of time in the weak topology of the space $L^a(\Omega)$.

We notice that the function $\varrho \mapsto H(\varrho)$ is a solution of the ordinary differential equation $\varrho H'(\varrho) - H(\varrho) = p(\varrho)$ with the constant of integration fixed such that $H(1) = 0$.

Note that the existence of weak solutions emanating from the finite energy initial data is well-known on bounded Lipschitz domains provided $\gamma > 3/2$, see Lions [38] for 'large' values of γ, Feireisl and coauthors [12] for $\gamma > 3/2$.

Proposition 2.2. *Suppose the $\Omega \subset R^3$ is a bounded domain of class C^3. Let r, $\mathbf{V}$ be a weak solution to problem (1.1)–(1.6) in $(0,T)\times\Omega$, originating from the initial data*

$$r_0 \in C^3(\overline{\Omega}), \ r_0 > 0 \ \textit{in } \overline{\Omega}, \tag{2.5}$$

$$\mathbf{V}_0 \in C^3(\overline{\Omega}; R^3), \tag{2.6}$$

satisfying the compatibility conditions

$$\mathbf{V}_0|_{\partial\Omega} = 0, \ \nabla_x p(r_0)|_{\partial\Omega} = \operatorname{div}_x \mathbb{S}(\nabla_x \mathbf{V}_0)|_{\partial\Omega}, \tag{2.7}$$

and such that

$$0 \leq r \leq \overline{r} \ \textit{a.a. in } (0,T)\times\Omega. \tag{2.8}$$

Then r, $\mathbf{V}$ is a classical solution satisfying the bounds:

$$\|1/r\|_{C([0,T]\times\overline{\Omega})} + \|r\|_{C^1([0,T]\times\overline{\Omega})} + \|\partial_t \nabla_x r\|_{C([0,T];L^6(\Omega;R^3))} + \|\partial_{t,t}^2 r\|_{C([0,T];L^6(\Omega))} \leq D, \tag{2.9}$$

$$\|\mathbf{V}\|_{C^1([0,T]\times\overline{\Omega};R^3)} + \|\mathbf{V}\|_{C([0,T];C^2(\overline{\Omega};R^3))} + \|\partial_t \nabla_x \mathbf{V}\|_{C([0,T];L^6(\Omega;R^{3\times 3}))} + \|\partial_{t,t}^2 \mathbf{V}\|_{L^2(0,T;L^6(\Omega))} \leq D, \tag{2.10}$$

where D depends on Ω, T, $\overline{r}$, and the initial data r_0, $\mathbf{V}_0$ (via $\|(r_0,\mathbf{V}_0)\|_{C^3(\overline{\Omega};R^4)}$ and $\min_{x\in\overline{\Omega}} r_0(x)$).

Proof. The proof will be carried over in several steps.

Step 1.

According to Cho *et al.* [4], problem (1.1)–(1.6) admits a strong solution unique in the class

$$r \in C([0,T_M); W^{1,6}(\Omega)), \ \partial_t r \in C([0,T_M); L^6(\Omega)), \ 1/r \in L^\infty(Q_T), \tag{2.11}$$

$$\mathbf{V} \in C([0,T_M]; W^{2,2}(\Omega;R^3)) \cap L^2(0,T_M; W^{2,6}(\Omega;R^3)), \ \partial_t \mathbf{V} \in L^2(0,T_M; W_0^{1,2}(\Omega;R^3)). \tag{2.12}$$

defined on a time interval $[0, T_M)$, where $T_M > 0$ is finite or infinite and depends on the initial data. Moreover, for any $T_M^* < T_M$, there is a constant $c = c(T_M^*)$ such that

$$\begin{aligned} &\|r\|_{L^\infty(0,T_M^*;W^{1,6}(\Omega))} + \|\partial_t r\|_{L^\infty(0,T_M^*;L^6(\Omega))} + \|1/r\|_{L^\infty(Q_T)} \\ &+\|\mathbf{V}\|_{L^\infty(0,T_M^*;W^{2,2}(\Omega;R^3))} + \|\mathbf{V}\|_{L^2(0,T_M^*;W^{2,6}(\Omega;R^3))} + \|\partial_t \mathbf{V}\|_{L^2(0,T_M^*;W^{1,2}(\Omega))} \\ &\leq c\left(\|r_0\|_{W^{1,6}(\Omega)} + \|\mathbf{V}_0\|_{W^{2,2}(\Omega)}\right). \end{aligned} \tag{2.13}$$

Step 2.
By virtue of the weak-strong uniqueness result stated in ([13], Thm. 4.1) (see also [14], Thm. 4.6), the weak solution r, $\mathbf{V}$ coincides on the time interval $[0, T_M)$ with the strong solution, the existence of which is claimed in the previous step. According to Sun *et al.* ([41], Thm. 1.3) if $T_M < \infty$ then

$$\limsup_{t \to T_M -} \|r(t)\|_{L^\infty(\Omega)} = \infty.$$

Since (2.8) holds, we infer that $T_M = T$. At this point we conclude that couple $(r, \mathbf{V})$ possesses regularity (2.11) and (2.12) and that that the bound (2.13) holds with c dependent solely on T.

Step 3.
Since the initial data enjoy the regularity and compatibility conditions stated in (2.5)–(2.7), a straightforward bootstrap argument gives rise to better bounds, specifically, the solution belongs to the Valli–Zajaczkowski (see [43], Thm. 2.5) class

$$r \in C([0,T]; W^{3,2}(\Omega)),\ \partial_t r \in L^2(0,T; W^{2,2}(\Omega)), \tag{2.14}$$

$$\mathbf{V} \in C([0,T]; W^{3,2}(\Omega)) \cap L^2(0,T; W^{4,2}(\Omega; R^3)),\ \partial_t \mathbf{V} \in L^2(0,T; W^{2,2}(\Omega; R^3)), \tag{2.15}$$

where, similarly to the previous step, the norms depend only on the initial data, $\overline{r}$, and T.

Step 4.
We write equation (1.2) in the form

$$\partial_t \mathbf{V} - \frac{1}{r}\mathrm{div}_x \mathbb{S}(\nabla_x \mathbf{V}) = -\mathbf{V} \cdot \nabla_x \mathbf{V} + \frac{1}{r}\nabla_x p(r), \tag{2.16}$$

where, by virtue of (2.15) and a simple interpolation argument, $\mathbf{V} \in C^{1+\nu}([0,T] \times \overline{\Omega}; R^{3\times 3})$, and, by the same token $r \in C^{1+\nu}([0,T] \times \overline{\Omega})$ for some $\nu > 0$. Consequently, by means of the standard theory of parabolic equations, see for instance Ladyzhenskaya *et al.* [37], we may infer that r, $\mathbf{V}$ is a classical solution,

$$\partial_t \mathbf{V},\ \nabla_x^2 \mathbf{V} \text{ Hölder continuous in } [0,T] \times \overline{\Omega}. \tag{2.17}$$

and, going back to (1.1),

$$\partial_t r \text{ Hölder continuous in } [0,T] \times \overline{\Omega}. \tag{2.18}$$

Step 5.
We write

$$\nabla_x \partial_t r = -\nabla_x \mathbf{V} \cdot \nabla_x r - \mathbf{V} \cdot \nabla_x^2 r - \nabla_x r \mathrm{div}_x \mathbf{V} - r \nabla_x \mathrm{div}_x \mathbf{V};$$

whence, by virtue (2.14), (2.17), (2.18), and the Sobolev embedding $W^{1,2} \hookrightarrow L^6$,

$$\partial_t r \in C([0,T]; W^{1,6}(\Omega)). \tag{2.19}$$

Next, we differentiate (2.16) with respect to t. Denoting $\mathbf{Z} = \partial_t \mathbf{V}$ we therefore obtain

$$\partial_t \mathbf{Z} - \frac{1}{r}\mathrm{div}_x \mathbb{S}(\nabla_x \mathbf{Z}) + \mathbf{V}\cdot\nabla_x \mathbf{Z} = \partial_t\left(\frac{1}{r}\right)\mathrm{div}_x \mathbb{S}(\nabla_x \mathbf{V}) - \partial_t \mathbf{V}\cdot\nabla_x \mathbf{V} + \partial_t\left(\frac{1}{r}\nabla_x p(r)\right), \tag{2.20}$$

where, in view of (2.19) and the previously established estimates, the expression on the right-hand side is bounded in $C([0,T];L^6(\Omega;R^3))$. Thus using the L^p-maximal regularity (see Denk *et al.* [5], Krylov [36] or Danchin [8], Thm. 2.2), we deduce that

$$\partial^2_{t,t}\mathbf{V} = \partial_t \mathbf{Z} \in L^2(0,T;L^6(\Omega;R^3)),\ \partial_t \mathbf{V} = \mathbf{Z} \in C([0,T];W^{1,6}(\Omega;R^3)). \tag{2.21}$$

Finally, writing

$$\partial^2_{t,t} r = -\partial_t \mathbf{V}\cdot\nabla_x r - \mathbf{V}\cdot\partial_t\nabla_x r - \partial_t r \mathrm{div}_x \mathbf{V} - r\partial_t \mathrm{div}_x \mathbf{V},$$

and using (2.19), (2.21), we obtain the desired conclusion

$$\partial^2_{t,t} r \in C([0,T];L^6(\Omega)). \qquad \square$$

Here and hereafter, we shall use notation $a \stackrel{<}{\sim} b$ and $a \approx b$. the symbol $a \stackrel{<}{\sim} b$ means that there exists $c = c(\Omega, T, \mu, \gamma) > 0$ such that $a \leq cb$; $a \approx b$ means $a \stackrel{<}{\sim} b$ and $b \stackrel{<}{\sim} a$.

2.2. Extension lemma

Lemma 2.3. *Under the hypotheses of Proposition 2.2, the functions r and $\mathbf{V}$ can be extended outside Ω in such a way that:*

(1) *The extended functions (still denoted by r and $\mathbf{V}$) are such that $\mathbf{V}$ is compactly supported in $[0,T]\times\mathbb{R}^3$ and $r \geq \underline{r} > 0$.*

(2)

$$\|\mathbf{V}\|_{C^1([0,T]\times R^3;R^3)} + \|\mathbf{V}\|_{C([0,T];C^2(R^3;R^3))} + \|\partial_t\nabla_x \mathbf{V}\|_{C([0,T];L^6(R^3;R^{3\times 3}))} + \|\partial^2_{t,t}\mathbf{V}\|_{L^2(0,T;L^6(R^3))} \tag{2.22}$$

$$\stackrel{<}{\sim} \|\mathbf{V}\|_{C^1([0,T]\times\overline{\Omega};R^3)} + \|\mathbf{V}\|_{C([0,T];C^2(\overline{\Omega};R^3))} + \|\partial_t\nabla_x \mathbf{V}\|_{C([0,T];L^6(\Omega;R^{3\times 3}))} + \|\partial^2_{t,t}\mathbf{V}\|_{L^2(0,T;L^6(\Omega))};$$

(3)

$$\|r\|_{C^1([0,T]\times R^3)} + \|\partial_t\nabla_x r\|_{C([0,T];L^6(R^3;R^3))} + \|\partial^2_{t,t} r\|_{C([0,T];L^6(R^3))} \tag{2.23}$$

$$\stackrel{<}{\sim} \|r\|_{C^1([0,T]\times\overline{\Omega})} + \|\partial_t\nabla_x r\|_{C([0,T];L^6(\Omega;R^3))} + \|\partial^2_{t,t} r\|_{C([0,T];L^6(\Omega))} +$$

$$\|\mathbf{V}\|_{C^1([0,T]\times\overline{\Omega};R^3)} + \|\mathbf{V}\|_{C([0,T];C^2(\overline{\Omega};R^3))}\| + \partial_t\nabla_x \mathbf{V}\|_{C([0,T];L^6(\Omega;R^{3\times 3}))} + \|\partial^2_{t,t}\mathbf{V}\|_{L^2(0,T;L^6(\Omega))};$$

(4)

$$\partial_t r + \mathrm{div}_x(r\mathbf{V}) = 0 \ \textit{in}\ (0,T)\times R^3. \tag{2.24}$$

Proof. We first construct the extension of the vector field $\mathbf{V}$. To this end, we follow the standard construction in the flat domain, see Adams ([1], Chap. 5, Thm. 5.22) and combine it with the standard procedure of 'flattening' of the boundary and the partition of unity technique, we get (2.22) Once this is done, we solve on the whole space the transport equation (2.24). It is easy to show that the unique solution r of this equation possesses regularity and estimates stated in (2.23). $\square$

Remark 2.4. Here and hereafter, we denote $X_T(\mathbb{R}^3)$ a subset of $L^2((0,T)\times\mathbb{R}^3)$ of couples $(r,\mathbf{V})$, $r>0$ with finite norm

$$\|(r,\mathbf{V})\|_{X_T(\mathbb{R}^3)} \equiv \|r\|_{C^1([0,T]\times R^3)} + \|\partial_t\nabla_x r\|_{C([0,T];L^6(R^3;R^3))} + \|\partial^2_{t,t} r\|_{C([0,T];L^6(R^3))} \tag{2.25}$$

$$\|\mathbf{V}\|_{C^1([0,T]\times R^3;R^3)} + \|\mathbf{V}\|_{C([0,T];C^2(R^3;R^3))} + \|\partial_t\nabla_x\mathbf{V}\|_{C([0,T];L^6(R^3;R^{3\times 3}))} + \|\partial^2_{t,t}\mathbf{V}\|_{L^2(0,T;L^6(R^3))}$$

We notice that if r, $\mathbf{V}$ are interrelated through (2.7), then the first component of the couple belonging to $X_T(\mathbb{R}^3)$ is always strictly positive on $[0,T]\times R^3$. We set

$$0<\underline{r}=\min_{(t,x)\in[0,T]\times\mathbb{R}^3} r(t,x), \quad \overline{r}=\max_{(t,x)\in[0,T]\times\mathbb{R}^3} r(t,x)<\infty \tag{2.26}$$

2.3. Physical domain, mesh approximation

The physical space is represented by a bounded domain $\Omega\subset R^3$ of class C^3. The numerical domains Ω_h are polyhedral domains,

$$\overline{\Omega}_h = \cup_{K\in\mathcal{T}} K, \tag{2.27}$$

where $\mathcal{T}$ is a set of tetrahedra which have the following property: If $K\cap L\neq\emptyset$, $K\neq L$, then $K\cap L$ is either a common face, or a common edge, or a common vertex. By $\mathcal{E}(K)$, we denote the set of the faces σ of the element $K\in\mathcal{T}$. The set of all faces of the mesh is denoted by $\mathcal{E}$; the set of faces included in the boundary $\partial\Omega_h$ of Ω_h is denoted by $\mathcal{E}_{\rm ext}$ and the set of internal faces (*i.e.* $\mathcal{E}\setminus\mathcal{E}_{\rm ext}$) is denoted by $\mathcal{E}_{\rm int}$.

Further, we ask

$$\mathcal{V}_h\in\partial\Omega_h \text{ a vertex } \Rightarrow\ \mathcal{V}_h\in\partial\Omega. \tag{2.28}$$

Furthermore, we suppose that each K is a tetrahedron such that

$$\xi[K]\approx \mathrm{diam}[K]\approx h, \tag{2.29}$$

where $\xi[K]$ is the radius of the largest ball contained in K.

The properties of this mesh needed in the sequel are formulated in the following lemma, whose proof is left to the reader, see Johnson and Nedelec [27] for the $2D$ case, and [26] for the general $3D$ case.

Lemma 2.5. *There exists a positive constant d_Ω depending solely on the geometric properties of $\partial\Omega$ such that*

$$\mathrm{dist}[x,\partial\Omega]\leq d_\Omega h^2,$$

for any $x\in\partial\Omega_h$. Moreover,

$$|(\Omega_h\setminus\Omega)\cup(\Omega\setminus\Omega_h)|\lesssim h^2.$$

We find important to emphasize that $\Omega_h\not\subset\Omega$, in general.

2.4. Numerical spaces

We denote by $Q_h(\Omega_h)$ the space of piecewise constant functions:

$$Q_h(\Omega_h)=\{q\in L^2(\Omega_h)\,|\ \forall K\in\mathcal{T},\ q_{|K}\in\mathbb{R}\}. \tag{2.30}$$

For a function v in $C(\overline{\Omega}_h)$, we set

$$v_K=\frac{1}{|K|}\int_K v\,\mathrm{d}x \text{ for } K\in\mathcal{T} \text{ and } \Pi_h^Q v(x)=\sum_{K\in\mathcal{T}} v_K 1_K(x),\ x\in\Omega. \tag{2.31}$$

Here and in what follows, 1_K is the characteristic function of K.

We define the Crouzeix–Raviart space with 'zero traces':

$$V_{h,0}(\Omega_h) = \{v \in L^2(\Omega_h),\ \forall K \in \mathcal{T},\ v_{|K} \in \mathbb{P}_1(K), \tag{2.32}$$

$$\forall \sigma \in \mathcal{E}_{\rm int},\ \sigma = K|L,\ \int_\sigma v_{|K}\,{\rm d}S = \int_\sigma v_{|L}\,{\rm d}S, \quad \forall \sigma' \in \mathcal{E}_{\rm ext},\ \int_{\sigma'} v\,{\rm d}S = 0\},$$

and 'with general traces'

$$V_h(\Omega_h) = \{v \in L^2(\Omega),\ \forall K \in \mathcal{T},\ v_{|K} \in \mathbb{P}_1(K),\ \forall \sigma \in \mathcal{E}_{\rm int},\ \sigma = K|L,\ \int_\sigma v_{|K}\,{\rm d}S = \int_\sigma v_{|L}\,{\rm d}S\}. \tag{2.33}$$

We denote by Π_h^V the standard Crouzeix–Raviart projection, and $\Pi_{h,0}^V$ the Crouzeix–Raviart projection with 'zero trace', specifically,

$$\Pi_h^V : C(\overline{\Omega}_h) \to V_h(\Omega_h),\ \int_\sigma \Pi_h^V[\phi]\ {\rm dS}_x = \int_\sigma \phi\ {\rm dS}_x \text{ for all } \sigma \in \mathcal{E},$$

$$\Pi_{h,0}^V : C(\overline{\Omega}_h) \to V_{h,0}(\Omega_h),\ \int_\sigma \Pi_{h,0}^V[\phi]\ {\rm dS}_x = \int_\sigma \phi\ {\rm dS}_x \text{ for all } \sigma \in \mathcal{E}_{\rm int}, \tag{2.34}$$

$$\int_\sigma \Pi_{h,0}^V[\phi]\ {\rm dS}_x = 0 \text{ whenever } \sigma \in \mathcal{E}_{\rm ext}.$$

If $v \in W^{1,1}(\Omega_h)$, we set

$$v_\sigma = \frac{1}{|\sigma|}\int_\sigma v {\rm d}S \text{ for } \sigma \in \mathcal{E}. \tag{2.35}$$

(See *e.g.* [9], Sect. 4.3) for the definition of traces of functions in $W^{1,1}$.)

Each element $v \in V_h(\Omega_h)$ can be written in the form

$$v(x) = \sum_{\sigma\in\mathcal{E}} v_\sigma \varphi_\sigma(x), \quad x \in \Omega_h, \tag{2.36}$$

where the set $\{\varphi_\sigma\}_{\sigma\in\mathcal{E}} \subset V_h(\Omega_h)$ is the classical Crouzeix–Raviart basis determined by

$$\forall (\sigma,\sigma') \in \mathcal{E}^2,\ \frac{1}{|\sigma'|}\int_{\sigma'} \varphi_\sigma\,{\rm d}S = \delta_{\sigma,\sigma'}. \tag{2.37}$$

Similarly, each element $v \in V_{h,0}(\Omega_h)$ can be written in the form

$$v(x) = \sum_{\sigma\in\mathcal{E}_{\rm int}} v_\sigma \varphi_\sigma(x), \quad x \in \Omega_h. \tag{2.38}$$

We first recall in Lemmas 2.6)–(2.10 the standard properties of the projection Π_h^V. The collection of their proofs in the requested generality can be found in the Appendix of [22] with exception of Lemma 2.11 and its Corollary 2.12. We refer to the monograph of Brezzi and Fortin [2], the Crouzeix's and Raviart's paper [6], Gallouet *et al.* [21] for the original versions of some of these proofs. We present the proof of Lemma 2.11 dealing with the comparison of projections Π_h^V and $\Pi_{h,0}^V$ that we did not find in the literature.

Lemma 2.6. *The following estimates hold true:*

$$\|\Pi_h^V[\phi]\|_{L^\infty(K)} + \|\Pi_{h,0}^V[\phi]\|_{L^\infty(K)} \stackrel{<}{\sim} \|\phi\|_{L^\infty(K)}, \tag{2.39}$$

for all $K \in \mathcal{T}$ and $\phi \in C(K)$;

$$\|\phi - \Pi_h^V[\phi]||_{L^p(K)} \lesssim h^s \|\nabla^s \phi\|_{L^p(K;\mathbb{R}^{d^s})},\ s = 1, 2,\ 1 \le p \le \infty, \tag{2.40}$$

and

$$||\nabla(\phi - \Pi_h^V[\phi])||_{L^p(K;\mathbb{R}^d)} \le c h^{s-1} \|\nabla^s \phi\|_{L^p(K;\mathbb{R}^{d^s})},\ s = 1, 2,\ 1 \le p \le \infty, \tag{2.41}$$

for all $K \in \mathcal{T}$ and $\phi \in C^s(K)$.

Lemma 2.7. *Let $1 \le p < \infty$. Then*

$$\sum_{\sigma \in \mathcal{E}} |\sigma| h |v_\sigma|^p \approx ||v||^p_{L^p(\Omega_h)}, \tag{2.42}$$

with any $v \in V_h(\Omega_h)$.

Lemma 2.8. *The following Sobolev-type inequality holds true:*

$$||v||^2_{L^6(\Omega_h)} \lesssim \sum_{K \in \mathcal{T}} \int_K |\nabla_x v|^2 \mathrm{d}x, \tag{2.43}$$

with any $v \in V_{h,0}(\Omega_h)$.

Lemma 2.9. *There holds:*

$$\sum_{K \in \mathcal{T}} \int_K q\ \operatorname{div} \Pi_h^V[\boldsymbol{v}]\, \mathrm{d}x = \int_\Omega q\ \operatorname{div} \boldsymbol{v}\, \mathrm{d}x, \tag{2.44}$$

for all $\boldsymbol{v} \in C^1(\overline{\Omega}_h, \mathbb{R}^d)$ and all $q \in Q_h(\Omega_h)$.

Lemma 2.10 (Jumps over faces in the Crouzeix–Raviart space)**.** *For all $v \in V_{h,0}(\Omega_h)$ there holds*

$$\sum_{\sigma \in \mathcal{E}} \frac{1}{h} \int_\sigma [v]^2_{\sigma, \boldsymbol{n}_\sigma}\, \mathrm{d}S \lesssim \sum_{K \in \mathcal{T}} \int_K |\nabla_x v|^2 \mathrm{d}x, \tag{2.45}$$

where $[v]_{\sigma,\boldsymbol{n}_\sigma}$ is a jump of v with respect to a normal $\boldsymbol{n}_\sigma$ to the face σ,

$$\forall x \in \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \quad [v]_{\sigma,\boldsymbol{n}_\sigma}(x) = \begin{cases} v|_K(x) - v|_L(x) \text{ if } \boldsymbol{n}_\sigma = \boldsymbol{n}_{\sigma,K} \\ v|_L(x) - v|_K(x) \text{ if } \boldsymbol{n}_\sigma = \boldsymbol{n}_{\sigma,L}, \end{cases}$$

($\boldsymbol{n}_{\sigma,K}$ is the normal of σ, that is outer w.r. to element K) and

$$\forall x \in \sigma \in \mathcal{E}_{\mathrm{ext}}, \quad [v]_{\sigma,\boldsymbol{n}_\sigma}(x) = v(x),\ \text{with } \boldsymbol{n}_\sigma \text{ an exterior normal to } \partial\Omega.$$

We will need to compare the projections Π_h^V and $\Pi_{h,0}^V$. Clearly they coincide on 'interior' elements meaning $K \in \mathcal{T}$, $K \cap \partial\Omega_h = \emptyset$. We have the following lemma for the tetrahedra with non void intersection with the boundary.

Lemma 2.11. *We have*

$$\|\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi]\|_{L^\infty(K)} + h\|\nabla_x(\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi])\|_{L^\infty(K;R^3)} \lesssim \sup_{\sigma \subset K \cap \partial\Omega_h} \|\phi\|_{L^\infty(\sigma)} \text{ if } K \in \mathcal{T},\ K \cap \partial\Omega_h \neq \emptyset, \tag{2.46}$$

for any $\phi \in C(K)$.

Proof. We recall the Crouzeix–Raviart basis (2.37) and the fact that Π_h^V and $\Pi_{h,0}^V$ differ only in basis functions corresponding to $\sigma \in \mathcal{E}_{\rm ext}$. We have

$$\|\Pi_h^V[\phi] - \Pi_{h,0}^v[\phi]\|_{L^\infty(K)} \leq \left\| \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm ext}} \varphi_\sigma \frac{1}{|\sigma|} \int_\sigma \phi \, {\rm d}S \right\|_{L^\infty(K)} \leq c(K) \cdot \sup_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm ext}} \|\phi\|_{L^\infty(\sigma)}, \tag{2.47}$$

and

$$\begin{aligned} h\|\nabla_x(\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi])\|_{L^\infty(K)} &\leq h \left\| \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm ext}} \nabla_x \varphi_\sigma \frac{1}{|\sigma|} \int_\sigma \phi \, {\rm d}S \right\|_{L^\infty(K)} \\ &\leq ch \sup_{\sigma \subseteq K \cap \partial\Omega_h} \|\phi\|_{L^\infty(\sigma)} \left\| \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm ext}} \nabla_x \varphi_\sigma \right\|_{L^\infty(K)}. \end{aligned}$$

The proof is completed by $\| \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm ext}} \nabla_x \varphi_\sigma \|_{L^\infty(K)} \leq c(K) h^{-1}$. □

In fact, in the derivation of the error estimates we will use the consequence of the above observations formulated in the following two corollaries.

Corollary 2.12. *Let $\phi \in C^1(R^3)$ such that $\phi|_{\partial\Omega} = 0$. Then we have,*

$$\|\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi]\|_{L^\infty(K)} = 0 \text{ if } K \in \mathcal{T},\ K \cap \partial_{\Omega_h} = \emptyset, \tag{2.48}$$

$$\|\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi]\|_{L^\infty(K)} + h\|\nabla_x(\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi])\|_{L^\infty(K;R^3)} \stackrel{<}{\sim} h^2 \|\nabla_x \phi\|_{L^\infty(R^3;R^3)}, \tag{2.49}$$

if $K \in \mathcal{T}_h$, $K \cap \partial\Omega_h \neq \emptyset$, $\partial K \not\subset \partial\Omega$.

Proof. Relation (2.48) follows immediately from (2.46), as there is an empty sum on the right hand side for 'interior' elements ($K \cap \partial\Omega_h = \emptyset$).

For any $x \in \partial\Omega_h$ there exists $y \in \partial\Omega$ (and thus $\phi(y) = 0$) such that

$$|\phi(x)| \leq {\rm dist}[x, y] \|\nabla_x \phi\|_{L^\infty(R^3;R^3)} \stackrel{<}{\sim} h^2 \|\nabla_x \phi\|_{L^\infty(R^3;R^3)}, \tag{2.50}$$

where we used Lemma 2.5 for the latter inequality. The proof is completed by taking supremum over $K \in \mathcal{T}_h$ and combining with (2.50). Note that the mesh regularity property (2.29) supplies a uniform estimate of constants $c(K)$ from the previous lemma, which enables to write the latter inequality in (2.50). □

Corollary 2.13. *For any $\phi \in C(R^3)$,*

$$\|\Pi_h^V[\phi] - \Pi_{h,0}^V[\phi]\|_{L^p(\Omega_h)} \stackrel{<}{\sim} h^{1/p} \|\phi\|_{L^\infty(\Omega_h)},\ 1 \leq p < \infty. \tag{2.51}$$

Proof. Apply inverse estimates (see *e.g.* [31], Lem. 2.9) to (2.46). □

We will frequently use the Poincaré, Sobolev and interpolation inequalities on tetrahedra reported in the following lemma.

Lemma 2.14.

(1) *We have,*

$$\|v - v_K\|_{L^p(K)} \lesssim h\|\nabla v\|_{L^p(K)}, \tag{2.52}$$

$$\forall \sigma \in \mathcal{E}(K),\ \|v - v_\sigma\|_{L^p(K)} \lesssim h\|\nabla v\|_{L^p(K)}, \tag{2.53}$$

for any $v \in W^{1,p}(K)$, *where* $1 \le p \le \infty$.

(2) *There holds*

$$\|v - v_K\|_{L^{p^*}(K)} \lesssim \|\nabla v\|_{L^p(K)}, \tag{2.54}$$

$$\forall \sigma \in \mathcal{E}(K),\ \|v - v_\sigma\|_{L^{p^*}(K)} \lesssim \|\nabla v\|_{L^p(K)}, \tag{2.55}$$

for any $v \in W^{1,p}(K)$, $1 \le p < d$, *where* $p^* = \frac{dp}{d-p}$.

(3) *We have,*

$$\|v - v_K\|_{L^q(K)} \le ch^\beta \|\nabla v\|_{L^p(K;\mathbb{R}^d)}, \tag{2.56}$$

$$\|v - v_\sigma\|_{L^q(K)} \le ch^\beta \|\nabla v\|_{L^p(K;\mathbb{R}^d)}, \tag{2.57}$$

for any $v \in W^{1,p}(K)$, $1 \le p < d$, *where* $\frac{1}{q} = \frac{\beta}{p} + \frac{1-\beta}{p^*}$, $p \le q \le p^*$.

We finish the section of preliminaries by recalling two algebraic inequalities: the 'imbedding' inequality

$$\left(\sum_{i=1}^{L} |a_i|^p\right)^{1/p} \le \left(\sum_{i=1}^{L} |a_i|^q\right)^{1/q}, \tag{2.58}$$

for all $a = (a_1, \dots, a_L) \in R^L$, $1 \le q \le p < \infty$ and the discrete Hölder inequality

$$\sum_{i=1}^{L} |a_i||b_i| \le \left(\sum_{i=1}^{L} |a_i|^q\right)^{1/q} \left(\sum_{i=1}^{L} |a_i|^p\right)^{1/p}, \tag{2.59}$$

for all $a = (a_1, \dots, a_L) \in R^L$, $b = (b_1, \dots, b_L) \in R^L$, $\frac{1}{q} + \frac{1}{p} = 1$.

3. Main result

Here and hereafter we systematically use the following abbreviated notation:

$$\hat{\phi} = \Pi_h^Q[\phi],\ \ \phi_h = \Pi_h^V[\phi],\ \ \phi_{h,0} = \Pi_{h,0}^V[\phi], \tag{3.1}$$

where projections Π_h^Q, Π_h^V and $\Pi_{h,0}^V$ are defined in (2.31) and (2.34). For a function $v \in C([0,T], L^1(\Omega))$ we set

$$v^n(x) = v(t_n, x), \tag{3.2}$$

where $t_0 = 0 < t_1 < \ldots < t_{n-1} < t_n < t_{n+1} < \ldots t_N = T$ is a partition of the interval $[0,T]$. Finally, for a function $v \in V_h(\Omega_h)$ we denote

$$\nabla_h v(x) = \sum_{K \in \mathcal{T}} \nabla_x v(x) 1_K(x),\ \ \mathrm{div}_h \mathbf{v}(x) = \sum_{K \in \mathcal{T}} \mathrm{div}_x \mathbf{v}(x) 1_K(x). \tag{3.3}$$

In order to ensure the positivity of the approximate densities, we shall use an upwinding technique for the density in the mass equation. For $q \in Q_h(\Omega_h)$ and $\boldsymbol{u} \in \mathbf{V}_{h,0}(\Omega_h; \mathbb{R}^3)$, the upwinding of q with respect to $\boldsymbol{u}$ is defined, for $\sigma = K|L \in \mathcal{E}_{\rm int}$ by:

$$q_\sigma^{\rm up} = \begin{cases} q_K \text{ if } \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{\sigma,K} > 0 \\ q_L \text{ if } \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{\sigma,K} \leq 0, \end{cases} \tag{3.4}$$

and we denote

$$\mathrm{Up}_K(q, \mathbf{u}) \equiv \sum_{\sigma \in \mathcal{E}(K) c \cap \mathcal{E}_{\rm int}} q_\sigma^{\rm up} \boldsymbol{u}_\sigma \cdot \mathbf{n}_{\sigma,K} = \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\rm int}} \left(q_K [\boldsymbol{u}_\sigma \cdot \mathbf{n}_{\sigma,K}]^+ + q_L [\boldsymbol{u}_\sigma \cdot \mathbf{n}_{\sigma,K}]^- \right),$$

where $a^+ = \max(a, 0)$, $a^- = \min(a, 0)$.

3.1. Numerical scheme

We consider a couple $(\varrho^n, \boldsymbol{u}^n) = (\varrho^{n,(\Delta t,h)}, \boldsymbol{u}^{n,(\Delta t,h)})$ *of (numerical) solutions of the following algebraic system (numerical scheme):*

$$\varrho^n \in Q_h(\Omega_h), \ \varrho^n > 0, \ \mathbf{u}^n \in V_{h,0}(\Omega_h; R^3), \quad n = 0, 1, \dots, N, \tag{3.5}$$

$$\sum_{K \in \mathcal{T}} |K| \frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} \phi_K + \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,{\rm up}} (\mathbf{u}_\sigma^n \cdot \mathbf{n}_{\sigma,K}) \phi_K = 0 \text{ for any } \phi \in Q_h(\Omega_h) \text{ and } n = 1, \dots, N, \tag{3.6}$$

$$\begin{aligned} &\sum_{K \in \mathcal{T}} \frac{|K|}{\Delta t} \left(\varrho_K^n \hat{\boldsymbol{u}}_K^n - \varrho_K^{n-1} \hat{\boldsymbol{u}}_K^{n-1} \right) \cdot \boldsymbol{v}_K + \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,{\rm up}} \hat{\boldsymbol{u}}_\sigma^{n,{\rm up}} [\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}] \cdot \boldsymbol{v}_K \\ &- \sum_{K \in \mathcal{T}} p(\varrho_K^n) \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \boldsymbol{v}_\sigma \cdot \mathbf{n}_{\sigma,K} + \mu \sum_{K \in \mathcal{T}} \int_K \nabla \boldsymbol{u}^n : \nabla \boldsymbol{v} \ \mathrm{d}x \\ &+ \frac{\mu}{3} \sum_{K \in \mathcal{T}} \int_K \mathrm{div} \boldsymbol{u}^n \mathrm{div} \boldsymbol{v} \ \mathrm{d}x = 0, \text{ for any } \boldsymbol{v} \in V_{h,0}(\Omega; R^3) \text{ and } n = 1, \dots, N. \end{aligned} \tag{3.7}$$

The numerical solutions depend on the size h of the space discretization and on the time step Δt. For the sake of clarity and in order to simplify notation we will always systematically write in all formulas $(\varrho^n, \boldsymbol{u}^n)$ instead of $(\varrho^{n,(\Delta t,h)}, \boldsymbol{u}^{n,(\Delta t,h)})$.

The numerical method (3.5)–(3.7) has been suggested in ([31], Def. 3.1); it is *strongly nonlinear and implicit.* It is therefore not a trivial question whether this (finite dimensional) problem admits a solution. The problem of the well posedness of this numerical scheme is investigated in Karper ([31], Prop. 3.3). Karper's result states that:

For each fixed $h > 0$, $\Delta t > 0$, *problem* (3.5)–(3.7) *admits a solution* $(\varrho_h^n, \mathbf{u}_h^n)$*:*

$$\varrho_h^n \in Q_h(\Omega_h), \ \mathbf{u}_h^n \in V_{h,0}(\Omega_h; R^3), \ n = 0, 1, \dots, N,$$

and $\varrho_h^n > 0$, $n = 1, \dots, N$, *provided* $\varrho_h^0 > 0$.

The proof uses topological degree theory in the spirit suggested in [20]. All its details are available in Section 11 of [31]. Notice that the above result does not guarantee the *uniqueness* of numerical solutions.

Remark 3.1. Throughout the paper, $q_\sigma^{\rm up}$ is defined in (3.4), where $\mathbf{u}$ is the numerical solution constructed in (3.5)–(3.7).

3.2. Error estimates

The main result of this paper is announced in the following theorem:

Theorem 3.2. *Let $\Omega \subset R^3$ be a bounded domain of class C^3 and let the pressure satisfy (1.4) with $\gamma \geq 3/2$. Let $\{\varrho^n, \mathbf{u}^n\}_{0\leq n\leq N}$ be a family of numerical solutions resulting from the scheme (3.5)–(3.7). Moreover, suppose there are initial data $[r_0, \mathbf{V}_0]$ belonging to the regularity class specified in Proposition 2.2 and giving rise to a weak solution $[r, \mathbf{V}]$ to the initial-boundary value problem (1.1)–(1.6) in $(0,T)\times\Omega$ satisfying*

$$0 \leq r(t,x) \leq \overline{r} \text{ a.a. in } (0,T)\times\Omega.$$

Then $[r, \mathbf{V}]$ is regular and there exists a positive number

$$C = C\Big(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1[\underline{r},\overline{r}]}, \|(\partial_t r, \nabla r, \boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V}, \nabla^2 \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{45})},$$
$$\|\partial_t^2 r\|_{L^1(0,T;L^{\gamma'}(\Omega))}, \|\partial_t \nabla r\|_{L^2(0,T;L^{6\gamma/5\gamma-6}(\Omega;\mathbb{R}^3))}, \|\partial_t^2 \boldsymbol{V}, \partial_t \nabla \boldsymbol{V}\|_{L^2(0,T;L^{6/5}(\Omega;\mathbb{R}^{12}))},\Big)$$

such that

$$\sup_{1\leq n\leq N} \int_{\Omega\cap\Omega_h} \left[\frac{1}{2}\varrho^n|\hat{\mathbf{u}}^n - \mathbf{V}(t_n,\cdot)|^2 + H(\varrho^n) - H'(r(t_n,\cdot))(\varrho^n - r(t_n,\cdot)) - H(r(t_n))\right] \,\mathrm{d}x \tag{3.8}$$
$$+\Delta t \sum_{1\leq n\leq N} \int_{\Omega\cap\Omega_h} |\nabla_h \mathbf{u}^n - \nabla_x \mathbf{V}(t_n,\cdot)|^2 \,\mathrm{d}x$$
$$\leq C\left(\sqrt{\Delta t} + h^a + \int_{\Omega\cap\Omega_h} \left[\frac{1}{2}\varrho^0|\hat{\mathbf{u}}^0 - \mathbf{V}_0|^2 + H(\varrho^0) - H'(r_0)(\varrho^0 - r_0) - H(r_0)\right] \,\mathrm{d}x\right),$$

where

$$a = \frac{2\gamma - 3}{\gamma} \text{ if } \tfrac{3}{2} \leq \gamma \leq 2, \quad a = \frac{1}{2} \text{ otherwise.} \tag{3.9}$$

Note that for $\gamma = 3/2$ Theorem 3.2 gives only uniform bounds on the difference of exact and numerical solution, not the convergence.

Remark 3.3. The constitutive assumptions for the pressure (1.4) in Theorem 3.2 require, in particular, $p'(0) > 0$. This condition excludes the isentropic pressure laws

$$p(\varrho) = \varrho^\gamma, \ \gamma > 1. \tag{3.10}$$

Nevertheless, Theorem 3.2 *holds under the same assumptions also for the isentropic pressure laws* (3.10). Here, we have adopted the more restrictive condition (1.4) (in particular $p'(0) > 0$) only for the sake of simplicity and clarity, in order to avoid some unnecessary technical difficulties. It allows to simplify proofs of some estimates: for example estimates (4.7), (4.10) are in this case immediate consequences of the energy inequality (4.2), while in the general case of pressure laws vanishing at 0, the derivation of the same estimates requires more effort (see [22], Cor. 4.1 and Lem. 4.2), where the proofs of these estimates are performed in the general case.

4. Uniform estimates

If we take $\phi = 1$ in formula (3.6) we get immediately the conservation of mass:

$$\forall n = 1, \dots N, \quad \int_{\Omega_h} \varrho^n \,\mathrm{d}x = \int_{\Omega_h} \varrho^0 \,\mathrm{d}x. \tag{4.1}$$

The next Lemma reports the standard energy estimates for the numerical scheme (3.5)–(3.7). The reader can consult Section 4.1 in Gallouet *et al.* ([22], Lem. 4.1) for its laborious but straightforward proof.

Lemma 4.1. *Let $(\varrho^n, \boldsymbol{u}^n)$ be a solution of the discrete problem (3.5)–(3.7) with the pressure p satisfying (1.4). Then there exist*

$$\overline{\varrho}^n_\sigma \in [\min(\varrho^n_K, \varrho^n_L), \max(\varrho^n_K, \varrho^n_L)],\ \sigma = K|L \in \mathcal{E}_{\rm int},\ n = 1, \dots, N,$$
$$\overline{\varrho}^{n-1,n}_K \in [\min(\varrho^{n-1}_K, \varrho^n_K), \max(\varrho^{n-1}_K, \varrho^n_K)],\ K \in \mathcal{T},\ n = 1, \dots, N,$$

such that

$$\sum_{K\in\mathcal{T}} |K| \left(\frac{1}{2} \varrho^m_K |\boldsymbol{u}^m_K|^2 + H(\varrho^m_K) \right) - \sum_{K\in\mathcal{T}} |K| \left(\frac{1}{2} \varrho^0_K |\boldsymbol{u}^0_K|^2 + H(\varrho^0_K) \right)$$
$$+ \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \left(\mu \int_K |\nabla_x \boldsymbol{u}^n|^2 \,\mathrm{d}x + \frac{\mu}{3} \int_K |\mathrm{div} \boldsymbol{u}^n|^2 \,\mathrm{d}x \right)$$
$$+ [D^{m,|\Delta \boldsymbol{u}|}_{\rm time}] + [D^{m,|\Delta \varrho|}_{\rm time}] + [D^{m,|\Delta \boldsymbol{u}|}_{\rm space}] + [D^{m,|\Delta \varrho|}_{\rm space}] = 0, \quad (4.2)$$

for all $m = 1, \dots, N$, where

$$[D^{m,|\Delta \boldsymbol{u}|}_{\rm time}] = \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| \varrho^{n-1}_K \frac{|\boldsymbol{u}^n_K - \boldsymbol{u}^{n-1}_K|^2}{2}, \quad (4.3a)$$

$$[D^{m,|\Delta \varrho|}_{\rm time}] = \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| H''(\overline{\varrho}^{n-1,n}_K) \frac{|\varrho^n_K - \varrho^{n-1}_K|^2}{2}, \quad (4.3b)$$

$$[D^{m,|\Delta \boldsymbol{u}|}_{\rm space}] = \Delta t \sum_{n=1}^{m} \sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} |\sigma| \varrho^{n,\rm up}_\sigma \frac{(\boldsymbol{u}^n_K - \boldsymbol{u}^n_L)^2}{2} |\boldsymbol{u}^n_\sigma \cdot \boldsymbol{n}_{\sigma,K}|, \quad (4.3c)$$

$$[D^{m,|\Delta \varrho|}_{\rm space}] = \Delta t \sum_{n=1}^{m} \sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} |\sigma| H''(\overline{\varrho}^n_\sigma) \frac{(\varrho^n_K - \varrho^n_L)^2}{2} |\boldsymbol{u}^n_\sigma \cdot \mathbf{n}_{\sigma,K}|. \quad (4.3d)$$

We have the following corollary of Lemma 4.1.

Corollary 4.2. *Under assumptions of Lemma 4.1, we have:*

(1) *There exists $c = c(M_0, E_0) > 0$ (independent of n, h and Δt) such that*

$$k \sum_{n=1}^{N} \int_K |\nabla_x \boldsymbol{u}^n|^2 \,\mathrm{d}x \le c, \quad (4.4)$$

$$k \sum_{n=1}^{N} \|\boldsymbol{u}^n\|^2_{L^6(\Omega_h;\mathbb{R}^3)} \le c, \quad (4.5)$$

$$\sup_{n=0,\dots N} \|\varrho^n |\hat{\boldsymbol{u}}^n|^2\|_{L^1(\Omega_h)} \le c. \quad (4.6)$$

(2)

$$\sup_{n=0,\dots N} \|\varrho^n\|_{L^\gamma(\Omega_h)} \le c, \quad (4.7)$$

(3) *If the pair $(r, \boldsymbol{U})$ belongs to the class (2.25) there is $c = c(M_0, E_0, \underline{r}, \overline{r}, \|\boldsymbol{U}, \nabla \boldsymbol{U}\|_{L^\infty(Q_T;\mathbb{R}^{12})}) > 0$ such that for all $n = 1, \dots, N$,*

$$\sup_{n=0,\dots N} \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n | \hat{r}(t_n), \hat{\mathbf{U}}(t_n)) \le c, \quad (4.8)$$

where

$$\mathcal{E}(\varrho, \boldsymbol{u} | z, \mathbf{v}) = \int_{\Omega_h} \left(\varrho |\boldsymbol{u} - \mathbf{v}|^2 + E(\varrho|z) \right) \mathrm{d}x, \quad E(\varrho|z) = H(\varrho) - H'(z)(\varrho - z) - H(z). \quad (4.9)$$

(4) *There exists* $c = c(M_0, E_0, \underline{r}, |p'|_{C[\underline{r},\overline{r}]}) > 0$ *such that*

$$\Delta t \sum_{n=1}^{m} \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} |\sigma|(\varrho_K^n - \varrho_L^n)^2 \Big[\frac{1_{\{\overline{\varrho}_\sigma^n \geq 1\}}}{[\max\{\varrho_K, \varrho_L\}]^{2-\gamma}} + 1_{\{\overline{\varrho}_\sigma^n < 1\}} \Big] |\boldsymbol{u}_\sigma^n \cdot \mathbf{n}_{\sigma,K}| \leq c \quad \text{if } \gamma \in [1,2), \tag{4.10}$$

$$\Delta t \sum_{n=1}^{m} \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} |\sigma|(\varrho_K^n - \varrho_L^n)^2 \, |\boldsymbol{u}_\sigma^n \cdot \mathbf{n}_{\sigma,K}| \leq c \quad \text{if } \gamma \geq 2$$

Items (1)–(3) of Corollary 4.2 are direct consequences of Lemma 4.1. Item (4) represents the convenient expression for the numerical dissipation (4.3d). The interested reader can consult Section 4.2 in (Gallouet *et al.* [22], Cor. 4.1, Lem. 4.2) for the detailed proofs of these estimates.

5. Discrete relative energy inequality

The starting point of our error analysis is the discrete relative energy inequality for the numerical scheme (3.5)–(3.7) formulated in the following lemma.

Lemma 5.1. *Let* $(\varrho^n, \boldsymbol{u}^n)$ *be a solution of the discrete problem* (3.5)–(3.7) *with the pressure* p *satisfying* (1.4). *Then there holds for all* $m = 1, \dots, N$,

$$\begin{aligned} &\sum_{K\in\mathcal{T}} \frac{1}{2}|K| \left(\varrho_K^m |\boldsymbol{u}_K^m - \boldsymbol{U}_K^m|^2 - \varrho_K^0 |\boldsymbol{u}_K^0 - \boldsymbol{U}_K^0|^2 \right) + \sum_{K\in\mathcal{T}} |K| \left(E(\varrho_K^m | r_K^m) - E(\varrho_K^0 | r_K^0) \right) \\ &\qquad + \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \left(\mu \int_K |\nabla_x (\boldsymbol{u}^n - \boldsymbol{U}^n)|^2 \,\mathrm{d}x + \frac{\mu}{3} \int_K |\mathrm{div}(\boldsymbol{u}^n - \boldsymbol{U}^n)|^2 \,\mathrm{d}x \right) \leq \sum_{i=1}^{6} T_i, \end{aligned} \tag{5.1}$$

for any $0 < r^n \in Q_h(\Omega_h)$, $\boldsymbol{U}^n \in V_{h,0}(\Omega_h; \mathbb{R}^3)$, $n = 1, \dots, N$, *where*

$$\begin{aligned} T_1 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \left(\mu \int_K \nabla_x \boldsymbol{U}^n : \nabla_x (\boldsymbol{U}^n - \boldsymbol{u}^n) \,\mathrm{d}x + \frac{\mu}{3} \int_K \mathrm{div}\boldsymbol{U}^n \mathrm{div}(\boldsymbol{U}^n - \boldsymbol{u}^n) \,\mathrm{d}x \right), \\ T_2 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| \varrho_K^{n-1} \frac{\boldsymbol{U}_K^n - \boldsymbol{U}_K^{n-1}}{\Delta t} \cdot \left(\frac{\boldsymbol{U}_K^{n-1} + \boldsymbol{U}_K^n}{2} - \boldsymbol{u}_K^{n-1} \right), \\ T_3 &= -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K) \\ \sigma=K|L}} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\frac{\boldsymbol{U}_K^n + \boldsymbol{U}_L^n}{2} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \boldsymbol{U}_K^n [\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}], \\ T_4 &= -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K) \\ \sigma=K|L}} |\sigma| p(\varrho_K^n) [\boldsymbol{U}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}], \\ T_5 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t} (r_K^n - \varrho_K^n) \left(H'(r_K^n) - H'(r_K^{n-1}) \right), \\ T_6 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K) \\ \sigma=K|L}} |\sigma| \varrho_\sigma^{n,\mathrm{up}} H'(r_K^{n-1}) [\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}]. \end{aligned} \tag{5.2}$$

Proof. Lemma 5.1 is proved in Section 5 in (Gallouet *et al.* [22], Thm. 5.1). We provide here the proof for the sake of completeness.

First, noting that the numerical diffusion represented by terms (4.3a)–(4.3d) in the energy identity (4.2) is positive, we infer

$$I_1 + I_2 + I_3 \leq 0, \tag{5.3}$$

with

$$I_1 := \sum_{K\in\mathcal{T}} \frac{1}{2}\frac{|K|}{\Delta t}\left(\varrho_K^n |\boldsymbol{u}_K^n|^2 - \varrho_K^{n-1}|\boldsymbol{u}_K^{n-1}|^2\right), \qquad I_2 := \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(H(\varrho_K^n) - H(\varrho_K^{n-1})\right),$$
$$I_3 := \sum_{K\in\mathcal{T}} \left(\mu \int_K |\nabla_x \boldsymbol{u}^n|^2 \,\mathrm{d}x + \frac{\mu}{3}\int_K |\mathrm{div}\boldsymbol{u}^n|^2 \,\mathrm{d}x\right).$$

Next, we consider the discrete continuity equation (3.6) with $\phi = \frac{1}{2}|\hat{\mathbf{U}}^n|^2$ as test function in order to obtain

$$I_4 := \sum_{K\in\mathcal{T}} \frac{1}{2}\frac{|K|}{\Delta t}(\varrho_K^n - \varrho_K^{n-1})|\boldsymbol{U}_K^n|^2 = -\sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} \frac{1}{2}|\sigma|\varrho_\sigma^{n,\mathrm{up}}[\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}]|\boldsymbol{U}_K^n|^2 := J_1. \tag{5.4}$$

In the next step, taking $-\boldsymbol{U}^n$ as test function $\mathbf{v}$ in the discrete momentum equation (3.7) one gets

$$I_5 = -\sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(\varrho_K^n \boldsymbol{u}_K^n - \varrho_K^{n-1}\boldsymbol{u}_K^{n-1}\right)\cdot \boldsymbol{U}_K^n = J_2 + J_3 + J_4,$$

with

$$J_2 = \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |\sigma|\varrho_\sigma^{n,\mathrm{up}}\hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \cdot \boldsymbol{U}_K^n [\boldsymbol{u}_\sigma^n \cdot \mathbf{n}_{\sigma,K}],$$
$$J_3 = \mu \sum_{K\in\mathcal{T}} \int_K \nabla \boldsymbol{u}^n : \nabla \boldsymbol{U}^n \,\mathrm{d}x + \frac{\mu}{3}\sum_{K\in\mathcal{T}} \int_K \mathrm{div}\boldsymbol{u}^n \mathrm{div}\boldsymbol{U}^n \,\mathrm{d}x$$

and

$$J_4 = -\sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |\sigma| p(\varrho_K^n)[\boldsymbol{U}_\sigma^n \cdot \mathbf{n}_{\sigma,K}].$$

We then consider the discrete continuity equation (3.6) with a test function $\phi = H'(r^{n-1})$ and obtain

$$-\sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}(\varrho_K^n - \varrho_K^{n-1})H'(r_K^{n-1}) = \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |\sigma|\varrho_\sigma^{n,\mathrm{up}}[\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}]H'(r_K^{n-1}).$$

Observing that $\varrho_K^n H'(r_K^n) - \varrho_K^{n-1}H'(r_K^{n-1}) = \varrho_K^n\left(H'(r_K^n) - H'(r_K^{n-1})\right) + (\varrho_K^n - \varrho_K^{n-1})H'(r_K^{n-1})$, we rewrite the last identity in the form

$$I_6 := -\sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(\varrho_K^n H'(r_K^n) - \varrho_K^{n-1}H'(r_K^{n-1})\right) = J_5 + J_6$$
$$\text{with } J_5 = -\sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\varrho_K^n\left(H'(r_K^n) - H'(r_K^{n-1})\right) \text{ and } J_6 = \sum_{K\in\mathcal{T}} \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |\sigma|\varrho_\sigma^{n,\mathrm{up}}[\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}]H'(r_K^{n-1}). \tag{5.5}$$

Finally, thanks to the convexity of the function H, we have

$$\begin{aligned} I_7 &:= \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\Big[\left(r_K^n H'(r_K^n) - H(r_K^n)\right) - \left(r_K^{n-1} H'(r_K^{n-1}) - H(r_K^{n-1})\right)\Big] \\ &= \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t} r_K^n \left(H'(r_K^n) - H'(r_K^{n-1})\right) - \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(H(r_K^n) - (r_K^n - r_K^{n-1})H'(r_K^{n-1}) - H(r_K^{n-1})\right) \\ &\le \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t} r_K^n \left(H'(r_K^n) - H'(r_K^{n-1})\right) := J_7. \end{aligned} \tag{5.6}$$

Now, we gather the expressions (5.3)–(5.6); this is performed in several steps.

Step 1: *Term* $I_1 + I_4 + I_5$. We obtain by direct calculation,

$$\begin{aligned} I_1 + I_4 + I_5 = &\sum_{K\in\mathcal{T}} \frac{1}{2}\frac{|K|}{\Delta t}\left(\varrho_K^n |\boldsymbol{u}_K^n - \boldsymbol{U}_K^n|^2 - \varrho_K^{n-1}|\boldsymbol{u}_K^{n-1} - \boldsymbol{U}_K^{n-1}|^2\right) \\ &- \sum_{K\in\mathcal{T}} |K|\varrho_K^{n-1}\frac{\boldsymbol{U}_K^n - \boldsymbol{U}_K^{n-1}}{\Delta t}\cdot\left(\frac{\boldsymbol{U}_K^{n-1} + \boldsymbol{U}_K^n}{2} - \boldsymbol{u}_K^{n-1}\right). \end{aligned} \tag{5.7}$$

Step 2: *Term* $J_1 + J_2$. Employing the definition (3.4) of the upwinding, one gets

$$J_1 + J_2 = -\sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma|\varrho_\sigma^{n,\mathrm{up}}\left(\frac{\boldsymbol{U}_K^n + \boldsymbol{U}_L^n}{2} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}\right)\cdot \boldsymbol{U}_K^n[\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}]. \tag{5.8}$$

Step 3: *Term* $I_3 - J_3$. This term can be written in the form

$$\begin{aligned} I_3 - J_3 = &\sum_{K\in\mathcal{T}}\left(\mu\int_K |\nabla_x(\boldsymbol{u}^n - \boldsymbol{U}^n)|^2\,\mathrm{d}x + \frac{\mu}{3}\int_K |\mathrm{div}(\boldsymbol{u}^n - \boldsymbol{U}^n)|^2\,\mathrm{d}x\right) \\ &- \sum_{K\in\mathcal{T}}\mu\int_K\left(\nabla\boldsymbol{U}^n : \nabla(\boldsymbol{U}^n - \boldsymbol{u}^n) + \frac{\mu}{3}\int_K \mathrm{div}\boldsymbol{U}^n\mathrm{div}(\boldsymbol{U}^n - \boldsymbol{u}^n)\right). \end{aligned} \tag{5.9}$$

Step 4: *Term* $I_2 + I_6 + I_7$. By virtue of (5.3), (5.5) and (5.6), we easily find that

$$I_2 + I_6 + I_7 = \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(E(\varrho_K^n\,|\,r_K^n) - E(\varrho_K^{n-1}\,|\,r_K^{n-1})\right), \tag{5.10}$$

where the function E is defined in (4.9).

Step 5: *Term* $J_5 + J_6 + J_7$. Coming back to (5.5) and (5.6), we deduce that

$$J_5 + J_6 + J_7 = \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}(r_K^n - \varrho_K^n)\left(H'(r_K^n) - H'(r_K^{n-1})\right) + \sum_{K\in\mathcal{T}}\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |\sigma|\varrho_\sigma^{n,\mathrm{up}}[\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}]H'(r_K^{n-1}). \tag{5.11}$$

Step 6: *Conclusion*

According to (5.3)–(5.6), we have

$$\sum_{i=1}^7 I_i \le \sum_{i=1}^7 J_i;$$

whence, writing this inequality by using expressions (5.7)–(5.11) calculated in steps 1-5, we get

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \frac{1}{2}\frac{|K|}{\Delta t}\left(\varrho_K^n|\boldsymbol{u}_K^n-\boldsymbol{U}_K^n|^2-\varrho_K^{n-1}|\boldsymbol{u}_K^{n-1}-\boldsymbol{U}_K^{n-1}|^2\right)+\sum_{K\in\mathcal{T}}\frac{|K|}{\Delta t}\left(E(\varrho_K^n|r_K^n)-E(\varrho_K^{n-1}|r_K^{n-1})\right)\\
&\qquad+\sum_{K\in\mathcal{T}}\left(\mu\int_K|\nabla_x(\boldsymbol{u}^n-\boldsymbol{U}^n)|^2\,\mathrm{d}x+\frac{\mu}{3}\int_K|\mathrm{div}(\boldsymbol{u}^n-\boldsymbol{U}^n)|^2\,\mathrm{d}x\right)\\
&\leq\sum_{K\in\mathcal{T}}\left(\mu\int_K\nabla_x\boldsymbol{U}_h^n:\nabla_x(\boldsymbol{U}^n-\boldsymbol{u}^n)\,\mathrm{d}x+\frac{\mu}{3}\int_K\mathrm{div}\boldsymbol{U}^n\mathrm{div}(\boldsymbol{U}^n-\boldsymbol{u}^n)\,\mathrm{d}x\right)\\
&\qquad+\sum_{K\in\mathcal{T}}|K|\varrho_K^{n-1}\frac{\boldsymbol{U}_K^n-\boldsymbol{U}_K^{n-1}}{\Delta t}\cdot\left(\frac{\boldsymbol{U}_K^{n-1}+\boldsymbol{U}_K^n}{2}-\boldsymbol{u}_K^{n-1}\right)\\
&\qquad-\sum_{K\in\mathcal{T}}\sum_{\sigma=K|L\in\mathcal{E}(K)}|\sigma|\varrho_\sigma^{n,\mathrm{up}}\left(\frac{\boldsymbol{U}_K^n+\boldsymbol{U}_L^n}{2}-\hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}\right)\cdot\boldsymbol{U}_K^n[\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}]\\
&\qquad-\sum_{K\in\mathcal{T}}\sum_{\sigma=K|L\in\mathcal{E}(K)}|\sigma|p(\varrho_K^n)[\boldsymbol{U}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}]+\sum_{K\in\mathcal{T}}\frac{|K|}{\Delta t}(r_K^n-\varrho_K^n)\left(H'(r_K^n)-H'(r_K^{n-1})\right)\\
&\qquad+\sum_{K\in\mathcal{T}}\sum_{\sigma=K|L\in\mathcal{E}_K}|\sigma|\varrho_\sigma^{n,\mathrm{up}}H'(r_K^{n-1})[\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}].
\end{aligned}\tag{5.12}$$

We obtain formula (5.1) by summing $(5.12)^n$ from $n=1$ to $n=m$ and multiplying the resulting inequality by Δt. □

6. Approximate discrete relative energy inequality

In this section, we transform the right hand side of the relative energy inequality (5.1) to a form that is more convenient for the comparison with the strong solution. This transformation is given in the following lemma.

Lemma 6.1 (Approximate relative energy inequality)**.** *Let $(\varrho^n,\boldsymbol{u}^n)$ be a solution of the discrete problem (3.5)–(3.7), where the pressure satisfies (1.4) with $\gamma\geq 3/2$. Then there exists*

$$\begin{aligned}
c=c\Big(&M_0,E_0,\underline{r},\overline{r},|p'|_{C^1[\underline{r},\overline{r}]},\|(\partial_t r,\nabla r,\boldsymbol{V},\partial_t\boldsymbol{V},\nabla\boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{18})},\\
&\|\partial_t^2 r\|_{L^1(0,T;L^{\gamma'}(\Omega))},\|\partial_t\nabla r\|_{L^2(0,T;L^{6\gamma/5\gamma-6}(\Omega;\mathbb{R}^3))}\Big)>0,
\end{aligned}$$

such that for all $m=1,\dots,N$, we have:

$$\begin{aligned}
&\int_{\Omega_h}\left(\varrho^m|\hat{\boldsymbol{u}}^m-\hat{\boldsymbol{V}}_{h,0}^m|^2+E(\varrho^m|\hat{r}^m)\right)\mathrm{d}x-\int_{\Omega_h}\left(\varrho^0|\hat{\boldsymbol{u}}^0-\hat{\boldsymbol{V}}_{h,0}^0|^2+E(\varrho^0|\hat{r}^0)\right)\mathrm{d}x\\
&+\Delta t\sum_{n=1}^m\sum_{K\in\mathcal{T}}\left(\mu\int_K|\nabla_x(\boldsymbol{u}^n-\boldsymbol{V}_{h,0}^n)|^2\,\mathrm{d}x+\frac{\mu}{3}\int_K|\mathrm{div}(\boldsymbol{u}^n-\boldsymbol{V}_{h,0}^n)|^2\,\mathrm{d}x\right)\leq\sum_{i=1}^6 S_i+R_{h,\Delta t}^m+G^m,
\end{aligned}\tag{6.1}$$

for any couple $(r, \mathbf{V})$ *belonging to the class* (2.25), *where*

$$
\begin{aligned}
S_1 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \left(\mu \int_K \nabla_x \boldsymbol{V}_{h,0}^n : \nabla_x (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x + \frac{\mu}{3} \int_K \mathrm{div} \boldsymbol{V}_{h,0}^n \mathrm{div}(\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x \right), \\
S_2 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| \varrho_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot \left(\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n \right), \\
S_3 &= \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n \right) \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \cdot \boldsymbol{n}_{\sigma,K}, \\
S_4 &= -\Delta t \sum_{n=1}^{m} \int_{\Omega_h} p(\varrho^n)\,\mathrm{div}\,\boldsymbol{V}^n\,\mathrm{d}x, \\
S_5 &= \Delta t \sum_{n=1}^{m} \int_{\Omega_h} (\hat{r}^n - \varrho^n) \frac{p'(\hat{r}^n)}{\hat{r}^n} [\partial_t r]^n\,\mathrm{d}x, \\
S_6 &= -\Delta t \sum_{n=1}^{m} \int_{\Omega_h} \frac{\varrho^n}{\hat{r}^n} p'(\hat{r}^n) \boldsymbol{u}^n \cdot \nabla r^n\,\mathrm{d}x,
\end{aligned}
\tag{6.2}
$$

and

$$
|G^m| \le c\,\Delta t \sum_{n=1}^{m} \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n \Big| \hat{r}^n, \hat{\boldsymbol{V}}^n), \quad |R_{h,\Delta t}^m| \le c(\sqrt{\Delta t} + h^a), \tag{6.3}
$$

with the power a *defined in* (3.9) *and with the functional* $\mathcal{E}$ *introduced in* (4.9). (*Recall that in agreement with the notation* (2.35), (3.1)–(3.3), $\mathbf{V}_{h,0}^n = \Pi_{h,0}^V[\mathbf{V}(t_n)]$, $\mathbf{V}_{h,0,K}^n = \Pi_h^Q \Pi_{h,0}^V \mathbf{V}(t_n)|_K$, $\mathbf{V}_{h,0,\sigma}^n = \frac{1}{|\sigma|}\int_\sigma \mathbf{V}_{h,0}^n$, $\hat{r}^n = \Pi_h^Q[r(t_n)]$, *where the projections* Π^Q, Π^V *are defined in* (2.31)) *and* (2.34).)

Proof. We take as test functions $\mathbf{U}^n = \mathbf{V}_{h,0}^n$ and $r^n = \hat{r}^n$ in the discrete relative energy inequality (5.1). We keep the left hand side and the first term (term T_1) at the right hand side as they stay. The transformation of the remaining terms at the right hand side (terms $T_2 - T_6$) is performed in the following steps:

Step 1: *Term* T_2. We have

$$
T_2 = T_{2,1} + R_{2,1} + R_{2,2}, \text{ with } T_{2,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| \varrho_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot \left(\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n \right), \tag{6.4}
$$

and

$$
R_{2,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} R_{2,1}^{n,K}, \quad R_{2,2} = \Delta t \sum_{n=1}^{m} R_{2,2}^n,
$$

where

$$
R_{2,1}^{n,K} = -\frac{|K|}{2} \varrho_K^{n-1} \frac{(\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1})^2}{\Delta t} = -\frac{|K|}{2} \varrho_K^{n-1} \frac{([\boldsymbol{V}^n - \boldsymbol{V}^{n-1}]_{h,0,K})^2}{\Delta t},
$$

and

$$
R_{2,2}^n = -\sum_{K\in\mathcal{T}} |K| \varrho_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot \left(\boldsymbol{u}_K^{n-1} - \boldsymbol{u}_K^n \right).
$$

We may write by virtue of the first order Taylor formula applied to function $t \mapsto \mathbf{V}(t,x)$,

$$\left|\frac{[\boldsymbol{V}^n - \boldsymbol{V}^{n-1}]_{h,0,K}}{\Delta t}\right| = \left|\frac{1}{|K|}\int_K \left[\frac{1}{\Delta t}\Big[\int_{t_{n-1}}^{t_n} \partial_t \mathbf{V}(z,x)\mathrm{d}z\Big]_{h,0}\right]\mathrm{d}x\right|$$
$$= \left|\frac{1}{|K|}\int_K \left[\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} [\partial_t \mathbf{V}(z)]_{h,0}(x)\mathrm{d}z\right]\mathrm{d}x\right| \leq \|[\partial_t \mathbf{V}]_{h,0}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^3))} \leq \|\partial_t \mathbf{V}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^3))},$$

where we have used the property (2.39) of the projection $\Pi^V_{h,0}$ on the space $V_{h,0}(\Omega_h)$. Therefore, thanks to the mass conservation (4.1), we get

$$|R^{n,K}_{2,1}| \leq \frac{M_0}{2}|K|\Delta t\|\partial_t \boldsymbol{V}\|^2_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^3))}. \tag{6.5}$$

To treat term $R^n_{2,2}$ we use the discrete Hölder inequality and identity (4.1) in order to get

$$|R^n_{2,2}| \leq \Delta t\, cM_0\|\partial_t \boldsymbol{V}\|^2_{L^\infty(0,T;W^{1,\infty}(\Omega;\mathbb{R}^3))} + cM_0^{1/2}\left(\sum_{K\in\mathcal{T}}|K|\varrho_K^{n-1}|\boldsymbol{u}_K^{n-1} - \boldsymbol{u}_K^n|^2\right)^{1/2}\|\partial_t \boldsymbol{V}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^3))};$$

whence, by virtue of estimate (4.2) for the upwind dissipation term (4.3a), one obtains

$$|R_{2,2}| \leq \sqrt{\Delta t}\, c(M_0, E_0, \|\partial_t \boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^3)}). \tag{6.6}$$

Step 2: *Term* T_3. Employing the definition (3.4) of upwind quantities, we easily establish that

$$T_3 = T_{3,1} + R_{3,1},$$
$$\text{with } T_{3,1} = \Delta t\sum_{n=1}^m\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\varrho^{n,\mathrm{up}}_\sigma\left(\hat{\boldsymbol{u}}^{n,\mathrm{up}}_\sigma - \hat{\boldsymbol{V}}^{n,\mathrm{up}}_{h,0,\sigma}\right)\cdot\boldsymbol{V}^n_{h,0,K}\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{\sigma,K},\quad R_{3,1} = \Delta t\sum_{n=1}^m\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}R^{n,\sigma}_{3,1},$$
$$\text{and } R^{n,\sigma}_{3,1} = |\sigma|\varrho^n_K\frac{|\boldsymbol{V}^n_{h,0,K} - \boldsymbol{V}^n_{h,0,L}|^2}{2}[\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{\sigma,K}]^+ + |\sigma|\varrho^n_L\frac{|\boldsymbol{V}^n_{h,0,L} - \boldsymbol{V}^n_{h,0,K}|^2}{2}[\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{\sigma,L}]^+,\ \forall\sigma = K|L\in\mathcal{E}_{\mathrm{int}}.$$

Writing

$$\boldsymbol{V}^n_{h,0,K} - \boldsymbol{V}^n_{h,0,L} = [\boldsymbol{V}^n_{h,0} - \boldsymbol{V}^n_h]_K + \boldsymbol{V}^n_{h,K} - \boldsymbol{V}^n_h + \boldsymbol{V}^n_h - \boldsymbol{V}^n_{h,\sigma}$$
$$+\boldsymbol{V}^n_{h,\sigma} - \boldsymbol{V}^n_h + \boldsymbol{V}^n_h - \boldsymbol{V}^n_{h,L} + [\boldsymbol{V}^n_h - \boldsymbol{V}^n_{h,0}]_L,\ \sigma = K|L\in\mathcal{E}_{\mathrm{int}},$$

and employing estimates (2.48) (if $K\cap\partial\Omega_h = \emptyset$), (2.49) (if $K\cap\partial\Omega_h\neq\emptyset$) to evaluate the L^∞-norm of the first term, (2.52) then $(2.41)_{s=1}$ and (2.53) after $(2.41)_{s=1}$ to evaluate the L^∞-norm of the second and third terms, and performing the same tasks at the second line, we get

$$\|\boldsymbol{V}^n_{h,0,K} - \boldsymbol{V}^n_{h,0,L}\|_{L^\infty(K\cup L;\mathbb{R}^3)} \leq ch\|\nabla\boldsymbol{V}\|_{L^\infty(K\cup L;\mathbb{R}^9)}; \tag{6.7}$$

consequently

$$|R^{n,\sigma}_{3,1}| \leq h^2\, c\|\nabla\boldsymbol{V}\|^2_{L^\infty((0,T)\times\Omega;\mathbb{R}^9)}|\sigma|(\varrho^n_K + \varrho^n_L)|\boldsymbol{u}^n_\sigma|,\ \forall\sigma = K|L\in\mathcal{E}_{\mathrm{int}},$$

whence

$$|R_{3,1}| \leq h\, c\|\nabla\boldsymbol{V}\|^2_{L^\infty((0,T)\times\Omega;\mathbb{R}^9)}\left(\sum_{K\in\mathcal{T}}\sum_{\sigma=K|L\in\mathcal{E}(K)}h|\sigma|(\varrho^n_K + \varrho^n_L)^{6/5}\right)^{5/6}$$
$$\times\left[\Delta t\sum_{n=1}^m\left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)}h|\sigma||\boldsymbol{u}^n_\sigma|^6\right)^{1/3}\right]^{1/2} \leq h\, c(M_0, E_0, \|\nabla\boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^9)}), \tag{6.8}$$

provided $\gamma \geq 6/5$, thanks to the discrete Hölder inequality, the equivalence relation (2.29), the equivalence of norms (2.42) and energy bounds listed in Corollary 4.2.

Clearly, for each face $\sigma = K|L \in \mathcal{E}_{\rm int}$, $\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K} + \boldsymbol{u}_\sigma^n \cdot \mathbf{n}_{\sigma,L} = 0$; whence, finally

$$T_{3,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} - \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,\sigma}^n \right) \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}. \tag{6.9}$$

Before the next transformation of term $T_{3,1}$, we realize that

$$\mathbf{V}_{h,0,K}^n - \mathbf{V}_{h,0,\sigma}^n = [\mathbf{V}_{h,0}^n - \mathbf{V}_h^n]_K + \mathbf{V}_{h,K}^n - \mathbf{V}_h^n + \mathbf{V}_h^n - \mathbf{V}_{h,\sigma}^n + [\mathbf{V}_h^n - \mathbf{V}_{h,0}^n]_\sigma;$$

whence by virtue of (2.48) and (2.49), (2.52) and (2.53) and $(2.41)_{s=1}$, similarly as in (6.7),

$$\|\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,\sigma}^n\|_{L^\infty(K;\mathbb{R}^3)} \leq ch \|\nabla_x \boldsymbol{V}\|_{L^\infty(0,T;L^\infty(\Omega:R^3))}, \quad \sigma \subset K. \tag{6.10}$$

Let us now decompose the term $T_{3,1}$ as

$$T_{3,1} = T_{3,2} + R_{3,2}, \text{ with } R_{3,2} = \Delta t \sum_{n=1}^{m} R_{3,2}^n,$$

$$T_{3,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n \right) \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \cdot \boldsymbol{n}_{\sigma,K}, \text{ and}$$

$$R_{3,2}^n = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n \right) (\boldsymbol{u}_\sigma^n - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}) \cdot \boldsymbol{n}_{\sigma,K}.$$

By virtue of discrete Hölder's inequality and estimate (6.10), we get

$$|R_{3,2}^n| \leq c \|\nabla \mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^9)} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma| \varrho_\sigma^{n,\mathrm{up}} \left| \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} - \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \right|^2 \right)^{1/2}$$
$$\times \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma| |\varrho_\sigma^{n,\mathrm{up}}|^{\gamma_0} \right)^{1/(2\gamma_0)} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma| \left| \boldsymbol{u}_\sigma^n - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right|^q \right)^{1/q},$$

where $\frac{1}{2} + \frac{1}{2\gamma_0} + \frac{1}{q} = 1$, $\gamma_0 = \min\{\gamma, 2\}$ and $\gamma \geq 3/2$. For the sum in the last term of the above product, we have

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma| \left| \boldsymbol{u}_\sigma^n - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right|^q \leq c \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma| |\boldsymbol{u}_\sigma^n - \boldsymbol{u}_K^n|^q$$

$$\leq c \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} \left(\|\boldsymbol{u}_\sigma^n - \boldsymbol{u}^n\|_{L^q(K;\mathbb{R}^3)}^q + \sum_{K\in\mathcal{T}} \|\boldsymbol{u}^n - \boldsymbol{u}_K^n\|_{L^q(K;\mathbb{R}^3)}^q \right) \right) \leq ch^{\frac{2\gamma_0-3}{2\gamma_0}q} \left(\sum_{K\in\mathcal{T}} \|\nabla_x \boldsymbol{u}^n\|_{L^2(K;\mathbb{R}^9}^2 \right)^{q/2},$$

where we have used the definition (3.4), the discrete Minkowski inequality, interpolation inequalities (2.56) and (2.57) and the discrete 'imbedding' inequality (2.58). Now we can go back to the estimate of $R_{3,2}^n$ taking into account the upper bounds (4.4), (4.7) and (4.8), in order to get

$$|R_{3,2}| \leq h^a \, c \left(M_0, E_0, \|\nabla \mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^9)} \right), \tag{6.11}$$

provided $\gamma \geq 3/2$, where a is given in (6.3).

Finally, we rewrite term $T_{3,2}$ as

$$
\begin{aligned}
&T_{3,2} = T_{3,3} + R_{3,3}, \text{ with } R_{3,3} = \Delta t \sum_{n=1}^{m} R_{3,3}^n, \\
&T_{3,3} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n \right) \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \cdot \boldsymbol{n}_{\sigma,K}, \text{ and} \\
&R_{3,3}^n = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \varrho_\sigma^{n,\mathrm{up}} \left(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} \right) \cdot \left(\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n \right) \left(\hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}} - \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \right) \cdot \boldsymbol{n}_{\sigma,K};
\end{aligned}
\tag{6.12}
$$

whence

$$
|R_{3,3}| \leq c(\|\nabla \boldsymbol{V}\|_{L^\infty(Q_T,\mathbb{R}^9)})\ \Delta t \sum_{n=1}^{m} \mathcal{E}(\varrho^n, \hat{\mathbf{u}}^n \,|\, \hat{r}^n, \hat{\boldsymbol{V}}_{h,0}^n). \tag{6.13}
$$

Step 3: *Term* T_4. Integration by parts over each $K \in \mathcal{T}$ gives

$$
T_4 = -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_{\mathrm{K}} p(\varrho_K^n) \mathrm{div}_x \boldsymbol{V}_{h,0}^n \,\mathrm{d}x.
$$

We may write

$$
\|\mathrm{div}_x(\boldsymbol{V}_{0,h}^n - \boldsymbol{V}_h^n)\|_{L^\infty(K)} \leq ch \|\nabla_x \boldsymbol{V}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^9))}, \tag{6.14}
$$

where we have used (2.48)–(2.49). Therefore, employing identity (2.44) we obtain

$$
T_4 = T_{4,1} + R_{4,1}, \quad T_{4,1} = -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_{\mathrm{K}} p(\varrho_K^n) \mathrm{div}_x \boldsymbol{V}^n \,\mathrm{d}x, \tag{6.15}
$$

$$
R_{4,1} = -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_{\mathrm{K}} p(\varrho_K^n) \mathrm{div}_x (\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_h^n) \,\mathrm{d}x.
$$

Due to (1.4) and (4.7), $p(\varrho^n)$ is bounded uniformly in $L^\infty(L^1(\Omega))$; employing this fact and (6.14) we immediately get

$$
|R_{4,1}| \leq h\ c(E_0, M_0, \|\nabla \mathbf{V}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^9))}). \tag{6.16}
$$

Step 4: *Term* T_5. Using the Taylor formula, we get

$$
H'(r_K^n) - H'(r_K^{n-1}) = H''(r_K^n)(r_K^n - r_K^{n-1}) - \frac{1}{2} H'''(\overline{r}_K^n)(r_K^n - r_K^{n-1})^2,
$$

where $\overline{r}_K^n \in [\min(r_K^{n-1}, r_K^n), \max(r_K^{n-1}, r_K^n)]$. We infer

$$
\begin{aligned}
&T_5 = T_{5,1} + R_{5,1}, \text{ with } T_{5,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| (r_K^n - \varrho_K^n) \frac{p'(r_K^n)}{r_K^n} \frac{r_K^n - r_K^{n-1}}{\Delta t},\ R_{5,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} R_{5,1}^{n,K}, \text{ and} \\
&R_{5,1}^{n,K} = \frac{1}{2} |K| H'''(\overline{r}_K^n) \frac{(r_K^n - r_K^{n-1})^2}{\Delta t} (\varrho_K^n - r_K^n).
\end{aligned}
$$

Consequently, by the first order Taylor formula applied to function $t \mapsto r(t,x)$ on the interval (t_{n-1}, t_n) and thanks to the mass conservation (4.1)

$$
|R_{5,1}| \leq \Delta t\ c(M_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}, \|\partial_t r\|_{L^\infty(Q_T)}). \tag{6.17}
$$

Let us now decompose $T_{5,1}$ as follows:

$$T_{5,1} = T_{5,2} + R_{5,2}, \text{ with } T_{5,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K (r_K^n - \varrho_K^n) \frac{p'(r_K^n)}{r_K^n} [\partial_t r]^n \mathrm{d}x,\ R_{5,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} R_{5,2}^{n,K}, \text{ and}$$
$$R_{5,2}^{n,K} = \int_K (r_K^n - \varrho_K^n) \frac{p'(r_K^n)}{r_K^n} \left(\frac{r_K^n - r_K^{n-1}}{\Delta t} - [\partial_t r]^n \right) \mathrm{d}x. \tag{6.18}$$

In accordance with (3.2), here and in the sequel, $[\partial_t r]^n(x) = \partial_t r(t_n, x)$. We write using twice the Taylor formula in the integral form and the Fubini theorem,

$$|R_{5,2}^{n,K}| = \frac{1}{\Delta t} \Big| p'(r_K^n) r_K^n (\varrho_K^n - r_K^n) \int_K \int_{t_{n-1}}^{t_n} \int_s^{t_n} \partial_t^2 r(z) \mathrm{d}z \mathrm{d}s \mathrm{d}x \Big|$$

$$\leq \frac{p'(r_K^n)}{r_K^n} \int_{t_{n-1}}^{t_n} \int_K |\varrho_K^n - r_K^n| \Big| \partial_t^2 r(z) \Big| \mathrm{d}x \mathrm{d}z \mathrm{d}s$$

$$\leq \frac{p'(r_K^n)}{r_K^n} \|\varrho^n - \hat{r}^n\|_{L^\gamma(K)} \int_{t_{n-1}}^{t_n} \|\partial_t^2 r(z)\|_{L^{\gamma'}(K)} \mathrm{d}z \mathrm{d}s.$$

Therefore, by virtue of Corollary 4.2, we have estimate

$$|R_{5,2}| \leq \Delta t\ c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}, \|\partial_t^2 r\|_{L^1(0,T;L^{\gamma'}(\Omega))}). \tag{6.19}$$

Step 5: *Term T_6.* We decompose this term as follows:

$$T_6 = T_{6,1} + R_{6,1}, \quad R_{6,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} R_{6,1}^{n,\sigma,K}, \text{ with}$$
$$T_{6,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| \varrho_K^n \left(H'(r_K^{n-1}) - H'(r_\sigma^{n-1}) \right) \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}, \text{ and}$$
$$R_{6,1}^{n,\sigma,K} = |\sigma| \left(\varrho_\sigma^{n,\mathrm{up}} - \varrho_K^n \right) \left(H'(r_K^{n-1}) - H'(r_\sigma^{n-1}) \right) \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}, \text{ for } \sigma = K|L \in \mathcal{E}_{\mathrm{int}}.$$

We will now estimate the term $R_{6,1}^{n,\sigma,K}$. We shall treat separately the cases $\gamma < 2$ and $\gamma \geq 2$. The 'simple' case $\gamma \geq 2$ is left to the reader. The more complicated case $\gamma < 2$ will be treated as follows: We first write

$$|R_{6,1}^{n,\sigma,K}| \leq \sqrt{h}\, \|\nabla H'(r)\|_{L^\infty(Q_T;\mathbb{R}^3)} |\sigma| |\varrho_\sigma^{n,\mathrm{up}} - \varrho_K^n| \left[\frac{1_{\{\overline{\varrho}_\sigma^n \geq 1\}}}{[\max\{\varrho_K, \varrho_L\}]^{(2-\gamma)/2}} + 1_{\{\overline{\varrho}_\sigma^n < 1\}} \right] \sqrt{|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}|}$$

$$\times \left[1_{\{\overline{\varrho}_\sigma^n \geq 1\}} [\max\{\varrho_K, \varrho_L\}]^{(2-\gamma)/2} + 1_{\{\overline{\varrho}_\sigma^n < 1\}} \right] \sqrt{h} \sqrt{|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}|},$$

where we have employed the first order Taylor formula applied to function $x \mapsto H'(r(t_{n-1},x)$. Consequently, the application of the discrete Hölder and Young inequalities yield

$$
\begin{aligned}
|R_{6,1}| &\leq \sqrt{h}\, c\|\nabla H'(r)\|_{L^\infty(Q_T;\mathbb{R}^3)} \\
&\quad \times \Delta t \sum_{n=1}^{m} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h \left[1_{\{\overline{\varrho}_\sigma^n \geq 1\}} [\max\{\varrho_K, \varrho_L\}]^{2-\gamma} + 1_{\{\overline{\varrho}_\sigma^n <1\}} \right] |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}| \right)^{1/2} \\
&\quad \times \left(\sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| h (\varrho_\sigma^{n,\mathrm{up}} - \varrho_K^n)^2 \left[\frac{1_{\{\overline{\varrho}_\sigma^n \geq 1\}}}{[\max\{\varrho_K, \varrho_L\}]^{2-\gamma}} + 1_{\{\overline{\varrho}_\sigma^n <1\}} \right] |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}| \right)^{1/2} \\
&\leq \sqrt{h}\, c\|\nabla H'(r)\|_{L^\infty(Q_T;\mathbb{R}^3)} \\
&\quad \times \Delta t \sum_{n=1}^{m} \left\{ \left[|\Omega_h|^{\frac{5}{6}} + \left(\sum_{K\in\mathcal{T}} |\sigma| h (\varrho_K^n)^{\frac{6}{5}(2-\gamma)} \right)^{\frac{5}{6}} \right] \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}|^6 \right)^{\frac{1}{6}} \right. \\
&\quad \left. \sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| h (\varrho_\sigma^{n,\mathrm{up}} - \varrho_K^n)^2 \left[\frac{1_{\{\overline{\varrho}_\sigma^n \geq 1\}}}{[\max\{\varrho_K, \varrho_L\}]^{2-\gamma}} + 1_{\{\overline{\varrho}_\sigma^n <1\}} \right] |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K} \right\}^{1/2} \\
&\leq \sqrt{h}\, c\|\nabla H'(r)\|_{L^\infty(Q_T;\mathbb{R}^3)} \left\{ \Delta t \sum_{n=1}^{m} \left[|\Omega_h|^{\frac{5}{6}} + \left(\sum_{K\in\mathcal{T}} |\sigma| h (\varrho_K^n)^{\frac{6}{5}(2-\gamma)} \right)^{\frac{5}{6}} \right] \left(\sum_{\sigma\in\mathcal{E}} |\sigma| h |\boldsymbol{u}_\sigma^n|^6 \right)^{1/6} \right. \\
&\quad \left. + \Delta t \sum_{n=1}^{m} \left[\sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| h (\varrho_\sigma^{n,\mathrm{up}} - \varrho_K^n)^2 \left[\frac{1_{\{\overline{\varrho}_\sigma^n \geq 1\}}}{[\max\{\varrho_K, \varrho_L\}]^{2-\gamma}} + 1_{\{\overline{\varrho}_\sigma^n <1\}} \right] |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}| \right] \right\} \\
&\leq \sqrt{h}\, c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C([\underline{r},\overline{r}])}, \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}),
\end{aligned}
$$

where, in order to get the last line, we have used the estimate (4.10) of the numerical dissipation to evaluate the second term, and finally equivalence of norms $(2.42)_{p=6}$ together with (4.5) and (4.7), under assumption $\gamma \geq 12/11$, to evaluate the first term.

Let us now decompose the term $T_{6,1}$ as

$$
\begin{aligned}
&T_{6,1} = T_{6,2} + R_{6,2}, \text{ with } T_{6,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| \varrho_K^n H''(r_K^{n-1})(r_K^{n-1} - r_\sigma^{n-1})[\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}], \\
&R_{6,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{K}} \sum_{\sigma\in\mathcal{E}(K)} R_{6,2}^{n,\sigma,K}, \text{ and} \\
&R_{6,2}^{n,\sigma,K} = |\sigma| \varrho_K^n \left(H'(r_K^{n-1}) - H'(r_\sigma^{n-1}) - H''(r_K^{n-1})(r_K^{n-1} - r_\sigma^{n-1}) \right) [\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{\sigma,K}].
\end{aligned}
$$

Therefore, by virtue of the second order Taylor formula applied to function H', the Hölder inequality, (2.42), and (4.5), (4.7) in Corollary 4.2, we have, provided $\gamma \geq 6/5$,

$$
\begin{aligned}
|R_{6,2}| &\leq hc \left(|H''|_{C([\underline{r},\overline{r}])} + |H'''|_{C([\underline{r},\overline{r}])} \right) \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)} \Delta t \sum_{n=1}^{m} \|\varrho^n\|_{L^\gamma(\Omega_h)} \|\boldsymbol{u}^n\|_{L^6(\Omega_h;\mathbb{R}^3)} \\
&\leq h\, c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}, \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}).
\end{aligned} \tag{6.20}
$$

Let us now deal with the term $T_{6,2}$. Noting that $\displaystyle\int_K \nabla r^{n-1}\,\mathrm{d}x = \sum_{\sigma\in\mathcal{E}(K)} |\sigma|(r_\sigma^{n-1} - r_K^{n-1})\boldsymbol{n}_{\sigma,K}$, we may write $T_{6,2} = T_{6,3} + R_{6,3}$, with

$$
\begin{aligned}
T_{6,3} &= -\Delta t \sum_{n=1}^{m}\sum_{K\in\mathcal{T}}\int_K \varrho_K^n H''(r_K^{n-1})\boldsymbol{u}^n\cdot\nabla r^{n-1}\,\mathrm{d}x,\\
R_{6,3} &= \Delta t \sum_{n=1}^{m}\sum_{K\in\mathcal{T}}\int_K \varrho_K^n H''(r_K^{n-1})(\boldsymbol{u}^n - \boldsymbol{u}_K^n)\cdot\nabla r^{n-1}\,\mathrm{d}x\\
&\quad + \Delta t \sum_{n=1}^{m}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\varrho_K^n H''(r_K^{n-1})(r_K^{n-1} - r_\sigma^{n-1})(\boldsymbol{u}_\sigma^n - \boldsymbol{u}_K^n)\cdot\boldsymbol{n}_{\sigma,K}.
\end{aligned}
$$

Consequently, by virtue of Hölder's inequality, interpolation inequality (2.56) (to estimate $\|\mathbf{u}^n - \mathbf{u}_K^n\|_{L^{\gamma_0'}(K;\mathbb{R}^3)}$ by $h^{(5\gamma_0-6)/(2\gamma_0)}\|\nabla_x\mathbf{u}^n\|_{L^2(K;\mathbb{R}^9)}$, $\gamma_0 = \min\{\gamma, 2\}$) in the first term, and by the Taylor formula applied to function $x \mapsto r(t_{n-1}, x)$, then Hölder's inequality and (2.56)–(2.57) (to estimate $\|\mathbf{u}_\sigma^n - \mathbf{u}_K^n\|_{L^{\gamma_0'}(K;\mathbb{R}^3)}$ by $h^{(5\gamma_0-6)/(2\gamma_0)}\|\nabla_x\mathbf{u}^n\|_{L^2(K;\mathbb{R}^9)}$), we get

$$
|R_{6,3}| \le h^b\, c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}\|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}),\quad b = \frac{5\gamma_0 - 6}{2\gamma_0},
\tag{6.21}
$$

provided $\gamma \ge 6/5$, where we have used at the end the discrete imbedding and Hölder inequalities (2.58) and (2.59) and finally estimates (4.4) and (4.7).

Finally we write $T_{6,3} = T_{6,4} + R_{6,4}$, with

$$
\begin{aligned}
T_{6,4} &= -\Delta t \sum_{n=1}^{m}\sum_{K\in\mathcal{T}}\int_K \varrho_K^n \frac{p'(r_K^n)}{r_K^n}\boldsymbol{u}^n\cdot\nabla r^n\,\mathrm{d}x,\\
R_{6,4} &= \Delta t \sum_{n=1}^{m}\sum_{K\in\mathcal{T}}\int_K \varrho_K^n \left(H''(r_K^n)\nabla r^n - H''(r_K^{n-1})\nabla r^{n-1}\right)\cdot\boldsymbol{u}^n\,\mathrm{d}x,
\end{aligned}
\tag{6.22}
$$

where by the same token as in (6.19),

$$
|R_{6,4}| \le \Delta t\, c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}, \|\nabla r, \partial_t r\|_{L^\infty(Q_T;\mathbb{R}^4)}, \|\partial_t\nabla r\|_{L^2(0,T;L^{6\gamma/(5\gamma-6)}(\Omega;\mathbb{R}^3))}),
\tag{6.23}
$$

provided $\gamma \ge 6/5$.

We are now in position to conclude the proof of Lemma 6.1: we obtain the inequality (6.1) by gathering the principal terms (6.4), (6.12), (6.15), (6.18), (6.22) and the residual terms estimated in (6.5), (6.6), (6.8), (6.11), (6.13), (6.17), (6.19), (6.20), (6.21), (6.23) at the right hand side $\sum_{i=1}^6 T_i$ of the discrete relative energy inequality (5.1). □

7. A discrete identity satisfied by the strong solution

This section is devoted to the proof of a discrete identity satisfied by any strong solution of problem (1.1)–(1.6) in the class (2.9)–(2.10) extended eventually to $\mathbb{R}^3$ according to Lemma 2.3. This identity is stated in Lemma 7.1 below. It will be used in combination with the approximate relative energy inequality stated in Lemma 6.1 to deduce the convenient form of the relative energy inequality verified by any function being a strong solution to the compressible Navier–Stokes system. This last step is performed in the next section.

Lemma 7.1 (A discrete identity for strong solutions)**.** *Let $(\varrho^n, \boldsymbol{u}^n)$ be a solution of the discrete problem (3.5)–(3.7) with the pressure satisfying (1.4), where $\gamma \geq 3/2$. There exists*

$$c = c\Big(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1[\underline{r},\overline{r}]}, \|(\partial_t r, \nabla r, \boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V}, \nabla^2 \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{45})},$$

$$\|\partial_t^2 r\|_{L^1(0,T;L^{\gamma'}(\Omega))}, \|\partial_t \nabla r\|_{L^2(0,T;L^{6\gamma/5\gamma-6}(\Omega;\mathbb{R}^3))}, \|\partial_t^2 \boldsymbol{V}, \partial_t \nabla \boldsymbol{V}\|_{L^2(0,T;L^{6/5}(\Omega;\mathbb{R}^{12}))}\Big) > 0,$$

such that for all $m = 1, \dots, N$, we have:

$$\sum_{i=1}^{6} \mathcal{S}_i + \mathcal{R}^m_{h,\Delta t} = 0, \tag{7.1}$$

where

$$\mathcal{S}_1 = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \left(\mu \int_K \nabla_x \boldsymbol{V}^n_{h,0} : \nabla_x (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x + \frac{\mu}{3} \int_K \mathrm{div} \boldsymbol{V}^n_{h,0} \mathrm{div}(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x \right),$$

$$\mathcal{S}_2 = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} |K| r_K^{n-1} \frac{\boldsymbol{V}^n_{h,0,K} - \boldsymbol{V}^{n-1}_{h,0,K}}{\Delta t} \cdot \left(\boldsymbol{V}^n_{h,0,K} - \boldsymbol{u}^n_K\right),$$

$$\mathcal{S}_3 = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| r^{n,\mathrm{up}}_\sigma \left(\hat{\boldsymbol{V}}^{n,\mathrm{up}}_{h,0,\sigma} - \hat{\boldsymbol{u}}^{n,\mathrm{up}}_\sigma\right) \cdot \left(\boldsymbol{V}^n_{h,0,\sigma} - \boldsymbol{V}^n_{h,0,K}\right) \hat{\boldsymbol{V}}^{n,\mathrm{up}}_{h,0,\sigma} \cdot \boldsymbol{n}_{\sigma,K}$$

$$\mathcal{S}_4 = -\Delta t \sum_{n=1}^{m} \int_{\Omega_h} p(\hat{r}^n)\,\mathrm{div}\,\boldsymbol{V}^n\,\mathrm{d}x,$$

$$\mathcal{S}_5 = 0,$$

$$\mathcal{S}_6 = -\Delta t \sum_{n=1}^{m} \int_{\Omega_h} p'(\hat{r}^n) \boldsymbol{u}^n \cdot \nabla r^n\,\mathrm{d}x,$$

and

$$|\mathcal{R}^m_{h,\Delta t}| \leq c\left(h^{5/6} + \Delta t\right),$$

for any couple $(r, \mathbf{V})$ belonging to (2.25) and satisfying the continuity equation (1.1) on $(0,T)\times\mathbb{R}^3$ and momentum equation (1.2) with boundary conditions (1.5) on $(0,T)\times\Omega$ in the classical sense. (Recall that in agreement with notation (2.35), (3.1)–(3.3), $\mathbf{V}^n_{h,0} = \Pi^V_{h,0}[\mathbf{V}(t_n)]$, $\mathbf{V}^n_{h,0,K} = [\mathbf{V}^n_{h,0}]_K$, $\mathbf{V}^n_{h,0,\sigma} = [\mathbf{V}^n_{h,0}]_\sigma$, $\hat{r}^n = \Pi^Q_h[r(t_n)]$, where projections Π^Q, Π^V are defined in (2.31) and (2.34)).

Before starting the proof we recall an auxiliary algebraic inequality whose straightforward proof is left to the reader, and introduce some notations.

Lemma 7.2. *Let p satisfies assumptions (1.4). Let $0 < a < b < \infty$. Then there exists $c = c(a,b) > 0$ such that for all $\varrho \in [0,\infty)$ and $r \in [a,b]$ there holds*

$$E(\varrho|r) \geq c(a,b)\left(1_{R_+\setminus[a/2,2b]}(\varrho) + \varrho^\gamma 1_{R_+\setminus[a/2,2b]}(\varrho) + (\varrho - r)^2 1_{[a/2,2b]}(\varrho)\right),$$

where $E(\varrho|r)$ is defined in (4.9).

If we consider Lemma 7.2 with $\varrho = \varrho^n(x)$, $r = \hat{r}^n(x)$, $a = \underline{r}$, $b = \overline{r}$ (where r is a function belonging to class (2.25) and $\underline{r}$, $\overline{r}$ are its lower and upper bounds, respectively), we obtain

$$E(\varrho^n(x)|\hat{r}^n(x)) \geq c(\underline{r},\overline{r})\left(1_{R_+\setminus[\underline{r}/2,2\overline{r}]}(\varrho^n(x)) + (\varrho^n)^\gamma(x) 1_{R_+\setminus[\underline{r}/2,2\overline{r}]}(\varrho^n(x)) + (\varrho^n(x) - \hat{r}^n(x))^2 1_{[\underline{r}/2,2\overline{r}]}(\varrho^n(x))\right). \tag{7.2}$$

Now, for fixed numbers $\underline{r}$ and $\overline{r}$ and fixed functions ϱ^n, $n = 0, \dots, N$, we introduce the residual and essential subsets of Ω (relative to ϱ^n) as follows:

$$N^n_{\rm ess} = \{x \in \Omega \,\Big|\, \frac{1}{2}\underline{r} \le \varrho^n(x) \le 2\overline{r}\},\ N^n_{\rm res} = \Omega \setminus N^n_{\rm ess}, \tag{7.3}$$

and we set

$$[g]_{\rm ess}(x) = g(x)1_{N^n_{\rm ess}}(x),\ [g]_{\rm res}(x) = g(x)1_{N^n_{\rm res}}(x),\ \ x \in \Omega,\ \ g \in L^1(\Omega).$$

Integrating inequality (7.2) we deduce

$$c(\underline{r}, \overline{r}) \sum_{K\in\mathcal{T}} \int_K \left(\Big[1\Big]_{\rm res} + \Big[(\varrho^n)^\gamma\Big]_{\rm res} + \Big[\varrho^n - \hat r^n\Big]^2_{\rm ess} \right) {\rm d}x \le \mathcal{E}(\varrho^n, \boldsymbol{u}^n \Big| \hat r^n, \boldsymbol{V}^n), \tag{7.4}$$

for any pair $(r, \boldsymbol{V})$ belonging to the class (2.25) and any $\varrho^n \in Q_h(\Omega_h)$, $\varrho^n \ge 0$.

We are now ready to proceed to the proof of Lemma 7.1.

Proof. Since $(r, \boldsymbol{V})$ satisfies (1.1) on $(0,T)\times\Omega$ and belongs to the class (2.25), equation (1.2) can be rewritten in the form

$$r\partial_t \boldsymbol{V} + r\boldsymbol{V}\cdot\nabla\boldsymbol{V} + \nabla p(r) - \mu\Delta\boldsymbol{V} - \mu/3\nabla\,{\rm div}\,\boldsymbol{V} = 0 \ \text{ in } (0,T)\times\Omega.$$

From this fact, we deduce the identity

$$\sum_{i=1}^5 \mathcal{T}_i = \mathcal{R}_0, \tag{7.5}$$

where

$$\mathcal{R}_0 = -\Delta t \sum_{n=1}^m \int_{\Omega_h\setminus\Omega} \Big(r^n[\partial_t \boldsymbol{V}]^n + r\boldsymbol{V}^n\cdot\nabla\boldsymbol{V}^n + \nabla p(r^n) - \mu\Delta\boldsymbol{V}^n - \frac{\mu}{3}\nabla\,{\rm div}\,\boldsymbol{V}^n\Big)\cdot(\boldsymbol{V}^n_{h,0} - \mathbf{u}^n){\rm d}x,$$

$$\mathcal{T}_1 = -\Delta t \sum_{n=1}^m \int_{\Omega_h} \Big(\mu\Delta\boldsymbol{V}^n + \frac{\mu}{3}\nabla\,{\rm div}\,\boldsymbol{V}^n\Big)\cdot(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,{\rm d}x, \quad \mathcal{T}_2 = \Delta t \sum_{n=1}^m \int_{\Omega_h} r^n[\partial_t\boldsymbol{V}]^n\cdot(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,{\rm d}x,$$

$$\mathcal{T}_3 = \Delta t \sum_{n=1}^m \int_{\Omega_h} r^n\boldsymbol{V}^n\cdot\nabla\boldsymbol{V}^n\cdot(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,{\rm d}x, \quad \mathcal{T}_4 = \Delta t \sum_{n=1}^m \int_{\Omega_h} \nabla p(r^n)\cdot\boldsymbol{V}^n_{h,0}\,{\rm d}x,$$

$$\mathcal{T}_5 = 0, \qquad \mathcal{T}_6 = -\Delta t \sum_{n=1}^m \int_{\Omega_h} \nabla p(r^n)\cdot\boldsymbol{u}^n\,{\rm d}x.$$

In the steps below, we deal with each of the terms $\mathcal{R}_0$ and $\mathcal{T}_i$.

Step 0: *Term* $\mathcal{R}_0$. By the Hölder inequality

$$|\mathcal{R}_0| \le |\Omega_h\setminus\Omega|^{5/6}\, c(\overline{r}, |p'|_{C[\underline{r},\overline{r}]}, \|(\partial_t r, \nabla r, \mathbf{V}, \nabla\mathbf{V}, \nabla^2\mathbf{V})\|_{L^\infty(Q_T;\mathbb{R}^{43})} \Delta t \sum_{n=1}^m (\|\boldsymbol{u}^n\|_{L^6(\Omega_h)} + \|\boldsymbol{V}^n_{h,0}\|_{L^6(\Omega_h)})$$

$$\le h^{5/3}\, c(M_0, E_0, \overline{r}, |p'|_{C[\underline{r},\overline{r}]}, \|(\partial_t r, \nabla r, \mathbf{V}, \nabla\mathbf{V}, \nabla^2\mathbf{V})\|_{L^\infty(Q_T;\mathbb{R}^{43})}, \tag{7.6}$$

where we have used (4.5), (2.48), (2.49) and (2.39).

Step 1: *Term* $\mathcal{T}_1$. Integrating by parts, we get:

$$
\begin{aligned}
&\mathcal{T}_1 = \mathcal{T}_{1,1} + \mathcal{R}_{1,1},\\
&\text{with } \mathcal{T}_{1,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K \left(\mu \nabla \boldsymbol{V}^n_{h,0} : \nabla(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) + \frac{\mu}{3} \operatorname{div} \boldsymbol{V}^n_{h,0} \operatorname{div}(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) \right) \mathrm{d}x,\\
&\text{and } \ \mathcal{R}_{1,1} = I_1 + I_2, \text{ with}\\
&I_1 = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K \left(\mu \nabla(\boldsymbol{V}^n - \boldsymbol{V}^n_{h,0}) : \nabla(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) + \frac{\mu}{3} \operatorname{div}(\boldsymbol{V}^n - \boldsymbol{V}^n_{h,0}) \operatorname{div}(\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) \right) \mathrm{d}x, \qquad (7.7)\\
&I_2 = -\Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} \int_\sigma \left(\mu \boldsymbol{n}_{\sigma,K} \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) + \frac{\mu}{3} \operatorname{div} \boldsymbol{V}^n (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n) \cdot \boldsymbol{n}_{\sigma,K} \right) \mathrm{d}S\\
&= -\Delta t \sum_{n=1}^{m} \sum_{\sigma\in\mathcal{E}} \int_\sigma \left(\mu \boldsymbol{n}_\sigma \cdot \nabla \boldsymbol{V}^n \cdot \left[\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n \right]_{\sigma,\mathbf{n}_\sigma} + \frac{\mu}{3} \operatorname{div} \boldsymbol{V}^n \left[\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n \right]_{\sigma,\mathbf{n}_\sigma} \cdot \boldsymbol{n}_\sigma \right) \mathrm{d}S,
\end{aligned}
$$

where in the last line $\mathbf{n}_\sigma$ is the unit normal to the face σ and $[\cdot]_{\sigma,\mathbf{n}_\sigma}$ is the jump over sigma (with respect to $\mathbf{n}_\sigma$) defined in Lemma 2.10.

To estimate I_1, we use the Cauchy–Schwartz inequality, decompose $\mathbf{V}^n - \mathbf{V}^n_{h,0} = \mathbf{V}^n - \mathbf{V}^n_h + \mathbf{V}^n_h - \mathbf{V}^n_{h,0}$ and employ estimates $(2.41)_{s=2}$, (2.48)–(2.49) to evaluate the norms involving $\nabla(\mathbf{V}^n - \mathbf{V}^n_{h,0})$, and decompose $\mathbf{V}^n_{h,0} = \mathbf{V}^n_{h,0} - \mathbf{V}^n_h + \mathbf{V}^n_h$ use (2.48)–(2.49), $(2.40)_{s=1}$, (4.4), the Minkowski inequality to estimate the norms involving $\nabla(\mathbf{V}^n_{h,0} - \mathbf{u}^n)$. We get

$$
|I_1| \leq h\, c(M_0, E_0, \|\nabla \mathbf{V}, \nabla^2 \boldsymbol{V}\|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^{36}))}).
$$

Since the integral over any face $\sigma \in \mathcal{E}_{\rm int}$ of the jump of a function from $V_{h,0}(\Omega_h)$ is zero, we may write

$$
\begin{aligned}
I_2 = \Delta t \sum_{n=1}^{m} \sum_{\sigma\in\mathcal{E}_{\rm int}} \int_\sigma \Big(&\mu \boldsymbol{n}_\sigma \cdot (\nabla \boldsymbol{V}^n - (\nabla \boldsymbol{V}^n)_\sigma) \cdot \left[\boldsymbol{u}^n - \boldsymbol{V}^n_{h,0} \right]_{\sigma,\mathbf{n}_\sigma}\\
&+ \frac{\mu}{3} (\operatorname{div} \boldsymbol{V}^n - (\operatorname{div} \boldsymbol{V}^n)_\sigma) \left[\boldsymbol{u}^n - \boldsymbol{V}^n_{h,0} \right]_{\sigma,\mathbf{n}_\sigma} \cdot \boldsymbol{n}_\sigma \Big) \mathrm{d}S;
\end{aligned}
$$

whence by using the first order Taylor formula applied to functions $x \mapsto \nabla \mathbf{V}^n(x)$ to evaluate the differences $\nabla \boldsymbol{V}^n - (\nabla \boldsymbol{V}^n)_\sigma$, $\operatorname{div} \boldsymbol{V}^n - [\operatorname{div} \boldsymbol{V}^n]_\sigma$, and Hölder's inequality,

$$
\begin{aligned}
|I_2| &\leq \Delta t\, h\, c \|\nabla^2 \boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^{27})} \sum_{n=1}^{m} \sum_{\sigma\in\mathcal{E}_{\rm int}} \sqrt{|\sigma|}\sqrt{h} \left(\frac{1}{\sqrt{h}} \left\| \left[\boldsymbol{u}^n - \boldsymbol{V}^n_{h,0} \right]_{\sigma,\mathbf{n}_\sigma} \right\|_{L^2(\sigma;\mathbb{R}^3)} \right)\\
&\leq \Delta t\, h\, c \|\nabla^2 \boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^{27})} \sum_{n=1}^{m} \sum_{\sigma\in\mathcal{E}_{\rm int}} \left(|\sigma| h + \frac{1}{h} \left\| \left[\boldsymbol{u}^n - \boldsymbol{V}^n_{h,0} \right]_{\sigma,\mathbf{n}_\sigma} \right\|^2_{L^2(\sigma;\mathbb{R}^3)} \right).
\end{aligned}
$$

Therefore,

$$
|\mathcal{R}_{1,1}| \leq h\, c(M_0, E_0, \|\mathbf{V}, \nabla \boldsymbol{V}, \nabla^2 \boldsymbol{V}\|_{L^\infty(Q_T,\mathbb{R}^{39})}), \qquad (7.8)
$$

where we have employed Lemma 2.10, (4.4) and (2.48)–(2.49), (2.40).

Step 2: *Term* $\mathcal{T}_2$. Let us now decompose the term $\mathcal{T}_2$ as

$$\mathcal{T}_2 = \mathcal{T}_{2,1} + \mathcal{R}_{2,1},$$
$$\text{with } \mathcal{T}_{2,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K r^{n-1} \frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t} \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x, \quad \mathcal{R}_{2,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \mathcal{R}^{n,K}_{2,1},$$
$$\text{and } \mathcal{R}^{n,K}_{2,1} = \int_K (r^n - r^{n-1})[\partial_t \mathbf{V}]^n \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x + \int_K r^{n-1} \left([\partial_t \boldsymbol{V}]^n - \frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t} \right) \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x.$$

The remainder $\mathcal{R}^{n,K}_{2,1}$ can be rewritten as follows

$$\mathcal{R}^{n,K}_{2,1} = \int_K \Big[\int_{t_{n-1}}^{t_n} \partial_t r(t,\cdot)\mathrm{d}t \Big] [\partial_t \mathbf{V}]^n \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x + \frac{1}{\Delta t} \int_K r^{n-1} \Big[\int_{t_{n-1}}^{t_n} \int_s^{t_n} \partial_t^2 \boldsymbol{V}(z,\cdot)\mathrm{d}z\mathrm{d}s \Big] \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x;$$

whence, by the Hölder inequality,

$$|\mathcal{R}^{n,K}_{2,1}| \le \Delta t \Big[(\|r\|_{L^\infty(Q_T)} + \|\partial_t r\|_{L^\infty(Q_T)}) (\|\partial_t \boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^3)} |K|^{5/6} (\|\boldsymbol{u}^n\|_{L^6(K)} + \|\boldsymbol{V}^n_{h,0}\|_{L^6(K)})$$
$$+ \|\partial_t^2 \boldsymbol{V}^n\|_{L^{6/5}(\Omega;\mathbb{R}^3))} (\|\boldsymbol{u}^n\|_{L^6(K)} + \|\boldsymbol{V}^n_{h,0}\|_{L^6(K)}) \Big].$$

Consequently, by the same token as in (6.19) or (6.23),

$$|\mathcal{R}_{2,1}| \le \Delta t\, c \left(M_0, E_0, \overline{r}, \|(\partial_t r, \boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{16})}, \|\partial_t^2 \boldsymbol{V}\|_{L^2(0,T;L^{6/5}(\Omega;\mathbb{R}^3))} \right), \tag{7.9}$$

where we have used the discrete Hölder and Young inequalities, the estimates (2.39), (2.48) and (2.49) and the energy bound (4.4) from Corollary 4.2.

Step 2a: *Term* $\mathcal{T}_{2,1}$. We decompose the term $\mathcal{T}_{2,1}$ as

$$\mathcal{T}_{2,1} = \mathcal{T}_{2,2} + \mathcal{R}_{2,2},$$
$$\text{with } \mathcal{T}_{2,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K r_K^{n-1} \frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t} \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x,\ \mathcal{R}_{2,2} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \mathcal{R}^{n,K}_{2,2},$$
$$\text{and } \mathcal{R}^{n,K}_{2,2} = \int_K (r^{n-1} - r_K^{n-1}) \frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t} \cdot (\boldsymbol{V}^n_{h,0} - \boldsymbol{u}^n)\,\mathrm{d}x;$$

therefore,

$$|\mathcal{R}^n_{2,2}| = |\sum_{K\in\mathcal{T}} \mathcal{R}^{n,K}_{2,2}| \le h\, c \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)} \|\partial_t \boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^3)} \|\boldsymbol{u}^n - \boldsymbol{V}^n_{h,0}\|_{L^6(\Omega;\mathbb{R}^3)}.$$

Consequently, by virtue of formula (4.5) for $\mathbf{u}^n$ and estimates (2.39), (2.48) and (2.49),

$$|\mathcal{R}_{2,2}| \le h\, c(M_0, E_0, \|(\nabla r, \boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{18})}). \tag{7.10}$$

Step 2b: *Term* $\mathcal{T}_{2,2}$. We decompose the term $\mathcal{T}_{2,2}$ as

$$\mathcal{T}_{2,2} = \mathcal{T}_{2,3} + \mathcal{R}_{2,3},$$

$$\text{with } \mathcal{T}_{2,3} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K r_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x,\ \mathcal{R}_{2,3} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \mathcal{R}_{2,3}^{n,K},$$

$$\text{and } \mathcal{R}_{2,3}^{n,K} = \int_K r_K^{n-1} \left(\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t} - \left[\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t}\right]_h \right) \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x$$

$$+ \int_K r_K^{n-1} \left(\left[\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t}\right]_h - \left[\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t}\right]_{h,K} \right) \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x$$

$$+ \int_K r_K^{n-1} \left(\left[\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t}\right]_{h,K} - \left[\frac{\boldsymbol{V}^n - \boldsymbol{V}^{n-1}}{\Delta t}\right]_{h,0,K} \right) \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}_n)\,\mathrm{d}x = I_1^K + I_2^K + I_3^K.$$

We calculate carefully

$$|I_3^K| = \frac{1}{\Delta t} r_K^{n-1} \int_K \left\{ \int_{t_{n-1}}^{t_n} \Big[[\partial_t \boldsymbol{V}(z)]_h - [\partial_t \boldsymbol{V}(z)]_{h,0} \Big]_K \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n) \mathrm{d}z \right\} \mathrm{d}x$$

$$\leq \frac{1}{\Delta t} r_K^{n-1} \int_{t_{n-1}}^{t_n} \Big\| \Big[[\partial_t \boldsymbol{V}(z)]_h - [\partial_t \boldsymbol{V}(z)]_{h,0} \Big]_K \Big\|_{L^{6/5}(K;\mathbb{R}^3)} \|\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n\|_{L^6(K;\mathbb{R}^3)} \mathrm{d}z.$$

Summing over polyhedra $K \in \mathcal{T}$ we get simply by using the discrete Sobolev inequality

$$\sum_{K\in\mathcal{T}} |I_3^K| \leq \frac{1}{\Delta t} r_K^{n-1} \int_{t_{n-1}}^{t_n} \left\{ \left(\sum_{K\in\mathcal{T}} \|\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n\|_{L^6(K;\mathbb{R}^3)}^6 \right)^{1/6} \left(\sum_{K\in\mathcal{T}} \Big\| [\partial_t \boldsymbol{V}(z)]_h - [\partial_t \boldsymbol{V}(z)]_{h,0} \Big\|_{L^{6/5}(K;\mathbb{R}^3)}^{6/5} \right)^{5/6} \right\} \mathrm{d}z$$

$$\leq \frac{1}{\Delta t} r_K^{n-1} \int_{t_{n-1}}^{t_n} \|\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n\|_{L^6(\Omega_h;\mathbb{R}^3)} \Big\| [\partial_t \boldsymbol{V}(z)]_h - [\partial_t \boldsymbol{V}(z)]_{h,0} \Big\|_{L^{6/5}(\Omega_h;\mathbb{R}^3)} \mathrm{d}z$$

$$\leq \frac{h^{5/6}}{\Delta t} \int_{t_{n-1}}^{t_n} \|\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n\|_{L^6(\Omega_h;\mathbb{R}^3)} \|\partial_t \boldsymbol{V}(z)\|_{L^\infty(\Omega_h;\mathbb{R}^3)} \mathrm{d}z,$$

where we have used estimate (2.51) to obtain the last line.

As far as the term I_2^K is concerned, we write

$$|I_2^K| = \frac{1}{\Delta t} r_K^{n-1} \Big| \int_K \left(\Big[\int_{t_{n-1}}^{t_n} \partial_t \boldsymbol{V}(z) \mathrm{d}z \Big]_h - \Big[\int_{t_{n-1}}^{t_n} \partial_t \boldsymbol{V}(z) \mathrm{d}z \Big]_{h,K} \right) \cdot (\boldsymbol{u}^n - \boldsymbol{V}_{h,0}^n)\,\mathrm{d}x \Big|$$

$$\leq \frac{h}{\Delta t} r_K^{n-1} \int_{t_{n-1}}^{t_n} \Big\| \nabla_x \Big[\partial_t \boldsymbol{V}(z) \Big]_h \Big\|_{L^{6/5}(K;\mathbb{R}^3)} \|\boldsymbol{u}^n - \boldsymbol{V}_{h,0}^n\|_{L^6(K;\mathbb{R}^3)},$$

where we have used the Fubini theorem, Hölder's inequality and (2.52), $(2.41)_{s=1}$. Further, employing the Sobolev inequality on the Crouzeix–Raviart space $V_{h,0}(\Omega_h)$ (2.43), the Hölder inequality and estimate $(2.41)_{s=1}$, we get

$$\sum_{K\in\mathcal{T}} |I_2^K| \leq \frac{h}{\Delta t} r_K^{n-1} \|\boldsymbol{u}^n - \boldsymbol{V}_{h,0}^n\|_{L^6(\Omega_h;\mathbb{R}^3)} \int_{t_{n-1}}^{t_n} \Big\| \nabla_x \partial_t \boldsymbol{V}(z) \Big\|_{L^{6/5}(\Omega_h;\mathbb{R}^3)} \mathrm{d}z.$$

We reserve the similar treatment to the term I_1^K. Resuming these calculations and summing over n from 1 to m we get by using Corollary 4.2 and estimates (2.48)–(2.49), (2.39),

$$|\mathcal{R}_{2,3}| \leq h^{5/6}\, c(M_0, E_0, \|(r, \boldsymbol{V}, \nabla \boldsymbol{V}, \partial_t \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{16})}, \|\partial_t \nabla \boldsymbol{V}\|_{L^2(0,T;L^{6/5}(\Omega;\mathbb{R}^9))}). \tag{7.11}$$

Step 2c: *Term* $\mathcal{T}_{2,3}$. We rewrite this term in the form

$$\begin{aligned}
&\mathcal{T}_{2,3} = \mathcal{T}_{2,4} + \mathcal{R}_{2,4},\ \mathcal{R}_{2,4} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \mathcal{R}_{2,4}^{n,K},\\
&\text{with } \mathcal{T}_{2,4} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K r_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot (\boldsymbol{u}_K^n - \boldsymbol{V}_{h,0,K}^n)\,\mathrm{d}x, \qquad (7.12)\\
&\text{and } \mathcal{R}_{2,4}^{n,K} = \int_K r_K^{n-1} \frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t} \cdot \big((\boldsymbol{u}^n - \boldsymbol{u}_K^n) - (\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,0,K}^n)\big)\,\mathrm{d}x.
\end{aligned}$$

First, we estimate the L^∞ norm of $\frac{\boldsymbol{V}_{h,0,K}^n - \boldsymbol{V}_{h,0,K}^{n-1}}{\Delta t}$ as in (6.5). Next, we decompose

$$\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,0,K}^n = \boldsymbol{V}_{h,0}^n - \boldsymbol{V}_h^n + \boldsymbol{V}_h^n - \boldsymbol{V}_{h,K}^n + [\boldsymbol{V}_h^n - \boldsymbol{V}_{h,0}^n]_K,$$

and use $(2.52)_{p=2}$ to estimate $\boldsymbol{u}^n - \boldsymbol{u}_K^n$, $(2.52)_{p=\infty}$, $(2.41)_{s=1}$ to estimate $\boldsymbol{V}_h^n - \boldsymbol{V}_{h,K}^n$ and (2.48)–(2.49) to evaluate $\|[\boldsymbol{V}_h^n - \boldsymbol{V}_{h,0}^n]_K\|_{L^\infty(K;\mathbb{R}^3)} \le \|\boldsymbol{V}_h^n - \boldsymbol{V}_{h,0}^n\|_{L^\infty(K;\mathbb{R}^3)}$. Thanks to the Hölder inequality and (4.4) we finally deduce

$$|\mathcal{R}_{2,4}| \le h\, c\left(M_0, E_0, \overline{r}, \|(\boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{15})}\right). \qquad (7.13)$$

Step 3: *Term* $\mathcal{T}_3$. Let us first decompose $\mathcal{T}_3$ as

$$\begin{aligned}
&\mathcal{T}_3 = \mathcal{T}_{3,1} + \mathcal{R}_{3,1},\\
&\text{with } \mathcal{T}_{3,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \int_K r_K^n \boldsymbol{V}_{h,0,K}^n \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n)\,\mathrm{d}x, \quad \mathcal{R}_{3,1} = \Delta t \sum_{n=1}^{m} \sum_{K\in\mathcal{T}} \mathcal{R}_{3,1}^{n,K},\\
&\text{and } \mathcal{R}_{3,1}^{n,K} = \int_K (r^n - r_K^n)\boldsymbol{V}^n \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x + \int_K r_K^n (\boldsymbol{V}^n - \boldsymbol{V}_{h,0}^n) \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x\\
&\qquad + \int_K r_K^n (\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,0,K}^n) \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\,\mathrm{d}x\\
&\qquad + \int_K r_K^n \boldsymbol{V}_{h,0,K}^n \cdot \nabla \boldsymbol{V}^n \cdot \big(\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,0,K}^n - (\boldsymbol{u}^n - \boldsymbol{u}_K^n)\big)\,\mathrm{d}x.
\end{aligned}$$

We have

$$\|r^n - r_K^n\|_{L^\infty(K)} \lesssim h \|\nabla r^n\|_{L^\infty(K)},$$

by the Taylor formula,

$$\|\boldsymbol{V}^n - \boldsymbol{V}_{h,0}^n\|_{L^\infty(K;\mathbb{R}^3)} \lesssim h \|\nabla \boldsymbol{V}^n\|_{L^\infty(K;\mathbb{R}^9)},$$

by virtue of $(2.40)_{s=1}$ and (2.48) and (2.49),

$$\begin{aligned}
\|\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,0,K}^n\|_{L^\infty(K;\mathbb{R}^3)} &\le \|\boldsymbol{V}_{h,0}^n - \boldsymbol{V}_{h,}^n\|_{L^\infty(K;\mathbb{R}^3)} + \|\boldsymbol{V}_h^n - \boldsymbol{V}_{h,K}^n\|_{L^\infty(K;\mathbb{R}^3)}\\
&+ \|[\boldsymbol{V}_h^n - \boldsymbol{V}_{h,0}^n]_K\|_{L^\infty(K;\mathbb{R}^3)} \lesssim h \|\nabla \boldsymbol{V}^n\|_{L^\infty(K;\mathbb{R}^9)}
\end{aligned}$$

by virtue of (2.52), $(2.40)_{s=1}$ $(2.41)_{s=1}$ and (2.48)–(2.49),

$$\|\boldsymbol{u}^n - \boldsymbol{u}_K^n\|_{L^\infty(K;\mathbb{R}^3)} \lesssim h \|\nabla \boldsymbol{u}^n\|_{L^\infty(K;\mathbb{R}^9)}.$$

Consequently by employing several times the Hölder inequality (for integrals over K) and the discrete Hölder inequality (for the sums over $K \in \mathcal{T}$), and using estimate (4.4), we arrive at

$$|\mathcal{R}_{3,1}| \leq h\, c(M_0, E_0, \overline{r}, \|(\nabla r, \boldsymbol{V}, \nabla \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{15})}). \tag{7.14}$$

Now we shall deal wit term $\mathcal{T}_{3,1}$. Integrating by parts, we get:

$$\begin{aligned}\int_K r_K^n \boldsymbol{V}_{h,0,K}^n \cdot \nabla \boldsymbol{V}^n \cdot (\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n)\,\mathrm{d}x &= \sum_{\sigma\in\mathcal{E}(K)} |\sigma| r_K^n [\boldsymbol{V}_{h,0,K}^n \cdot \mathbf{n}_{\sigma,K}] \boldsymbol{V}_\sigma^n \cdot (\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n) \\ &= \sum_{\sigma\in\mathcal{E}(K)} |\sigma| r_K^n [\boldsymbol{V}_{h,0,K}^n \cdot \mathbf{n}_{\sigma,K}] (\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n) \cdot (\boldsymbol{V}_{h,K}^n - \boldsymbol{u}_K^n),\end{aligned}$$

thanks to the the fact that $\sum_{\sigma\in\mathcal{E}(K)} \int_\sigma \boldsymbol{V}_{h,K}^n \cdot \mathbf{n}_{\sigma,K} \mathrm{d}S = 0$.

Next we write

$$\mathcal{T}_{3,1} = \mathcal{T}_{3,2} + \mathcal{R}_{3,2}, \quad \mathcal{R}_{3,2} = \Delta t \sum_{n=1}^m \mathcal{R}_{3,2}^n,$$

$$\mathcal{T}_{3,2} = \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \hat{r}_\sigma^{n,\mathrm{up}} [\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \cdot \mathbf{n}_{\sigma,K}] (\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n) \cdot (\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}), \tag{7.15}$$

$$\begin{aligned}\text{and } \mathcal{R}_{3,2}^n = &\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| (r_K^n - \hat{r}_\sigma^{n,\mathrm{up}}) [\boldsymbol{V}_{h,0,K}^n \cdot \mathbf{n}_{\sigma,K}] (\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n) \cdot (\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n) \\ &+ \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \hat{r}_\sigma^{n,\mathrm{up}} \Big[\Big(\boldsymbol{V}_{h,0,K}^n - \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \Big) \cdot \mathbf{n}_{\sigma,K} \Big] (\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n) \cdot (\boldsymbol{V}_{h,K}^n - \boldsymbol{u}_K^n) \\ &+ \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \hat{r}_\sigma^{n,\mathrm{up}} [\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} \cdot \mathbf{n}_{\sigma,K}] (\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n) \cdot \Big((\boldsymbol{V}_{h,0,K}^n - \hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}}) - (\boldsymbol{u}_K^n - \hat{\boldsymbol{u}}_{h,\sigma}^{n,\mathrm{up}}) \Big).\end{aligned}$$

We may write

$$\mathbf{V}_\sigma^n - \mathbf{V}_{h,0,K}^n = \mathbf{V}_\sigma^n - \mathbf{V}^n + \mathbf{V}^n - \mathbf{V}_h^n + \mathbf{V}_h^n - \mathbf{V}_{h,K}^n + [\mathbf{V}_h^n - \mathbf{V}_{h,0}^n]_K,$$

and use several times the Taylor formula along with $(2.40)_{s=1}$, (2.52), $(2.41)_{s=1}$, (2.48)–(2.49) (in order to estimate $r_K^n - \hat{r}_\sigma^{n,\mathrm{up}}$, $\mathbf{V}_\sigma^n - \mathbf{V}_{h,0,K}^n$, $\mathbf{V}_{h,K}^n - \hat{\mathbf{V}}_{h,\sigma}^{n,\mathrm{up}}$) to get the bound

$$|\mathcal{R}_{3,2}^n| \leq h\, c \|r\|_{W^{1,\infty}(\Omega)} \Big(1 + \|\boldsymbol{V}\|_{W^{1,\infty}(Q_T;\mathbb{R}^3)} \Big)^3 \sum_{K\in\mathcal{T}} h|\sigma||\boldsymbol{u}_K^n|$$

$$+ c\|r\|_{W^{1,\infty}(\Omega)} \Big(1 + \|\boldsymbol{V}\|_{W^{1,\infty}(Q_T;\mathbb{R}^3)} \Big)^2 \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} h|\sigma||\boldsymbol{u}_K^n - \boldsymbol{u}_\sigma^n|.$$

We have by the Hölder inequality

$$\sum_{K\in\mathcal{T}} h|\sigma||\boldsymbol{u}_K^n| \le c\left(\sum_{\sigma\in\mathcal{T}} h|\sigma||\boldsymbol{u}_K^n|^6\right)^{1/6} \le c\left[\left(\sum_{K\in\mathcal{T}} \|\boldsymbol{u}^n-\boldsymbol{u}_K^n\|^6_{L^6(K;\mathbb{R}^3)}\right)^{1/6}\right.$$
$$\left.+\left(\sum_{K\in\mathcal{T}} \|\boldsymbol{u}^n\|^6_{L^6(K;\mathbb{R}^3)}\right)^{1/6}\right] \le c\left(\sum_{K\in\mathcal{T}} \|\nabla\boldsymbol{u}_n\|^2_{L^2(K;\mathbb{R}^9)}\right)^{1/2},$$
$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} h|\sigma||\boldsymbol{u}_K^n-\boldsymbol{u}_\sigma^n| \le c\left[\left(\sum_{K\in\mathcal{T}} \|\boldsymbol{u}^n-\boldsymbol{u}_K^n\|^2_{L^2(K;\mathbb{R}^3)}\right)^{1/2}\right.$$
$$\left.+\left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} \|\boldsymbol{u}^n-\boldsymbol{u}_\sigma^n\|^2_{L^2(K;\mathbb{R}^3)}\right)^{1/2}\right] \le h\,c\left(\sum_{K\in\mathcal{T}} \|\nabla\boldsymbol{u}_n\|^2_{L^2(K;\mathbb{R}^9)}\right)^{1/2},$$

where we have used $(2.54)_{p=2}$, (2.52)–$(2.53)_{p=2}$. Consequently, we may use (4.4) to conclude

$$|\mathcal{R}_{3,2}| \le h\,c\left(M_0, E_0, \overline{r}, \|\nabla r, \boldsymbol{V}, \nabla\boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^{15})}\right). \tag{7.16}$$

Finally, we replace in $\mathcal{T}_{3,2}$ $\boldsymbol{V}_\sigma^n - \boldsymbol{V}_{h,K}^n$ by $\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n$. We get

$$\mathcal{T}_{3,2} = \mathcal{T}_{3,3} + \mathcal{R}_{3,3}, \quad \mathcal{R}_{3,3} = \Delta t\sum_{n=1}^m \mathcal{R}_{3,3}^n,$$

$$\mathcal{T}_{3,3} = \Delta t\sum_{n=1}^m\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\hat{r}_\sigma^{n,\mathrm{up}}[\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}}\cdot\mathbf{n}_{\sigma,K}](\boldsymbol{V}_{h,0,\sigma}^n - \boldsymbol{V}_{h,0,K}^n)\cdot(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}), \tag{7.17}$$

and

$$\mathcal{R}_{3,3}^n = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\hat{r}_\sigma^{n,\mathrm{up}}\boldsymbol{V}_{h,0,K}^n\cdot\mathbf{n}_{\sigma,K}\left([\boldsymbol{V}^n - \boldsymbol{V}_{h,0}]_\sigma^n - [\boldsymbol{V}_h^n - \boldsymbol{V}_{h,0}^n]_K\right)\cdot(\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}),$$

committing error

$$|\mathcal{R}_{3,3}^n| \le h\,c\left(M_0, E_0, \overline{r}, \|\nabla r, \boldsymbol{V}, \nabla\boldsymbol{V}\|_{L^\infty(Q_T;\mathbb{R}^{15})}\right), \tag{7.18}$$

as in the previous step.

Step 4: *Terms* $\mathcal{T}_4$ We write

$$\mathcal{T}_4 = \mathcal{T}_{4,1} + \mathcal{R}_{4,1}, \quad \mathcal{T}_{4,1} = -\int_{\Omega_h} \nabla p(r^n)\cdot\mathbf{V}^n\,\mathrm{d}x,$$

$$\mathcal{R}_{4,1} = \int_{\Omega_h} \nabla p(r^n)\cdot(\mathbf{V}^n - \boldsymbol{V}_{h,0}^n)\,\mathrm{d}x;$$

whence

$$|\mathcal{R}_{4,1}| \le hc\left(\overline{r}, |p'|_{C[\underline{r},\overline{r}]}, \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}\right), \tag{7.19}$$

by virtue of $(2.40)_{s=1}$, (2.48)–(2.49).

Next, employing the integration by parts

$$\mathcal{T}_{4,2} = \mathcal{T}_{4,2} + \mathcal{R}_{4,2}, \quad \mathcal{T}_{4,2} = \int_{\Omega_h} p(r^n) \operatorname{div} \mathbf{V}^n \,\mathrm{d}x,$$

$$\mathcal{R}_{4,2} = -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K),\sigma\in\partial\Omega_h} \int_\sigma p(r^n)\mathbf{V}^n \cdot \mathbf{n}_{\sigma,K}\mathrm{d}S = -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K),\sigma\in\partial\Omega_h} \int_\sigma p(r^n)\left(\mathbf{V}^n - \boldsymbol{V}^n_{h,0,\sigma}\right)\cdot \mathbf{n}_{\sigma,K}\mathrm{d}S.$$

Writing

$$\mathbf{V}^n - \boldsymbol{V}^n_{h,0,\sigma} = \mathbf{V}^n - \boldsymbol{V}^n_h + \mathbf{V}^n_h - \boldsymbol{V}^n_{h,\sigma} + [\boldsymbol{V}^n_h - \boldsymbol{V}^n_{h,0}]_\sigma,$$

we deduce by using $(2.40)_{s=1}$, $(2.41)_{s=1}$, $(2.53)_{p=\infty}$, (2.48), (2.49),

$$\|\mathbf{V}^n - \boldsymbol{V}^n_{h,0,\sigma}\|_{L^\infty(K;\mathbb{R}^3)} \stackrel{<}{\sim} h\|\nabla \boldsymbol{V}^n\|_{L^\infty(K;\mathbb{R}^3)}, \ \sigma\in K.$$

Now, we employ the fact that

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K),\sigma\in\partial\Omega_h} \int_\sigma \mathrm{d}S \approx 1;$$

whence

$$|\mathcal{R}_{4,2}| \leq hc(\overline{r}, |p|_{C[\underline{r},\overline{r}]}, \|\nabla\mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^9)}) \tag{7.20}$$

Finally,

$$\mathcal{T}_{4,2} = \mathcal{T}_{4,3} + \mathcal{R}_{4,3}, \quad \mathcal{T}_{4,3} = \int_{\Omega_h} p(\hat{r}^n) \operatorname{div} \mathbf{V}^n \,\mathrm{d}x, \quad \mathcal{R}_{4,3} = \int_{\Omega_h} (p(r^n) - p(\hat{r}^n)) \operatorname{div} \mathbf{V}^n \,\mathrm{d}x; \tag{7.21}$$

whence

$$|\mathcal{R}_{4,3}| \leq hc(|p'|_{C[\underline{r},\overline{r}]}, \|(\nabla r, \nabla\mathbf{V})\|_{L^\infty(Q_T;\mathbb{R}^{12})}). \tag{7.22}$$

Step 5: *Term* $\mathcal{T}_6$ We decompose $\mathcal{T}_6$ as

$$\begin{aligned} &\mathcal{T}_6 = \mathcal{T}_{6,1} + \mathcal{R}_{6,1}, \text{with } \mathcal{T}_{6,1} = -\Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K p'(\hat{r}^n)\boldsymbol{u}^n \cdot \nabla r^n \,\mathrm{d}x, \\ &\mathcal{R}_{6,1} = \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K (p'(\hat{r}^n) - p'(r^n)) \cdot \boldsymbol{u}^n \cdot \nabla r^n \,\mathrm{d}x. \end{aligned} \tag{7.23}$$

Consequently, by the Taylor formula, Hölder inequality and estimate (4.5),

$$|\mathcal{R}_{6,1}| \leq h\, c(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1([\underline{r},\overline{r}])}, \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}). \tag{7.24}$$

Gathering the formulae (7.7), (7.12), (7.17), (7.21), (7.23) and estimates for the residual terms (7.8), (7.9)–(7.13), (7.14)–(7.18), (7.19), (7.20), (7.22), (7.24) concludes the proof of Lemma 7.1. □

8. A Gronwall inequality

In this section we put together the relative energy inequality (6.1) and the identity (7.1) derived in the previous section. The final inequality resulting from this manipulation is formulated in the following lemma.

Lemma 8.1. *Let $(\varrho^n, \boldsymbol{u}^n)$ be a solution of the discrete problem (3.5)–(3.7) with the pressure satisfying (1.4), where $\gamma \geq 3/2$. Then there exists a positive number*

$$c = c\Big(M_0, E_0, \underline{r}, \overline{r}, |p'|_{C^1[\underline{r},\overline{r}]}, \|(\partial_t r, \nabla r, \boldsymbol{V}, \partial_t \boldsymbol{V}, \nabla \boldsymbol{V}, \nabla^2 \boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{45})},$$
$$\|\partial_t^2 r\|_{L^1(0,T;L^{\gamma'}(\Omega))}, \|\partial_t \nabla r\|_{L^2(0,T;L^{6\gamma/5\gamma-6}(\Omega;\mathbb{R}^3))}, \|\partial_t^2 \boldsymbol{V}, \partial_t \nabla \boldsymbol{V}\|_{L^2(0,T;L^{6/5}(\Omega;\mathbb{R}^{12}))}\Big),$$

such that for all $m = 1, \dots, N$, there holds:

$$\mathcal{E}(\varrho^m, \boldsymbol{u}^m | \hat{r}^m, \hat{\boldsymbol{V}}^m_{h,0}) + \Delta t \frac{\mu}{2} \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K |\nabla_x (u^n - \mathbf{V}^n_{h,0})|^2 \mathrm{d}x$$
$$\leq c\Big[h^a + \sqrt{\Delta t} + \mathcal{E}(\varrho^0, \boldsymbol{u}^0 | \hat{r}(0), \hat{\boldsymbol{V}}_{h,0}(0))\Big] + c\,\Delta t \sum_{n=1}^m \mathcal{E}(\varrho^n, \boldsymbol{u}^n | \hat{r}^n, \hat{\boldsymbol{V}}^n_{h,0}),$$

with any couple $(r, \mathbf{V})$ belonging to (2.25) and satisfying the continuity equation (1.1) on $(0,T)\times\mathbb{R}^3$ and momentum equation (1.2) with boundary conditions (1.5) on $(0,T)\times\Omega$ in the classical sense, where a is defined in (3.9) and $\mathcal{E}$ is given in (4.9).

Proof. We observe that

$$S_6 - \mathcal{S}_6 = \Delta t \sum_{n=1}^m \int_{\Omega_h} p'(\hat{r}^n) \frac{\hat{r}^n - \varrho^n}{\hat{r}^n} \mathbf{V}^n \cdot \nabla r^n \,\mathrm{d}x + \Delta t \sum_{n=1}^m \int_{\Omega_h} p'(\hat{r}^n) \frac{\hat{r}^n - \varrho^n}{\hat{r}^n} (\mathbf{u}^n - \mathbf{V}^n) \cdot \nabla r^n \,\mathrm{d}x.$$

Gathering the formulae (6.1) and (6.2), one gets

$$\mathcal{E}(\varrho^m, \boldsymbol{u}^m \Big| \hat{r}^m, \hat{\boldsymbol{V}}^m_{h,0}) - \mathcal{E}(\varrho^0, \boldsymbol{u}^0 \Big| \hat{r}(0), \hat{\boldsymbol{V}}_{h,0}(0)) + \mu \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \Big| \nabla(\boldsymbol{u}^n - \boldsymbol{V}^n_{0,h}) \Big|^2_{L^2(K;\mathbb{R}^3)} \leq \sum_{i=1}^4 \mathcal{P}_i + \mathcal{Q}, \tag{8.1}$$

where

$$\mathcal{P}_1 = \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} |K| (\varrho_K^{n-1} - r_K^{n-1}) \frac{\boldsymbol{V}^n_{h,0,K} - \boldsymbol{V}^{n-1}_{h,0,K}}{\Delta t} \cdot \left(\boldsymbol{V}^n_{h,0,K} - \boldsymbol{u}^n_K\right),$$
$$\mathcal{P}_2 = \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \sum_{\sigma = K|L\in\mathcal{E}_K} |\sigma| \left(\varrho^{n,\mathrm{up}}_\sigma - \hat{r}^{n,\mathrm{up}}_\sigma\right) \left(\hat{\boldsymbol{V}}^{n,\mathrm{up}}_{h,0,\sigma} - \hat{\boldsymbol{u}}^{n,\mathrm{up}}_\sigma\right) \cdot \left(\boldsymbol{V}^n_{h,0,\sigma} - \boldsymbol{V}^n_{h,0,K}\right) \boldsymbol{V}^{n,\mathrm{up}}_{h,0,\sigma} \cdot \boldsymbol{n}_{\sigma,K},$$
$$\mathcal{P}_3 = -\Delta t \sum_{n=1}^m \int_{\Omega_h} \left(p(\varrho^n) - p'(\hat{r}^n)(\varrho^n - \hat{r}^n) - p(\hat{r}^n)\right) \mathrm{div}\mathbf{V}^n,$$
$$\mathcal{P}_4 = \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K p'(\hat{r}^n) \frac{\hat{r}^n - \varrho^n}{\hat{r}^n} (\mathbf{u}^n - \mathbf{V}^n) \cdot \nabla r^n \,\mathrm{d}x,$$
$$\mathcal{Q} = \mathcal{R}^m_{h,\Delta t} + R^m_{h,\Delta t} + G^m.$$

Now, we estimate conveniently the terms $\mathcal{P}_i$, $i = 1, \dots, 4$ in four steps.

Step 1: *Term* $\mathcal{P}_1$. We estimate the L^∞ norm of $\frac{\mathbf{V}^n_{h,0,K}-\mathbf{V}^{n-1}_{h,0,K}}{\Delta t}$ by L^∞ norm of $\partial_t \mathbf{V}$ in the same manner as in (6.5). According to Lemma 7.2, $|\varrho - r|^\gamma 1_{R_+\setminus[\underline{r}/2,2\overline{r}]}(\varrho) \leq c(p)E^p(\varrho|r)$, with any $p \geq 1$; in particular,

$$|\varrho - r|^{6/5} 1_{R_+\setminus[\underline{r}/2,2\overline{r}]}(\varrho) \leq cE(\varrho|r) \tag{8.2}$$

provided $\gamma \geq 6/5$.

We get by using the Hölder inequality,

$$\left| \sum_{K\in\mathcal{T}} |K|(\varrho_K^{n-1} - r_K^{n-1}) \frac{\mathbf{V}^n_{h,0,K} - \mathbf{V}^{n-1}_{h,0,K}}{\Delta t} \cdot \left(\mathbf{V}^n_{h,K} - \mathbf{u}^n_K\right) \right| \leq c\|\partial_t \mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^3)}$$
$$\times \left[\left(\sum_{K\in\mathcal{T}} |K||\varrho_K^{n-1} - r_K^{n-1}|^2 1_{[\underline{r}/2,2\overline{r}]}(\varrho_K) \right)^{1/2} + \left(\sum_{K\in\mathcal{T}} |K||\varrho_K^{n-1} - r_K^{n-1}|^{6/5} 1_{R_+\setminus[\underline{r}/2,2\overline{r}]}(\varrho_K) \right)^{5/6} \right]$$
$$\times \left(\sum_{K\in\mathcal{T}} |K| \left|\mathbf{V}^n_{h,0,K} - \mathbf{u}^n_K\right|^6 \right)^{1/6} \leq c\left(\|(\partial_t \mathbf{V})\|_{L^\infty(Q_T;\mathbb{R}^3)} \right) \left(\mathcal{E}^{1/2}\left(\varrho^{n-1}, \hat{\boldsymbol{u}}^{n-1}|\hat{r}^{n-1}, \hat{\mathbf{V}}^{n-1}_{h,0}\right) \right.$$
$$\left. + \mathcal{E}^{5/6}\left(\varrho^{n-1}, \hat{\boldsymbol{u}}^{n-1}|\hat{r}^{n-1}, \hat{\mathbf{V}}^{n-1}_{h,0}\right) \right) \left(\sum_{K\in\mathcal{T}} \|\mathbf{V}^n_{h,0,K} - \mathbf{u}^n_K\|^6_{L^6(K;\mathbb{R}^3)} \right)^{1/6},$$

where we have used (8.2) and estimate (4.8) to obtain the last line. Now, we write $\mathbf{V}^n_{h,0,K} - \mathbf{u}^n_K = ([\mathbf{V}^n_{h,0} - \boldsymbol{u}^n]_K - (\mathbf{V}^n_{h,0} - \boldsymbol{u}^n)) + (\mathbf{V}^n_{h,0} - \boldsymbol{u}^n)$ and use the Minkowski inequality together with formulas (2.54), (2.43) to get

$$\left(\sum_{K\in\mathcal{T}} \|\mathbf{V}^n_{h,0,K} - \mathbf{u}^n_K\|^6_{L^6(K;\mathbb{R}^3)} \right)^{1/6} \leq \left(\sum_{K\in\mathcal{T}} \|\nabla(\mathbf{V}^n_{h,0} - \boldsymbol{u}^n)\|^2_{L^2(K;\mathbb{R}^3)} \right)^{1/2}.$$

Finally, employing Young's inequality, and estimate (4.8), we arrive at

$$|\mathcal{P}_1| \leq c\left(\delta, M_0, E_0, \underline{r}, \overline{r}, \|(\mathbf{V}, \nabla\mathbf{V}, \partial_t\mathbf{V})\|_{L^\infty(Q_T,\mathbb{R}^{15})}\right)$$
$$\times \left(\Delta t \mathcal{E}(\varrho^0, \hat{\boldsymbol{u}}^0|\hat{r}^0, \hat{\mathbf{V}}^0_{h,0}) + \Delta t \sum_{n=1}^m \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n|\hat{r}^n, \hat{\mathbf{V}}^n_{h,0}) \right) + \delta \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \|\nabla(\mathbf{V}^n_{h,0} - \boldsymbol{u}^n)\|^2_{L^2(K;\mathbb{R}^3)}, \tag{8.3}$$

with any $\delta > 0$.

Step 2: *Term* $\mathcal{P}_2$. We rewrite $\mathbf{V}^n_{h_0,\sigma} - \mathbf{V}^n_{h_0,K} = \mathbf{V}^n_{h,\sigma} - \mathbf{V}^n_{h,K} + [\mathbf{V}^n_{h,0} - \mathbf{V}^n_h]_\sigma + [\mathbf{V}^n_{h,0} - \mathbf{V}^n_h]_K$ and estimate the L^∞ norm of this expression by $h\|\nabla\mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^9)}$ by virtue of (2.48)–(2.49), (2.52)–(2.53), $(2.41)_{s=1}$. Now we write $\mathcal{P}_2 = \Delta t \sum_{n=1}^m \mathcal{P}_2^n$ where Lemma 7.2 and the Hölder inequality yield, similarly as in the previous step,

$$|\mathcal{P}_2^n| \leq c(\underline{r}, \overline{r}, \|\nabla\mathbf{V}\|_{L^\infty(Q_T;\mathbb{R}^9)})$$
$$\times \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h \left(E^{1/2}(\varrho^{n,\mathrm{up}}_\sigma|\hat{r}^{n,\mathrm{up}}_\sigma) + E^{2/3}(\varrho^{n,\mathrm{up}}_\sigma|\hat{r}^{n,\mathrm{up}}_\sigma) \right) |\hat{\mathbf{V}}^{n,\mathrm{up}}_{h,0,\sigma}| \, |\hat{\mathbf{V}}^{n,\mathrm{up}}_{h,0,\sigma} - \hat{\boldsymbol{u}}^{n,\mathrm{up}}_\sigma|$$
$$\leq c(\underline{r}, \overline{r}, \|(\mathbf{V}, \nabla\mathbf{V})\|_{L^\infty(Q_T;\mathbb{R}^{12})}) \left[\left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h \left(E(\varrho^{n,\mathrm{up}}_\sigma|\hat{r}^{n,\mathrm{up}}_\sigma) \right) \right)^{1/2} \right.$$
$$\left. + \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h E(\varrho^{n,\mathrm{up}}_\sigma|\hat{r}^{n,\mathrm{up}}_\sigma) \right)^{2/3} \right] \times \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h \left| \hat{\mathbf{V}}^{n,\mathrm{up}}_{h,0,\sigma} - \hat{\boldsymbol{u}}^{n,\mathrm{up}}_\sigma \right|^6 \right)^{1/6},$$

provided $\gamma \geq 3/2$. Next, we observe that the contribution of the face $\sigma = K|L$ to the sums $\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|hE(\varrho_\sigma^{n,\mathrm{up}}|\hat{r}_\sigma^{n,\mathrm{up}})$ and $\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|h|\hat{\boldsymbol{V}}_{h,0,\sigma}^{n,\mathrm{up}} - \hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}|^6$ is less or equal than $2|\sigma|h(E(\varrho_K^n|\hat{r}_K^n) + E(\varrho_L^n|\hat{r}_L^n))$, and than $2|\sigma|h(|\boldsymbol{V}_{h,0,K}^n - \boldsymbol{u}_K^n|^6 + |\boldsymbol{V}_{h,0,L}^n - \boldsymbol{u}_L^n|^6)$, respectively. Consequently, we get by the same reasoning as in the previous step, under assumption $\gamma \geq 3/2$,

$$|\mathcal{P}_2| \leq c(\delta, M_0, E_0, \underline{r}, \overline{r}, \|(\boldsymbol{V}, \nabla\boldsymbol{V})\|_{L^\infty(Q_T;\mathbb{R}^{12})})\, \Delta t \sum_{n=1}^m \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n|\hat{r}^n, \hat{\boldsymbol{V}}_{h,0}^n) + \delta\Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \|\nabla(\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\|_{L^2(K;\mathbb{R}^3)}^2. \tag{8.4}$$

Step 3: *Term* $\mathcal{P}_3$. We realize that

$$p(\varrho_K^n) - p'(r_K^n)(\varrho_K^n - r_K^n) - p(r_K^n) \leq c(\underline{r}, \overline{r})E(\varrho_K|r_K),$$

by virtue of Lemma 7.2 in combination with assumption (1.4). Consequently,

$$|\mathcal{P}_3| \leq c\|\operatorname{div}\boldsymbol{V}\|_{L^\infty(Q_T)}\Delta t \sum_{n=1}^m \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n|\hat{r}^n, \hat{\boldsymbol{V}}_{h,0}^n). \tag{8.5}$$

Step 4: *Term* $\mathcal{P}_4$. We write $\boldsymbol{u}^n - \boldsymbol{V}^n$ as the sum $(\boldsymbol{u}^n - \boldsymbol{V}_{h,0}^n) + (\boldsymbol{V}_{h,0}^n - \boldsymbol{V}^n)$ accordingly splitting $\mathcal{P}_4$ into two terms

$$\Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K p'(\hat{r}^n)\frac{\hat{r}^n - \varrho^n}{\hat{r}^n}(\mathbf{u}^n - \boldsymbol{V}_{h,0}^n)\cdot\nabla r^n \,\mathrm{d}x \quad \text{and} \quad \Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K p'(\hat{r}^n)\frac{\hat{r}^n - \varrho^n}{\hat{r}^n}(\boldsymbol{V}_{h,0}^n - \mathbf{V}^n)\cdot\nabla r^n \,\mathrm{d}x.$$

Reasoning similarly as in Step 2, we get

$$\begin{aligned}|\mathcal{P}_4| \leq{}& h^2\, c(\delta, M_0, E_0, \underline{r}, \overline{r}, |p'|_{C([\underline{r},\overline{r}])}\|(\nabla r, \nabla\boldsymbol{V})\|_{L^\infty(\Omega;\mathbb{R}^9)})\\ &+ c(\delta, \|\underline{r}, \overline{r}, |p'|_{C([\underline{r},\overline{r}])}\|\nabla r\|_{L^\infty(\Omega;\mathbb{R}^3)})\, \Delta t \sum_{n=1}^m \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n|\hat{r}^n, \hat{\boldsymbol{V}}_{h,0}^n) + \delta\,\Delta t \sum_{n=1}^m \sum_{K\in\mathcal{T}} \|\nabla(\boldsymbol{V}_{h,0}^n - \boldsymbol{u}^n)\|_{L^2(K;\mathbb{R}^3)}^2.\end{aligned} \tag{8.6}$$

Gathering the formulae (8.1) and (8.3)–(8.6) with δ sufficiently small (with respect to μ), we conclude the proof of Lemma 8.1. □

9. End of the proof of the error estimate (Thm. 3.2)

Finally, Lemma 8.1 in combination with the bound (4.8) yields

$$\mathcal{E}(\varrho^m, \hat{\boldsymbol{u}}^m|\hat{r}^m, \hat{\boldsymbol{V}}_{h,0}^m) \leq c\Big[h^A + \sqrt{\Delta t} + \Delta t + \mathcal{E}(\varrho^0, \hat{\boldsymbol{u}}^0|\hat{r}(0), \hat{\boldsymbol{V}}_{h,0}(0))\Big] + c\,\Delta t \sum_{n=1}^{m-1} \mathcal{E}(\varrho^n, \hat{\boldsymbol{u}}^n|\hat{r}^n, \hat{\boldsymbol{V}}_{h,0}^n);$$

whence by the discrete standard version of the Gronwall lemma one gets at the first step

$$\mathcal{E}(\varrho^m, \hat{\boldsymbol{u}}^m|\hat{r}^m, \hat{\boldsymbol{V}}_{h,0}^m) \leq c\Big[h^a + \sqrt{\Delta t} + \mathcal{E}(\varrho^0, \hat{\boldsymbol{u}}^0|\hat{r}(0), \hat{\boldsymbol{V}}_{h,0}(0))\Big].$$

Going with this information back to Lemma 8.1, one gets finally

$$\mathcal{E}(\varrho^m, \hat{\boldsymbol{u}}^m|\hat{r}^m, \hat{\boldsymbol{V}}_{h,0}^m) + \Delta t\frac{\mu}{2}\sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K |\nabla_x(\mathbf{u}^n - \mathbf{V}_{h,0}^n)|^2 \mathrm{dx} \leq c\Big[h^a + \sqrt{\Delta t} + \mathcal{E}(\varrho^0, \hat{\boldsymbol{u}}^0|\hat{r}(0), \hat{\boldsymbol{V}}_{h,0}(0))\Big]. \tag{9.1}$$

Now, we write

$$\varrho_K^n(\boldsymbol{u}_K^n - \mathbf{V}_{h,0,K}^n)^2 = \varrho_K^n(\boldsymbol{u}_K^n - \mathbf{V}^n)^2 + 2\varrho_K^n\mathbf{V}^n(\boldsymbol{u}_K^n - \mathbf{V}_{h,0,K}^n) + \varrho_K^n(\mathbf{V}^n - \mathbf{V}_{h,0,K}^n)^2,$$

where

$$\|\mathbf{V}^n - \mathbf{V}_{h,0,K}^n\|_{L^\infty(K;\mathbb{R}^3)} \stackrel{<}{\sim} \|\mathbf{V}^n - \mathbf{V}_h^n\|_{L^\infty(K;\mathbb{R}^3)} + \|\mathbf{V}_h^n - \mathbf{V}_{h,K}^n\|_{L^\infty(K;\mathbb{R}^3)} + \|[\mathbf{V}_h^n - \mathbf{V}_{h,0}^n]_K\|_{L^\infty(K;\mathbb{R}^3)}$$

$$\stackrel{<}{\sim} h\left(\|\nabla_x\mathbf{V}^n\|_{L^\infty(K;\mathbb{R}^9)} + \|\nabla_x\mathbf{V}_h^n\|_{L^\infty(K;\mathbb{R}^9)} + \|\mathbf{V}_h^n - \mathbf{V}_{h,0}^n\|_{L^\infty(K;\mathbb{R}^3)}\right) \stackrel{<}{\sim} h\|\nabla\mathbf{V}^n\|_{L^\infty(K;\mathbb{R}^9)}.$$

In the above calculation we have employed formula (2.40) to estimate the first term, estimates $(2.52)_{s=1}$, $(2.41)_{s=1}$ to estimate the second term, and formulas (2.48) and (2.49) for $K \cap \partial\Omega_h = \emptyset$ and $K \cap \partial\Omega_h \neq \emptyset$, respectively, to evaluate the last term. We conclude that

$$\sum_{K\in\mathcal{T}} \frac{1}{2}|K|\left(\varrho_K^m|\boldsymbol{u}_K^m - \boldsymbol{V}_{h,0,K}^m|^2 - \varrho_K^0|\boldsymbol{u}_K^0 - \boldsymbol{V}_{h,0,K}^0|^2\right) \geq \int_{\Omega\cap\Omega_h} \varrho^m(\hat{\boldsymbol{u}}^m - \mathbf{V}^m)^2\mathrm{d}x - \int_{\Omega\cap\Omega_h} \varrho^0(\hat{\boldsymbol{u}}^0 - \mathbf{V}^0)^2\mathrm{d}x + L_1, \tag{9.2}$$

where

$$|L_1| \stackrel{<}{\sim} h\, M_0\|\nabla_x\mathbf{V}\|_{L^\infty((0,T)\times\Omega;\mathbb{R}^9)}.$$

Similarly, we find with help of (4.8),

$$\|E(\varrho_K^n|\hat{r}^n) - E(\varrho_K^n, r^n)\|_{L^\infty(K)} \leq h\, c(M_0, \underline{r}, \overline{r}, |p|_{C^1[\underline{r},\overline{r}]}\|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)});$$

whence

$$\sum_{K\in\mathcal{T}} |K|\left(E(\varrho_K^n|\hat{r}^n) - E(\varrho_K^0|\hat{r}^0)\right) \geq \int_{\Omega\cap\Omega_h} E(\varrho^m|r^m)\mathrm{d}x - \int_{\Omega\cap\Omega_h} E(\varrho^0|r^0)\mathrm{d}x + L_2, \tag{9.3}$$

where

$$|L_2| \leq h\, c(M_0, \underline{r}, \overline{r}, |p|_{C^1[\underline{r},\overline{r}]}, \|\nabla r\|_{L^\infty(Q_T;\mathbb{R}^3)}).$$

Finally, by virtue of (2.48)–(2.49) and $(2.41)_{s=2}$

$$\|\nabla(\boldsymbol{V}_{h,0}^n - \boldsymbol{V}^n)\|_{L^2(K;\mathbb{R}^3)} \stackrel{<}{\sim} h\|(\nabla\boldsymbol{V}^n, \nabla^2\boldsymbol{V}^n)\|_{L^\infty(K;\mathbb{R}^{12})};$$

whence

$$\Delta t\sum_{n=1}^m \sum_{K\in\mathcal{T}} \int_K |\nabla_x(\mathbf{u}^n - \mathbf{V}_{h,0}^n)|^2\mathrm{d}\mathrm{x} \geq \Delta t\sum_{n=1}^m \int_{\Omega\cap\Omega_h} |(\nabla_h\mathbf{u}^n - \nabla_x\mathbf{V}^n)|^2\mathrm{d}\mathrm{x} + L_3, \tag{9.4}$$

where

$$|L_3| \leq h^2 c(\|(\nabla\boldsymbol{V}^n, \nabla^2\boldsymbol{V}^n)\|_{L^\infty(K;\mathbb{R}^{12})}).$$

Theorem 3.2 is a direct consequence of estimate (9.1) and identities (9.2)–(9.4). Theorem 3.2 is thus proved.

10. Concluding remarks

In the convergence proofs one usually needs to complete the numerical scheme by stabilizing terms, so that the new numerical scheme reads

$$\sum_{K\in\mathcal{T}_h} |K|\frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t}\phi_K + \sum_{K\in\mathcal{T}_h}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\varrho_\sigma^{n,\mathrm{up}}(\mathbf{u}_\sigma^n\cdot\mathbf{n}_{\sigma,K})\phi_K + T_c(\phi) = 0, \tag{10.1}$$

for any $\phi \in Q_h(\Omega_h)$ and $n = 1, \dots, N$,

$$\sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\left(\varrho_K^n \boldsymbol{u}_K^n - \varrho_K^{n-1}\boldsymbol{u}_K^{n-1}\right)\cdot \boldsymbol{v}_K + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\varrho_\sigma^{n,\mathrm{up}}\hat{\boldsymbol{u}}_\sigma^{n,\mathrm{up}}[\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{\sigma,K}]\cdot\boldsymbol{v}_K \tag{10.2}$$

$$- \sum_{K\in\mathcal{T}} p(\varrho_K^n)\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\boldsymbol{v}_\sigma\cdot\mathbf{n}_{\sigma,K} + \mu\sum_{K\in\mathcal{T}}\int_K \nabla\boldsymbol{u}^n : \nabla\boldsymbol{v}\ \mathrm{d}x$$

$$+\frac{\mu}{3}\sum_{K\in\mathcal{T}}\int_K \mathrm{div}\boldsymbol{u}^n\mathrm{div}\boldsymbol{v}\ \mathrm{d}x + T_m(\phi) = 0,\ \text{for any } \boldsymbol{v}\in V_{h,0}(\Omega;R^3) \text{ and } n = 1,\dots,N,$$

where

$$T_c(\phi) = h^{1-\varepsilon}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|\sigma|[\varrho^n]_{\sigma,\boldsymbol{n}_\sigma}[\phi]_{\sigma\boldsymbol{n}_\sigma},\quad T_m(\phi) = \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|\sigma|[\varrho^n]_{\sigma,\boldsymbol{n}_\sigma}\{\hat{\boldsymbol{u}}^n\}_\sigma[\hat{\phi}]_{\sigma,\boldsymbol{n}_\sigma},\ \varepsilon\in[0,1),$$

(see [20, 30]). These terms are designed to provide the supplementary positive term

$$h^{1-\varepsilon}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|\sigma|[\varrho^n]^2_{\sigma,\boldsymbol{n}_\sigma},$$

to the left hand side of the discrete energy identity (4.2). They contribute to the right hand side of the discrete relative energy (5.1) by supplementary terms whose absolute value is bounded from above by

$$h^{(1-\varepsilon)/2}\,c\left(M_0, E_0, \sup_{n=0,\dots,N}\|r^n, \mathbf{U}^n, \nabla\mathbf{U}^n\|_{L^\infty(\Omega_h;\mathbb{R}^{13})}, \sup_{n=0,\dots,N}\sup_{\sigma\in\mathcal{E}\mathrm{int}}[r^n]_{\sigma,\boldsymbol{n}_\sigma}/h\right).$$

Consequently, they give rise to the contributions at the right hand side of the approximate relative energy inequality (6.1) whose bound is

$$h^{(1-\varepsilon)/2}\,c\left(M_0, E_0, \|r, \nabla r, \mathbf{U}, \nabla\mathbf{U}\|_{L^\infty(Q_T;\mathbb{R}^{16})}\right).$$

Similar estimates are true, if we replace in the numerical scheme everywhere classical upwind formula (3.4)

$$\mathrm{Up}_K(q,\mathbf{u}) = \sum_{\sigma\in\mathcal{E}(K)} q_\sigma^{\mathrm{up}}\boldsymbol{u}_\sigma\cdot\mathbf{n}_{\sigma,K} = \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}}\left(q_K[\boldsymbol{u}_\sigma\cdot\mathbf{n}_{\sigma,K}]^+ + q_L[\boldsymbol{u}_\sigma\cdot\mathbf{n}_{\sigma,K}]^-\right),$$

by the modified upwind suggested in [15]:

$$\mathrm{Up}_K(q,\mathbf{u}) = \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}}\frac{q_K}{2}\left([\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{\sigma,K} + h^{1-\varepsilon}]^+ + [\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{\sigma,K} - h^{1-\varepsilon}]^+\right)$$

$$+\frac{q_L}{2}\left([\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{\sigma,K} + h^{1-\varepsilon}]^- + [\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{\sigma,K} - h^{1-\varepsilon}]^-\right), \tag{10.3}$$

where $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$. We will finish by formulating the error estimate for the numerical problem (3.5), (10.1), (10.2) or for (3.5), (3.6), (3.7) with modified upwind (10.3).

Theorem 10.1. *Let Ω, p, $[r_0, \mathbf{V}^0]$, $[r, V]$ satisfy assumptions of Theorem 3.2. Let $(\varrho^n, \mathbf{u}^n)_{n=0,\dots,N}$ be a family of numerical solutions to the scheme (3.5), (10.1), (10.2) or to the scheme (3.5), (3.6), (3.7) with modified upwind (10.3), where $\varepsilon\in[0,1)$. Then error estimate (3.8)* holds true with the exponent

$$a = \min\left\{\frac{2\gamma-3}{\gamma}, \frac{1-\varepsilon}{2}\right\} \text{ if } \tfrac{3}{2}\le\gamma<2,\quad a = \frac{1-\varepsilon}{2} \text{ if } \gamma\ge 2.$$

Finally, a natural question arises as top what extent the obtained error estimates are optimal. In the light of the results obtained in [28, 29], it may seem we loose, in particular in terms of the spatial discretization parameter h for $\gamma\to 3/2$. On the other hand, however, it is worth noting we do not make any extra assumption concerning boundedness of the numerical solutions in contrast with [28].

References

[1] R.A. Adams, *Sobolev spaces*. Academic Press, New York (1975).

[2] F. Brezzi and M. Fortin, Mixed and hybrid finite elements methods. In vol. 15 of *Springer series in computational mathematics* (1991).

[3] C. Cancès, H. Mathis and N. Seguin, Relative entropy for the finite volume approximation of strong solutions to systems of conservation laws. HAL: hal-00798287 (2013).

[4] Y. Cho, H.J. Choe and H. Kim, Unique solvability of the initial boundary value problems for compressible viscous fluids. *J. Math. Pures. Appl.* **83** (2004) 243–275.

[5] R. Denk, M. Hieber and J. Prüss. Optimal $L^p - L^q$-estimates for parabolic boundary value problems with inhomogenous data. *Math. Z.* **257** (2007) 193–224.

[6] M. Crouzeix and P. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO: Anal. Numer.* **7** (1973) 33–75.

[7] C. Dafermos, The second law of thermodynamics and stability. *Arch. Rational Mech. Anal.* **70** (1979) 167–179.

[8] R. Danchin, On the solvability of the compressible Navier–Stokes system in bounded domains. *Nonlinearity* **23** (2010) 383–407.

[9] L.C. Evans and R.F. Gariepy, *Measure theory and fine properties of functions*. CRC Press (1992)

[10] R. Eymard, T. Gallouët, M. Ghilani and R. Herbin, Error estimates for the approximate solutions of a non-linear hyperbolic equation given by a finite volume scheme, *IMA J. Numer. Anal.* **18** (1998) 563–594.

[11] E. Feireisl and A. Novotný, Weak-strong uniqueness property for the full Navier–Stokes-Fourier system. *Arch. Rational Mech. Anal.* **204** (2012) 683–706.

[12] E. Feireisl, A. Novotný and H. Petzeltová, On the existence of globally defined weak solutions to the Navier–Stokes equations. *J. Math. Fluid Mech.* **3** (2001) 358–392.

[13] E. Feireisl, B.J. Jin and A. Novotný, Relative entropies, suitable weak solutions and weak-strong uniqueness for the compressible Navier-Stokes system. *J. Math. Fluid Mech.* **14** (2012) 717–730.

[14] E. Feireisl, A. Novotný and Y. Sun, Suitable weak solutions to the Navier-Stokes equations of compressible viscous fluids. *Indiana Univ. Math. J.* **60** (2011) 611–631.

[15] E. Feireisl, M. Michálek and T.K. Karper, Convergence of a numerical method for the compressible Navier–Stokes system on general domains. *Inst. Math. Cz. Acad. Sci.* **57** (2014).

[16] M. Feistauer, Analysis in compressible fluid mechanics. *ZAMM* **78** (1998) 579–596.

[17] M. Feistauer, J. Felcman and V. Dolejší, Numerical simulation of compressible viscous flow through cascades of profiles. *ZAMM* **76** (1996) 297–300.

[18] M. Feistauer, J. Felcman and M. Lukáčová-Medviďova, Combined finite element – finite volume solution of compressible flow. *J. Comput. Appl. Math.* **63** (1995) 179–199.

[19] M. Feistauer, J. Felcman and I. Straškraba, Mathematical and Computational Methods for Compressible Flow. Clarendon Press, Oxford (2003)

[20] T. Gallouët, L. Gastaldo, R. Herbin and J.-C. Latché, An unconditionally stable pressure correction scheme for the compressible barotropic Navier-Stokes equations. *ESAIM: M2AN* **42** (2008) 303–331.

[21] T. Gallouët, R. Herbin and J.-C. Latché, A convergent finite element-finite volume scheme for the compressible Stokes problem. I. The isothermal case. *Math. Comp.* **78** (2009) 1333–1352.

[22] T. Gallouët, R. Herbin, D. Maltese and A. Novotný, Error estimate for a numerical approximation to the compressible barotropic Navier–Stokes equations. *IMA J. Numer. Anal.* **36** (2016) 543–592.

[23] L. Gastaldo, R. Herbin, W. Kheriji, C. Lapuerta and J.C. Latché, Staggered discretizations, pressure correction schemes and all speed barotropic flows, *Finite volumes for complex applications. VI. Problems and perspectives, Vols. 1, 2*. Vol. 4 of *Springer Proc. Math.* Springer, Heidelberg (2011) 839–855.

[24] L. Gastaldo, R. Herbin, J.-C. Latché and N. Therme, Consistency result of an explicit staggered scheme for the Euler equations. Preprint (2014).

[25] R. Herbin, W. Kheriji and J.C. Latché, On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *ESAIM: M2AN* **48** (2014) 1807–1857.

[26] R. Hošek, Strongly regular families of boundary-fitted tetrahedral meshes of bounded C^2 domains. *Preprint Inst. Math. Cz. Acad. Sci.* **3** (2016).

[27] C. Johnson and J.C. Nedelec, On the coupling of boundary integral and finite element methods. *Math. Comp.* **35** (1980) 1063–1079.

[28] V. Jovanović, An error estimate for a numerical scheme for the compressible Navier–Stokes system. *Kragujevac J. Math.* **30** (2007) 263–275.

[29] V. Jovanović and C. Rohde, Finite volume schemes for Friedrichs systems in multiple space dimensions: a priori and a posteriori estimates. *Numer. Methods Partial Differ. Equ.* **21** (2005) 104–131.

[30] K.H. Karlsen and T.K. Karper, A convergent nonconforming finite element method for compressible Stokes flow. *SIAM J. Numer. Anal.* **48** (2010) 1846–1876.

[31] T.K. Karper, A convergent FEM-DG method for the compressible Navier–Stokes equations. *Numer. Math.* **125** (2013) 441–510

[32] D. Kröner, Directionally adapted upwind schemes in 2-D for the Euler equations. Finite approximations in fluid mechanics. Vol. 25 of *Notes Numer. Fluid Mech.* Friedr. Vieweg, Braunschweig (1989) 249–263.

[33] D. Kröner, Numerical schemes for the Euler equations in two space dimensions without dimensional splitting. Nonlinear hyperbolic equations – theory, computation methods, and applications (Aachen, 1988). Vol. 24 of *Notes Numer. Fluid Mech.* Friedr. Vieweg, Braunschweig (1989) 342–352.

[34] D. Kröner, M. Rokyta and M. Wierse, A Lax - Wendroff type theorem for upwind finite volume schemes in 2-D. *East-West J. Numer. Math.* **4** (1996) 279–292.

[35] D. Kröner and M. Ohlberger, A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions. *Math. Comp.* **69** (2000) 25–39.

[36] N.V. Krylov, Parabolic equations with VMO coefficients in Sobolev spaces with mixed norms. *J. Funct. Anal.* **250** (2007) 521–558.

[37] O.A. Ladyzhenskaya, V.A. Solonnikov and N.N. Uralceva, Linear and qusilinear equations of parabolic type. Vol. 23 of *AMS, Trans. Math. Monograph.* Providence (1968).

[38] P.-L. Lions, Mathematical topics in fluid mechanics. Vol. 2, Compressible models, Oxford Science Publications. Vol. 10 of *Oxford Lect. Ser. Math. Appl.* The Clarendon Press, Oxford University Press, New York (1998).

[39] B. Liu, The analysis of a finite element method with streamline diffusion for the compressible Navier–Stokes equations. *SIAM J. Numer. Anal.* **38** (2000) 1–16.

[40] B. Liu, On a finite element method for three-dimensional unsteady compressible viscous flows. *Numer. Methods Partial Differ. Eq.* **20** (2004) 432–449.

[41] Y. Sun, C. Wang and Z. Zhang, A Beale-Kato-Majda blow-up criterion for the 3-D compressible Navier–Stokes equations *J. Math. Pures Appl.* **95** (2011) 36–47.

[42] R. Temam, Navier-Stokes equations, Theory and numerical analysis, With an appendix by F. Thomasset. Vol. 2 of *Stud. Math. Appl.* North-Holland Publishing Co., Amsterdam, 3rd edition (1984).

[43] A. Valli and M. Zajaczkowski, Navier–Stokes equations for compressible fluids: Global existence and qualitative properties of the solutions in the general case. *Commun. Math. Phys.* **103** (1986) 259–296.

[44] J.P. Villa and P. Villedieu, Convergence of an explicite finite volume scheme for first order symmetric systems. *Numer. Math.* **94** (2003) 573–602.

Appendix D

R. H.: Face-to-face partition of 3D space with identical well-centered tetrahedra.

60 (2015) APPLICATIONS OF MATHEMATICS No. 6, 637–651

FACE-TO-FACE PARTITION OF 3D SPACE WITH IDENTICAL WELL-CENTERED TETRAHEDRA

Radim Hošek, Praha

(Received April 16, 2015)

Abstract. The motivation for this paper comes from physical problems defined on bounded smooth domains Ω in 3D. Numerical schemes for these problems are usually defined on some polyhedral domains Ω_h and if there is some additional compactness result available, then the method may converge even if $\Omega_h \to \Omega$ only in the sense of compacts. Hence, we use the idea of meshing the whole space and defining the approximative domains as a subset of this partition.

Numerical schemes for which quantities are defined on dual partitions usually require some additional quality. One of the used approaches is the concept of *well-centeredness*, in which the center of the circumsphere of any element lies inside that element. We show that the one-parameter family of Sommerville tetrahedral elements, whose copies and mirror images tile 3D, build a well-centered face-to-face mesh. Then, a shape-optimal value of the parameter is computed. For this value of the parameter, Sommerville tetrahedron is invariant with respect to reflection, i.e., 3D space is tiled by copies of a single tetrahedron.

Keywords: rigid mesh; well-centered mesh; approximative domain; single element mesh; Sommerville tetrahedron

MSC 2010: 65N30, 65N50

1. Introduction

One of the widely accepted full models of a compressible, viscous and heat conducting fluid is the Navier-Stokes-Fourier system. For a convergence proof to a numerical method for this system in a smooth bounded domain $\Omega \subset \mathbb{R}^3$, developed recently in [2], we are looking for a family of approximative closed polyhedral domains Ω_h,

The research of R. Hošek leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ ERC Grant Agreement 320078.

$h \to 0$, admitting a mesh $\mathcal{T}_h$ consisting of compact convex tetrahedral elements that have diameters of order h, with the following properties.

(M1) The mesh is face-to-face, i.e., any face of any element $T \in \mathcal{T}_h$ is either a subset of $\partial\Omega_h$ or a face of another element $T' \in \mathcal{T}_h$.

(M2) The approximative domains Ω_h converge to Ω in the following sense

$$\Omega \subset \overline{\Omega} \subset \Omega_h \subset \{x \in \mathbb{R}^3 \colon \operatorname{dist}(x, \Omega) < h\}. \tag{1.1}$$

(M3) In every element $T \in \mathcal{T}_h$ there exists a point $x_T \in \operatorname{int} T$ such that for T, T' sharing a common face σ we have that the segment $x_T x_{T'}$ is orthogonal to σ and

$$d_\sigma := |x_T - x_{T'}| \geqslant ch > 0, \tag{1.2}$$

with $c > 0$ a universal constant independent of T and T'.

For the method developed in [2] we succeeded to relax the condition (1.2) to $d_\sigma > 0$. Anyway, some works discussed later require the stronger condition (1.2). Therefore, we will construct approximative domains and mesh satisfying the conditions (M1)–(M3) listed above.

Note that the usual convergence $\partial\Omega_h \to \partial\Omega$ in $W^{1,1}$ is substituted by a weaker condition (1.1) thanks to an additional result on compactness obtained.

The property (M3) emanates from the need of dealing with the Neumann boundary condition for the temperature and was introduced by Eymard et al. [1], Definition 3.6. The easiest way to ensure $d_\sigma > 0$ is to guarantee that the center of the circumsphere (also called circumcenter) of any element building the mesh lies strictly inside that element. This property is called d-well-centeredness, where d denotes the dimension. A special structure of the mesh will then imply also the existence of $c > 0$ such that $d_\sigma \geqslant ch > 0$.

The concept of well-centeredness has been extensively studied by VanderZee et al., see [10] and [11]. However, to our knowledge, there are so far only few applications, moreover without ambitions on a rigorous proof of convergence of the method.

Hirani, a coauthor of VanderZee in [10] and [11], with his colleagues uses well-centered elements in [5] for modelling the equations of Darcy's flow model. It describes the flow of a viscous incompressible fluid in a porous medium, with pressure being defined in the circumcenters of the elements. They point out that for *good quality* Delaunay mesh their method works well, and the use of a well-centered mesh is therefore not necessary.

Sazonov et al. use well-centered elements in [7] for a co-volume method for Maxwell's equations. Electric and magnetic fields are defined on mutually orthogonal meshes. As the time step has to be proportional to d_σ, it is necessary to keep it as large as possible. Therefore, well-centered mesh is used. See [7] for details.

In order to satisfy the above requirements for domains Ω_h and their meshes $\mathcal{T}_h$, we construct a 3-well-centered face-to-face mesh that covers $\mathbb{R}^3$, whose elements have radius comparable to h. Then for any $\Omega \in C^{0,1}$ given, we simply define Ω_h as a union of elements having nonempty intersection with Ω.

We will mesh the whole 3-dimensional space with an element of one type and its mirror image. This enables us to compute the exact distance of circumcenters of two neighbouring elements, but it also may reduce both memory demands and computational time.

Obviously, in 2D it is possible to tile the whole space with regular simplices, which are equilateral triangles. In 3D it is not that easy, the regular tetrahedra do not tile 3D, see e.g. [8]. However, there have been shown many tilings of 3D so far. Sommerville in 1923 ([9], page 56) introduced a one-parameter family of elements that tile an infinite prism with equilateral-triangular base (see also Goldberg [4]). We will deal with these *Sommerville II type* elements and show the range of the parameter for which they build a 3-well-centered mesh. Such mesh will then fulfil (M1)–(M3). Moreover, we compute in a sense an *ideal value* of the parameter which will guarantee that all tetrahedra in the mesh are identical.

2. Notation

We work in $\mathbb{R}^3$, a 3-dimensional space endowed with Euclidean coordinates. Then for $m \leqslant 3$, σ^m or τ^m will denote a simplex, which is a convex hull of $m+1$ affinely independent points in $\mathbb{R}^3$. We recall that points $\{P_0, P_1, \dots, P_m\}$ are affinely independent if

$$\left(\sum_{i=0}^{m} c_i P_i = 0 \quad \& \quad \sum_{i=0}^{m} c_i = 0\right) \Rightarrow c_i = 0 \quad \forall\, i \in \{0, \dots, m\}.$$

Analogously, every simplex σ^m determines an m-dimensional affine space.

We introduce the following list of the used notation.

$A, B, C, \dots$	points in $\mathbb{R}^3$
σ^m, τ^m or also $P_0P_1 \dots P_m$	m-dimensional simplex
$\mathrm{aff}(\sigma^m)$	affine space determined by (vertices of) σ^m
S_{σ^m}	circumcenter of σ^m
Σ_{σ^m}	incenter of σ^m (center of the inscribed sphere of σ^m)
R_{σ^m}	radius of the circumsphere of σ^m
ϱ_{σ^m}	radius of the inscribed sphere of σ^m

Note that the above notation can be used independently of the dimension. We will use also the following dimension-dependent notation.

$A = [A^x, A^y, A^z]$	point with its Euclidean coordinates
$\mathbf{n}_{ABC}$	normal vector of the plane ABC
o_{AB}	axial plane of the segment AB
$\mathbf{o}_{AB(C)}$	axis of the segment AB in the plane ABC

3. 3-well-centered mesh of 3-dimensional space

3.1. Elements. Following [9], we define the tetrahedron $\tau^3(p)$ depending on a positive parameter p with the following Euclidean coordinates of its vertices:

$$
\begin{aligned}
&\tau^3(p) := (ADEF)(p), \quad p > 0,\\
&A = [0, 0, 0],\\
&D = [0, 0, 3p],\\
&E = [1, 0, p],\\
&F = \Big[\frac{1}{2}, \frac{\sqrt{3}}{2}, 2p\Big],
\end{aligned}
\tag{3.1}
$$

see Figure 1. All the vertices and also further derived quantities depend on p, which will be often omitted in the notation for the sake of brevity.

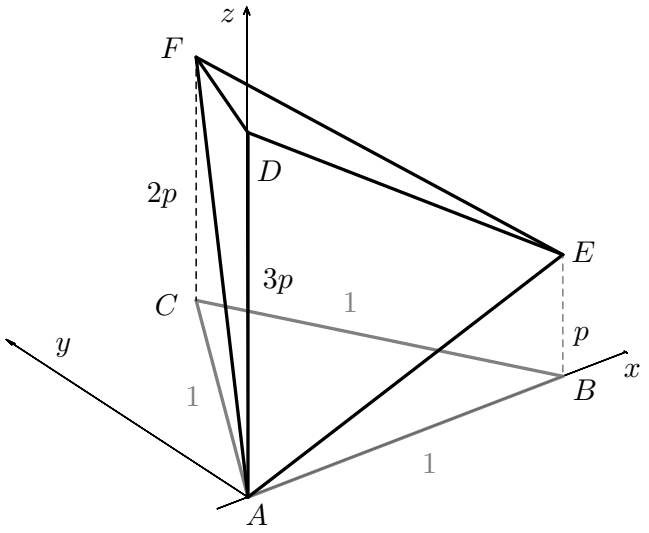


Figure 1. Element $\tau^3(p)$ defined in (3.1).

3.2. Tiling the space. Consider tetrahedra $ADEF(p)$, $DEFE'(p)$, $DE'FF'(p)$, where

$$E' = E + 3p \cdot \vec{e}_3,$$
$$F' = F + 3p \cdot \vec{e}_3,$$

see Figure 2. They are identical and build a skew prism with an equilateral triangle as its base. Repeating the structure periodically in the z direction, we can fill the whole infinite triangular prism. It is obvious that with copies and reflections of those prisms we can tile the whole 3-dimensional space, which follows from the tiling of 2D with equilateral triangles. The task is to show that we can tile in such way that the elements build a face-to-face mesh.

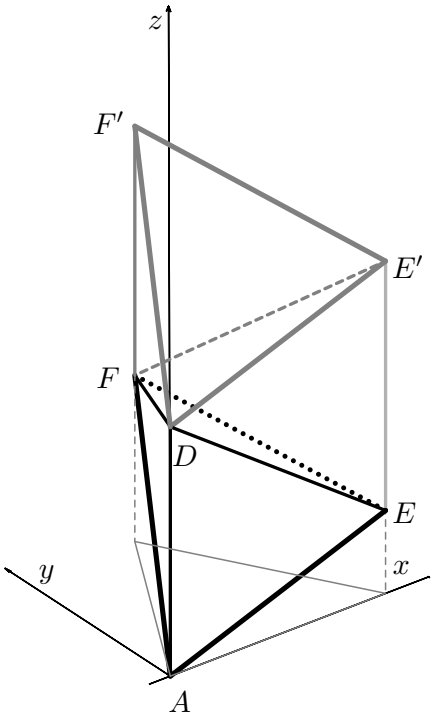


Figure 2. Three copies of element $\tau^3(p)$ arranged in a prism with equilateral-triangular base.

Lemma 3.1. *It is possible to create a face-to-face partition of $\mathbb{R}^3$ with copies of the tetrahedron $\tau^3(p)$ and its mirror images.*

P r o o f. After previous discussion it suffices to show that infinite prisms built with elements $\tau^3(p)$ can be arranged such that the elements' edges on the prism surfaces meet. Note that each infinite prism is a convex hull of three vertical lines of three different types, each of them having vertices of elements in the height $3k + r$, $k \in \mathbb{Z}$, for $r = 0, 1, 2$. Projecting the whole situation into xy-plane, it suffices to show that an equilateral triangulation of $\mathbb{R}^2$ is a 3-vertex-colorable graph. As neighbouring

triangles in $\mathbb{R}^2$ share an edge, their preimages share an infinite strip where the edges (and thus also the facets) of elements coincide.

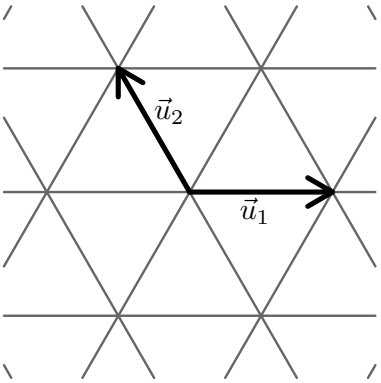


Figure 3. Illustration to the proof of Lemma 3.1: xy-plane with the basis $\vec{u}_1$, $\vec{u}_2$.

Employing the basis $\vec{u}_1 = (1,0)$, $\vec{u}_2 = \frac{1}{2}(-1,\sqrt{3})$, any vertex v of equilateral triangulation of xy plane has unique coordinates, i.e., $\vec{v} = c_1\vec{u}_1 + c_2\vec{u}_2$, with *integer values* of c_1, c_2, see Figure 3. Then for vertex v we define its color $\xi(v)$ equal to

$$\xi(v) = c_1 + c_2 \mod 3.$$

Note that for any neighbouring vertices v, w we have

$$\vec{v} - \vec{w} = d_1\vec{u}_1 + d_2\vec{u}_2,$$

with $(d_1, d_2) \in \{(1,0),(1,1),(0,1),(-1,0),(-1,-1),(0,1)\}$. Hence, we conclude that $\xi(v) \neq \xi(w)$, i.e., ξ is indeed a vertex coloring. □

An alternative proof is suggested in [6]. Reflecting the triplet of elements shown in Figure 2 with respect to the point $P = (D + E)/2$, we obtain a parallelepiped. Its copies tile the 3-dimensional space and it can be checked that the face-to-face property of the mesh is not violated.

Note that so far we do not restrict the value of p, i.e., copies and reflections of $\tau^3(p)$ tile $\mathbb{R}^3$ for any $p > 0$.

3.3. Well-centeredness. We introduce the concept of well-centeredness by the definition of VanderZee, see [10], page 5.

Definition 3.2. Let $0 \leqslant k \leqslant n \leqslant d$. Let $\sigma^n := \{V_0V_1\dots V_n\}$ be an n-dimensional simplex. A k-dimensional face of σ^n is a simplex $\sigma^k := \{U_0U_1\dots U_k\}$ with U_i being distinct vertices of σ^n. We say that

(1) σ^n is n-well-centered if its circumcenter lies in the interior of σ^n,
(2) for $1 \leqslant k < n$, σ^n is k-well-centered if all its k-dimensional faces are k-well centered,
(3) σ^n is well-centered if it is k-well centered for all $k \in \{1, \dots, n\}$.

Note that any simplex is 1-well-centered, as the midpoint of any segment lies strictly inside the segment. In $\mathbb{R}^2$, a triangle is well-centered if and only if it is acute.

VanderZee et al. in [10] prove the following characterization for n-well-centeredness of an n-dimensional simplex.

Theorem 3.3 (VanderZee)**.** *The n-dimensional simplex $\sigma_n = V_0V_1 \dots V_n$ is n-well centered if and only if for each $i = 0, \dots, n$ the vertex V_i lies outside the circumsphere $B_i^n := B(V_0, V_1, \dots, V_{i-1}, V_{i+1}, \dots, V_n)$, which is the smallest ball in $\mathbb{R}^n$ which contains the $(n-1)$-dimensional circumsphere of the simplex $V_0V_1 \dots V_{i-1}V_{i+1} \dots V_n$.*

Theorem 3.3 will be our tool for proving the following main Theorem 3.4.

Theorem 3.4. *The tetrahedron $\tau^3(p) = ADEF(p)$ defined by* (3.1) *is 3-well-centered if and only if*

$$p < \frac{\sqrt{2}}{2}. \tag{3.2}$$

P r o o f. The proof is a simple but laborious computation based on the result of Theorem 3.3, from which we will get the desired restriction on p. Let K, L, M, N be affinely independent points in $\mathbb{R}^3$ and let the circumsphere of the triangle LMN have the radius r_{LMN} and center S_{LMN}. The goal is to determine the value of p for which

$$|K - S_{LMN}| > r_{LMN} \tag{3.3}$$

is valid for all vertices A, D, E, F alternating in the role of K. We have all necessary ingredients for the computation since we can compute

$$S_{LMN} = \mathbf{o}_{LM(N)} \cap \mathbf{o}_{LN(M)}, \tag{3.4}$$

where

$$\begin{aligned} \mathbf{o}_{LM(N)} &= S_{LM} + t \cdot \mathbf{n}_{LMN} \times \overrightarrow{LM}, \quad t \in \mathbb{R}, \\ \mathbf{o}_{LN(M)} &= S_{LN} + t \cdot \mathbf{n}_{LMN} \times \overrightarrow{LN}, \quad t \in \mathbb{R}, \\ \mathbf{n}_{LMN} &= \overrightarrow{LM} \times \overrightarrow{LN}, \end{aligned} \tag{3.5}$$

for given points K, L, M, N.

1. Vertex D

Substituting the ordered quadruplet $[D, A, E, F]$ for $[K, L, M, N]$ in (3.3), (3.4), and (3.5), and performing the computations, we get

$$\begin{aligned} \mathbf{n}_{AEF} &= \Big(-\frac{\sqrt{3}}{2}p, -\frac{3}{2}p, \frac{\sqrt{3}}{2}\Big), \\ \mathbf{o}_{AE(F)} &= \Big[\frac{1}{2}, 0, \frac{p}{2}\Big] + u\Big(-\frac{3}{2}p^2, \frac{\sqrt{3}}{2}(1+p^2), \frac{3}{2}p\Big), \quad u \in \mathbb{R}, \\ \mathbf{o}_{AF(E)} &= \Big[\frac{1}{4}, \frac{\sqrt{3}}{4}, p\Big] + v\Big(-\frac{3}{4} - 3p^2, \frac{\sqrt{3}}{4} + \sqrt{3}p^2, 0\Big), \quad v \in \mathbb{R}, \end{aligned} \tag{3.6}$$

from which we obtain

$$S_{AEF} = \Big[\frac{1}{2}(1-p^2), \frac{\sqrt{3}}{6}(1+p^2), p\Big].$$

To conclude for which values of p it holds that $|D - S_{AEF}| > r_{AEF} = |A - S_{AEF}|$, it is sufficient to compare the third component of both expressions only, since A and D differ only in that one. We get

$$|\vec{e}_3 \cdot (S_{AEF} - A)| < |\vec{e}_3 \cdot (S_{AEF} - D)|$$

for any $p > 0$, i.e., condition (3.3) holds for $K = D$, $LMN = AEF$, $p > 0$.

2. Vertex F

Using elementary analytic geometry in $\mathbb{R}^2$ (ADE lies in the xz-plane), we obtain the parametric equations of the axes,

$$\begin{aligned} \mathbf{o}_{AD(E)} &= \Big[0, 0, \frac{3}{2}p\Big] + u(1, 0, 0), \quad u \in \mathbb{R}, \\ \mathbf{o}_{AE(D)} &= \Big[\frac{1}{2}, 0, \frac{1}{2}p\Big] + v(p, 0, -1), \quad v \in \mathbb{R}, \end{aligned}$$

and their intersection

$$S_{ADE} = \Big[\frac{1}{2} - p^2, 0, \frac{3}{2}p\Big]. \tag{3.7}$$

We want to obtain a bound on p such that

$$|S_{ADE} - F|^2 - r_{ADE}^2 = |S_{ADE} - F|^2 - |S_{ADE} - A|^2 > 0.$$

Substituting from (3.1) and (3.7) we get from the inequality above that

$$p < \sqrt{\frac{1}{2}} = \frac{\sqrt{2}}{2}. \tag{3.8}$$

3. Vertex E

Substituting the quadruplet $[E, A, D, F]$ for $[K, L, M, N]$ into the scheme (3.3), (3.4), and (3.5), one can compute

$$\begin{aligned}
\mathbf{n}_{ADF} &= \Big(-\frac{3\sqrt{3}}{2}p, \frac{3}{2}p, 0\Big),\\
\mathbf{o}_{AD(F)} &= \Big[0, 0, \frac{3}{2}p\Big] + u\Big(\frac{9}{2}p^2, \frac{9\sqrt{3}}{2}p^2, 0\Big), \quad u \in \mathbb{R},\\
\mathbf{o}_{AF(D)} &= \Big[\frac{1}{4}, \frac{\sqrt{3}}{4}, p\Big] + v(-3p^2, 3\sqrt{3}p^2, -3p), \quad v \in \mathbb{R},
\end{aligned}$$

from which we obtain

$$S_{ADF} = \Big[\frac{1}{4} + \frac{1}{2}p^2, \frac{\sqrt{3}}{4} + \frac{\sqrt{3}}{2}p^2, \frac{3}{2}p\Big]. \tag{3.9}$$

Again, we want to get a bound on p for which

$$|S_{ADF} - E|^2 - r_{ADF}^2 = |S_{ADF} - E|^2 - |S_{ADF} - A|^2 > 0.$$

Substituting from (3.9), we arrive at

$$p < \sqrt{\frac{2}{3}},$$

which is a weaker requirement than already obtained in (3.8) and therefore does not affect the result.

4. Vertex A

Finally, taking $[K, L, M, N] = [A, D, E, F]$ and performing the computations, we get

$$\begin{aligned}
\mathbf{n}_{DEF} &= \Big(\sqrt{3}p, 0, \frac{\sqrt{3}}{2}\Big),\\
\mathbf{o}_{DE(F)} &= \Big[\frac{1}{2}, 0, 2p\Big] + u\Big(0, \frac{\sqrt{3}}{2} + 2\sqrt{3}p^2, 0\Big), \quad u \in \mathbb{R},\\
\mathbf{o}_{DF(E)} &= \Big[\frac{1}{4}, \frac{\sqrt{3}}{4}, \frac{5}{2}p\Big] + v\Big(-\frac{3}{4}, \frac{\sqrt{3}}{4} + \sqrt{3}p^2, \frac{3}{2}p\Big), \quad v \in \mathbb{R},
\end{aligned} \tag{3.10}$$

which gives

$$S_{DEF} = \Big[\frac{1}{2}, \frac{\sqrt{3}}{6} - \frac{\sqrt{3}}{3}p^2, 2p\Big].$$

By the same token as in the first case, $|\vec{e}_3 \cdot (S_{DEF} - A)| > |\vec{e}_3 \cdot (S_{DEF} - D)|$ for any value of $p > 0$, which implies that $|A - S_{DEF}| > r_{DEF} = |D - S_{DEF}|$ for any $p > 0$. □

Corollary 3.5. *The tetrahedron $\tau^3(p)$ is well-centered if and only if*

$$p \in \Big(0, \frac{\sqrt{2}}{2}\Big).$$

P r o o f. Using the characterization of an acute triangle (i.e., $a^2 + b^2 > c^2$, where $c \leqslant b \leqslant a$), one can check that for $\tau^3(p), p \in (0, \sqrt{2}/2)$ all faces are 2-well-centered. The tetrahedron $\tau^3(p)$ is 3-well-centered for $p \in (0, \sqrt{2}/2)$ by virtue of Theorem 3.4. □

VanderZee et al. introduced also a sufficient condition of n-well-centeredness, the so called Prism Condition, [11], Proposition 8, which applied to $\tau^{n-1} = AED$ and $v = F$ gives the condition $p < 1/2$. This is more restrictive than the condition (3.2), which we get by the equivalence criterion in Theorem 3.3.

We state the following corollary.

Corollary 3.6. *Let $\Omega \subset \mathbb{R}^3$ be a smooth (at least Lipschitz) bounded domain. Then there exists a family of polyhedral domains $\{\Omega_h\}_{h\to 0}$, such that any Ω_h admits a face-to-face mesh $\mathcal{T}_h$, satisfying the conditions* (1.1) *and* (1.2).

P r o o f. For $h > 0$ and $p \in (0, \frac{1}{2}\sqrt{2})$ arbitrary take the tetrahedron $\tau_h^3(p) := \frac{1}{2}h \cdot \tau^3(p)$ and mesh the whole $\mathbb{R}^3$ in the way described in Section 3.2. Denoting the whole mesh with $\widetilde{\mathcal{T}_h}$ and defining the set $\mathcal{T}_h := \{T \in \widetilde{\mathcal{T}_h};\ T \cap \Omega \neq \emptyset\}$, we put

$$\Omega_h := \bigcup_{T\in\mathcal{T}_h} T.$$

The face-to-face property follows from Lemma 3.1. Convergence in the sense of (1.1) is guaranteed, since for $T \in \mathcal{T}_h$ we have

$$\operatorname{diam} \tau_h^3(p) \leqslant \frac{h}{2}\sqrt{1 + (2p)^2} \leqslant h\frac{\sqrt{3}}{2} < h.$$

Finally, the property (1.2) is satisfied by virtue of Theorem 3.4 and the fact that the mesh is build by elements with equal radius of the inscribed sphere, i.e., $d_\sigma > h\varrho(\tau^3(p))$. The value of $\varrho(\tau^3(p))$ will be specified in the next section, see Proposition 4.1. □

4. Shape optimization

Notice that we have a criterion for the well-centeredness of our elements in a form of an open interval $p \in (0, \sqrt{1/2})$. We would like to get an *optimal value* from the computational point of view, which we expect to be *far enough* especially from the singular value $p = 0$. One of the criteria used (see [3] or [6]) is the so-called *normalized shape ratio*. Using the notation introduced in Section 2, we define the normalized shape ratio of tetrahedron σ^3 by

$$\eta(\sigma^3) := \frac{3\varrho(\sigma^3)}{R(\sigma^3)}. \tag{4.1}$$

The maximal value of (4.1) is $\eta = 1$ for the regular tetrahedron. In what follows we use a shorter notation $\varrho(p) := \varrho(\tau^3(p))$, analogously also for R and η. Next we compute the radii in dependence on p.

Proposition 4.1. *The radius $\varrho(p)$ of the inscribed sphere of the tetrahedron $\tau^3(p)$ equals*

$$\varrho(p) = \frac{3}{4\sqrt{3} + 2\sqrt{4 + 1/p^2}}. \tag{4.2}$$

P r o o f. Note that having tetrahedron $\tau^3(p)$ placed in Euclidean coordinates, we have $\varrho(p) = \Sigma^y$, where $\Sigma = [\Sigma^x, \Sigma^y, \Sigma^z]$ are the coordinates of the center of the inscribed sphere.

As the faces ADE and ADF are vertical, orthogonal projection of τ^3 and its inscribed sphere into xy-plane is an equilateral triangle ABC and a circle that touches both segments AB and AC (see Figure 4). The center of the circle $P(\Sigma) = [\Sigma^x, \Sigma^y, 0]$

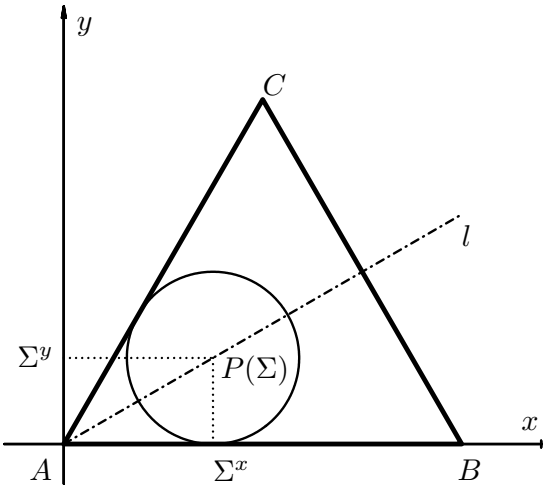


Figure 4. Projection of $\tau^3(p)$ and its inscribed sphere into the xy-plane.

must lie on a bisector of the 60° angle BAC. Hence,

$$\Sigma^x = \sqrt{3}\Sigma^y. \tag{4.3}$$

Further, the center Σ must lie on α, an axial plane of the dihedral angle of the planes aff(AEF) and aff(DEF). Recalling $\mathbf{n}_{AEF}$ and $\mathbf{n}_{DEF}$ from $(3.6)_1$ and $(3.10)_1$ respectively, and realizing that their lengths are equal, we can compute

$$\alpha\colon\ \mathbf{n}_\alpha \cdot \mathbf{x} + d = 0, \tag{4.4}$$

with $\mathbf{n}_\alpha = \frac{1}{2}(\mathbf{n}_{AEF} + \mathbf{n}_{DEF})$. Then d is determined by substituting $\mathbf{x} = E$ into (4.4) and we get

$$\alpha\colon\ \frac{\sqrt{3}}{4}px - \frac{3}{4}py + \frac{\sqrt{3}}{2}z - \frac{3\sqrt{3}}{4}p = 0. \tag{4.5}$$

Substituting $(\Sigma^x, \Sigma^y, \Sigma^z)$ into (4.5) and using (4.3) leads to conclusion that $\Sigma^z = \frac{3}{2}p$. Our problem reduces to finding a point

$$\Sigma = \Sigma(p) = \Big[\sqrt{3}\varrho(p), \varrho(p), \frac{3}{2}p\Big], \tag{4.6}$$

such that $\operatorname{dist}(AEF, \Sigma(p)) = \varrho(p)$. Such point Σ lies in a plane given by a normal vector $\mathbf{n}_{AEF}$ and point $\varrho(p)\mathbf{n}_{AEF}/|\mathbf{n}_{AEF}|$. The general equation of this plane can be expressed as

$$\mathbf{n}_{AEF} \cdot (x, y, z)^T - \varrho(p)\frac{|\mathbf{n}_{AEF}|^2}{|\mathbf{n}_{AEF}|} = 0,$$

which is

$$-\frac{\sqrt{3}}{2}px - \frac{3}{2}py + \frac{\sqrt{3}}{2}z - \varrho(p)\sqrt{3p^2 + \frac{3}{4}} = 0. \tag{4.7}$$

Substituting (4.6) to (4.7) yields the final result. □

Proposition 4.2. *The radius of the circumsphere to tetrahedron $\tau^3(p)$ is given by*

$$R(p) = \sqrt{\frac{4}{3}p^4 + \frac{11}{12}p^2 + \frac{1}{3}}. \tag{4.8}$$

P r o o f. For the radius we have that $R = |S - A| = |S|$. Hence, only the center $S = [S^x, S^y, S^z]$ of the circumsphere is of our interest. We proceed in two steps.

First, $|SD| = |SA| = |SE|$ suffices to determine both S^x and S^z. The point S must lie on a line which is a cross-section of axial planes o_{AE} and o_{DE},

$$o_{AE}\colon \left[\frac{1}{2}, 0, \frac{p}{2}\right] + r(0,1,0) + s(-p,0,1), \quad r,s \in \mathbb{R},$$
$$o_{DE}\colon \left[\frac{1}{2}, 0, 2p\right] + r(0,1,0) + t(-2p,0,-1), \quad r,t \in \mathbb{R}.$$

From this we easily conclude that

$$S \in (o_{AE} \cap o_{DE}) = (S^x, 0, S^z) + r(0,1,0), \quad r \in \mathbb{R}, \tag{4.9}$$

where further computation gives $S^x = \frac{1}{2} - p^2$ and $S^z = \frac{3}{2}p$.

Second, we determine S^y by computing the appropriate value of parameter r in (4.9) from the equality $|SA| = |SF|$, we get

$$S = \left[\frac{1}{2} - p^2, \frac{1}{\sqrt{3}}\left(\frac{1}{2} - p^2\right), \frac{3}{2}p\right].$$

We finish the proof with computing $R = |S|$, which gives (4.8). □

Theorem 4.3. *Let $\tau^3(p)$, $p \in (0, \sqrt{2}/2)$ be a one-parameter family of tetrahedra defined in* (3.1)*. Let $\varrho(p)$ be the radius of its inscribed sphere and $R(p)$ the radius of its circumsphere. Then $\eta(p)$ defined by* (4.1) *is maximal for*

$$p = p^\star = \sqrt{\frac{1}{8}}.$$

P r o o f. Both $\varrho(p), R(p)$ being continuously differentiable, one can search for the optimum as a point of vanishing derivative. If we obtain one critical point in $\mathbb{R}^+$, it has to be maximum since $\eta(p) > 0$ and

$$\lim_{p\to 0^+} \eta(p) = \lim_{p\to\infty} \eta(p) = 0. \tag{4.10}$$

The relations in (4.10) are derived using basic algebra of limits from

$$\lim_{p\to 0^+} \varrho(p) = 0, \qquad \lim_{p\to 0^+} R(p) = \frac{\sqrt{3}}{3},$$

and

$$\varrho(p) < 1 \text{ for all } p > 0, \qquad \lim_{p\to\infty} R(p) = \infty.$$

Solving the equation $\eta'(p) = 0$ leads to searching for roots of

$$32\left(2 + \sqrt{3}\cdot\sqrt{\frac{1}{p^2} + 4}\right)p^6 + \left(30 + 11\sqrt{3}\cdot\sqrt{\frac{1}{p^2} + 4}\right)p^4 - 2 = 0,$$

which, employing a new variable $b = p^2$, can be shown to have unique solution in positive real half-axis, which is $b^\star = 1/8$, therefore $p^\star = \sqrt{1/8}$. □

Note that $\tau^3(p^\star)$ is unique in the family of Sommerville II type tetrahedra having the property that they are identical with their mirror image. Therefore, for $p = p^\star$, we get a mesh that is build by copies of a single element. Moreover, $\tau^3(p^\star)$ has all faces identical—isosceles triangles with the ratio of the leg to the base equal to $\sqrt{3}/2$. Dihedral angles of $\tau(p^\star)$ are equal to 90° at the longer edges and 60° at the shorter ones. Naylor in [6] calls $\tau(p^\star)$ an *isotet*, or it is called simply the *Sommerville tetrahedron*. Substituting $p^\star$ into (4.2) and (4.8) gives

$$\eta(p^\star) = \frac{3\varrho(p^\star)}{R(p^\star)} = \sqrt{\frac{9}{10}} \approx 0.949.$$

As for Naylor (see [6]), this is a maximal value of η for meshing 3-dimensional space with a single element type.

R e m a r k 4.4. Analogously, it can be shown that the value $p = p^\star$ is ideal also in the sense of maximizing the ratio of the inscribed sphere to the diameter of an element. Note that $\operatorname{diam}\tau^3(p) = \sqrt{1+4p^2}$. One can compute that

$$\kappa(\tau^3(p^\star)) := \frac{\varrho(p^\star)}{\operatorname{diam}\tau^3(p^\star)} = \frac{\sqrt{3}/8}{\sqrt{3/2}} = \frac{\sqrt{2}}{8}.$$

We summarize the above discussion in the following corollary. If we use the construction of the approximative domain and mesh introduced in the proof of Corollary 3.6 with the choice $p = p^\star = \sqrt{1/8}$, it is possible to get a family of approximative domains admitting meshing by tetrahedra of one type.

Corollary 4.5. *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with Lipschitz boundary. Then there exists a family of polyhedral domains $\{\Omega_h\}_{h\to 0}$ such that any Ω_h admits a face-to-face mesh $\mathcal{T}_h$, satisfying conditions* (1.1) *and* (1.2) *and such that all the elements in $\mathcal{T}_h$ are identical.*

References

[1] *R. Eymard, T. Gallouët, R. Herbin*: Finite volume methods. Handbook of Numerical Analysis. Vol. 7: Solution of Equations in $\mathbb{R}^n$ (Part 3). Techniques of Scientific Computing (Part 3) (P. Ciarlet et al., eds.). North-Holland/Elsevier, Amsterdam, 2000, pp. 713–1020. zbl MR

[2] *E. Feireisl, R. Hošek, M. Michálek*: A convergent numerical method for the full Navier-Stokes-Fourier system in smooth physical domains. Submitted to SIAM J. Numer. Anal. (2015). Available as preprint IM-2015-3 at http://math.cas.cz.

[3] *D. A. Field, W. D. Smith*: Graded tetrahedral finite element meshes. Int. J. Numer. Methods Eng. *31* (1991), 413–425. zbl

[4] *M. Goldberg*: Three infinite families of tetrahedral space-fillers. J. Comb. Theory, Ser. A *16* (1974), 348–354. zbl MR

[5] *A. N. Hirani, K. B. Nakshatrala, J. H. Chaudhry*: Numerical method for Darcy flow derived using discrete exterior calculus. ArXiv:0810.3434 [math.NA] (2008).
[6] *D. J. Naylor*: Filling space with tetrahedra. Int. J. Numer. Methods Eng. *44* (1999), 1383–1395. zbl MR
[7] *I. Sazonov, O. Hassan, K. Morgan, N. P. Weatherill*: Yee's scheme for the integration of Maxwell's equation on unstructured meshes. Proceedings of the European Conference on Computational Fluid Dynamics (ECCOMAS CFD 2006) (P. Wesseling, et al., eds.). TU Delft, The Netherlands, 2006.
[8] *M. Senechal*: Which tetrahedra fill space? Math. Mag. *54* (1981), 227–243. zbl MR
[9] *D. Sommerville*: Space-filling tetrahedra in Euclidean space. Proc. Edinburgh Math. Soc. *41* (1923), 49–57.
[10] *E. VanderZee, A. N. Hirani, D. Guoy, E. A. Ramos*: Well-centered triangulation. SIAM J. Sci. Comput. *31* (2010), 4497–4523. zbl MR
[11] *E. VanderZee, A. N. Hirani, D. Guoy, V. Zharnitsky, E. A. Ramos*: Geometric and combinatorial properties of well-centered triangulations in three and higher dimensions. Comput. Geom. *46* (2013), 700–724. zbl MR

Author's address: *Radim Hošek*, Institute of Mathematics, Czech Academy of Sciences, Žitná 25, CZ-115 67 Praha 1, Czech Republic, e-mail: `hosek@math.cas.cz`.

Appendix E

R. H.: Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains.

61 (2016) APPLICATIONS OF MATHEMATICS No. 3, 233–251

STRONGLY REGULAR FAMILY OF BOUNDARY-FITTED TETRAHEDRAL MESHES OF BOUNDED C^2 DOMAINS

Radim Hošek, Praha

(Received January 8, 2016)

Abstract. We give a constructive proof that for any bounded domain of the class C^2 there exists a strongly regular family of boundary-fitted tetrahedral meshes. We adopt a refinement technique introduced by Křížek and modify it so that a refined mesh is again boundary-fitted. An alternative regularity criterion based on similarity with the Sommerville tetrahedron is used and shown to be equivalent to other standard criteria. The sequence of regularities during the refinement process is estimated from below and shown to converge to a positive number by virtue of the convergence of q-Pochhammer symbol. The final result takes the form of an implication with an assumption that can be obviously fulfilled for any bounded C^2 domain.

Keywords: boundary fitted mesh; strongly regular family; Sommerville tetrahedron; Sommerville regularity ratio; mesh refinement; tetrahedral mesh

MSC 2010: 65N30, 65N50

1. Introduction

In numerical schemes approximating PDE problems, smooth domains Ω are often approximated by polyhedral domains Ω_h that are split into tetrahedral meshes. Each such mesh is characterized by a discretization parameter h, bounding from above the size of elements. For convergence proofs, we need this parameter to decrease to zero, usually by decomposition of every element into several smaller ones. Using this process we create a new, finer mesh. However, during this process we need to control the quality of the mesh, mainly the *shape regularity*, excluding the occurrence of extremely flat or prolonged elements, see [3], Section 14.

The research of R. Hošek leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement 320078.

Creating such *strongly regular refinement of the mesh* is elementary in 2D, the technique for 3D case was shown by Křížek in [9]. In our work we will have special requirement on the mesh: The vertices of the mesh that lie on the boundary of the polyhedral domain $\partial\Omega_h$ should lie also on the boundary of the smooth domain $\partial\Omega$. We call such mesh *boundary-fitted*. The proof of existence of such a refinement for 2D can be found in [8], for 3D we bring the result in this paper.

The motivation for this work emanates from [5], where the authors define a numerical method for compressible Navier-Stokes equations in a strongly regular family of boundary-fitted meshes.

We start with the following three definitions and state the main result afterwards.

Definition 1. Let $\Omega \subset \mathbb{R}^3$ be a bounded domain of the class C^2. We denote by $r_\Omega \in \mathbb{R}^+$ the minimal radius of an osculation sphere of $\partial\Omega$ and set $h_0 := \min\{\frac{1}{2}r_\Omega, \frac{1}{2}\alpha\}$, where α is a lower bound for the mutual distance of two parts of the boundary $\partial\Omega$.

For the exact definition of α we refer to the standard Evans' PDE textbook [4], page 626.

Definition 2. We say that a couple $(\Omega_h, \mathcal{T}_h)$ is an *approximative domain with a boundary-fitted mesh* of Ω, if $\partial\Omega_h$ consists of triangles, vertices of these triangles belong to $\partial\Omega$ and $\mathcal{T}_h$ is a mesh consisting of closed tetrahedral elements K satisfying the following conditions:

- ▷ For any element $K \in \mathcal{T}_h$, any of its faces is either a face of another element $L \in \mathcal{T}_h$, or a face of the polyhedron Ω_h,
- ▷ $\operatorname{diam} K \leqslant h \leqslant h_0$ for any $K \in \mathcal{T}_h$,
- ▷ $\bigcup\limits_{K\in\mathcal{T}_h} K = \overline{\Omega}_h$.

Further, we denote by $\varrho(K)$ the radius of the largest ball contained in the element K.

Definition 3. We say that the infinite sequence $\{\mathcal{T}_h\}_{h\to 0}$ is a *family of boundary-fitted meshes* if for any $\varepsilon > 0$ there exists $h \in (0, \varepsilon)$ such that $\mathcal{T}_h$ is a boundary-fitted mesh in the sense of Definition 2.

In addition, if there exists $\theta_0 > 0$ independent of h such that for any $\mathcal{T}_h$ and any $K \in \mathcal{T}_h$ we have

$$\theta(K) := \frac{\varrho(K)}{\operatorname{diam} K} \geqslant \theta_0,$$

we say that $\{\mathcal{T}_h\}_{h\to 0}$ is a *strongly regular* family.

There are several equivalent definitions of strong regularity, see [2]. We introduce a different regularity criterion and use it later in this work.

Having introduced the basic definitions, we can state the main theorem.

Theorem 1. *Let Ω be a bounded domain in $\mathbb{R}^3$ of the class C^2. Suppose that for some $h_1 \leqslant h_0$ there exists an approximative domain $(\Omega_{h_1}, \mathcal{T}_{h_1})$ with boundary-fitted mesh and let*

$$\theta(K) \geqslant \frac{4b\sqrt{2}}{r_\Omega} \operatorname{diam} K \tag{1}$$

for any $K \in \mathcal{T}_{h_1}$, where

$$b > b_0 = \frac{8}{\sqrt{3}}(2+\sqrt{5}). \tag{2}$$

Then there exists a strongly regular family of boundary-fitted meshes $\{\mathcal{T}_h\}_{h\to 0}$.

Moreover, there exists a constant $d_\Omega > 0$ depending solely on the geometric properties of $\partial\Omega$ such that for all $x \in \partial\Omega_h$,

$$\operatorname{dist}[x, \partial\Omega] \leqslant d_\Omega h^2. \tag{3}$$

R e m a r k 1. Note that (2) implies

$$\frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)} < 1. \tag{4}$$

The rest of the paper is devoted to the proof of Theorem 1.

2. Distance of approximative domain

We start with proving the latter part of Theorem 1 concerning the size of the gap between Ω_h and Ω.

Lemma 1. *Let Ω, r_Ω, h_0 be as in Definition 1. Then for any $h \leqslant h_0$ and for any $x \in \Omega_h$, where Ω_h is an approximative domain from Definition 2, the following inequality holds:*

$$\operatorname{dist}[x, \partial\Omega] \leqslant \frac{(\operatorname{diam} E_h^j)^2}{r_\Omega} \tag{5}$$

if $x \in E_h^j$, where E_h^j is an edge of $\partial\Omega_h$, and

$$\operatorname{dist}[x, \partial\Omega] \leqslant 2\frac{(\operatorname{diam} T_h^j)^2}{r_\Omega} \tag{6}$$

if $x \in T_h^j$, where T_h^j is a boundary triangle of $\partial\Omega_h$.

P r o o f. From the definition of a C^2-domain we have $\partial\Omega = \bigcup_{i=1}^{M} \partial\Omega^i$, where $\partial\Omega^i$ are manifolds that are graphs of C^2 functions from subsets of $\mathbb{R}^2$ to $\mathbb{R}$. Let us denote these functions by G_i, $i = 1, \dots, M$. Then clearly $r_\Omega = (\max_i \|\nabla^2 G_i\|_\infty)^{-1}$.

Take any approximative domain Ω_h. From Definition 2, $\partial\Omega_h = \bigcup_j T_h^j$, where T_h^j are triangles with diameter not exceeding h. Take an arbitrary $x \in \partial\Omega_h$. Then there is a triangle $T_h^j\colon\ x \in T_h^j$. Without loss of generality, $T_h^j \subset G_i^{-1}(\partial\Omega^i)$ for some $i = i(j)$. (Actually, it is true up to a rotation and shift of coordinates.)

If x is a vertex, then $\operatorname{dist}[x, \partial\Omega] = 0$ by the assumption and both (5), (6) hold.

Let $x \in T_h^j \setminus \{v_1, v_2, v_3\}$ for some boundary triangle T_h^j, where v_1, v_2, v_3 are its vertices. Define g as the restriction of G_i to the line v_1x. Then the Taylor expansion gives

$$g(y) = g'(v_1)(y - v_1) + \frac{1}{2} g''(\widetilde{y})(y - v_1)^2 \tag{7}$$

for any y on the line and some $\widetilde{y} \in T_h^j$. Note that $g(v_r) = 0$, $r \in \{1, 2, 3\}$, as by the assumption $v_r \in \partial\Omega$. Further,

$$|g''(\widetilde{y})| \leqslant \|\nabla^2 G_i\|_\infty \leqslant \frac{1}{r_\Omega}. \tag{8}$$

Let x lie on the edge E_h^j of $T_h^j \subset \partial\Omega_h$. Then we can use (7) twice, for $y = x$ and $y = v_2$, which together with estimate (8) gives

$$|g(x)| \leqslant |g'(v_1)(x - v_1)| + \frac{(\operatorname{diam} E_h^j)^2}{2r_\Omega}, \quad |g'(v_1)(v_2 - v_1)| \leqslant \frac{(\operatorname{diam} E_h^j)^2}{2r_\Omega},$$

from which we infer $|g(x)| \leqslant r_\Omega^{-1}(\operatorname{diam} E_h^j)^2$.

Let $x \in \operatorname{int} T_h^j$. Then we use (7) twice, for $y = x$ and $y = e$, where e is the intersection of the line v_1x with the edge v_2v_3. With help of (8) we get

$$|g(x)| \leqslant |g'(v_1)(x - v_1)| + \frac{1}{2r_\Omega}(\operatorname{diam} T_h^j)^2,$$
$$|g'(v_1)(e - v_1)| \leqslant |g(e)| + \frac{1}{2r_\Omega}(\operatorname{diam} T_h^j)^2.$$

As we already have $|g(e)| \leqslant r_\Omega^{-1}(\operatorname{diam} T_h^j)^2$ for an edge point e, we can infer $|g(x)| \leqslant 2r_\Omega^{-1}(\operatorname{diam} T_h^j)^2$. The proof is concluded by realizing that $\operatorname{dist}[x, \partial\Omega] \leqslant \operatorname{dist}[x, g(x)] = |g(x)|$. □

Lemma 1 implies the following corollary.

Corollary 1 (h^2-property)**.** *Let Ω, r_Ω, h_0 be as in Definition* 1*. Then there exists $d_\Omega > 0$ depending solely on the geometrical properties of Ω such that for any $h \leqslant h_0$, Ω_h from Definition* 2*, and for any $x \in \partial\Omega_h$,*

$$\operatorname{dist}[x, \partial\Omega] \leqslant d_\Omega h^2.$$

P r o o f. Set $d_\Omega := 2r_\Omega^{-1}$ in (5) and (6) and recall that $\operatorname{diam} E_h^j \leqslant \operatorname{diam} T_h^j \leqslant \operatorname{diam} K \leqslant h$. □

Note that in this section we worked only with the approximative domain, no requirements on the mesh were needed.

3. Preliminaries

To prove the existence of a strongly regular family of boundary-fitted meshes, we will use a decomposition of a tetrahedron into eight tetrahedra which inherit the regularity estimate. However, it is not the strong regularity condition introduced in Definition 3 that is being preserved. Therefore, we introduce an alternative criterion of regularity.

Before that, we recall some properties of affine transformations that play a crucial role throughout this paper. Some tetrahedra established by the refinement process need to be modified (boundary vertices should be shifted to the smooth boundary) so that their union satisfies the definition of a boundary-fitted mesh (Definition 2). The shift is performed using affine transformations.

The final part of this section is devoted to the so-called q-Pochhammer symbols, which will finally ensure the existence of a lower bound on the regularity ratio θ_0 in (3).

3.1. Affine transformations and singular values. An affine transformation F is a one-to-one mapping of a linear vector space to itself, preserving linearity and the *ratio of division*, see e.g. [1], Proposition 2.8. Endowing the three-dimensional space with Euclidean coordinates, we can represent an affine transformation F by a 3×3 nonsingular matrix Q and a shift vector q:

$$F(x) = Qx + q.$$

In what follows, we will be mainly interested in the effects to the geometric properties of the objects undergoing the transformation. As the translation vector q cannot affect the shape change, we focus on the properties of the matrix Q.

Lemma 2 (Singular Value Decomposition)**.** *Let $Q \in \mathbb{R}^{3\times 3}$ be a nonsingular matrix. Then there exist matrices U, Σ, V satisfying $Q = U\Sigma V^T$, where $U^TU = I$, $V^TV = I$, and Σ is a diagonal matrix of the so-called singular values $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \sigma_3)$, where all three σ_i are positive.*

Moreover, Q transforms the unit sphere into an ellipsoid with semi-axes of the lengths σ_i, $i = 1, 2, 3$.

The proof of the above assertion can be found in any linear algebra textbook, see for instance [6], Section 7.3.

From the above lemma we will use mainly $\sigma_{\min} := \min\{\sigma_1, \sigma_2, \sigma_3\}$ and $\sigma_{\max} := \max\{\sigma_1, \sigma_2, \sigma_3\}$, the maximal shrinking and prolongation factors, respectively. In the sequel, we write $\sigma_{\min}(F)$ (and $\sigma_{\max}(F)$) for the minimal (maximal) singular value of the affine transformation F, referring to the minimal (maximal) singular value of its matrix Q.

The following lemma provides a tool for estimating singular values of a composition of affine mappings.

Lemma 3. *Let A and B be affine transformations. Then we have*

$$\sigma_{\min}(A \circ B) \geqslant \sigma_{\min}(A) \cdot \sigma_{\min}(B)$$

and

$$\sigma_{\max}(A \circ B) \leqslant \sigma_{\max}(A) \cdot \sigma_{\max}(B).$$

3.2. Sommerville regularity ratio. An alternative regularity criterion, introduced in this section, measures the similarity of a general tetrahedron to a reference tetrahedron, which is in our case the *Sommerville tetrahedron*, introduced in 1923 in [10].

Definition 4 (Sommerville tetrahedron)**.** Sommerville tetrahedron is any tetrahedron similar to the unit tetrahedron $\widetilde{K}$, which is defined through Euclidean coordinates of its vertices:

$$\widetilde{A} = \left[\tfrac{1}{2}, 0, 0\right]^\top, \quad \widetilde{B} = \left[-\tfrac{1}{2}, 0, 0\right]^\top, \quad \widetilde{C} = \left[0, \tfrac{1}{2}, \tfrac{1}{2}\right]^\top, \quad \widetilde{D} = \left[0, -\tfrac{1}{2}, \tfrac{1}{2}\right]^\top.$$

The unit Sommerville tetrahedron $\widetilde{K}$ (see Figure 1) has two opposite edges of length 1, the other four of length $\sqrt{3}/2$ and dihedral angles attain the values $60°$ and $90°$. For further use we will need the following characterization of $\widetilde{K}$:

$$\operatorname{diam}\widetilde{K} = 1, \quad e(\widetilde{K}) = \frac{\sqrt{3}}{2}, \quad \widetilde{\varrho} = \theta(\widetilde{K}) = \frac{\sqrt{2}}{8}, \quad m(\widetilde{K}) = \frac{\sqrt{2}}{2}, \tag{9}$$

where $e(\widetilde{K})$ is the length of the shortest edge, $\widetilde{\varrho} = \varrho(\widetilde{K})$ is the radius of an inscribed sphere and $m(\widetilde{K})$ is the shortest median of a face of the Sommerville tetrahedron. For detailed computations, see [7].

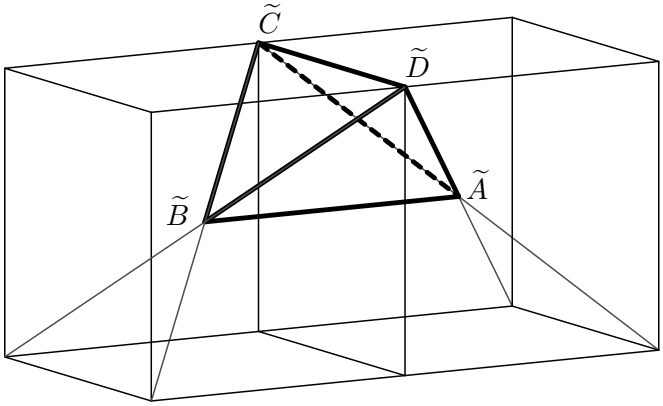


Figure 1. The unit Sommerville tetrahedron $\widetilde{K}$ inscribed in two auxilliary cubes. (Axes are omitted for the sake of brevity.)

Note that for any tetrahedron $K = \mathrm{co}(ABCD)$, there exists a unique affine transformation F_K that maps the Sommerville tetrahedron $\widetilde{K} = \mathrm{co}(\widetilde{A}\widetilde{B}\widetilde{C}\widetilde{D})$ onto K, i.e.

$$F_K(\widetilde{x}) = Q_K\widetilde{x} + q_K, \tag{10}$$

determined by $F_K(\widetilde{A}) = A, F_K(\widetilde{B}) = B, F_K(\widetilde{C}) = C, F_K(\widetilde{D}) = D$. It can be easily shown that $Q_K = [A - B, C - D, C + D - A - B]$ and $q_K = \frac{1}{2}(A + B)$.

However, as we get a different transformation just by relabelling the vertices of the tetrahedron K, we must be careful with employing the following alternative regularity criterion.

Definition 5. Let $K = \mathrm{co}(ABCD)$ be a tetrahedron, let

$$\mathcal{A}_K := \{F_K; F_K \text{ is an affine transformation}, F_K(\widetilde{K}) = K\} \tag{11}$$

be a set of all affine transformations mapping Sommerville tetrahedron $\widetilde{K}$ onto K. Then we define the *Sommerville regularity ratio* of the tetrahedron K as

$$\kappa(K) = \max_{F_K \in \mathcal{A}_K} \frac{\sigma_{\min}(F_K)}{\sigma_{\max}(F_K)}, \tag{12}$$

where $\sigma_{\min}(F_K)$, $\sigma_{\max}(F_K)$ are the minimal and maximal singular values of F_K, respectively.

Note that κ attains its maximum of 1 for the Sommerville tetrahedron, while the minimal value of 0 would be attained for a degenerate tetrahedron. Consequently, κ plays the role of a regularity measure.

R e m a r k 2. Taking the regular tetrahedron as the reference one, we could leave out the maximization in (12). However, we prefer the Sommerville tetrahedron, as its copies tile the three-dimensional space, see [7], [10], while the regular tetrahedron does not.

Analogously to other standard regularity ratios, also the Sommerville regularity ratio (12) can be used to formulate a criterion for strong regularity. We show its equivalence to a standard regularity criterion in a form of two lemmas that we use directly in the next section.

Lemma 4. *Let $\kappa_0 > 0$ and let there exist a sequence $h_n \to 0$ such that $\{\mathcal{T}_{h_n}\}_{n\in\mathbb{N}}$ is a family of boundary-fitted meshes satisfying*

$$\kappa(K) \geqslant \kappa_0 > 0$$

for any $n \in \mathbb{N}$ and any $K \in \mathcal{T}_{h_n}$.

Then $\{\mathcal{T}_{h_n}\}_{n\in\mathbb{N}}$ is a strongly regular family of boundary-fitted meshes.

The proof is strongly based on ideas of Křížek, see [9].

P r o o f. We take an arbitrary $n \in \mathbb{N}$, an arbitrary element $K \in \mathcal{T}_{h_n}$, and consider the affine function F_K from (11). We denote by $\widetilde{\mathcal{S}}(\widetilde{x_0}, \widetilde{\varrho})$ the inscribed sphere of $\widetilde{K}$. Then $F_K(\widetilde{\mathcal{S}}) =: \mathcal{E} \subset K$ is an ellipsoid. Let us label its center with x_0. Take $r(K)$ as the shortest semi-axis of $\mathcal{E}$. Then the sphere $\mathcal{S}(x_0, r(K))$ is contained in K and therefore $\varrho(K) \geqslant r(K)$.

From the properties of the singular values of an affine transformation we get the estimates $r(K) = \sigma_{\min}(F_K) \cdot \widetilde{\varrho}$ and $\operatorname{diam} K \leqslant \sigma_{\max}(F_K) \cdot \operatorname{diam} \widetilde{K}$. Hence, we can write

$$\theta(K) = \frac{\varrho(K)}{\operatorname{diam} K} \geqslant \frac{r(K)}{\operatorname{diam} K} \geqslant \frac{\sigma_{\min}(F_K) \cdot \widetilde{\varrho}}{\sigma_{\max}(F_K) \cdot \operatorname{diam} \widetilde{K}} = \kappa(K)\theta(\widetilde{K}), \tag{13}$$

where the last equality holds assuming we take an appropriate F_K that realizes the maximum in (12). By the assumption, $\kappa(K) \geqslant \kappa_0$ and using (13), we can conclude

$$\theta(K) \geqslant \kappa_0\theta(\widetilde{K}) = \frac{\sqrt{2}}{8}\kappa_0 =: \theta_0$$

for any K in the family of meshes. □

Lemma 5. *Let $s > 0$ and let K be a tetrahedron satisfying $\theta(K) \geqslant s$. Then*

$$\kappa(K) \geqslant \frac{\sqrt{2}}{8} s.$$

P r o o f. Setting K into coordinates in such a way that its shortest edge belongs to the line parallel to the longest edge of the Sommerville tetrahedron, we can write $\varrho(K) \leqslant \sigma_{\min}(F_K) \cdot \operatorname{diam} \widetilde{K}$. Further, the mapping F_K transforms the inscribed sphere of $\widetilde{K}$ onto an inscribed ellipsoid of K, hence $\operatorname{diam} K \geqslant \sigma_{\max}(F_K)\widetilde{\varrho}$. Therefore,

$$s \leqslant \theta(K) = \frac{\varrho(K)}{\operatorname{diam} K} \leqslant \frac{\sigma_{\min}(F_K) \cdot \operatorname{diam} \widetilde{K}}{\sigma_{\max}(F_K) \cdot \varrho(\widetilde{K})} \leqslant \kappa(K) \frac{8}{\sqrt{2}}.$$

□

We conclude this part with the following corollary of Lemma 3.

Corollary 2. *Let K, K' be two tetrahedra, and let S be an affine transformation that maps K onto K'. Then we have*

$$\kappa(K') \geqslant \kappa(K) \frac{\sigma_{\min}(S)}{\sigma_{\max}(S)}.$$

3.3. q-Pochhammer symbol. Further, we prove some properties of the so-called *q-Pochhammer symbol*, which will be the final tool used for showing the existence of a lower bound κ_0.

Definition 6. Let $n \in \mathbb{N}$ and $a, q \in [0, 1]$. The product

$$(a; q)_n := \prod_{j=0}^{n-1} (1 - aq^j)$$

is called the *q-Pochhammer symbol*.

Lemma 6. *Let $a \in (0, 1)$ and $q \in (0, 1)$. Then there exists $P(a, q) > 0$ such that for any $n \in \mathbb{N}$,*

$$(a; q)_n > \lim_{n \to \infty} (a; q)_n = P(a, q).$$

P r o o f. As $(a;q)_{n+1} = (1-aq^n)\cdot(a;q)_n$, the sequence is monotonically decreasing. To prove the existence of a positive limit of $(a;q)_n$, it suffices to find its positive lower bound. Consider

$$s_n := \sum_{k=0}^{n-1} \log(1-aq^k).$$

Clearly $(a;q)_n = \exp s_n$ and using $\log(1-az) > -7az \geqslant -7z$ for $z \in (0,1]$, $a \in (0, 1-\varepsilon]$, where $\varepsilon < 10^{-3}$, we can estimate

$$s_n > -7\sum_{k=0}^{n-1} q^k = -7\frac{1-q^n}{1-q}. \tag{14}$$

Combining (14) with the monotonicity of both the exponential function and the partial sums of the geometric series, we get

$$(a;q)_n = \exp s_n > \exp\Big(-7\frac{1-q^n}{1-q}\Big) > \exp\Big(\frac{-7}{1-q}\Big) > 0.$$

Note that for ε smaller it is only necessary to increase the multiplicative constant in estimate (14). □

4. Mesh refinement

In 1982, Křížek proved the following result, see [9].

Theorem 2 ([9], Theorem 3.2)**.** *For any polyhedron there exists a strongly regular family of decompositions into tetrahedra.*

For our purpose it is not possible to use this result directly, because the decomposition in [9] creates a mesh that is no longer boundary-fitted, as new vertices on the boundary of the polyhedral domain are created and do not lie on $\partial\Omega$, in general. Our idea is to use this decomposition and to modify (i.e. affinely transform) the tetrahedra in the boundary layer to put all boundary vertices to $\partial\Omega$. By virtue of Lemma 1 we will show that this change is small in comparison with the diameter of the element, and the strong regularity is therefore preserved.

4.1. Decomposition of a tetrahedron. We start with the first step, from the proof of Theorem 2 we extract the following lemma.

Lemma 7. *Let $\mathcal{T}_h$ be a mesh of Ω_h. Then for any $K \in \mathcal{T}_h$ there exists its decomposition $\mathcal{D}(K) = \{K_i\}_{i=1}^8$ into eight face-to-face tetrahedra such that the vertices of K_i are either vertices of K or midpoints of its edges, and for all $i = 1, \dots, 8$ we have that*

$$\operatorname{diam} K_i \leqslant \frac{1}{2} \operatorname{diam} K \quad \text{and} \quad \kappa(K_i) \geqslant \kappa(K). \tag{15}$$

P r o o f. The unit Sommerville tetrahedron $\widetilde{K}$ can be decomposed into eight tetrahedra similar to $\widetilde{K}$—cutting all six edges at their midpoints creates four tetrahedra and one octahedron which can be decomposed into four identical tetrahedra, see Figure 2 and [9], proof of Theorem 3.2 or [11], Theorem 4.3. We denote the decomposition by $\widetilde{\mathcal{D}} = \{\widetilde{K}_i\}_{i=1}^8$ and it follows that $\operatorname{diam} \widetilde{K}_i = \frac{1}{2}$. Then we take the affine transformation F_K that realizes $\kappa(K)$. We observe that

$$F_K(\widetilde{\mathcal{D}}) = \{F_K(\widetilde{K}_i), \widetilde{K}_i \in \widetilde{\mathcal{D}}\}_{i=1}^8$$

is a decomposition of K.

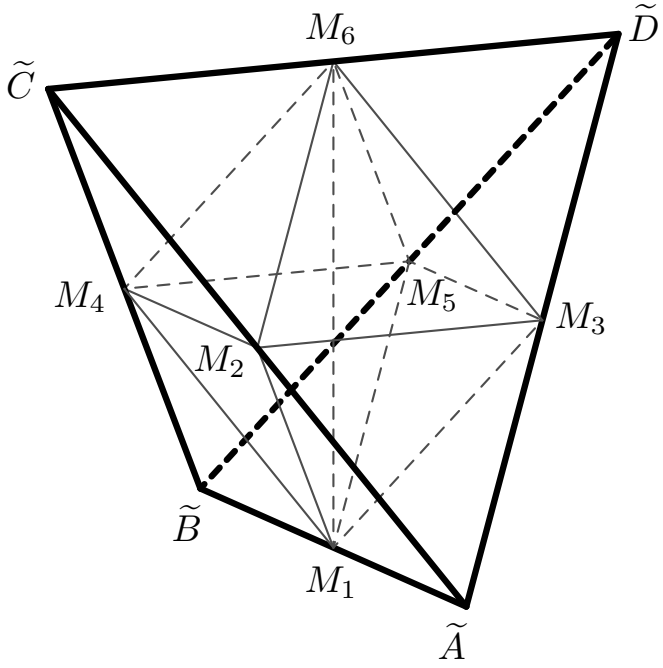


Figure 2. The sketch of Křížek's decomposition of the Sommerville tetrahedron $\widetilde{K}$. Reproduction from [9].

The key idea is that $\widetilde{K_i}$ are also Sommerville tetrahedra and F_K transforms $\widetilde{K_i}$ into K_i, which implies $\kappa(K_i) \geqslant \kappa(K)$ for any $K_i \in \mathcal{D}(K)$, since F_K does not have to be the mapping realizing the maximum in $\kappa(K_i)$. The first part of (15) is a consequence of the ratio of division being invariant w.r.t. an affine transformation. □

4.2. Correction of the decomposition. The tetrahedra $K_i \in \mathcal{D}(K)$, $K \in \mathcal{T}_h$, do not create a boundary-fitted mesh (according to Definition 2) as new vertices were created on the boundary of the polyhedral domain Ω_h that do not belong to the boundary of the smooth domain Ω. To fix that, we apply an affine shift to these vertices. We set the domain of vertices that must be shifted in order to obtain a boundary-fitted mesh:

$$V(\mathcal{T}_h) := \{x \text{ is a vertex of some } K_i \in \mathcal{D}(K),\ K \in \mathcal{T}_h \text{ and } x \in \partial\Omega_h \setminus \partial\Omega\}.$$

For any $x \in V(\mathcal{T}_h)$ we choose one $y(x) \in \partial\Omega$ such that

$$\operatorname{dist}[x, \partial\Omega] = \operatorname{dist}[x, y(x)]. \tag{16}$$

Then for any $K_i \in \mathcal{D}(K)$ of a given $K \in \mathcal{T}_h$, we consider an affine shift function S_{K_i} defined uniquely by the images of four vertices of the tetrahedron K_i:

$$S_{K_i}(v) = \begin{cases} y(v) & \text{for } v \in V(\mathcal{T}_h), v \text{ a vertex of } K_i, \\ v & \text{for } v \notin V(\mathcal{T}_h), v \text{ a vertex of } K_i. \end{cases} \tag{17}$$

From Lemma 1 we have an upper bound on the size of this shift. We have to prove that under the assumptions given in Theorem 1, the shift of vertices does not damage the topology of the finer mesh.

Lemma 8. *Let Ω, Ω_h, $\mathcal{T}_h$ be as in Definitions 1 and 2. Let v_1, v_2 be distinct vertices of the refined mesh, i.e. v_i, $i = 1, 2$, is either a vertex or a midpoint of an edge of some tetrahedron in $\mathcal{T}_h$. Let*

$$\{tv_1 + (1-t)v_2,\ t \in (0, t_1)\} \subset K \in \mathcal{T}_h,$$
$$\{tv_1 + (1-t)v_2,\ t \in (t_2, 1)\} \subset L \in \mathcal{T}_h,$$

for some $t_1, t_2 \in (0,1)$, $t_1 \leqslant t_2$, and $K, L \in \mathcal{T}_h$ not necessarily distinct. Then

$$\operatorname{dist}[v_1, v_2] \geqslant \frac{\sqrt{3}}{8}(\sigma_{\min}(F_K) + \sigma_{\min}(F_L)). \tag{18}$$

P r o o f. Let $K = L$. Then the segment v_1v_2 is either half of an edge, a midsegment of a face triangle, an edge itself, the median of a face, or a median of a tetrahedron (both v_1, v_2 are midpoints of the edges of tetrahedron K). For the first three options, we clearly have $\operatorname{dist}[v_1, v_2] \geqslant \frac{1}{2}e(K) \geqslant \frac{1}{4}\sqrt{3}\,\sigma_{\min}(F_K)$. For a median of a triangle we have $\operatorname{dist}[v_1, v_2] \geqslant m(K) \geqslant \frac{1}{2}\sqrt{2}\,\sigma_{\min}(F_K)$, as an affine mapping

maps median onto median. The same estimate applies to the last option. In both cases we used (9).

Let $K \neq L$. If v_1 is a vertex of K, then we denote by Γ_K the face of K opposite to v_1. Then $\operatorname{dist}[v_1, \Gamma_K] \geqslant \sigma_{\min}(F_K) \cdot \frac{1}{2}\sqrt{2}$, where the last fraction is the (minimal) distance of a vertex from the opposite face in the Sommerville tetrahedron.

In the case of v_1 being the midpoint of an edge of K, we denote by Γ_K^1, Γ_K^2 the faces of K that do not contain v_1. Then

$$\min_{i=1,2} \operatorname{dist}[v_1, \Gamma_K^i] \geqslant \sigma_{\min}(F_K)\frac{\sqrt{2}}{4},$$

where the last fraction is the minimal value of such distance in the Sommerville tetrahedron.

Taking the minimum over the above listed possibilities, we conclude that (18) holds. □

Lemma 9. *For any $h \leqslant h_0$, let every $K \in \mathcal{T}_h$ satisfy the so-called minimal regularity condition*

$$\kappa(K) \geqslant b\frac{\operatorname{diam} K}{r_\Omega}, \qquad \text{where } b > b_0 = \frac{8}{\sqrt{3}}(2+\sqrt{5}). \tag{19}$$

Then for any vertices v_1, v_2 of $K_i \in \mathcal{D}(K)$, $L_j \in \mathcal{D}(L)$, respectively, we have that

$$\operatorname{dist}[v_1, v_2] > \operatorname{dist}[v_1, S_{K_i}(v_1)] + \operatorname{dist}[v_2, S_{L_j}(v_2)],$$

i.e. the shift above does not damage the topological properties of the mesh.

P r o o f. By construction, if $v_i \in V(\mathcal{T}_h)$, then it is the midpoint of an edge of some boundary triangle T_j^h. By virtue of Lemma 1, in particular from (5), together with (16) and (17) we obtain

$$\frac{1}{r_\Omega}\big((\operatorname{diam} K)^2 + (\operatorname{diam} L)^2\big) \geqslant \operatorname{dist}[v_1, S_{K_i}(v_1)] + \operatorname{dist}[v_2, S_{L_j}(v_2)]. \tag{20}$$

Lemma 8 gives

$$\operatorname{dist}[v_1, v_2] \geqslant \frac{\sqrt{3}}{8}(\sigma_{\min}(F_K) + \sigma_{\min}(F_L)), \tag{21}$$

where F_K and F_L realize the maxima in $\kappa(K)$ and $\kappa(L)$, respectively. From the definition of κ and Lemma 2 we have

$$\sigma_{\min}(F_K) = \kappa(K)\sigma_{\max}(F_K) \geqslant \kappa(K)\operatorname{diam} K\frac{2}{\sqrt{3}} > \kappa(K)\operatorname{diam} K. \tag{22}$$

Using the assumption (19), we can rewrite (22) as

$$\sigma_{\min}(F_K) + \sigma_{\min}(F_L) \geqslant b\frac{(\operatorname{diam} K)^2 + (\operatorname{diam} L)^2}{r_\Omega}. \tag{23}$$

Substituting (23) into (21), we get

$$\operatorname{dist}[v_1, v_2] > \frac{b\sqrt{3}}{8r_\Omega}\big((\operatorname{diam} K)^2 + (\operatorname{diam} L)^2\big), \tag{24}$$

which, combined with (20), completes the proof, since $\frac{1}{8}b\sqrt{3} > 1$. □

Having defined the shift, we focus on the bounds of the singular values of the affine shift, which will be needed in a moment.

Lemma 10. *Let $K \in \mathcal{T}_h$ be a tetrahedron, let $K_i \in \mathcal{D}(K)$ and let the affine shift S_{K_i} be defined by* (17)*. Then for its singular values we have*

$$\sigma_{\min}(S_{K_i}) \geqslant 1 - \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}, \tag{25}$$

$$\sigma_{\max}(S_{K_i}) \leqslant 1 + \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}, \tag{26}$$

and the regularity criterion for the new tetrahedra satisfies the estimate

$$\kappa(S_{K_i}) \geqslant \frac{1 - \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}}{1 + \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}}\kappa(K) \geqslant \left(1 - \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}\right)^2\kappa(K). \tag{27}$$

P r o o f. The maximal singular value of S_{K_i} represents the maximal relative prolongation, which can be achieved at the shortest edge of K_i, i.e. $e(K_i) = \frac{1}{2}e(K)$ by moving the vertices from each other with the maximal radius, i.e.

$$\sigma_{\max}(S_{K_i}) \leqslant \frac{\frac{1}{2}e(K) + 2r_\Omega^{-1}(\operatorname{diam} K)^2}{\frac{1}{2}e(K)} = 1 + 4\frac{(\operatorname{diam} K)^2}{e(K)r_\Omega}. \tag{28}$$

Using $e(K) \geqslant e(\widetilde{K}) \cdot \sigma_{\min}(F_K)$ and $\operatorname{diam} K \leqslant \operatorname{diam} \widetilde{K} \cdot \sigma_{\max}(F_K)$, where F_K realizes the maximum in the definition of κ, we can deduce that

$$e(K) \geqslant \kappa(K) \cdot \operatorname{diam} K\frac{e(\widetilde{K})}{\operatorname{diam}\widetilde{K}} = \kappa(K) \cdot \operatorname{diam} K\frac{\sqrt{3}}{2}. \tag{29}$$

Using estimate (29) in (28), we conclude (26). The same steps prove the inequality (25). Then by virtue of Corollary 2 we can estimate

$$\kappa(S_{K_i}(K_i)) \geqslant \frac{\sigma_{\min}(S_{K_i})}{\sigma_{\max}(S_{K_i})}\kappa(K). \tag{30}$$

The last relation (27) is obtained from (30) using the estimates (25), (26), and the inequality $(1+z)^{-1} \geqslant 1 - z$, $z \in \mathbb{R}^+$. □

Next we show that shifting the new vertices to the smooth boundary does not disturb the uniform decrease of the discretization parameter.

Lemma 11. *Let $h \leqslant h_0$ and let $\mathcal{T}_h$ be a boundary-fitted mesh. Let a tetrahedron $K \in \mathcal{T}_h$ satisfy the minimal regularity condition* (19) *with some admissible b. Then there exists a number $\mu(b) \in (0,1)$ such that for any $K_i \in \mathcal{D}(K)$ we have*

$$\operatorname{diam} S_{K_i}(K_i) \leqslant \mu(b) \cdot \operatorname{diam} K. \tag{31}$$

P r o o f. From Lemma 7 we recall $\operatorname{diam} K_i \leqslant \frac{1}{2} \operatorname{diam} K$. From the construction it follows that

$$\operatorname{diam} S_{K_i}(K_i) \leqslant \frac{\sigma_{\max}(S_{K_i})}{2} \operatorname{diam} K. \tag{32}$$

Substituting the minimal regularity condition (19) into the upper bound (26) for $\sigma_{\max}(S_{K_i})$, we get the estimate

$$\sigma_{\max}(S_{K_i}) \leqslant 1 + \frac{8}{b\sqrt{3}}. \tag{33}$$

Then, combining (32) and (33), we conclude that

$$\operatorname{diam} S_{K_i}(K_i) \leqslant \Big(\frac{1}{2} + \frac{4}{b\sqrt{3}}\Big) \operatorname{diam} K =: \mu(b) \cdot \operatorname{diam} K.$$

The factor $\mu(b)$ belongs to $(0,1)$, as clearly $b > 8/\sqrt{3}$. □

Corollary 3. *Let $h \leqslant h_0$ and let $\mathcal{T}_h$ be a boundary-fitted mesh. Let every $K \in \mathcal{T}_h$ satisfy the minimal regularity condition* (19) *with some admissible b. Then*

$$\mathcal{T}_k := \{S_{K_i}(K_i), K_i \in \mathcal{D}(K), K \in \mathcal{T}_h\}$$

is a boundary-fitted mesh in the sense of Definition 2 *with*

$$k < \Big(\frac{1}{2} + \frac{4}{b\sqrt{3}}\Big) h. \tag{34}$$

P r o o f. The construction together with condition (19) ensures that $\mathcal{T}_k$ is a boundary-fitted mesh. Even if every element is transformed by a different affine function, still the common faces (and edges) of two neighbouring elements are transformed identically for both elements, hence the face-to-face property is preserved.

We define k to be the maximal diameter of an element in $\mathcal{T}_k$, say L. But clearly this L was created by splitting and shifting some tetrahedron $M \in \mathcal{T}_h$. Then it follows from Lemma 11 that

$$k = \operatorname{diam} L < \mu(b) \cdot \operatorname{diam} M \leqslant \mu(b) \cdot h = \Big(\frac{1}{2} + \frac{4}{b\sqrt{3}}\Big)h.$$

□

R e m a r k 3. Notice that so far it has been sufficient that $b \geqslant 8/\sqrt{3}$. For the next lemma we need the stronger condition (19), indeed.

Next, we need to show that in the process of refinement, the newly established elements do not violate the minimal regularity condition (19) with given b, which is necessary to allow the repetition of the refinement process.

Lemma 12. *Let K be such that $\kappa(K)$ satisfies condition* (19) *with some admissible b and let $K_i \in \mathcal{D}(K)$. Then $S_{K_i}(K_i)$ also satisfies* (19) *with b.*

P r o o f. We know from (27) that

$$\kappa(S_{K_i}(K_i)) \geqslant \frac{1 - \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}}{1 + \frac{8}{\sqrt{3}r_\Omega}\frac{\operatorname{diam} K}{\kappa(K)}}\kappa(K), \tag{35}$$

and from (19) that

$$\kappa(K) \geqslant \frac{b}{r_\Omega}\operatorname{diam} K. \tag{36}$$

Substituting (36) into (35), we get

$$\kappa(S_{K_i}(K_i)) \geqslant \frac{1 - \frac{8}{b\sqrt{3}}}{1 + \frac{8}{b\sqrt{3}}}\frac{b}{r_\Omega}\operatorname{diam} K. \tag{37}$$

Finally, (34) implies

$$\operatorname{diam} K \geqslant \frac{2}{1 + \frac{8}{b\sqrt{3}}}\operatorname{diam} S_{K_i}(K_i),$$

which substituted into (37) together with inequality (4) from Remark 1 recovers (19) with b also for $S_{K_i}(K_i)$. □

Theorem 3 (Existence of family)**.** *Let Ω, h_0 be as in Definition* 1 *and for some $h_1 \leqslant h_0$ let there exist a boundary-fitted mesh $\mathcal{T}_{h_1}$ of Ω such that every tetrahedron $K \in \mathcal{T}_{h_1}$ satisfies* (19) *with some admissible b. Then there exists a family of boundary-fitted meshes $\{\mathcal{T}_{h_n}\}_{n\in\mathbb{N}}$ with $h_n \to 0$.*

P r o o f. We proceed via mathematical induction. By assumption, for h_1 there exists a boundary-fitted mesh $\mathcal{T}_{h_1}$ with elements satisfying (19) with b.

Corollary 3 gives the following implication: If for h_n there exists a boundary-fitted mesh $\mathcal{T}_{h_n}$ with elements satisfying regularity condition (19) with some b, then there exists $h_{n+1} \leqslant \mu(b)h_n$ such that there exists a boundary-fitted mesh $\mathcal{T}_{h_{n+1}}$. By virtue of Lemma 12 all elements of this finer mesh satisfy (19) with b.

The proof is completed, as we have proven the property for h_1 as well as the induction step. □

4.3. Proof of the Sommerville strong regularity.

Theorem 4. *Let Ω, h_0 be as in Definition* 1*. For $h_1 \leqslant h_0$ let there exist $\mathcal{T}_{h_1}$ a boundary-fitted mesh of Ω, whose every element satisfies* (19) *with some admissible b. Then the family $\{\mathcal{T}_{h_n}\}_{n\in\mathbb{N}}$ of boundary-fitted meshes obtained through Theorem* 3 *is Sommerville strongly regular, i.e. there exists $\kappa_0 > 0$ such that for any $n \in \mathbb{N}$, any $K \in \mathcal{T}_{h_n}$ we have that $\kappa(K) \geqslant \kappa_0$.*

P r o o f. Consider the family of elements $\{L_n\}_{n\in\mathbb{N}\cup\{0\}}$ such that $L_0 \in \mathcal{T}_{h_1}$, and for any $n \in \mathbb{N}$, $L_n \in \mathcal{T}_{h_{n+1}}$ and $L_n := S_{K_i}(K_i)$, where $K_i \in \mathcal{D}(L_{n-1})$.

Thanks to Lemma 10 we have

$$\kappa(L_{n+1}) \geqslant \left(1 - \frac{8}{\sqrt{3}}\frac{\operatorname{diam} L_n}{r_\Omega \kappa(L_n)}\right)^2 \kappa(L_n). \tag{38}$$

Further, we have from Lemma 11 that

$$\operatorname{diam} L_n \leqslant \frac{1}{2}\Big(1 + \frac{8}{b\sqrt{3}}\Big) \operatorname{diam} L_{n-1}, \tag{39}$$

and from Lemma 10 combined with (19) also

$$\kappa(L_n) \geqslant \frac{1 - \frac{8}{b\sqrt{3}}}{1 + \frac{8}{b\sqrt{3}}}\kappa(L_{n-1}). \tag{40}$$

Combining (39) and (40), we get

$$\frac{\operatorname{diam} L_n}{\kappa(L_n)} \leqslant \frac{1}{2}\frac{\big(1 + \frac{8}{b\sqrt{3}}\big)^2}{1 - \frac{8}{b\sqrt{3}}}\frac{\operatorname{diam} L_{n-1}}{\kappa(L_{n-1})},$$

i.e.

$$\frac{\operatorname{diam} L_n}{\kappa(L_n)} \leqslant \left(\frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)} \right)^n \frac{\operatorname{diam} L_0}{\kappa(L_0)}.$$

As the condition (19) holds also for L_0, we have

$$\frac{\operatorname{diam} L_n}{\kappa(L_n)} \leqslant \left(\frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)} \right)^n \frac{r_\Omega}{b}. \tag{41}$$

Then, substituting (41) to (38), we get

$$\kappa(L_{n+1}) \geqslant \left(1 - \frac{8}{b\sqrt{3}} \left(\frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)} \right)^n \right) \kappa(L_n).$$

Hence, we can explicitly estimate

$$\kappa(L_{n+1}) \geqslant \prod_{i=0}^{n} \left(1 - \frac{8}{b\sqrt{3}} \left(\frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)} \right)^i \right) \kappa(L_0). \tag{42}$$

The product on the right-hand side of (42) is a q-Pochhammer symbol with parameters

$$a = \frac{8}{b\sqrt{3}}, \qquad q = \frac{\left(1+\frac{8}{b\sqrt{3}}\right)^2}{2\left(1-\frac{8}{b\sqrt{3}}\right)}.$$

Assumption (19) guarantees that $q \in (0,1)$, see Remark 1, and also $a \in (0,1)$. Therefore, we have from Lemma 6 that the right-hand side of (42) has a positive limit $P(a,q) > 0$ for $n \to \infty$ and hence also

$$\kappa(L_n) \geqslant (a;q)_n \cdot \kappa(L_0) > P(a,q) \cdot \kappa(L_0).$$

We recall that $L_0 \in \mathcal{T}_{h_1}$ and set

$$\kappa_0 := P(a,q) \cdot \min_{L \in \mathcal{T}_{h_1}} \kappa(L),$$

which completes the proof. □

5. Proof of Theorem 1

The final step of the proof is a simple bridging of the main Theorem 1 and Theorem 4.

P r o o f. By virtue of Lemma 5, the conditions (1), (2) can be transformed to the minimal regularity condition (19). Then we apply Theorem 4 to get the existence of a family of boundary-fitted meshes satisfying $\kappa(K) \geqslant \kappa_0 > 0$ for all tetrahedral elements K in the family of meshes. Then by virtue of Lemma 4 we conclude the strong regularity of the family.

The estimate (3) is ensured by Corollary 1. □

References

[1] *M. Audin*: Geometry. Universitext, Springer, Berlin, 2003, 2012. zbl MR

[2] *J. Brandts, S. Korotov, M. Křížek*: On the equivalence of ball conditions for simplicial finite elements in $\mathbb{R}^d$. Appl. Math. Lett. *22* (2009), 1210–1212. zbl MR

[3] *H. Edelsbrunner*: Triangulations and meshes in computational geometry. Acta Numerica *9* (2000), 133–213. zbl MR

[4] *L. C. Evans*: Partial Differential Equations. Graduate Studies in Mathematics 19, American Mathematical Society, Providence, 1998. zbl MR

[5] *E. Feireisl, R. Hošek, D. Maltese, A. Novotný*: Error estimates for a numerical method for the compressible Navier-Stokes system on sufficiently smooth domains. To appear in ESAIM, Math. Model. Numer. Anal. (2016). Preprint IM-2015-46, available at `http://math.cas.cz/fichier/preprints/IM_20150826112605_92.pdf`, 2015.

[6] *R. A. Horn, C. R. Johnson*: Matrix Analysis. Cambridge University Press, Cambridge, 2013. zbl MR

[7] *R. Hošek*: Face-to-face partition of 3D space with identical well-centered tetrahedra. Appl. Math., Praha *60* (2015), 637–651. zbl MR

[8] *S. Korotov, M. Křížek, P. Neittaanmäki*: On the existence of strongly regular families of triangulations for domains with a piecewise smooth boundary. Appl. Math., Praha *44* (1999), 33–42. zbl MR

[9] *M. Křížek*: An equilibrium finite element method in three-dimensional elasticity. Apl. Mat. *27* (1982), 46–75. zbl MR

[10] *D. M. Y. Sommerville*: Space-filling tetrahedra in Euclidean space. Proc. Edinburgh Math. Soc. *41* (1923), 49–57.

[11] *S. Zhang*: Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. Houston J. Math. *21* (1995), 541–556. zbl MR

Author's address: *Radim Hošek*, Institute of Mathematics, Czech Academy of Sciences, Žitná 25, 115 67 Praha 1, Czech Republic, e-mail: `hosek@math.cas.cz`.

Appendix F

R. H.: Construction and shape optimization of simplical meshes in d-dimensional space.

CONSTRUCTION AND SHAPE OPTIMIZATION OF SIMPLICAL MESHES IN d-DIMENSIONAL SPACE

RADIM HOŠEK

Abstract. We provide a constructive proof of a face-to-face simplical partition of a d-dimensional space for arbitrary d by generalizing the idea of Sommerville, used to create space-filling tetrahedra out of triangular base, to any dimension. Each step of construction that increases the dimension is determined up to a positive parameter, d-dimensional simplical partition is therefore parametrized by d parameters. We show the shape optimal value of those parameters and reveal that the shape optimal partition of d-dimensional space is constructed over the shape optimal partition of $(d-1)$-dimensional space.

Key words: simplical tessellation, simplical mesh, high dimension, Sommerville tetrahedron, Sommerville simplex, mesh regularity, shape optimization.
Subj. AMS Class.: 51M20, 51M04, 51M09, 65N50.

1. Introduction

Tessellations, periodically repeated patterns of disjoint d-dimensional shapes that tile the complete d-dimensional space, have been investigated, at least for $d = 2$, for thousands of years and therefore a vast number of results are available. However, with growing dimension the number of results decreases rapidly. There exist only few results concerning general dimension, we mention those of Brandts et al. on existence of a decomposition of a d-dimensional cube into simplices, see [2], [5].

If the tessellations are created by polytopes and satisfy the *face-to-face* constraint, i.e. a facet of some polytopic element of the tessellation is a facet of another one, a suitable part of such tessellation can be used as a *computational mesh* for various numerical methods. A majority of today's computations take place in two or three spatial dimensions while those in higher dimension still occur rather rarely. However, some elliptic problems are treated in more dimension, see e.g. [21] for such example emanating from stochastic analysis. Beside that, for problems represented by evolutionary partial differential equations of the hyperbolic type in three spatial dimensions, one can understand time as fourth variable and use a mesh in four-dimensional space, see e.g. the practical examples [11] and [16].

There has been created a huge apparatus for automatic mesh generating for domains of various geometries while the procedure that we introduce creates only very rigid meshes. However, some theoretical results suggest that for numerical methods to be convergent, the numerical domain and target domain do not necessarily have to coincide and that is where our meshes might find their use. Two different approaches can be found in works of Feireisl et al. [8], [9] and of Angot et al. [1], [14]. Another advantage of our mesh are its low memory demands, as the vertices of the elements are distributed in a periodic pattern.

Our result is strongly based on the (almost 100 years old) construction developed by Sommerville, which uses a regular triangle as a base for building a one-parametric family of tetrahedral elements that tile the three-dimensional space, see [10], [12] or the original Sommerville's article [23]. We start the construction from one-dimensional simplices, i.e. segments, to increase the dimension repeatedly and build a d-parametrical family of simplical tessellations of d-dimensional space. Its existence is stated in Theorem 2.1 and its proof covers Section 2. Then, in Section 3 we determine the shape-optimizing vector of parameters with the result summarized in Theorem 3.1. Section 4 introduces some concluding remarks and open questions.

2. Construction of the tesselation

We start with stating the existence result in the first of two central theorems of this article.

Theorem 2.1. *For any d-dimensional space there exists a d-parametric family of simplical tessellations $\mathcal{T}_d(\mathbf{p}), \mathbf{p} = (p_1, p_2, \dots, p_d), p_i > 0$. For $\boldsymbol{p}$ fixed, all elements $K \in \mathcal{T}_d(\boldsymbol{p})$ have the same d-dimensional measure equal to*

$$\operatorname{meas}_d K = \prod_{i=1}^{d} p_i. \tag{2.1}$$

Moreover, every connected compact subset of the tessellation builds a face-to-face mesh.

We start with introducing the original Sommerville's construction (see [10] or [23]) which creates a tesselation of an infinite triangular prism over an equilateral triangle (which tessellates the two dimensional space). In the construction, new vertices are created above (and below) the three vertices of the triangle in the heights $\dots, 0, 3p, 6p, \dots$; $\dots, p, 4p, 7p, \dots$ and $\dots, 2p, 5p, \dots$, respectively, with a positive parameter p. Ordering these vertices with respect to their *height* (i.e. third component), tetrahedra are defined as convex hulls of four consequent vertices. A sketch of this construction is given by Figure 1, with the notation given by upcoming Lemma 2.2, which is the key ingredient of Theorem 2.1.

Lemma 2.2 (Induction Step)**.** *Let $d \geq 2$ and $\mathcal{T}_{d-1} = \{K^k_{d-1}\}_{k \in \mathbb{Z}^{d-1}}$ be a simplical tessellation of $(d-1)$-dimensional space such that the graph constructed from vertices and edges of $\mathcal{T}_{d-1}$ is a d-vertex-colorable graph.*

Then

- *there exists $\mathcal{T}_d = \{L^l_d\}_{l \in \mathbb{Z}^d}$ a simplical tesselation of d-dimensional space with additional shape parameter p_d,*
- *any connected compact subset of $\mathcal{T}_d$ is a face-to-face mesh,*
- *$\mathcal{T}_d$ is a $(d+1)$-vertex-colorable graph.*

Proof. Take an element $K^k_{d-1} \in \mathcal{T}_{d-1}$, $K^k_{d-1} = \operatorname{co}\{A_0, A_1, \dots A_{d-1}\}$. Thanks to the d-vertex-colorability we can assume that the labels of vertices represent their color. Let $A_i = [A_{i,1}, A_{i,2}, \dots A_{i,d-1}]$ be the coordinates of A_i in $d-1$ dimensional space.

We define the following points in d-dimensional space:

$$B_j = [A_{i(j),1}, A_{i(j),2}, \dots, A_{i(j),d-1}, jp_d], \quad j \in \mathbb{Z},$$

where $i(j) \equiv j \mod d$ and $p_d > 0$ is a parameter. Denote

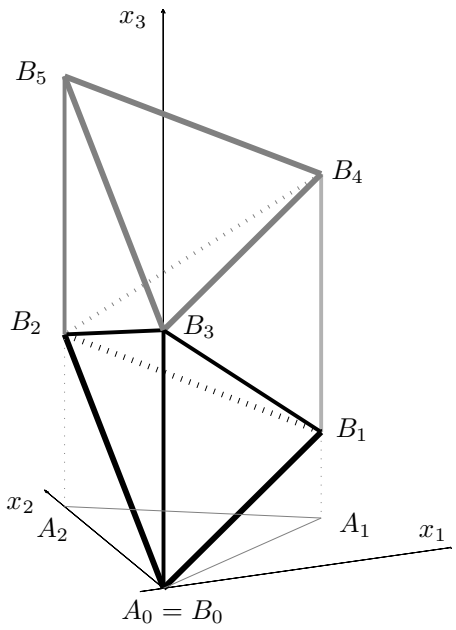


FIGURE 1. Illustration of Sommerville's original construction creating a three-dimensional face-to-face mesh above unilateral triangular mesh. For the sake of clarity, only elements $K_2^{0,0}$ and $L_3^{0,0,0}, L_3^{0,0,1}, L_3^{0,0,2}$ are shown.

$$L_d^{k,z} = \mathrm{co}\{B_z, B_{z+1}, \ldots, B_{z+d+1}\}, \tag{2.2}$$

the d-simplex as a convex hull of $d+1$ consequent vertices. Then $\{L_d^{k,z}\}_{z\in\mathbb{Z}}$ is a tessellation of an infinite d-dimensional prism with the cross-section K_{d-1}^k, see Figures 1 and 2 for illustration. As $\mathcal{T}_{d-1} = \{K_{d-1}^k\}_{k\in\mathbb{Z}^{d-1}}$ is a tessellation of $(d-1)$-dimensional space, then the set $\mathcal{T}_d := \{L_d^{k,z}\}_{(k,z)\in\mathbb{Z}^{d-1}\times\mathbb{Z}^d}$ forms a tessellation of d-dimensional space.

The construction uses the colors from the previous tessellation. Thus it is ensured that from any vertex A_j, that is shared by more simplices in $\mathcal{T}_{d-1}$, we create new vertices V_z of only one type; having the last coordinate of the form

$$V_{z,d}\frac{1}{p_d} \equiv c_{d-1}(A_j) \mod d = j. \tag{2.3}$$

This implies the face-to-face property, i.e. the facet of a simplex in tessellation $\mathcal{T}_d$ is a facet of another simplex.

Finally, we define the new coloring with

$$c_d(B_j) \equiv j \mod d+1, \qquad \text{for } B_j = [A_{i(j)}, jp_d]. \tag{2.4}$$

Such mapping is a vertex coloring, since edges of the graph are only edges in simplices and vertices in any simplex $L_d^{k,z}$ have a different last component, but the 'height' difference of two vertices connected by an edge does not exceed dp_d. □

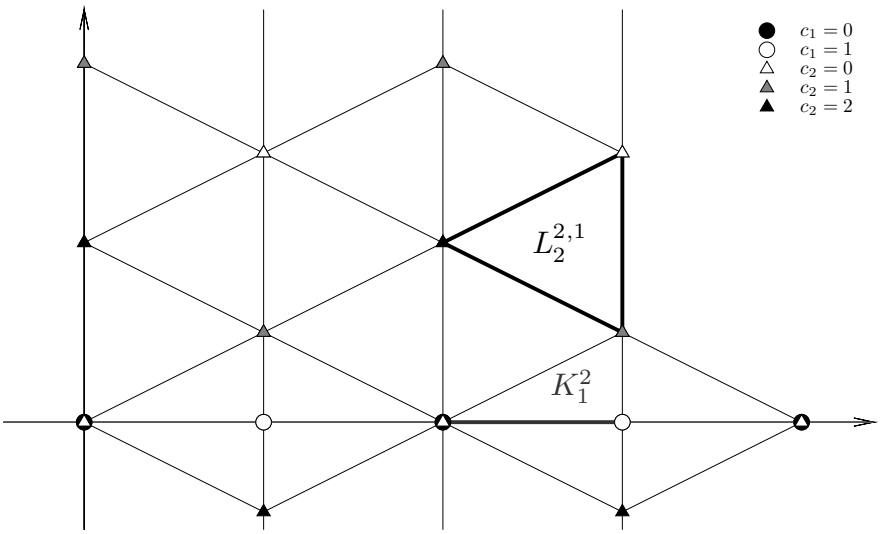


FIGURE 2. Illustration of creating a simplical face-to-face mesh of two dimensional space out of the one-dimensional one, with the parameters $p_1 = 1, p_2 = \frac{1}{2}$. The simplices K_1^2 and $L_2^{2,1}$ are marked in bold to clarify the notation defined by (2.2). For general values of the parameters there are two candidates for diameter of $L_2^{k,z}$, equal to $\sqrt{p_1^2 + p_2^2}$ and $2p_2$. Notice also the vertex coloring, assigned through (2.4).

The part that proves the face-to-face property based on vertex coloring of a graph was used already in [12]. Lemma 2.2 supplies the induction step, to complete the proof of Theorem 2.1, we show the initial step.

Proof of Theorem 1. A 1-dimensional Euclidean space (a line) can be divided into intervals of the length p_1. The color of a border point $A_z \in \{zp_1\}_{z\in\mathbb{Z}}$ is given by

$$c_1(A_z) \equiv z \mod 2.$$

The assumptions of Lemma 2.2 are satisfied, hence we have the initial step and the induction step, which finishes the proof. □

In general the created simplices are not identical. However, the following proposition shows that all elements of the tessellation $\mathcal{T}_d(\mathbf{p})$ have the same volume, i.e. the d-dimensional measure.

Proposition 2.3 (Equal Volume of the Elements)**.** *Let $\mathcal{T}_d(\boldsymbol{p})$ be the tessellation constructed by the procedure introduced in Proof of Lemma 2.2, with parameter vector $\mathbf{p} = (p_1, p_2, \ldots, p_d)$. Then for every simplex $L \in \mathcal{T}_d(\boldsymbol{p})$ we have*

$$\text{meas}_d L = \prod_{i=1}^{d} p_i. \tag{2.5}$$

Proof. In one-dimensional space, the situation is obvious; points $zp_1, z \in \mathbb{Z}$ divide a line into segments of the same length p_1. We prove the induction step. Let us assume that there exists $M_{d-1} > 0$ such that $\text{meas}_{d-1} K = M_{d-1}$ for any $K \in \mathcal{T}_{d-1}$.

According to the construction, an element $L \in \mathcal{T}_d$ is determined by the points

$$\begin{aligned} &B_z = [A_0, zp_d]; \quad B_{z+1} = [A_1, (z+1)p_d]; \quad \dots \\ &B_{z+d-1} = [A_{d-1}, (z+d-1)p_d]; \quad B_{z+d} = [A_0, (z+d)p_d], \end{aligned} \tag{2.6}$$

where $\text{co}(A_0, A_1, \dots, A_{d-1}) = K \in \mathcal{T}_{d-1}$.

The d-dimensional measure of a simplex is determined by the determinant of a matrix composed of the vectors that build the simplex, more precisely by the $(d!)^{-1}$ multiple of its absolute value. We use (2.6) and performing operations that do not affect the value of the determinant we obtain

$$\begin{aligned} meas_d L = \frac{1}{d!} \left| \det \begin{pmatrix} A_1 - A_0 & p_d \\ A_2 - A_0 & 2p_d \\ \vdots & \vdots \\ A_{d-1} - A_0 & (d-1)p_d \\ 0 & dp_d \end{pmatrix} \right| = \frac{dp_d}{d!} \left| \det \begin{pmatrix} A_1 - A_0 \\ A_2 - A_0 \\ \vdots \\ A_{d-1} - A_0 \end{pmatrix} \right| = \\ = p_d \cdot \text{meas}_{d-1} K. \end{aligned} \tag{2.7}$$

The proof is concluded by repeated use of (2.7) up to $d = 1$, which yields (2.5). □

Remark 2.4. *We consider only positive values of p_i, shortly we write $\boldsymbol{p} \in \mathbb{R}^d_+$, where $\mathbb{R}^d_+ = \{\mathbf{v} = (v_1, \dots, v_d) \in \mathbb{R}^d; v_i \geq 0, \forall i \in \{1, \dots, d\}\}$. It is rather a technical constraint, in fact one could allow $p_i \in \mathbb{R} \setminus \{0\}$. However, negative parameters affect only the orientation of the elements, not their shape characteristics. Therefore for the regularity optimization we can restrict ourselves to $\boldsymbol{p} \in \mathbb{R}^d_+$ which also simplifies the process.*

3. Regularity optimization

We have constructed a d-parametric family of tessellations in d-dimensional space, where the values of parameters $p_i, i = 1, \dots, d$ influence their shape. We find a vector of parameters $\mathbf{p}^\star = (p_1^\star, \dots, p_d^\star)$ for which the simplical elements are *shape optimal*. There are several regularity ratios with respect to which we might optimize. Some of them have been shown to be equivalent in the sense of the *strong regularity* even in general dimension, see [4], but not in the sense of their maximization.

For convenient calculation we use the following ratio

$$\vartheta(K) = \frac{\text{meas}_d K}{(\text{diam } K)^d}, \qquad d \geq 2, \tag{3.1}$$

where meas_d is the d-dimensional Lebesgue measure and diam K is the maximal distance of two points in K. The ratio $\vartheta(K)$ can be interpreted as a similarity of K to an equilateral simplex. In other words, we find $\mathbf{p}^\star$ and $K^\star$ which realize

$$\sup_{\mathbf{p}\in\mathbb{R}^d_+} \min_{K\in\mathcal{T}_d(\mathbf{p})} \vartheta(K). \tag{3.2}$$

As the simplices in $\mathcal{T}_d(\mathbf{p})$ are not identical, the optimization focuses on the *worst* simplex only. Since we proved by Proposition 2.3 that all elements in $\mathcal{T}_d(\mathbf{p})$ have the same d-measure, this *worst case* in the sense of (3.1) occurs when the diameter is maximal.

One can think through that the Sommerville's construction enables us to rewrite (3.2) using (2.5) as

$$\sup_{\mathbf{p}\in\mathbb{R}^d_+} \min_{\mathbf{w}\in W_d} \frac{\prod_{i=1}^d p_i}{\left(\sum_{i=1}^d w_i^2 p_i^2\right)^{\frac{d}{2}}}, \tag{3.3}$$

where

$$W_d := \left\{ \mathbf{w}\in(\mathbb{N}\cup\{0\})^d \,\middle|\, \exists k\in\{1,\dots,d\} : \begin{cases} w_k = k, \\ w_i = 0, \quad \text{for } 1\le i<k, \\ w_j = j-1, \ \text{for } k<j\le d \end{cases} \right\}. \tag{3.4}$$

We already reduced the set of possible vectors $\mathbf{w}$, as some of them are obviously dominated by those in W_d. Since $|W_d| = d$, we can also label its elements as $\mathbf{w}_j = (w_{j,1}, w_{j,2}, \dots, w_{j,d})$ where j is its first nonzero coordinate. We also define

$$D_j(\mathbf{p}) = \sqrt{\sum_{i=1}^d w_{j,i}^2 p_i^2}, \qquad \text{and} \qquad D(\mathbf{p}) = \max_{j\in\{1,\dots,d\}} D_j(\mathbf{p}), \tag{3.5}$$

so that (3.3) can be rewritten as

$$\sup_{\mathbf{p}\in\mathbb{R}^d_+} \min_{k\in\{1,\dots,d\}} \frac{\prod_{i=1}^d p_i}{D_k(\mathbf{p})^d}. \tag{3.6}$$

For illustration, we write out the 'diameter candidates' D_j explicitly,

$$\begin{array}{lcccccccc}
D_1(\mathbf{p})^2 &=& p_1^2 &+& p_2^2 &+& 4p_3^2 &+\dots& +(d-1)^2p_d^2, \\
D_2(\mathbf{p})^2 &=& & & 4p_2^2 &+& 4p_3^2 &+\dots& +(d-1)^2p_d^2, \\
D_3(\mathbf{p})^2 &=& & & & & 9p_3^2 &+\dots& +(d-1)^2p_d^2, \\
\vdots & & & & & & & & \\
D_{j-1}(\mathbf{p})^2 &=& (j-1)^2p_{j-1}^2 &+& (j-1)^2p_j^2 &+& j^2p_{j+1}^2 &+\dots& +(d-1)^2p_d^2, \\
D_j(\mathbf{p})^2 &=& & & j^2p_j^2 &+& j^2p_{j+1}^2 &+\dots& +(d-1)^2p_d^2, \\
\vdots & & & & & & & & \\
D_d(\mathbf{p})^2 &=& & & & & & & d^2p_d^2.
\end{array} \tag{3.7}$$

Now we can state the central theorem.

Theorem 3.1 (Optimal Parameters)**.** *Let $d \geq 2$ and let $\mathcal{T}_d(\boldsymbol{p})$ be a tessellation constructed through the procedure in Section 2. Then there exists a unique one-dimensional vector half-space*

$$P^\star = \Bigg\{ \boldsymbol{p}^\star_\kappa \in \mathbb{R}^d_+ | \boldsymbol{p}^\star_\kappa = \kappa \boldsymbol{p}^\star, \kappa > 0, \boldsymbol{p}^\star = (p^\star_1, \dots, p^\star_d), \\ p^\star_1 = 1, p^\star_2 = \frac{1}{\sqrt{3}}, p^\star_j = \frac{1}{j-1}\sqrt{\frac{2}{3}}, j \in \{3, \dots, d\} \Bigg\}, \tag{3.8}$$

of optimal parameters that realize

$$\sup_{\boldsymbol{p} \in \mathbb{R}^d_+} \min_{K \in \mathcal{T}_d(\boldsymbol{p})} \frac{\mathrm{meas}_d K}{(\mathrm{diam}\ K)^d}. \tag{3.9}$$

The rest of this section is devoted to the proof of Theorem 3.1, which consists of three main steps. First, we prove the existence of the maximizer $\mathbf{p}^\star$, then we show the particular form of the largest possible diameter that corresponds to the 'most deformed' simplex in $\mathcal{T}_d(\mathbf{p}^\star)$ and conclude the proof with determining the values of the components of $\mathbf{p}^\star$ through constrained optimization.

We would like to recall that we have three equivalent formulations of the optimization problem; (3.3), (3.6) and (3.9).

Lemma 3.2 (Existence of the Maximizer)**.** *Let $d \geq 2$ and let $\mathcal{T}_d(\boldsymbol{p})$ be a tessellation constructed through the procedure in Section 2. Then there exists a one-dimensional vector half-space*

$$P^\star = \left\{ \boldsymbol{p}^\star_\kappa \in \mathbb{R}^d_+ | \boldsymbol{p}^\star_\kappa = \kappa \boldsymbol{p}^\star, \kappa > 0 \right\}, \tag{3.10}$$

of optimal parameters that satisfy

$$\vartheta(K^\star(\boldsymbol{p}^\star_\kappa)) = \sup_{\boldsymbol{p} \in \mathbb{R}^d_+} \min_{K \in \mathcal{T}_d(\boldsymbol{p})} \frac{\mathrm{meas}_d K}{(\mathrm{diam}\ K)^d}, \tag{3.11}$$

for any $\boldsymbol{p}^\star_\kappa \in P^\star$ and some $K^\star \in \mathcal{T}_d(\boldsymbol{p}^\star_\kappa)$.

Proof. As for the above discussion, (3.9) is equivalent to (3.3). We observe that the ratio in (3.9) is 0-homogeneous, thus without loss of generality we may fix $p_1 = 1$. We continue with denoting the parametric vector by $\mathbf{p} \in \mathbb{R}^d_+$, keeping in mind that due to its first component being fixed, $\mathbf{p}$ may be considered as $(p_2, \dots, p_d) \in \mathbb{R}^{d-1}_+$. Defining

$$F(\mathbf{p}) := \min_{\mathbf{w} \in W_d} \frac{\prod_{i=2}^d p_i}{\left(\sum_{i=1}^d w_i^2 p_i^2 \right)^{\frac{d}{2}}},$$

we can rewrite (3.11) as $\sup_{\mathbf{p} \in \mathbb{R}^d_+} F(\mathbf{p})$ and we observe that

$$\lim_{p_j \to 0^+} F(\mathbf{p}) = 0, \qquad \lim_{p_j \to \infty} F(\mathbf{p}) = 0,$$

for any $j \in \{2, \dots, d\}$. Moreover, $F \in C(\mathbb{R}^{d-1}_+)$ and $F > 0$. Thus we infer that for any (sufficiently small) ε the set $\Omega_\varepsilon := \{F(\mathbf{p}) \geq \varepsilon\}$ is a non-empty, bounded and

closed subset of $\mathbb{R}^{d-1}_+$ and due to the continuity of F, it must attain its maximum in Ω_ε which necessarily coincides with the maximum of F in $\mathbb{R}^{d-1}_+$. □

In the next step we show which element of W_d in (3.3) or equivalently which D_k in (3.6) realizes the maximal diameter.

Lemma 3.3. *Let $\boldsymbol{p}^\star = (1, p_2^\star, \dots, p_d^\star)$ be the maximizer of (3.6). Then it holds that*

$$D(\boldsymbol{p}^\star) := \max_{k\in\{1,\dots,d\}} D_k(\boldsymbol{p}^\star) = D_1(\boldsymbol{p}^\star).$$

Proof. We proceed via contradiction. Let $D_1(\mathbf{p}^\star) < D_k(\mathbf{p}^\star) = D(\mathbf{p}^\star)$ for some $k \in \{2, \dots, d\}$. Then we define $\mathbf{p}' = (p'_1, \dots, p'_d)$ with

$$p'_1 = 1, \qquad p'_j = p_j^\star \cdot \frac{1}{1+\delta}, \quad j \in \{2, \dots, d\}, \tag{3.12}$$

where $\delta > 0$ is chosen small enough to ensure $D_1(\mathbf{p}') < D_k(\mathbf{p}') = D(\mathbf{p}')$. Then it holds that

$$D(\mathbf{p}') = D_k(\mathbf{p}') = D_k(\mathbf{p}^\star)\frac{1}{1+\delta} = D(\mathbf{p}^\star)\frac{1}{1+\delta}. \tag{3.13}$$

Substitution from (3.12) and (3.13) into (3.3) yields

$$\vartheta(K(\mathbf{p}')) = \frac{\prod_{i=1}^d p'_i}{D_j(\mathbf{p}')^d} = \frac{\prod_{i=1}^d p_i^\star}{D_j(\mathbf{p}^\star)^d} \cdot \frac{(1+\delta)^d}{(1+\delta)^{d-1}} = (1+\delta)\vartheta(K(\mathbf{p}^\star)),$$

which contradicts the assumption of the maximality of $\mathbf{p}^\star$. □

By virtue of Lemma 3.3, the maximization problem (3.6), which is equivalent to (3.9), reduces to the optimization of a C^1 function with inequality constraints,

$$\max\left\{ \frac{\prod_{i=1}^d p_i}{D_1(\mathbf{p})^d} \,\middle|\, \mathbf{p} \in \mathbb{R}^d_+, p_1 = 1, D_1(\mathbf{p})^2 \geq D_j(\mathbf{p})^2, \text{for all } j \in \{2, \dots, d\} \right\}. \tag{3.14}$$

To prove Theorem 3.1 it suffices to show that problem (3.14) has a unique solution, which is $\mathbf{p}^\star$ in (3.8). By virtue of Lemma 3.3 the optimization problem (3.14) is equivalent to (3.3) and further to the original problem (3.9), hence Lemma 3.2 guarantees it has a solution.

The function

$$F_1(\mathbf{p}) = F_1(p_2, \dots, p_d) = \frac{\prod_{i=2}^d p_i}{D_1(1, p_2, \dots, p_d)^d}, \tag{3.15}$$

is continuously differentiable in $\mathbb{R}^{d-1}_+$, hence its constrained maximizer $\mathbf{p}^\star$ satisfies the *necessary* Karush-Kuhn-Tucker conditions. They read as follows,

$$\frac{\partial}{\partial p_j} F_1(\mathbf{p}) = \sum_{i=2}^d \mu_i \frac{\partial}{\partial p_j} \left(D_i(\mathbf{p})^2 - D_1(\mathbf{p})^2 \right), \tag{3.16}$$

$$\mu_j \left(D_j(\mathbf{p})^2 - D_1(\mathbf{p})^2 \right) = 0, \tag{3.17}$$

$$\mu_j \geq 0, \qquad D_j(\mathbf{p}) \leq D_1(\mathbf{p}), \tag{3.18}$$

for $j = \{2, \dots, d\}$.

Let us focus on the right hand side of (3.16). Recalling (3.7) with $p_1 = 1$, one can express

$$\frac{\partial}{\partial p_j}(D_i(\mathbf{p})^2 - D_1(\mathbf{p})^2) = \begin{cases} -2(j-1)^2 p_j & \text{for } j < i, \\ 2(2j-1)p_j & \text{for } j = i, \\ 0 & \text{for } j > i. \end{cases} \tag{3.19}$$

Then, by virtue of (3.15) and (3.4) with (3.5) and just derived (3.19), we can rewrite (3.16) as

$$\begin{aligned} &\frac{\prod_{i=2}^{d} p_i}{D_1(\mathbf{p})^{2d}} \left(\frac{1}{p_j} D_1(\mathbf{p})^d - d(j-1)^2 D_1(\mathbf{p})^{d-2} p_j \right) \\ &\quad - 2\mu_j(2j-1)p_j + 2(j-1)^2 p_j \sum_{i=j+1}^{d} \mu_i = 0, \qquad j \in \{2, \dots, d\}. \end{aligned} \tag{3.20}$$

It is not obvious how to get a solution of (3.16–3.18) or its equivalent (3.17, 3.18, 3.20), nor its uniqueness. At the end, we show that $\mu_j = 0$ for $j \in \{3, \dots, d\}$ and $\mu_2 > 0$ which is then enough to determine uniquely the solution. To get this, we proceed in three steps. We show that

- there *exists* $k \in \{2, \dots, d\}$ such that $\mu_k > 0$,
- this k is *unique*,
- $k = 2$.

We introduce three lemmas, each corresponding to one of the items at the above list.

Lemma 3.4 (Existence of an Active Constraint)**.** *Let $d \geq 2$ and $\boldsymbol{p}^\star$ be the maximizer of (3.14). Then $\boldsymbol{p}^\star$ is a solution of (3.16–3.18) with $(\mu_2, \dots, \mu_d) \neq \mathbf{0}$, i.e. there exists $k \in \{2, \dots, d\}$ such that $\mu_k > 0$.*

Proof. We proceed via contradiction. Assume that $\mu_j = 0$ for all $j \in \{2, \dots, d\}$. In such case (3.20), which is a consequence of (3.16), implies

$$p_j^\star = \frac{D_1(\mathbf{p}^\star)}{(j-1)\sqrt{d}}, \quad j \in \{2, \dots, d\},$$

which substituted into $D_2(\mathbf{p})^2$ yields

$$D_2(\mathbf{p}^\star)^2 = \frac{4}{d} D_1(\mathbf{p}^\star)^2 + \sum_{i=3}^{d} \frac{D_1(\mathbf{p}^\star)^2}{d} = \frac{d+2}{d} D_1(\mathbf{p}^\star)^2 > D_1(\mathbf{p}^\star)^2,$$

which contradicts (3.18). Thus there is some $k \in \{2, \dots, d\}$ for which $\mu_k > 0$. □

For $d = 2$ Lemma 3.4 implies directly that $k = 2$. For $d \geq 3$ we supply the following lemma.

Lemma 3.5 (One Active Constraint)**.** *Let $d \geq 3$ and $\boldsymbol{p}^\star$ be a maximizer in (3.14) which satisfies (3.16–3.18) with $\mu_k > 0$ for some $k \in \{3, \dots, d\}$. Then $\mu_j = 0$ for all $j \in \{2, \dots, k-1, k+1, \dots, d\}$ and $\boldsymbol{p}^\star = (1, p_2^\star, \dots, p_d^\star)$ fulfills*

$$p_j^\star = \begin{cases} \dfrac{\sqrt{2}D_k(\boldsymbol{p}^\star)}{(j-1)\sqrt{dk}}\sqrt{\dfrac{2k-1}{k-1}}, & \text{for } j \in \{2,\dots,k-1\}, \\ \dfrac{D_k(\boldsymbol{p}^\star)}{\sqrt{dk}}, & \text{for } j = k, \\ \dfrac{D_k(\boldsymbol{p}^\star)}{(j-1)\sqrt{d}}, & \text{for } j \in \{k+1,\dots,d\}. \end{cases} \tag{3.21}$$

Proof. Let us take the largest $k \in \{3,\dots,d\}$ for which $\mu_k > 0$. Then for $j \in \{k+1,\dots,d\}$ we have $\mu_j = 0$. This enables us to deduce directly from (3.20) that

$$p_j^\star = \frac{D_1(\mathbf{p}^\star)}{(j-1)\sqrt{d}}, \quad j \in \{k+1,\dots,d\}. \tag{3.22}$$

And as $D_1 = D_k$ (this follows from the assumption $\mu_k > 0$ and (3.17)) we can use (3.22) for computing $p_k^\star$. The computation

$$D_k(\mathbf{p}^\star)^2 = k^2(p_k^\star)^2 + \sum_{j=k+1}^{d}(j-1)^2(p_j^\star)^2 = k^2(p_k^\star)^2 + \frac{d-k}{d}D_k(\mathbf{p}^\star)^2, \tag{3.23}$$

yields

$$p_k^\star = \frac{D_k(\mathbf{p}^\star)}{\sqrt{dk}}. \tag{3.24}$$

Notice that (3.24) holds even if $k = d$ and the summation in (3.23) is void.

Since $D(\mathbf{p}^\star) = D_1(\mathbf{p}^\star) = D_k(\mathbf{p}^\star)$, then the constrained maximization problem (3.14) is equivalent to a constrained optimization, where D_k is taken as the diameter, i.e.

$$\max\left\{ \frac{\prod_{i=1}^d p_i}{D_k(\mathbf{p})^d} \,\middle|\, \mathbf{p} \in \mathbb{R}_+^d, p_1 = 1, D_k(\mathbf{p})^2 \geq D_j(\mathbf{p})^2, \text{for all } j \in \{1,\dots,d\} \right\}. \tag{3.25}$$

Arguing as before, the maximizer in (3.25) exists and fulfills the following necessary Karush-Kuhn-Tucker conditions,

$$\frac{\partial}{\partial p_j}\frac{\prod_{i=2}^d p_i}{D_k(\mathbf{p})^d} = \sum_{\substack{i=1 \\ i\neq k}}^{d} \nu_i \frac{\partial}{\partial p_j}(D_i(\mathbf{p})^2 - D_k(\mathbf{p})^2), \tag{3.26}$$

for $j \in \{2,\dots,d\}$ and

$$\nu_i(D_i(\mathbf{p})^2 - D_k(\mathbf{p})^2) = 0, \tag{3.27}$$

$$\nu_i \geq 0, \qquad D_i(\mathbf{p}) \leq D_k(\mathbf{p}), \tag{3.28}$$

for $i \in \{1,\dots,k-1,k+1,\dots,d\}$ and moreover we know that $D_1(\mathbf{p}) = D_k(\mathbf{p})$. As we already settled $j \in \{k+1,\dots,d\}$, we need to focus on $j \in \{2,\dots,k-1\}$ only, hence we consider only those.

We know that

$$\frac{\partial}{\partial p_j}\frac{\prod_{i=2}^d p_i}{D_k(\mathbf{p})^d} = \frac{\prod_{i=2}^d p_i}{D_k(\mathbf{p})^d}\frac{1}{p_j}, \qquad j \in \{2,\dots,k-1\}, \tag{3.29}$$

and using (3.7) we compute the right-hand side of (3.26) for $j \in \{2,\dots,k-1\}$ as

$$\frac{\partial}{\partial p_j}(D_i(\mathbf{p})^2 - D_k(\mathbf{p})^2) = \begin{cases} 2(j-1)^2 p_j & \text{for } i < j, \\ 2j^2 p_j & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases} \tag{3.30}$$

Collecting (3.29–3.30) together with $\nu_i = 0$ for $i > k$ (as $D_i(\mathbf{p}) < D_k(\mathbf{p})$ by assumption), we can rewrite (3.26) in the form

$$\frac{\prod_{i=2}^d p_i}{D_k(\mathbf{p})^d}\frac{1}{p_j} = 2\nu_j j^2 p_j + 2(j-1)^2 p_j \sum_{i=1}^{j-1}\nu_i, \quad j \in \{2,\dots,k-1\}. \tag{3.31}$$

Take any $j \in \{2,\dots,k-1\}$, we have either $\nu_j = 0$ or $\nu_j > 0$.
Let first assume $\nu_j = 0$. Then, from (3.31) we deduce

$$p_j^2 = p_{j,u}^2 = \frac{\prod_{i=2}^d p_i}{2D_k(\mathbf{p})^d}\frac{1}{(j-1)^2\sum_{i=1}^{j-1}\nu_i}. \tag{3.32}$$

If $\nu_j > 0$, then

$$p_j^2 = p_{j,c}^2 = \frac{\prod_{i=2}^d p_i}{2D_k(\mathbf{p})^d}\frac{1}{j^2\nu_j + (j-1)^2\sum_{i=1}^{j-1}\nu_i}.$$

We observe that $p_{j,c} < p_{j,u}$ and $\mathbf{p}^\star$ is supposed to maximize $\prod_{i=2}^d p_i \cdot (D_k(\mathbf{p}))^{-d}$, where $D_k(\mathbf{p})$ is independent of p_j for $j \in \{1,\dots,k-1\}$. Thus $p_j^\star$ needs to maximize only $\prod_{i=2}^d p_i$, i.e. only its value. That is why we choose its *unconstrained* version $p_{j,u}^\star$ from (3.32), i.e. $\nu_j = 0$ for any $j \in \{2,\dots,k-1\}$. This enables to rewrite (3.32) into

$$p_j^2 = p_{j,u}^2 = \frac{\prod_{i=2}^d p_i}{2\nu_1 D_k(\mathbf{p})^d}\frac{1}{(j-1)^2}. \tag{3.33}$$

Computing (3.26) also for $j = k$, one gets

$$\frac{1}{D_k(\mathbf{p})^{2d}}\left(\prod_{i=2}^d p_i D_k(\mathbf{p})^d\frac{1}{p_k} - d\prod_{i=2}^d p_i D_k(\mathbf{p})^{d-2}k^2 p_k\right) = \nu_1 2(-2k+1)p_k, \tag{3.34}$$

and after substituting $p_k^\star$ from (3.24) into (3.34) we can express ν_1 as

$$\nu_1 = \frac{dk\prod_{i=2}^d p_i}{2D_k(\mathbf{p})^{d+2}}\frac{k-1}{2k-1}. \tag{3.35}$$

Collecting (3.22), (3.24) and substituting from (3.35) into (3.33) we get (3.21), which concludes the proof. □

Lemmas 3.4 and 3.5 give rise to the following corollary.

Corollary 3.6. *Let $d \geq 3$ and $\boldsymbol{p}^\star$ be a maximizer in (3.14). Then there exists a unique $k \in \{2, \dots, d\}$ such that $D_k(\boldsymbol{p}^\star) = D_1(\boldsymbol{p}^\star) = D(\boldsymbol{p}^\star)$ and (3.21) holds.*

Proof. Lemma 3.4 together with (3.17) gives existence of $k \in \{2, \dots, d\}$ such that $D_k(\mathbf{p}^\star) = D_1(\mathbf{p}^\star) = D(\mathbf{p}^\star)$. For $k \geq 3$, Lemma 3.5 gives uniqueness of such k and also (3.21).

Let further $D_j(\mathbf{p}^\star) < D_1(\mathbf{p}^\star)$ for all $j \in \{3, \dots, d\}$, then by Lemma 3.4 and (3.17) necessarily $D_2(\mathbf{p}^\star) = D_1(\mathbf{p}^\star) = D(\mathbf{p}^\star)$. Using the procedure from the beginning of the proof of Lemma 3.5, one recovers (3.21) also for $k = 2$. □

Finally, we show that k from the previous lemma is equal to 2 which will enable us to determine also the values of $p_i^\star$.

Lemma 3.7. *Let $d \geq 2$ and $\boldsymbol{p}^\star$ be a maximizer in (3.14). Then it holds that*

$$D(\boldsymbol{p}^\star) = D_1(\boldsymbol{p}^\star) = D_2(\boldsymbol{p}^\star), \tag{3.36}$$

and

$$p_2^\star = \sqrt{\frac{1}{3}}, \qquad p_j^\star = \sqrt{\frac{2}{3}}\frac{1}{j-1}, \quad j \in \{3, \dots, d\}. \tag{3.37}$$

Proof. Let $d = 2$. Then Lemma 3.4 implies (3.36), which can be written explicitly as $1 + (p_2^\star)^2 = 4(p_2^\star)^2$. Thus we infer $p_2^\star = 3^{-1/2}$.

Let further $d \geq 3$. Then from Corollary 3.6 we get a unique existence of some $k \in \{2, \dots, d\}$ for which $D(\mathbf{p}^\star) = D_1(\mathbf{p}^\star) = D_k(\mathbf{p}^\star)$ and the relation (3.21) for $\mathbf{p}^\star$.

We prove $k = 2$ via contradiction. Let us assume that $k \geq 3$. Then, $D(\mathbf{p}^\star) = D_1(\mathbf{p}^\star) = D_k(\mathbf{p}^\star) > D_2(\mathbf{p}^\star)$. Writing out $D_2(\mathbf{p}^\star)$ explicitly using (3.21), we get

$$D(\mathbf{p}^\star)^2 > D_2(\mathbf{p}^\star)^2 = \frac{D(\mathbf{p}^\star)^2}{d}\left(2\frac{4(2k-1)}{k(k-1)} + 2\frac{(k-2)(2k-1)}{k(k-1)} + \frac{(k-1)^2}{k} + (d-k)\right). \tag{3.38}$$

Direct computation simplifies inequality (3.38) into

$$\frac{2k^2 + 9k - 5}{k(k-1)} < 0,$$

which is not true for any $k \in \mathbb{N}$, a contradiction. Therefore $k = 2$, and from (3.21) we get

$$p_2^\star = \frac{D(\mathbf{p}^\star)}{\sqrt{2d}}, \qquad p_j^\star = \frac{D(\mathbf{p}^\star)}{(j-1)\sqrt{d}}, \quad j \in \{3, \dots, d\}, \tag{3.39}$$

which we substitute into $D_1(\mathbf{p})^2$ to get

$$D(\mathbf{p}^\star)^2 = D_1(\mathbf{p}^\star)^2 = 1 + \frac{D(\mathbf{p}^\star)^2}{2d} + (d-2)\frac{D(\mathbf{p}^\star)^2}{d}. \tag{3.40}$$

From (3.40) we deduce $D(\mathbf{p}^\star)^2 = \frac{2}{3}d$ which, substituted into (3.39) yields (3.37). □

4. Concluding remarks

We conclude with five remarks on various topics.

4.1. **Optimization at each step.** Notice that the optimal values of parameters (3.37) are independent of the dimension d. This can be interpreted that the most regular partition of d-dimensional space is constructed above the most regular partition of $(d-1)$-dimensional space. As a consequence, the shape optimization we performed is equivalent to the shape optimization at every dimension, which gives a sequence of one-dimensional optimization problems that is technically much less demanding.

4.2. **Integer sequence for OEIS.** One can easily see that for suitable κ it is possible to express the squares of the components of $\mathbf{p}^\star_\kappa$ from (3.8) as fraction with unit numerator and integer denominator. Largest such κ, yielding the smallest possible integers in those fractions, is $\kappa = 2^{-1/2}$. For this value, the denominators give the following values: `2, 6, 12, 27, 48, 75, 108, 147, 192, 243, 300...`, having the formula for j-th item $a_j = 3(j-1)^2$ for $j \geq 3$. This sequence has not been indexed in Sloane's database of integer sequences [22].

4.3. **Shape optimality of the partition.** It is not obvious whether there exists any better simplical mesh that cannot be constructed by our method. However, in 2D there is no triangle with better ratio ϑ than the equilateral one. Similarly, in 3D, our method gives the standard Sommerville tetrahedron (see [13, Figure 2]), which as for Naylor [20] is the best one among space-filling tetrahedra when considering the regularity ratio ϑ.

4.4. **The four-dimensional case.** Křížek in [15] states that the question of the existence of a partition of 4-dimensional space into acute simplices is open while Brandts et al. [3, Conjecture 1.] expect that there will be no such partition. Using the partition

$$\mathcal{T}_4(\mathbf{p}^\star) = \mathcal{T}_4\left(1, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{6}}, \sqrt{\frac{2}{27}}\right), \tag{4.1}$$

one can verify, that the partition consists of simplices of a single type (and its reflections), whose largest dihedral angle equals $\pi/2$. Therefore we have a *non-obtuse* partition of $\mathbb{R}^4$, leaving the question of existence of an acute one open.

4.5. **Non-euclidean geometries.** We devote the last remark to the fact that the construction as well as the optimization is independent of the underlying geometry and thus might be used also for computations in non-euclidean spaces. More on tessellations of hyperbolic spaces can be found in works of Coxeter [6] or [7], and Margenstern [17], [18], [19]. As Margenstern points out, these works might find their use in computational problems of theory of relativity or cosmological research, but such results had not been published before 2003 and to the best author's knowledge not even since these days.

References

[1] P. Angot, C.-H. Bruneau, and P. Fabrie. A penalization method to take into account obstacles in incompressible viscous flows. *Numerische Mathematik*, 81(4):497–520, 1999.

[2] J. Brandts, S. Korotov, and M. Křížek. Simplicial finite elements in higher dimensions. *Applications of Mathematics*, 52(3):251–265, 2007.

[3] J. Brandts, S. Korotov, M. Křížek, and J. Šolc. On nonobtuse simplicial partitions. *SIAM review*, 51(2):317–335, 2009.

[4] J. Brandts, S. Korotov, and M. Křížek. On the equivalence of ball conditions for simplicial finite elements in $\mathbf{R}^d$. *Appl. Math. Lett.*, 22(8):1210–1212, 2009.
[5] J. Brandts and M. Křížek. Gradient superconvergence on uniform simplicial partitions of polytopes. *IMA Journal of Numerical Analysis*, 23(3):489–505, 2003.
[6] H. Coxeter. *Non-Euclidean Geometry*. MAA spectrum. Mathematical Association of America, 1998.
[7] H. S. M. Coxeter and G. J. Whitrow. World-structure and non-euclidean honeycombs. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 201(1066):417–437, 1950.
[8] E. Feireisl, R. Hošek, D. Maltese, and A. Novotný. Error estimates for a numerical method for the compressible Navier-Stokes system on sufficiently smooth domains. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2016. To appear. Available as preprint IM-2015-46 at http://math.cas.cz.
[9] E. Feireisl, T. Karper, and M. Michálek. Convergence of a numerical method for the compressible Navier-Stokes system on general domains. *Submitted to Numerische Mathematik*, 2014. To appear. Available as preprint IM-2014-47 at http://math.cas.cz.
[10] M. Goldberg. Three infinite families of tetrahedral space-fillers. *J. Comb. Theory, Ser. A*, 16(3):348–354, 1974.
[11] P. Groth, H. Mårtensson, and L.-E. Eriksson. Validation of a 4d finite volume method for blade flutter. In *ASME 1996 International Gas Turbine and Aeroengine Congress and Exhibition*, pages V005T14A052–V005T14A052. American Society of Mechanical Engineers, 1996.
[12] R. Hošek. Face-to-face partition of 3D space with identical well-centered tetrahedra. *Appl. Math.*, 60(6):637–651, 2015.
[13] R. Hošek. Strongly regular family of boundary-fitted tetrahedral meshes of bounded C^2 domains. *Appl. Math.*, 61(3):233–251, 2016.
[14] K. Khadra, P. Angot, S. Parneix, and J.-P. Caltagirone. Fictitious domain approach for numerical modelling of Navier–Stokes equations. *International Journal for Numerical Methods in Fluids*, 34(8):651–684, 2000.
[15] M. Křížek. There is no face-to-face partition of R5 into acute simplices. *Discrete & Computational Geometry*, 36(2):381–390, 2006.
[16] B. Laumert, H. Mårtensson, and T. H. Fransson. Simulation of rotor/stator interaction with a 4d finite volume method. In *ASME Turbo Expo 2002: Power for Land, Sea, and Air*, pages 1045–1054. American Society of Mechanical Engineers, 2002.
[17] M. Margenstern. Cellular automata and combinatoric tilings in hyperbolic spaces. a survey. In *Discrete Mathematics and Theoretical Computer Science*, pages 48–72. Springer, 2003.
[18] M. Margenstern. About an algorithmic approach to tilings p,q of the hyperbolic plane. *J.UCS*, 12(5):512–550, may 2006.
[19] M. Margenstern. Coordinates for a new triangular tiling of the hyperbolic plane. *CoRR*, abs/1101.0530, 2011.
[20] D. J. Naylor. Filling space with tetrahedra. *Int. J. Numer. Methods Eng.*, 44(10):1383–1395, 1999.
[21] C. Schwab, E. Süli, and R. A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 42(5):777–819, 2008.
[22] N. Sloane. The on-line encyclopaedia of integer sequences. `http://www.oeis.org`.
[23] D. Sommerville. Space-filling tetrahedra in Euclidean space. *Proc. Edinburgh Math. Soc.*, 41:49–57, 1923.

R.H., Department of Mathmatics, University of West Bohemia, 306 14 Plzen, Czech Republic
E-mail address: `radhost@kma.zcu.cz`